Prominent microblog users prediction during crisis events : using phase-aware and temporal modeling of users behavior

Imen Bizid

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE S2IM

THÈSE présentée par :

Imen BIZID

préparée aux : **Laboratoire Informatique, Image et Interaction (L3i)**
&
Laboratoire de Télédétection et Systèmes d'Information à Référence Spatiale (LTSIRS)

soutenue le : **13 Décembre 2016**
pour obtenir le grade de : **Docteur**

Discipline : **Informatique et applications**

Prédiction des utilisateurs primordiaux des microblogs durant les situations de crise : Modélisation temporelle des comportements des utilisateurs en fonction des phases des évènements

---

JURY :

Djamel BEN SLIMANE	Professeur, Université Lyon 1, Rapporteur
Ouajdi KORBAA	Professeur, Université de Sousse, Rapporteur
Mohand BOUGHANEM	Professeur, Université Toulouse 3, Examinateur
Imed Riadh FARAH	Professeur, Université de La Mannouba, Examinateur
Antoine DOUCET	Professeur, Université de La Rochelle, Examinateur
Sami FAIZ	Professeur, Université de La Mannouba, Directeur de thèse
Patrice BOURSIER	Professeur, Université de La Rochelle, Directeur de thèse
Nibal NAYEF	Chercheuse Post-doc, Université de La Rochelle, Co-encadrante

UNIVERSITY OF LA ROCHELLE

S2IM DOCTORAL SCHOOL

THESIS by :

Imen Bizid

carried out at : **Laboratoire Informatique, Image et Interaction (L3i)**
&
Laboratoire de Télédétection et Systèmes d'Information à Référence Spatiale (LTSIRS)

defended on : **13 December 2016**
for the award of the degree of : **Doctor of Philosophy**

Discipline : **Computer science and applications**

Prominent Microblog Users Prediction during Crisis Events : Using Phase-aware and Temporal Modeling of Users Behavior

---

JURY :

Djamel Ben Slimane	Professor, University of Lyon 1, Reviewer
Ouajdi Korbaa	Professor, University of Sousse, Reviewer
Mohand Boughanem	Professor, University of Toulouse 3, Examiner
Imed Riadh Farah	Professor, University of Mannouba, Examiner
Antoine Doucet	Professor, University of La Rochelle, Examiner
Sami Faiz	Professor, University of La Mannouba, Thesis director
Patrice Boursier	Professor, University of La Rochelle, Thesis director
Nibal Nayef	Postdoctoral researcher, University of La Rochelle, Co-supervisor

This thesis is dedicated to :

my parents

my husband

my sister

my brothers

my best friends
and everyone who loves me

Acknowledgments

First of all I would like to thank the members of the jury Professors Mohand Boughanem, Djamel Ben Slimane, Ouajdi Korbaa, Imed Riadh Farah and Antoine Doucet for having accepted to assess my thesis.

I would like also to express my heartfelt gratitude to my supervisor, Doctor Nibal Nayef, for the continuous academic and emotional support, her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I can not forget her hard times reviewing my thesis progress, giving me her valuable suggestions and made corrections. Her unflinching courage and conviction will always inspire me, and I hope to continue to work with her noble thoughts. My sincere gratitude goes to my advisors, Professor Patrice Boursier, and Professor Sami Faiz for the guidance, encouragement, and support. This thesis was made possible due to their constant encouragement and belief in me. My thanks are also due to my collaborators Frédéric Rousseaux, Jacques Morcos, and Antoine Doucet with whom I enjoyed intellectually stimulating discussions. All of these people have broadened my perspective on research.

A special acknowledgment goes to the Poitou-Charentes region for this thesis funding. I would like to thank all contributors for their generous support. A heartfelt thanks to supportive colleagues at the L3i lab and all the members of the LTSIRS lab, for their instant help and kindness. Many thanks go to everyone who participated in this research study, volunteers who have participated in conducted user evaluation.

I also wish to thank my family especially my parents for believing in me and supporting me through this process. Special thanks goes to my husband Mehdi Chaibi for his invaluable support, both emotional as well as intellectual. He is the main motivation for me to do this PhD.

I would like to thank my best friends Mariem and Ameni for their support, and encouragement during difficult times. A lot of thanks and gratitude to my 123 open space friends Sophea, Maroua, Marcela, Chloé, Sovan, and Damien for those special moments that they have provided during these three years. It was a real pleasure to work with them in the same office. Special thanks go to Wafa, Fairouz and my cousin Wiem for their help during difficult times.

Abstract

Real-time information retrieval from microblogs during crisis events is hindered by many challenges such as : streaming data analysis in real time, the variety of information format and language processing, large crisis events datasets, and extracting relevant and fresh information from a huge amount of outdated and redundant data. Existing methods which perform the information retrieval task during crisis events are either based on a user-centric retrieval approach or a content-based retrieval approach. However, state-of-the-art content-based retrieval approaches are sensitive to the complexity of the analyzed content in terms of format, language and freshness. This sensitivity makes these approaches unsuitable to the information retrieval task in the context of crisis events where any format of relevant information need to be considered.

This dissertation explores user-centric approaches for information retrieval in the context of crisis events. The relevance and freshness of event-related information is associated with the prominence of their producers. Prominent microblog users in this thesis context refer to key users who are susceptible to share relevant and exclusive information during a specific crisis event. Accessing the shared information of these users in real time would help emergency teams to have knowledge about the situation in the threatened and affected areas. Identifying prominent users is achieved using a set of novel methods evaluating users according to their behavior during the analyzed crisis event. Those methods have performed significantly better than state-of-the-art methods. An overview of the key contributions of this dissertation is given in the following :

First, this dissertation presents a multi-agent system composed of two main modules : data collection module and user tracking module. The data collection module insures the collection of the different information shared by users interested in the specific analyzed event. This module has been used to collect two crisis events datasets relative to the 2014 Herault and the 2015 Alpes-Maritimes flooding events. Our system also integrates a users tracking module which supports the integration of any prominent users identification approach and insures the tracking of the selected prominent users.

Novel approaches for real-time prominent users identification in the context of crisis events are also proposed in this dissertation. These approaches focus on three key aspects of prominent users identification. Firstly, we have studied the efficiency of state-of the art and new proposed raw features for characterizing user behavior during crisis events. Based on the selected features, we have designed several engineered features qualifying user activities by considering both their on-topic and off-topic shared information. Secondly, we have proposed a phase-aware user modeling approach taking into account the user behavior change

according to the event evolution over time. This user modeling approach comprises the following new novel aspects (1) modeling microblog users behavior evolution by considering the different event phases, (2) characterizing users activity over time through a temporal sequence representation, and (3) time-series-based selection of the most discriminative features characterizing users at each event phase. Thirdly, based on this proposed user modeling approach, we train various prediction models to learn to differentiate between prominent and non-prominent users behavior during crisis events based on prior events data. The learning task has been performed using SVM and MoG-HMMs supervised machine learning algorithms. Learning the different models based on prior events data makes the prediction process computationally feasible in real time during new real-world crisis events cases.

The two collected datasets were used to evaluate the performance of our resulted identification models. One dataset was used for learning the model and the other one for testing. We have experimentally shown that the best prediction results were obtained while we represent and evaluate user behavior based on the following dimensions : (1) topical activities dimension by considering both on- and off-topic user activities specially during the red alert phase of the analyzed crisis event, (2) temporal dimension by characterizing the user behavior evolution over time, and (3) event phases dimension by highlighting the user behavior and prominence evolution at each event phase. Based on this user behavior representation, the learned MoG-HMMs models have succeeded to point out the particularities of prominent and non-prominent users behavior during crisis events. These phase-aware MoG-HMMs have outperformed state-of-the-art prediction models in terms of prediction, classification and ranking performance. Most of prominent users have been identified at an early stage of each phase of the analyzed crisis event.

Overall, these contributions could be considered as important steps in the right direction of the research of information retrieval from microblogs during crisis events. The hope is that, such contributions could insure better situation awareness for emergency teams during crisis events.

Résumé

La recherche d'information dans les microblogs durant les situations de crise est entravée par plusieurs défis tels que : l'analyse des flux d'informations partagées en temps réel, la variété des formats (i.e. texte, image, lien et vidéo) et des langues utilisés dans les microblogs, le grand volume de données partagées durant ses évènements et l'extraction des informations pertinentes et fraiches du grand volume d'informations redondantes et obsolètes. Il existe dans la littérature deux principales approches de recherche d'information pour faire face à ces défis : les approches basées sur le contenu et les approches centrées sur l'utilisateur. Cependant, les approches basées sur le contenu sont sensibles au format, à la langue et à la fraicheur du contenu analysé. Cette sensibilité rend ces approches inadaptées pour la recherche d'information en temps réel durant les situations de crise où tout type d'information doit être considéré.

Cette thèse explore les approches centrées utilisateurs pour la recherche d'information dans les microblogs durant les situations de crise. La pertinence et l'exclusivité des informations partagées par rapport au sujet de l'évènement sont associées à l'importance de l'utilisateur qui les a partagées. Dans le cadre de cette thèse, les utilisateurs primordiaux sont définis comme étant les utilisateurs clés qui sont susceptibles de partager des informations pertinentes et exclusives au sujet des évènements en question. L'accès en temps réel aux informations partagées par ces utilisateurs permettra aux équipes intervenant en cas d'urgence d'avoir une vue globale sur ce qui se passe dans les zones affectées et/ou menacées par l'évènement. L'identification de ces utilisateurs est assurée par un ensemble de nouvelles méthodes évaluant chaque utilisateur selon son comportement durant l'évènement. Ces méthodes se sont avérées plus performantes que celles proposées dans la littérature. Nous détaillons les principales contributions de cette thèse ci-dessous.

Cette thèse propose en premier lieu un système multi-agents composé de deux modules : un module chargé de la collecte des données et un module chargé de traquer les utilisateurs primordiaux. Le premier module assure la collecte de toute information partagée par les utilisateurs intéressés par l'évènement en question. Ce module nous a permis de collecter deux collections de données relatives aux innondations qui ont eu lieu dans l'Hérault en 2014 et les Alpes-Maritimes en 2015. Quant au module de suivi des utilisateurs, il a été conçu pour supporter l'intégration de toute approche d'identification d'une catégorie d'utilisateurs bien déterminée tout en assurant l'accès à leurs informations en temps réel.

Dans cette thèse, nous explorons des nouvelles approches d'identification des utilisateurs primordiaux en temps réel. Ces approches sont centrées sur trois principaux aspects. Nous avons tout d'abord étudié l'efficacité de différentes catégories de mesures issues de

la littérature et proposées dans cette thèse. Ces mesures décrivent principalement le comportement des utilisateurs des microblogs au fil du temps. En nous basant sur les mesures pertinentes résultant de cette étude, nous concevons des nouvelles caractéristiques permettant de mettre en évidence la qualité des informations partagées par les utilisateurs selon leurs comportements. Le deuxième aspect consiste à proposer une approche de modélisation du comportement de chaque utilisateur se basant sur les critères suivants (1) la modélisation des utilisateurs selon l'évolution de l'évènement, (2) la modélisation de l'évolution des activités des utilisateurs au fil du temps à travers une représentation sensible au temps, et (3) la sélection des caractéristiques les plus discriminantes à chaque phase de l'évènement. En nous basant sur cette approche de modélisation, nous entraînons différents modèles de prédiction en utilisant les collections de données recueillies durant des évènements antérieurs. Ces modèles apprennent à différencier les comportements des utilisateurs importants de ceux qui ne le sont pas durant les situations de crise. Les algorithmes d'apprentissage supervisés SVM et MOG-HMMs ont été utilisés durant la phase d'apprentissage. Apprendre les différents modèles en se basant sur les données des événements antérieurs assure l'exécution du modèle de prédiction en temps réel durant les évènements à venir.

Pour évaluer la performance de nos modèles, les deux collections de données ont été utilisées pour les phases d'apprentissage et de test. Les différents tests de ces modèles ont prouvé l'efficacité de notre approche de modélisation utilisateur intégrant les dimensions suivantes : (1) la dimension thématique représentant les utilisateurs par rapport à leurs positions vis-à-vis de la thématique liée à l'évènement d'une part et vis-à-vis de toute autre thématique d'autre part, (2) la dimension temporelle qui est représentée par la modélisation de l'évolution du comportement des utilisateurs au fil du temps, et (3) la dimension événementielle qui met en évidence l'évolution du comportement et de l'importance de l'utilisateur à chaque phase de l'évènement. En nous basant sur cette modélisation, les modèles MoG-HMMs ont réussi à distinguer les particularités des utilisateurs primordiaux par rapport à ceux qui ne le sont pas et vice versa. Les modèles de prédictions résultants ont été plus performants que les modèles de prédiction présentés dans la littérature en termes de classification, prédiction et classement. La plupart des utilisateurs primordiaux a été prédit à un stade avancé de chaque phase de l'évènement.

Globalement, ces contributions peuvent être considérées comme des étapes importantes incitant à explorer d'avantage les approches centrées utilisateurs pour la recherche d'information en temps réel. Nous espérons que ces contributions peuvent assurer une meilleure connaissance de la situation pour les équipes d'urgence durant les situations de crise.

List of Publications

Journal papers

1. **I. Bizid**, N. Nayef, P. Boursier, S. Faiz, and A. Doucet, Modeling the Behavior of Microblog Users for Prominent Users Detection over Crisis Events Phases. TOIS, 2016 [under review].

Books parts

1. **I. Bizid**, N. Nayef, S. Faiz and P. Boursier, Microblog Information Retrieval for Disaster Management: Identification of Prominent Microblog Users in a Disaster Context. *Chapter in Handbook of Research on Geographic Information Systems Applications and Advancements*, 2016 [to be published].

International Conference papers

1. **I. Bizid**, N. Nayef, P. Boursier, S. Faiz, and J. Morcos, Real-time Detection of Prominent Users during Specific Events by learning On- and Off-topic Features of User Activities. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, ACM, pages 500-503, Paris, France, 2015.

2. **I. Bizid**, N. Nayef, P. Boursier, S. Faiz, and A. Doucet, Identification of Microblogs Prominent Users during Events by Learning Temporal Sequences of Features. *In Proceedings of the* 24^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, ACM, pages 1715-1718, Melbourne, Australia, 2015.

3. **I. Bizid**, N. Nayef, O. Naoui, P. Boursier and S. Faiz, A Comparative Study of Microblogs Features Effectiveness for the Identification of Prominent Microblog Users During Unexpected Disasters. *In proceeding of the* 2^{nd} *International Conference of Information Systems for Crisis Management in Mediterranean Countries (ISCRAM-med)*, Springer, pages 15-26, Tunis, Tunisia, 2015.

4. **I. Bizid**, N. Nayef, P. Boursier, S. Faiz, and J. MORCOS, MASIR: A Multi-agent System for Real-Time Information Retrieval from Microblogs During Unexpected

Events. In Proceedings of the 9^{th} KES International Conference on Agents and Multi-Agent Systems: Technologies and Applications (KES-AMSTA), Springer, pages 3-13, Sorrento, Italy, 2015.

5. **I. Bizid**, N. Nayef, P. Boursier, S. Faiz, and J. MORCOS, A Classification Model for the Identification of Prominent Microblogs Users during a Disaster. In Proceedings of the 11^{th} International Conference on Information Systems for Crisis Response and Management (ISCRAM), Kristiansand, Norway, 2015.

International Workshop papers

1. **I. Bizid**, S. Faiz, P. Boursier and F. Rousseaux, The Role of Spatial, Temporal and Sociological Features in the Distribution of Tweets during a Disaster case. *In Proceedings of International Workshop on Artificial Intelligence Technologies for Spatial Risk Prediction* (AITSRP), pages 221-227, Le Caire, Egypte, 2013.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The surge of crisis events which might threaten us at any moment, has become a major population concern in many regions of our planet. These crisis events strike under various new dangerous characteristics which have been rarely observed in previous decades. The complexity of such events is increasing through years. Nowadays natural disasters, diseases or even human-made disasters like terrorist attacks differ substantially from the standard disasters, attacks or diseases that we already have known before.

With the evolution and the change of such crisis events characteristics, the already collected data are not anymore sufficient to deal with these unanticipated patterns. Thus, researchers from various domains seek to collect and access relevant data in order to analyze and decrypt the different hidden aspects behind these events. Optimal hardware or/and intelligent software are not anymore the main preoccupation of science. Data is now seen as the main key that can decrypt the nature and human secrets that surrounds us. Keeping track of valuable data which make sense is currently one of the big challenges which face man-kind.

The rapid development of interactive and collaborative communication platforms, especially microblogging platforms, has revolutionized data collection strategies, especially during crisis events. These platforms offer a direct access to rich information which was not previously accessible through traditional communication technologies. The effectiveness and ease-of-use of supported microblogging platforms – especially Twitter – have marked a significant change on the communication habits in our society. Microblog users unconsciously play the role of voluntary sensors by providing situational information in real time. Rich with such information, microblogs have become indispensable within everyday life and have been significantly implicated in several domains, particularly for crisis events management.

Any user can quickly and conveniently post and get information with the latest news. These platforms are accessible through websites or cellphone applications allowing users to instantly post relevant information about what they are seeing, hearing and experiencing around them. In a disaster case, such platforms provide valuable information shared voluntary to inform or alert a wide range of connected people about what is really happening on the threatened or affected areas. The need for information hunters and gatherers to go on the event area, risking their lives, diminishes greatly. Exploring such platforms during

crisis events is nowadays indispensable to get fresh information out from witnesses' users in a quick and efficient manner.

During major crisis events such as Boston Attack[1] and Colorado Floods[2], it has been observed that exclusive information are generally shared in microblogging platforms before their official announcement in media channels outlets or by disaster management organizations. During the Boston Attack, microbloggers have shown great independent efforts to identify the bombers, even before the FBI had singled out any images of potential suspects, by posting videos, images taken before or after the attack. Microbloggers were conducting their own "investigation" in parallel with law enforcement efforts.

An early perception of such voluntary shared situational information is nowadays inevitable to save as many lives as possible and speed up the ongoing investigations and the intervention plans (Deng & Jaitly 2014). However, retrieving relevant and exclusive information from the huge amount of shared data in these microblogs remains complex. Manually looking at this shared information through microblogging platforms interfaces and judging the relevance of their content cannot scale to handle the size of these extremely active networks. Data shared in these microblogs falls into the category of big data. The main challenges to effectively explore this data lies in finding retrieving techniques coping with the 4 V components characterizing microblogs data (i.e. volume, velocity, variety, and veracity).

The success and richness of these platforms, for example Twitter, is behind its 200 million active users producing more than 700 million tweets daily[3]. Extracting event-related tweets from the huge amount of streaming data shared in the same time cannot be easily processed. Twitter Application Programming Interfaces (APIs) provide a limited access for both streaming and historic data. Moreover, not all event-related tweets are necessarily valuable. The majority of these tweets are generally non-valuable, redundant, outdated or incredible. The same tweet content can be expressed differently using different formats (i.e. texts, images, links, and videos) and/or different languages. Thirty four languages are supported by Twitter while Facebook supports eighty three languages.

Analyzing each event-related information content for relevant information retrieval in real time is challenging. Traditional information retrieval techniques applied within the field of crisis events have mainly relied on mining information containing text. However, such techniques cannot be efficient enough, they systematically neglect any information that does not include any text. The additional difficulties in valuable information retrieval is information veracity checking. Disinformation floods the microblogging platforms during crisis events. In such situation, it is not rational to refer to official organization sites

[1] http://www.govtech.com/public-safety/Social-Media-Big-Lessons-from-the-Boston-Marathon-Bombing.html

[2] https://storify.com/nbcnews/social-media-covers-the-flash-flooding-in-colorado

[3] http://www.internetlivestats.com

for checking the veracity of microblogging event-related information content. The main targeted pieces of information in such situations are the exclusive ones that have not yet been confirmed in official media.

Indeed, the primary difficulty for real-time information retrieval from microblogs is to extract and sift out the exclusive situational information from the tens of thousands pieces of information shared in microblogs. This problem has been generally addressed by using standard retrieval techniques appropriated for each content type such as text mining, image and video analysis (MacEachren et al. 2011, Starbird & Stamberger 2010). However, such techniques are generally computationally expensive and do not take into account the characteristics of both the event and information providers (Pal & Counts 2011).

Associating the quality and the relevance of event-related tweets with their authors' prominence regarding the analyzed crisis event could be interesting in these situations. The strategy of identifying and tracking relevant information providers has been widely explored in the field of microblog information retrieval. Different categories of information providers have been targeted in a general information retrieval context such as domain experts, topical authorities and influencers. However, none of these defined microblog users categories and their identification methods are appropriated to the targeted users category in the context of crisis events.

In this dissertation, we aim to explore new methods to identify key microblog users in real time during crisis events. We define key microblog users in this thesis context as *prominent microblog users* who are susceptible to share relevant and exclusive information during the analyzed crisis event regardless of their popularity and their domain of expertise in the platform. We choose Twitter as an example of microblogging platforms as it is the main platform sought during these events. We focus at first on exploring new ways to cope with data collection and users tracking restrictions imposed by Twitter APIs. We have been then interested in proposing an efficient microblog user modeling approach that well reflects the realistic behavior of microblog users during crisis events. Based on this modeling approach, we aim to build a prediction model highlighting prominent users against the non-prominent ones. These highlighted prominent users need to be tracked in real time to access the required exclusive information. Our goal in exploring this research is to help crisis events management authorities to have a real time access to exclusive and relevant information describing what is happening on the ground during such unpredictable events.

1.1 Research Problem Statement

The issue of key microblog users identification has been raised with the emergence of microblogging platforms. This identification problem is generally cast into a microblog users

ranking problem where users sharing the required information in a specific context need to be ranked at the top. In the following, we briefly describe how this issue has been solved for the identification of different categories of key users in a general context on one hand and in a crisis events context on the other hand.

In a general context, targeted microblog users categories in the literature mainly refer to influencers, domain experts or/and topical authorities. Such standard targeted users do not typically provide the required information during crisis events. As detailed in the following, prominent microblog users in the context of crisis events have their own specificities that can be neither covered by these standard categories nor identified using standard key users identification and modeling techniques:

- *Influencers*, such as CNN and T.V shows stars, cannot be systematically categorized as prominent users even if they are extremely active regarding the analyzed crisis event. These users typically report outdated information that have been already diffused in the microblogging platform. The used ranking techniques employed to identify influencers cannot result in a high accuracy for prominent users' identification. The evaluated users are represented through a social graph describing their followorship connections. PageRank and HITS algorithms are used for the users ranking process (Cappelletti & Sastry 2012, Romero et al. 2011). Such ranking strategy is sensitive to well-connected users who are generally evaluated as non-prominent in crisis events context.

- *Topical authorities* and *domain experts*, such as government organizations, may be evaluated as prominent users as presented in this thesis context. However, such user categories does not cover ordinary users who are neither domain experts nor topical authorities and who may provide their testimony from the ground. The few proposed domain experts identification systems presented in the literature rely on real-time user behavior ranking algorithms (Pal & Counts 2011, Xianlei et al. 2014). These algorithms rank users according to their behavior represented by a feature vector composed of a set of textual, microblogging and/or social network structure features. Such user modeling approach can neither realistically nor accurately represent the evolution of user behavior over time during an event. Based on this modeling strategy users would be evaluated according the quantity of their produced information independently of its quality. This yields weaker performance of detection and ranking algorithms which learn to distinguish behavioral differences among different users.

In the context of crisis events, targeted key microblog users were defined in the literature either as witnesses who are geo-located in the crisis event area or as central users in specific communities (Gupta et al. 2012, Starbird et al. 2011). While these targeted users categories

are prominent during crisis events, they did not essentially cover all the prominent users targeted in this work. As detailed in the following, these targeted users and their adopted identification techniques are not well suited to detect most of prominent users in the context of crisis events:

- *Witnesses* refer to on-the-ground Twitterers reporting what is happening around them. To differentiate between microblog users who are on the ground and those who are not, a variety of user activities features were explored in the literature (Gupta et al. 2012). Such features represent the evaluated users according to their interest in the event and their current location. However, the geolocation information on which these approaches are based are rarely provided. Around 98% of tweets shared during crisis events are not attached to any geolocation (Imran et al. 2015). Thus, referring mainly to geolocation information to identify prominent users is not sufficient.
- *Specific communities influencers* such as journalists, official organizations, are typically prominent during crisis events. However, ordinary prominent users are not covered in these categories of users. Identifying specific communities influencers in the context of crisis events is mainly processed using similar algorithms as those explored for influencers detection in a general context. Such algorithms are strongly criticized due to their sensitivity to popular users reporting what have been already shared in the network (Hemant et al. 2014).

While some of the already targeted microblog users in the literature can be defined as prominent ones, the identification approaches defined for their detection are not suitable for the identification of prominent users in crisis events context. Such inadaptability is mainly due to the modeling approaches selected for the evaluated users representation. Many components highlighting the differences between the targeted users and those that have to be rejected have been neglected in the literature. Existing user characterization approaches would diminish the identification algorithm performance even if their ranking strategy is efficient. This is due to the following problems:

- On-topical characterization of users: Practically, such characterization represents users only according to the quantity of their activities related to the targeted topic independently of the other topics. Such strategy extremely promotes active users, such as news outlets, toggling between several topics and sharing several outdated or irrelevant information in the microblog and penalizes those sharing few pieces of information but extremely relevant and exclusive.
- Quantitative characterization of users: Users sharing the same quantity of information are typically represented similarly using state-of-the-art user representation models.

Such user characterization does not efficiently reflect user behavior evolution over time. Users sharing various information at an early stage of the event are represented similarly as those sharing the same information at its end. The temporal distribution of user activities is neglected which do not help to distinguish users sharing exclusive information from those sharing outdated ones.

- 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. Assume the case of a flooding disaster, the behavior of users during the orange alert phase (i.e. prevention phase) would not be the same like during the red alert phase (i.e. response phase) or once the red alert was disabled (i.e. recovery phase). Users may act differently according to the event phase. It is thus not rational to characterize users uniformly during the whole event period.
- Overall user prominence evaluation over the whole 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. There are some particular microblog users who tend to be prominent only in the last phase. Thus, they have not to be penalized regarding their absence during the first phases.

Moreover, there is no adapted Twitter data collection techniques that can provide the needed data for understanding user behavior during crisis events. Twitter APIs have many restrictions limiting the access to its users data. Tracking and collecting a wide range of Twitter users' information is typically afforded to a small number of public and private institutions. Researches conducted on the key users identification fields usually uses some samples of data covering a small set of user behavior criteria. There is no available techniques or collections that can be adapted to any key users identification goals and methods.

This work aimed to alleviate the mentioned shortcomings by proposing a set of methods within a system detecting prominent microblog users in real time during crisis events. Two main problems are being addressed:

1. How to gain real time access to relevant information shared in Twitter?
   (i) Which are the crawling limits and loopholes of microblogs for information extraction and user tracking in real time?
   (ii) How the loopholes of microblogs APIs can be exploited for boosting the number of tracked users and insure real time user behavior analysis?
2. How to highlight the different particularities of microblog users' behavior during crisis events to identify prominent users at an early stage of an event?

(i) What are the features which could efficiently characterize microblog users' behavior during crisis events and could be computed in real time?

(ii) How to integrate the time factor for microblog user modeling to highlight the change in users behavior over time?

(iii) How to consider crisis events specificities while modeling and identifying microblog users over time?

(iv) How can we learn to distinguish prominent users' behavior from the non-prominent ones in real time?

1.2 Contributions and Significance of this Dissertation

Based on the discussion presented in the previous section, there is a need to develop more accurate, more efficient and more robust solutions for prominent users identification in the context of crisis events. Various aspects need to be considered for developing these solutions. These aspects refer to the identification model feasibility in real time, the user modeling approach adaptability to the context of crisis event and the accessibility to prominent users information in real time. We summarize the contributions of this dissertation in the following, whereas the detailed contributions along with the experiments and evaluations necessary to prove them are discussed in the rest of the chapters.

1. *A Multi-Agent System for Users Information Extraction and Tracking:* We process users information extraction and tracking through a new proposed multi-agent system named MASIR (Bizid et al. 2015a). This system copes with the limits imposed by Twitter APIs. It explores loopholes of these APIs in order to be able to collect most of the required information for key users identification on one hand and to provide a real-time access to key microblog users profiles on the second hand. This system collects in a first step historic data characterizing users interested in the event and then analyzes this data in order to identify and track key users. This system sits on a parallel processing multi-agent architecture boosting the number of tracked user and crawled profiles. In this architecture, a three-layered structure is proposed to accommodate all the agents. Multiple tracker agents are proposed to manage information extracted across different hosts and different agents connections. Three categories of agents are proposed with different management and extraction roles.

2. *Studying Raw User Features Effectiveness in the Context of Crisis Events:* We extract different high level features categories characterizing user's activity during the analyzed event. Most of the extracted features are derived from the state-of-the-art. New topical and geographical features that have not been explored in the literature

are also proposed. Our rational behind the new proposed topical raw features is to characterize each active user by considering both his on- and off-topic activities. Based on those features, we would be able to differentiate between users focusing only on the event under consideration and those toggling among several topics. Geographical features are explored in order to highlight users who are or have been geo-located in the area of the crisis event. We conduct a comparison study in order to select the most relevant raw features for users characterization in the context of crisis events (Bizid et al. 2015f). We use SVM and ANN learning-based-approaches for the selection process.

3. *Microblog User's Behavior Modeling in the Context of Crisis Events:* In order to reflect the real users' behavior by taking into account both the crisis event evolution and the user's prominence change over time. We propose three complementary user behavior modeling approaches:

   (i) *A Qualification of the Quantified Raw Features Approach.* In order to point out the quality of the different raw features characterizing microblog users, we propose a new set of engineered features exploiting the different combinations between the already selected efficient raw features (i.e. topical and geographical features) (Bizid et al. 2015e). The novelty of these features lies in representing the topical dimension of user activities. User's on-topic raw features describing users' activities of same nature are combined and adjusted by the corresponding off-topic ones. The conducted experiments confirmed the importance of these proposed features in order to highlight the real prominent users over the non-prominent ones.

   (ii) *A Temporal Sequence Representation Approach.* We propose to integrate the user activities temporal dimension while modeling the user behavior (Bizid et al. 2015c). This dimension would point out the evolution of the user behavior over time. This highlights the temporal behavior of prominent users regarding the other non-prominent ones. The temporal dimension is integrated by representing the evaluated users as a temporal sequence of feature vectors characterizing their behavior during the analyzed crisis event. The obtained experimental results using this user modeling approach confirmed the importance of characterizing user behavior evolution over time.

   (iii) *A Temporal Phase-aware User Modeling Approach.* We consider: -Event evolution over time, and -User behavioral change over event phases and over time while modeling user behavior during crisis events (Bizid et al. 2016d). We assume that as event characteristics and level of importance change according to each event phase, the user interest and behavior regarding a particular event

would be different from one phase to another. We also assume that characterizing users behavior at each new phase independently of the prior analyzed phases insures a fair evaluation of the different microblog users over time. These hypotheses were validated by our different conducted experiments. This modeling approach integrating the topical, temporal and phase-aware dimensions has yield to promising identification results.

4. *Real-time Prominent Users Prediction:* We propose different phase-aware prediction models learning to differentiate between prominent and non-prominent users behavior. The learning task has been performed using ergodic Mixture of Gaussians Hidden Markov Models (MoG-HMMs). These models are learned *a priori* using prior events data and processed in real time for prominent users identification during new similar crisis events. Learning the different models based on prior events data makes the prediction process computationally feasible in real time during new real-world crisis events cases. Based on the temporal phase-aware user modeling approach, these models have proved experimentally their efficiency and efficacy to point out the particularities of prominent and non-prominent users behavior during crisis events.

1.3 Thesis Organization

This thesis is organized into a number of chapters, each of which pursues a distinct research goal. Each of these goals strengthens our understanding of prominent users behavior during crisis events and enables us to build mechanisms to effectively characterize and identify the targeted microblog users category in real time.

In chapter 1, we present the important role of social media in crisis management, which is the motivation behind this work. Research questions and main contributions are also presented in this chapter.

In chapter 2, we present an overview of related work that focuses on issues related to those highlighted in this dissertation. This literature review chapter is organized in three main parts. First, different approaches for data acquisition and extraction from microblogs are discussed. Second, we review the existing information retrieval techniques in microblogs during crisis events. Finally, we discuss in the third part of this chapter the different existing key microblog users identification techniques in a general context and their adaptability to the context of crisis events.

In chapter 3, we present a modular Multi-Agent System for Information extraction and Retrieval (MASIR). First, we describe the MASIR extraction module designed for boosting historic Twitter data access. Second, we present the MASIR tracking module for real-time

identification and tracking of key microblog users. Finally, we evaluate the performance of these modules in real-world cases.

In chapter 4, we conduct a comparative study evaluating the effectiveness of different raw features that can characterize microblog users. First, we describe the role of these features and how they can be extracted from user's profile and timeline. Second, we list most of the existing raw features presented in prior work and a few new proposed features. Finally, we conduct various experiments to select the most appropriated features categories for microblog users characterization in the context of crisis events.

In chapter 5, we design new efficient engineered features derived from the already selected raw ones in chapter 4. These features mainly focus on highlighting on- and off-topical user behavior by penalizing users toggling between several topics. The performance of these features is experimented and compared with state-of-the-art raw and engineered features.

In chapter 6, we propose a prominent users prediction model evaluating microblog users according to the temporal distribution of their topical activities over time. We thus enrich the user modeling approach presented in chapter 5 by modeling users behavior using a temporal sequence of feature vectors. The user behavior classification and ranking is processed using ergodic MoG-HMM probabilistic models. The performance of this time-sensitive user modeling approach is evaluated and compared with standard state-of-the-art modeling approaches and the prior proposed approaches.

In chapter 7, we present a user characterization and identification approach considering the analyzed events evolution over time and its impact on users behavior. First, an event phase-aware user characterization approach is described. Subsequently, a phase-aware prominent users prediction model is proposed to identify the targeted users in real time. This model is compared with the prior presented phase-unaware models in both this dissertation and the state-of-the-art.

In chapter 8, we conclude this dissertation by discussing our findings and then outlining some possible directions for future work.

Chapter 2

Background and Relevant Literature

Contents

2.1 Introduction

Sharing and accessing content on the web is nowadays easily accessible to everyone within few seconds. According to internet live stats website[1], 90% of information shared daily on the web is essentially provided from microblogging platforms. Twitter microblogging platform[2] is ranked in the second position with around 700 million of daily shared tweets. Microblogs data includes various updates regarding different topics and events. Such updates are generally not provided by search engines websites as they are usually not yet indexed and thus available only through the microblogging platform search tools. Moreover, exclusive information shared in microblogs, specially during crisis events, are extensively spread in these platforms before their official announcement in news outlet channels. For example, before the intervention of emergency responders, only users geo-located in the threatened or affected disaster area would share various valuable information describing what is really happening in real time.

Accessing and analyzing this voluntarily shared data in microblogs is nowadays indispensable to have the last news regarding a specific topic or event. To access the required information, many organizations opted to employ certified persons to continuously follow the information shared about a given topic in real time using microblogs web interface. However, this practice is tedious and infeasible in real time, especially when there is a surge of updates shared in a short period of time. This data has to be accessed and analyzed automatically in order to gain a real time access to the relevant and exclusive information. Data access is limited by microblogging platform. Gaining an unlimited access is fairly costly. These data access limits have always been a constraint for researchers aiming to evaluate and learn new information retrieval models appropriated to specific research problems.

Information retrieval within microblogs also differs from regular information retrieval from the web. Microblogs data content has its specific format, syntax and motivation. While queries submitted in web search are generally performed for informational, transactional or navigational purposes, search queries within microblogs are mostly performed for informational purpose. Various specific factors are considered while executing such queries. These factors refer to targeted user's activity, the analyzed event or topic specificities, the freshness of user's shared information, the user's position regarding microblogs communities interested in the analyzed topic and many other factors. Relevant and exclusive information retrieval from microblogs can be processed using various techniques. The efficiency of these techniques depends on the targeted information, the analyzed event or topic and whether the retrieved results have to be provided in real time or not.

[1] http://www.internetlivestats.com/
[2] http://www.Twitter.com/

In this chapter, we study the existing microblog information retrieval approaches for valuable information extraction during crisis events. This study is organized in three main sections. In Section 2.2, the different existing techniques for data acquisition from microblogs are detailed and discussed. We then detail information retrieval techniques from microblogs within crisis events context in Section 2.3. Both content-based and user-based information retrieval approaches are discussed in the context of crisis events. In Section 2.4, we describe the different proposed user-based information retrieval approaches in a general context and we discuss the adaptability of these techniques to the context of crisis events.

2.2 Microblogs Data Acquisition and Extraction

Various research works have been proposed for data acquisition and extraction from microblogging platforms. In this literature review, we focus on analyzing extraction and acquisition techniques adapted to the microblogging platform Twitter. This platform is one of the most popular microblog platforms affording a large set of rich information shared publicly regarding various topics. The wealth of information shared in Twitter is attracting an increasing attention of researchers in many fields especially knowledge discovery and data mining. The different information shared in Twitter represent a new gold mine in the new social science. Exploring such information qualitatively and quantitatively could lead to understand and propose new powerful models learning the particularities of human behavior and interests.

In order to be able to learn or/and test new models and new Twitter data analysis methods, researchers need to access this real world microblogs data. The accessed data has to be appropriated to both their proposed research approaches and their goals. The design of any model highly depends on the type of information that have to be analyzed. Targeting and analyzing all types of information shared in these platforms is complex. Research models generally require as input particular data composed of a subset of information relevant to specific queries. The format and nature of these inputs differ according to the analysis approaches integrated in each model. For example, to identify microblog influencers, most of researchers acquire only microblogs social graph data in order to be able to analyze users' relationships graph (Smailovic et al. 2014, Romero et al. 2011). For this same influential retrieval problem, other authors have referred to topical-related tweets data (i.e. Twitter timeline data) to identify influencers according to their topical activity (Pal & Counts 2011, Weng et al. 2010). Thus, microblogs data needs differ according to research goals and the targeted model specificities. It is hard to find a single collection of data that can fit to all the research fields' needs.

Collecting the required research data from microblogging platforms is a challenging task. Such data is exceedingly protected by Twitter as it is considered as the main financial resource of the company. Even acquiring a privileged data access for research institutes is not any more easily accessible. In order to provide a direct access to the needed research data for some research communities, both researchers and organizations have sought to find a compromise solution respecting Twitter policies (Abdulrahman et al. 2011, McCreadie et al. 2012). However, the proposed direct access methods did not cover all the researchers needs (Chau et al. 2007). To deal with such problem, many researchers have implemented their own platforms integrating new extraction techniques dealing with the Twitter interfaces restrictions.

In the following sub-sections, we detail : (1) Twitter particularities and its different provided data which interests researchers, (2) the main existing direct data acquisition techniques, (3) the different advanced data collection techniques that were proposed in the literature and (4) a comparative study summarizing the advantages and drawbacks of each data acquisition technique followed by a discussion highlighting the remaining challenges.

2.2.1 Twitter Microblogging Platform

Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass and launched in July 2006. This service enables users worldwide to publish messages, known as *tweets*, expressed with no more than 140 characters via SMS or web or/and mobile applications. Nowadays, Twitter has gained a huge popularity and is used in our daily life to comment on any news and discuss trending topics. It has integrated richer characteristics by enabling users to publish various data content formats such as texts, images, links and videos. Recently, Twitter has announced an upcoming set of changes that will be available in the next months. The main announced change consists of enabling Twitter users, known as *Twitterrers*, to express tweets with more than 140 characters. This change will ensure the share of richer contents without scarifying characters to meet Twitter text restrictions. In the following, we briefly describe the specificities of information shared in this platform and the different data natures that could be extracted for research.

2.2.1.1 Twitter Specificities

Twitter is generally perceived as a social network composed of a huge number of users connected via “following” and “followees” links. These connections allow users to have a direct access to the publications of their followed Twitterers. Each user followees and followers lists are accessible from the user profile as illustrated in box 1 of Figure 2.1. User tweets are by default shared publicly without any restrictions. The origin of these tweets

can be recognized by the user identifier following this notation @username. Users can insert their own biography describing their domain of interest, location or any other information they judge relevant to describe them. They can be also subscribed to some topical lists by other users who judged that their profile is adequate to the interests described through the specific list. As shown in box 1, @ImenBizid was added in 5 lists.

As mentioned in the timeline of our stated example in Figure 2.1, different nature of tweets can be published in a user profile. Tweets included in box 3 refer to *original tweets* shared by @ImenBizid. Those in box 2 and 4, known as *retweets*, refer to tweets shared originally by someone else and the profile owner has chosen to share them with his/her own followers. The retweet in box 2 is originally a *mention tweet* from @dataiku to @hugolsqm. The total number of an original message retweets and likes are always mentioned below the original tweet such is the case in box 2. Dataiku tweet was retweeted 16 times and liked 9 times. Such metadata is largely used in the literature to evaluate the relevance of tweets. In addition to retweets and likes, users can interact regarding a specific tweet by commenting on it using specific tweets known as *replies*. Such replies are detectable via the user identifier @username mentioned in the beginning of the tweet.

2.2.1.2 Data of Interest

The Twitter data targeted by researchers differs according to the defined research problem. This data is generally divided into the following five categories :

1. *Profile Data* refers to personal data composed of two types of user's information :
   - Information uploaded manually by each profile owner such as full name, pseudo-name, photo, detailed or/and brief biography and country.
   - Information generated automatically by the microblogging platform which mainly summarizes the statistics of the user activity since the creation of his/her Twitter account. Such information includes the total number of tweets shared by each user, the number of favorite posts, the number of the user followers and followees and the number of lists in which the user is subscribed as described in box 1 of Figure 2.1.
2. *User Generated Content Data* refers to any content added by the user such as tweets, retweets, replies and mentions. The metadata joined to these contents is also recorded to give more details about the shared information. Such metadata are either automatically generated by Twitter or set manually by the user. The metadata automatically shared by Twitter denotes the time and date of the content publication, the number of retweets of this content, the tweet's language, the tweet time zone, the number

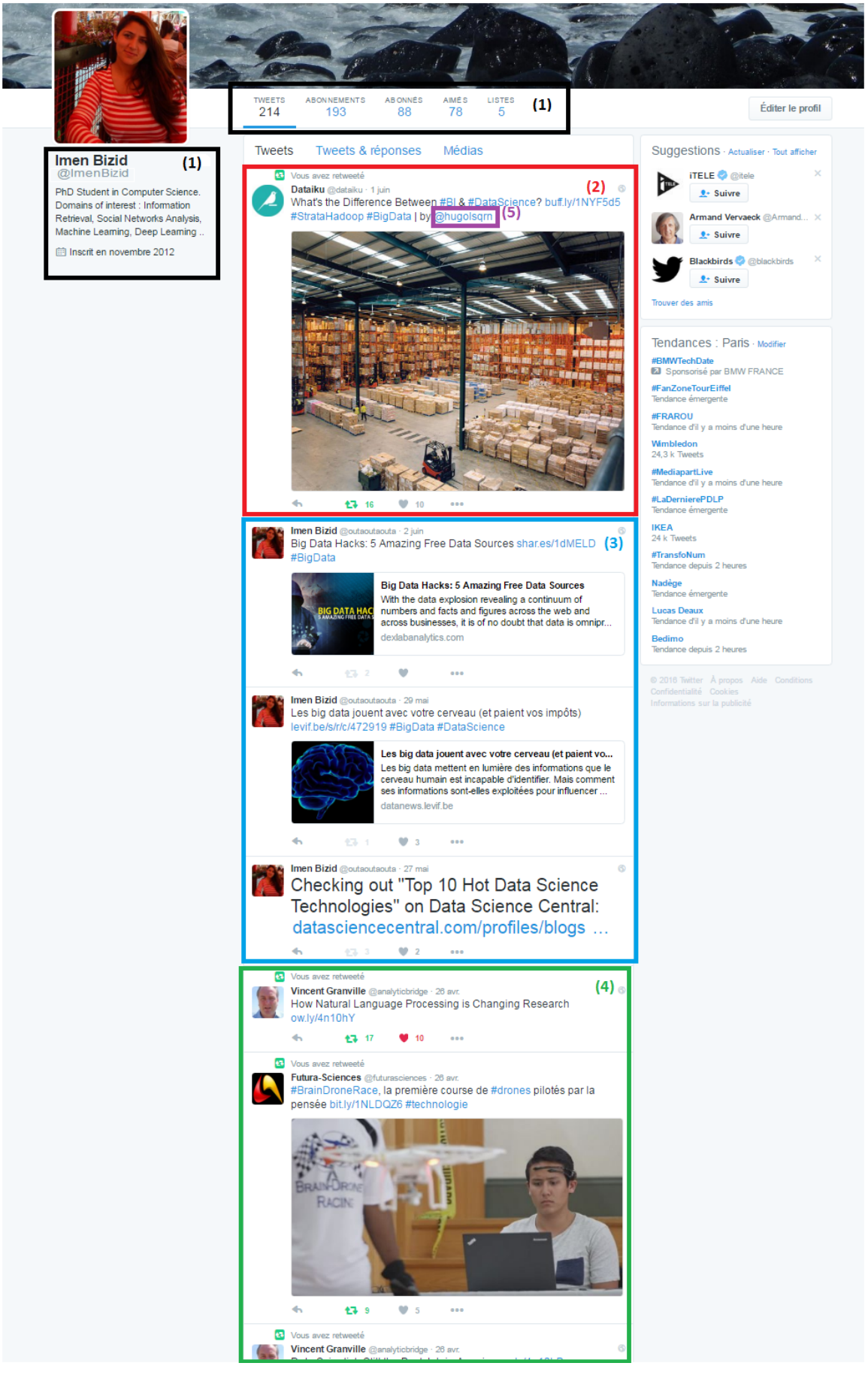

FIGURE 2.1: An example of a Twitter user profile content

of collected favorites and the user geolocation at that time if the user has enabled this option. The additional metadata that can be manually joined to the user shared content is the user geolocation if this option is disabled by default.

3. *Social Graph Data* refers to the different relationships among Twitter users where users are considered as nodes and the friendship relations as edges. Such information are extracted from each user's list of followers and followees available on their Twitter profile. There is a wide variety of interesting research work exploring such social graph data : identifying popular users and influencers (Romero et al. 2011), predicting future social links (Liben-Nowell & Kleinberg 2003), detecting communities of similar users (Purohit et al. 2014) and many other applications.

4. *Interaction Graph Data* refers to any nature of interaction relating Twitter users. Such interactions refer to a retweeting or/and mentioning or/and replying activities. The retweeting graph is generally extracted to record Twitter users' tweets diffusion in the network where users are denoted as nodes and the retweeting activity of tweets as edges. Other mentioning graph data is explored to report the mentioning activity between different users where in this case edges refer to the user's mentions. This graph data is required in various research fields such as rumors (Seo et al. 2012) or spams propagation analysis (Wang 2010), information sources and popular microblog users' identification (Weng et al. 2010), etc.

5. *Twitter Timeline Data* refers to any information, matching certain search criteria, shared in Twitter. Such search criteria refer generally to :

   - a list of keywords or/and hashtags included in the tweets' text,
   - two time boundaries delimiting the tweets shared in a specific period of time,
   - a specific gelocation area delimiting the geo-located tweets shared from this area.

   Such timeline data is generally required for tweets sentiment analysis (Beigi et al. 2016), witnessers detection during disasters (Morstatter et al. 2014), Twitter user's behavior analysis (Pal & Counts 2011), trending topics and sub-events detection on Twitter (Pohl et al. 2012),etc.

2.2.2 Direct Data Access

Microblogging platforms are making information extraction more difficult by using access-control mechanisms and limiting the number of accessed information through their platforms. Given these limitations, many research organizations sought to offer a direct access to Twitter data in order to encourage the different communities to explore these new platforms (McCreadie et al. 2012). Directly accessed data is generally provided either for free

or charged by public or/and private organizations. In this subsection, we split the direct data access techniques into the following four broad categories : *data access via research data collections*, *via data resellers*, *via US congress library* and *via data grants*.

2.2.2.1 Data Access via Research Data Collections

Since the conducted experiments by Cleverdom in 1997 presenting the efficiency of the Cranfield paradigm (Cleverdon 1997) for information retrieval systems evaluation, test collections have become an unavoidable element in the information retrieval research field. Such collections are shared in open access for several goals :

- to encourage researchers in the information retrieval community to gain an open access to large test collections.
- to easily compare the efficiency of the different proposed information retrieval systems using the same provided test collections and recommended evaluation metrics.
- to increase the communication among industry, academia and government through providing test collections adapted to the main common challenges which interest these different organizations.
- to easily connect theoretical aspects with practical examples and speed the transfer of innovative systems from research labs into commercial products.

Test collections are generally composed of a set of documents related to one or various topics suited to specific information needs. These documents are generally labeled through a conducted relevance judgment study classifying a document as either relevant or irrelevant to a specific topical query. In following, we list most of the main standard test collections that have been publicly shared by information retrieval community through organized competitions and workshops. We focus particularly on test collections for information retrieval from microblogs.

TREC Text Retrieval Conference (TREC) collections (McCreadie et al. 2012). TREC[3] is co-sponsored by the U.S. National Institute of Standards and Technology (NIST) and U.S. Department of Defense. Since 1992, it has provided a large test collections. The best known test collection adapted to the problematic of information retrieval from microblogs is the dataset known as Tweets2011 corpus distributed in TREC 2011 microblog track. This corpus contains of approximately 1% of tweets (after spam removal) posted from January 23rd to February 8th 2011 in Twitter. The resulting Tweets2011 corpus is composed of 16 million tweets. The chosen period of time covers two of the major 2011 events including

[3] http://trec.nist.gov/

the Arab spring revolutions in Egypt and the Super Bowl in United States. The tweets collection was distributed as a set of tweet identifiers and tweet crawling tool for downloading the different identified tweets. This crawling system was developed through collaboration between Twitter and TREC in order to legally collect and distribute Tweets2011 Corpus.

CrisisLex collections. Crisislex[4] is a platform sharing various crisis-related social media collections and tools. This platform was initially created by Olteanu et al. (2014) in order to share lexicons of disaster-related terms. The different provided data in this platform were collected using two data acquisition techniques, *keyword-based search* and *location-based search*, using Twitter streaming APIs. Most of the location-based samples were obtained through external data providers mainly GNIP[5] and Topsy[6]. A total of 7 crisis collections are distributed in this platform. *ChileEarthquakeT1*, *CrisisLexT6* and *SoSItalyT4* collections are suited to evaluate disaster-related information retrieval systems as tweets are labeled according to their relevance. *BlackLivesMatterUT1* collection offers a variety range of new research possibilities as it covers the specificities of users interested in such blacks problem.

Stanford Large Network Project (SNAP) Collections(Leskovec & Krevl 2014). The SNAP[7] library was developed since 2004 in the Stanford university. More than 50 large network datasets from tens of thousands of nodes and edges to tens of millions of nodes and edges were published in open access. These collections offer a wide variety of network data having different natures and purposes such as social networks, web graphs, road networks, product co-purchasing networks, citation networks, location-based online social networks, and communication networks. These various open collections, especially microblogs social graph ones, can be explored in the field of graph-based-information retrieval from microblogs.

Kaggle Collections. Kaggle[8] is a platform organizing various predictive modeling and analytic competitions. This platform contains various collections made publicly available by competitions hosters, companies, researchers and staticians. These collections cover different research domains appropriated to various research problems. The most recent Twitter collection shared in this platform comes from the Crowdflower[9] library. This collection targets the US Airline travelers feelings analysis problem.

Other considerable collections were published in academic research institutions websites such as the disaster-related tweets collections shared by the Indian institute of technology Khragupur (Rudra et al. 2015). More general collections are also distributed in open access by NII Test Collections for IR Systems (NTCIR[10]), REUTERS and CLEF[11].

[4] http://crisislex.org/
[5] https://gnip.com/
[6] https://topsy.com/
[7] https://snap.stanford.edu/
[8] https://www.kaggle.com/
[9] https://www.crowdflower.com/
[10] http://research.nii.ac.jp/ntcir
[11] http://www.clef-initiative.eu/

2.2.2.2 Data Access via Data Resellers

Twitter has many certified product partners (e.g. GNIP and Datasift[12]) having access to the Twitter Firehose API. Compared to public available APIs (i.e. REST API and Streaming API), this API guaranties the extraction of 100% of tweets responding to a specific query by removing a lot of usage restrictions imposed by Twitter. However, this API is fairly costly, especially for individual users. It is generally handled by certified Twitter partners who sell access to the Firehose through commercial tools offering full access to Twitter data. Due to the Firehose costs involved, the vast majority of these tools provide access to 1 or 2 years Twitter data.

A special mention goes to Brandwatch[13] Twitter partner proposing a historical tweets data extraction tool covering all the tweets shared since Twitter's inception in 2006. This tool is highly recommended for research looking to evaluate the evolution of user's behavior over time or analyze historical events. Few research institutions have used such tools for research data collection due to their exorbitant cost (Ashktorab et al. 2014). Such tools are mainly used by industrial organizations in order to improve their marketing strategy in Twitter.

2.2.2.3 Data Access via US Congress Library

Since April 2010, the Library of Congress[14] has announced its intention to archive public historic tweets for conservation and research. Such announcement was an official recognition of the historical and cultural values communicated through these new digital short information that may serve even as references in the future. The idea of archiving such electronic records was a historic announcement especially for researchers who need to access such information in order to gain better understanding of microblog users behavior. This announcement was published after the agreement signing by Twitter and the US library providing the library an archive covering all public tweets shared from the Twitter inception in 2006 through the date of the agreement April 2010. The Twitter partner Gnip has managed the transfer of tweets to this archive. The resulted 2006-2010 archive contains approximately 170 billion tweets including more than 50 million tweets per day shared from people around the world.

While such announcement has been the first initiative to provide a free data access to the research community, over six years since this announcement, even the 2006-2010 archive remains unavailable. The US library congress has received more than 400 requests from

[12] http://datasift.com/
[13] https://www.brandwatch.com/
[14] https://www.loc.gov/

researchers to be able to access to the tweets archive. These requests have been denied so far due to technical challenges which could be organized into two categories :

- challenges involving practice, such as how to organize the huge amount of tweets which is growing day per day, how to provide a useful search engine answering to the different researchers requests, how to physically store all this data.
- challenges involving policy, such as how to manage access controls to the archive, is it better to make some restrictions, how to manage tweets that threat some users privacy, how to ensure data update in response to the users requirement who wish to delete some of their own public tweets.

2.2.2.4 Data Access via Data Grants

Inspired by the big technological companies including Facebook[15] and Google[16], which frequently make collaboration with public or private institutions to tackle big research problems, Twitter has followed this same strategy in a more highly formalized way. In February 2014, Twitter has introduced its Data Grants project accepting applications from any member of research institutions to access to the needed historical and public information required in their research studies.

Attracted by both the wealth of the expensive data that can be provided for free and the wide range of research possibilities in the field of online social media information analysis, more than 1,300 proposals have been submitted to Twitter Data Grants call in 2014. These proposals were received from more than 60 different countries, with more than half of them belong to research institutions located outside the United States. This remarkable interest on Twitter data explains the lack of available Twitter data for research and the growing researchers need to obtain valuable historical data collections appropriated to their research problem. Only six institutions were selected to receive datasets appropriated to their research problem needs. Such provided datasets have not been shared yet.

2.2.3 Data Access via Ad-hoc Applications

As available open Twitter data does not cover all the researchers' needs, many researchers have explored new ways for automatically extracting their own required tweets data. In the following subsection, we detail the different extraction methods from standard techniques using Twitter APIs to more advanced ones using distributed data crawlers.

[15] http://www.facebook.com/
[16] http://www.google.com/

2.2.3.1 Data Access via Public Twitter APIs

As the most famous microblogging platforms, Twitter has its own Application Programming Interfaces (APIs). These APIs are the main microblogging platform door for Twitter data extraction. Twitter allows researchers to easily extract the required data via three different kinds of data extraction APIs: two REST APIs and a Streaming one.

- **REST APIs** include two distinct APIs, the *RESTFul API* and the *Search* one. These APIs are based on the REST architecture now popularly used for designing web APIs which use the pull strategy for data retrieval. To collect information a user must explicitly request it. We present below the two distinct APIs :
  - **RESTful API** (Representational State Transfer) enables researchers to access to information and resources using a simple HTTP invocation. This API provides automated functions for things which could be manually carried out through Twitter web interface (e.g. access to a specific user's timeline, automatic search of specific information related to a specific topic, filter tweets based on certain criteria and display those tweets in your blog or website). This API is intended for developers of websites/blogs or web applications.
  - **Search API** offers different techniques to interact with Twitter search and trends data. Unlike REST API which enables the access to core data, the Search API provides access to historic data. This API accepts words as queries (e.g. full name, company name, location, or other criteria) and hashtags referring to topical interest. Using this interface, multiple queries can be combined using a comma separated list to process results matching more than a single search criterion. The search tool provided by the search API integrates similar functions as those available through the Twitter web search tool with some limitations on the returned results. This API allows the extraction of historic data only dating from a week before the time of the query. The other older historic data are only accessible through the search tool of the main Twitter website. This API was widely used by researchers to collect historical data meeting their research goals. Rudra et al. (2015) used this API to collect historical data –matching a set of defined keywords– relative to four disaster events in order to explore how situational information can be identified during disasters. McCreadie et al. (2012) used this API in order to collect tweets shared between two defined dates. The collected data was shared publically during the TREC microblog Track.
- **The Streaming API** gives access to tweets shared in real-time. This API uses the push strategy for data retrieval rather than the pull one processed by the REST API. Once a request for information is set-up, the Streaming API provides a continuous

stream of updates with no further input from the user. Using this API, it is possible to search for tweets matching a set of defined keywords, hashtags, user ids, and geographic bounding boxes from current data as it is being posted. The filter function integrated in this API facilitates the streaming search and provides a continuous stream of tweets matching the search criteria. The Streaming API has different capabilities and limitations with respect to what and how much information can be retrieved. It has the following three types of endpoints processing the required streaming data with different data restrictions :

- *Public streams :* These are streams containing the public tweets on Twitter.
- *User streams :* These are single-user streams, providing access to user tweets.
- *Site streams :* These are multi-user streams and intended for applications which access tweets from multiple users.

This API is generally used by researchers to collect real-time data matching a wide set of keywords. Olteanu et al. (2014) have collected 6 disaster datasets that have occurred between October 2012 and July 2013 using principally this API. This data was extracted using a keyword-based search approach. A large set of keywords were executed for data filtering and extraction during each disaster. Kumar et al. (2011) has also used the Public Stream reader to obtain real-time tweets filtered using keywords, hashtags and geolocations search criteria.

Comparing the two major Twitter APIs as described in Table 1, REST APIs are intended for the extraction of tweets posted in the past few days. However, the Streaming API is used to collect the recent ones. The two APIs require authentication, the REST ones necessitate one log for each user connected to the application and the Streaming ones can use a single connection. The extraction of tweets could be performed using one of these APIs. To search in real-time a high volume of tweets –sent by specific accounts, or within a geographic area– using more than 250 keywords, the Streaming API would be more efficient in this case. Otherwise, to search for tweets using multiple requests based on location origin, language and various other measures per minute, Search APIs are recommended in this case.

2.2.3.2 Data Access via Crawling Techniques

Crawling is the most popular data acquisition technique in microblogging platforms. This technique consists of traversing across users' profiles in order to collect the required information. Such technique was mainly used for crawling the microblog social graph for acquiring publicly available information about users. Crawling may take one of these three forms : a distributed crawler, a parallel crawler and a sequential crawler. These different

Table 2.1: Rest APIs vs. Streaming API for tweets collection.

	REST APIs	**Streaming API**
Targeted tweets	Past (7days)	Recent
Authentification	One log for each user connected to the application	A single Streaming API connection
Rate limit	250 keyword/ minute using 15 requests; 100 tweets/ search	400 keyword, 5,000 accounts, 25 geographic areas; 3000 tweets/ min
Data Format	JSON	JSON
Type of queries	Location of sender, language and various other measures	Words, user and geographic area

crawlers generally adapt their crawling methods according to the data acquisition functions offered by Twitter public APIs or/and web pages.

Crawling Principle

Crawling is generally processed using the typical graph structure of microblogs. During such process, the microblogs graph is divided into interconnected nodes and edges referring respectively to users and any relations that link these users. The crawling process of these graphs differs according to the data targeted by researchers. The proposed crawling systems process specific algorithms at each crawling step to access their targeted data. The effectiveness of the different crawling strategies generally depends on the following choices :

- *Initial seed nodes choice.* The choice of the node where the crawler has to start the data collection is very important. For example, for researchers aiming to crawl the social graph data relative to the whole microblog. Choosing a list of users who are not well-connected as seed nodes is not a rational choice as the crawler will not be able to reach most of the microblog users. For social graph data collection, it is better to select popular users as seeds. Bošnjak et al. (2012) collected data relative to users belonging to the Portuguese community by selecting popular users in that community as seeds. The seed nodes are generally chosen according to their potential to continuously expand the social graph by discovering new microblog users belonging to the community of interest.

- *Crawling algorithm choice.* The chosen crawling algorithm has to be appropriated to the data acquisition purpose. This algorithm has to define the visiting order of the next selected users for graph crawling. The crawling starts from the seed nodes and proceeds to the next nodes at each step following the chosen crawling algorithm strategy. The most popular graph crawling algorithms such as Breadth-Search-First (BSF), Greedy, Lottery and Hypothetical greedy were widely used for social graph data acquisition from microblogs (Ye et al. 2010). Others crawling algorithms which fit better to the context of graph sampling were also explored for this task. Leskovec

& Krevl (2014) explored the Jump and Walk algorithms in the context of sampling to avoid the crawling of useless nodes.

- *Focused crawling choice.* While the crawling algorithm expands the list of tentative users identified through the crawling process, the focused crawling approach has to identify the nodes that have to be monitored by the system. This approach would orient the crawling system to focus more on nodes matching certain selection criteria. Bošnjak et al. (2012) analyzed the profiles and tweets languages relative to the users represented by the expanded nodes in order to extract only the data relative to Portuguese microblog users. Valkanas et al. (2014) analyzed users location during the crawling process in order to only focus on users geo-located in the specified 2D bounding box. Saroop & Karnik (2011) focused their crawling process to only access user profiles that are judged relevant to a pre-defined topic.

- *Stopping criteria choice.* crawling the entire microblog graph is not generally essential in the case of specific research data targeting. The crawling process has to continue its nodes expansion until some criteria are met. By default for social graph crawling, the stopping criterion represents either a constant number of samples that has to be reached or a fixed number of iteration that has to be processed during the expansion of new discovered nodes. The number of samples is generally estimated by experts or computed automatically according to the search criteria.

Crawling Architecture

As the data crawling process is further delayed by the countermeasures deployed by the Twitter APIs to block any extensive data access (see Section 1.2), many crawling systems attempted to propose a convenient architecture enabling an extensive crawling in reasonable periods of time. The architectural solutions proposed in the literature have dealt with the stated microblogs APIs issues below as follows :

- *Slow data collection process.* To speed up the data collection process, many online social networks crawling systems parallelized the graph exploration. Following this strategy, many nodes can be expanded in the same time using parallel crawlers. However, the issue that rises in this case is how these crawlers can be managed in parallel. Chau et al. (2007) managed their parallel crawlers using a centralized coordinator and a data master server managing the sub-list of users queue that has to be processed by each crawler. Canali et al. (2011) integrated a centralized engine module coordinating between the different parallel crawling tasks by exploiting the MapReduce programming paradigm.

- *IP banning.* To avoid IP banning problem caused by Twitter APIs to limit data extraction, many crawling systems distribute their crawling process in several machines. Such strategy has been employed largely for extensive data collection from online social networks. Planetlab project (Spring et al. 2006) proposed a data collection system that can be adapted to any microblogging platform or/and online social network. Their system was deployed in an open platform for accessing planetary-scale network services. This platform currently consists of 1,333 nodes distributed at 634 locations across the world.

- *Limited access for user's profiles and tweets.* As the Twitter APIs only return a small number number of the most recent tweets shared by each user, some researchers implemented a web crawler relying on Twitter APIs in order to extract the most recent tweets relative to specific users. Through this web crawler strategy, Wang (2010) extracts the 20 most recent tweets relative to some non-protected users based on their IDs. McCreadie et al. (2012) also implemented a web crawler system to allow the participants for the special microblog track of TREC 2011 to download their tweets collection even if they do not have access to the non-restricted REST API. Tweets that have to be crawled through this system were already pre-identified using a common set of tweets (user-name, tweet id pairs) distributed for all participants.

2.2.4 Discussion

In this subsection, we discuss the different data acquisition techniques according to their advantages and drawbacks. These techniques are summarized in Table 2.2. By comparing these techniques, we can conclude that :

- Direct data access techniques – except research data collections— provide rich Twitter data suited to any research domain. However, accessing such data is very expensive for academic research institutions. Granted data is provided to a limited number of projects or potential collaborators from academic or industrial institutions. Research data collections voluntarily shared to encourage scientists to deal with the trending scientific challenges do not fit all research needs and goals. For example, there is no available collection adapted to test the different proposed approaches dealing with the problem of key microblog users identification during events. Moreover, the content of these collections becomes inaccessible over time. They have to be downloaded at the same period of time of their publication. They are also sensitive to user's privacy state change or tweets removal over time.

- Ad-hoc applications for Twitter data access provide an open data access to any person or institution. This access is limited and has to respect Twitter APIs defined

restrictions. Ad-hoc applications based on a distributed and parallel crawling can circumvent these data limits by providing further data. However, these Ad-hoc applications are implemented to target specific communities or/and specific tweets. The proposed Ad-hoc systems in the literature differ according to their data selection criteria and crawling algorithms. Such systems were mainly used for social graph or specific tweets data collection. There is no standard data acquisition system that can respond to any data needs.

According to these comparisons, we can conclude that direct access methods are affordable for a constrained number of researchers. In the case of specific data needs it would be more convenient to collect data using Ad-hoc application. The distributed and parallel crawling technique seems to be the most efficient when a huge amount of data is targeted. However, applications following such crawling technique have mainly targeted Twitter timeline or social graph data. Many efforts are still needed in order to propose new ad-hoc applications targeting the different specific data of interest using varied search criteria. To the best of our knowledge, there is no available modular ad-hoc applications adapted for both a real time extraction and analysis of various forms of Twitter data during events. Through this thesis, we explore the ad-hoc data access approach for the extraction of new data collections adapted to test and learn any key microblog users identification model. The wealth of distributed and parallel crawling techniques is also explored for building a new extraction and tracking system enabling both the identification and the tracking of prominent microblog users in real time during specific events.

2.3 Information Retrieval from Microblogs during Crisis Events

Urgent information describing the real-time situation of regions threatened by crisis is voluntarily shared in microblogging platforms. Disaster-related information is shared and spread voluntarily in microblogs without any external incitation. These platforms have become the most popular communication and fresh information provider tool. They are continuously consulted in order to follow and share the last event news in real time. Many disaster management organizations have investigated the particularities of these microblogs in order to efficiently manage emergency situations (Theodore 2013, MacEachren et al. 2011). In this section, we focus on (1) describing how microblogs are explored during crisis events, (2) detailing how microblogs are used to ensure situation awareness by listing the main information retrieval techniques explored in this context.

Table 2.2: State-of-the-art microblogs data acquisition techniques : Advantages and drawbacks. Examples of collected collections by each existing technique are also specified.

Data Acquisition techniques	Description and Examples	Advantages	Drawbacks
Direct Data Access			
Research data Collections	Research data collections refer to pre-collected data shared in open access to encourage scientists to conduct further researchs exploring the shared information in microblogging platforms. **Examples :** Twitter2011 Collection (McCreadie et al. 2012), CrisisLex Collections (Olteanu et al. 2014)	-Direct access to data. -Easier comparison of the different information retrieval models during scientific competitions by reposing on the same collections and the same evaluation metrics.	-The content of these collections degrades over time. -Comparing different models using the same collection is only effective when the collection is downloaded in the same time. -Such data is not generally suitable to all the proposed research models in the targeted research field.
Granted data	Granted data are personalized data afforded to some research institutions. This data responds to specific research data needs defined in the project proposal submitted to the Twitter Data Grant program. **Examples :** NICT granted data for disaster information analysis	-Access to personalized data according to each project needs and goals. -There is no missing data that can bias the research models results.	-2% of the submitted projects for data grants are accepted. -calls for data grants are rare. There was only one Twitter Grants call until to date.
Data sellers	Data provided by data sellers covers 100% of tweets responding to a specific executed query. Any required public data can be provided by these Twitter partners without any restriction. **Examples :** Gnip and DataSift	-Unlimited data access. -All historic Twitter data are accessible.	-Acquiring data from data sellers is fairly costly. -Accessing to Twitter Firehose API is not afforded for public institutions.
US library Congress Data	US library congress data contains all the historic tweets which were provided for free by Twitter. This data is detonated for research and tweets archiving as historical and cultural data.	-Suitable to all the researchers data needs in different research fields.	-This data sharing project is still in progress since 2010. Data is not yet accessible by researchers. -Managing the access to the rich data provided by the US library remains challenging until to date.

Access through Ad-hoc Applications			
Access through Twitter APIs	Twitter APIs are the open public door for Twitter data acquisition. These APIs provide a limited data access to both the historic Twitter data relative to the 7 past days and the streams of real time shared public data in the platform. **Examples :** the collected data in (Olteanu et al. 2014, Rudra et al. 2015) were acquired using these APIs.	-Researchers can implement their own personalized data extraction system by respecting these APIs restrictions.	- Limited access to Twitter data (typically around 1% of the tweets matching the researchers search criteria can be extracted). - Sensitive to IP banning.
Access through Crawling Techniques			
Sequential Crawling	The sequential crawling technique consists of collecting data relative to researchers' requests by looping the list of users or tweets responding to a particular search criteria. Such crawling process can be processed through Twitter APIs or/and web interfaces. **Examples :** TREC2011 Twitter timeline data collection (McCreadie et al. 2012)	-Access to further historic Twitter data which are not accessible through standard APIs. (in the case of using web crawling) -Personalize the Twitter search according to the researchers' data needs and goals.	-Limited access to Twitter data. -Time consuming. -Sensitive to IP banning.
Parallel Crawling	The parallel crawling executes different crawlers in parallel. Similarly, such crawling technique can be processed through Twitter APIs or/and web interfaces. **Examples :** Social graph data collection (Canali et al. 2011), Users generated content data collection (Wang 2010).	-Rapid data extraction. -Outperforms the sequential crawling techniques.	-Limited access to Twitter data. -Sensitive to Twitter data limits updates. (In the case of managing the crawlers according to outdated Twitter restrictions) -Sensitive to IP banning.
Distributed Crawling	This crawling technique has the same characteristics as the parallel one. However, such technique distributes the parallel crawlers in different interconnected machines. **Examples :** Social graph data collection using Planetlab (Spring et al. 2006)	-Rapid data extraction. -Twitter APIs limits bypass. -IP banning risk minimization.	-Sensitive to Twitter data limits and access controls updates.

2.3.1 Microblogs Role during Crisis Events

In the following, we describe how emergency teams can benefit from such platforms to enhance crisis event management. We review the different explored approaches to effectively manage different event phases. These approaches fall into three categories : *alert dissemination*, *crisis events detection* and *situation awareness*.

2.3.1.1 Alert Dissemination

Many official emergency departments and government agencies disseminate real-time disaster alerts through microblogging platforms before their official announcement on news outlet channels. For instance, the Global Disaster Alert and Coordination System (GDACS) publishes disaster-related alerts and updates through their Twitter account @GDACS. The Boston Police also adapted this strategy during the Boston marathon attack by providing official information in real time during the prosecution. Similarly, the United States Geological Survey (USGS) is used to share frequent updates of earthquakes magnitudes in the following Twitter accounts by referring to two categories of earthquakes :

- *@USGSBigQuakes* (USGS Big Quakes) diffuses detailed alerts for earthquakes worldwide which have magnitudes greater than 5.5.
- *@USGSted* (USGS Tweet Earthquake Dispatch) shares news about the different earthquakes with magnitudes under 5.5.

The availability of active official agencies and government departments in Twitter eases the dissemination of the valuable information on one hand and helps to reduce rumors that can be spread during unexpected disasters on the other hand. However, there are no available functions in Twitter that can ensure the banning of outdated information sharing. Microblog users may share outdated information–that have been already updated–, and the original spreaders have no control over the retweeting process of their shared information.

2.3.1.2 Event Detection

The first step for efficiently exploring microblogging platforms data in the context of crises situation management is detecting crisis events at an early stage. Crisis events are either unexpected such as earthquakes or predictable such as some tornados or storms events. The occurrence of such events provokes a consistent increase of tweets influx in a short period of time. Tweets expressed in such situations have specific characteristics that can be explored in order to make the detection of such events automatic. Through the instant

detection of such tweets influx, emergency teams would be able to intervene at the right time and to speed up the emergency management process. Many methods have been proposed in order to automatically detect events in the context of crisis and emergencies. Most of proposed methods are an adaptation of the detection approaches that have proved their efficiency in other more general detection tasks such as news detection. Yin et al. (2012) proposed a burst-detection module extracting and analyzing burst words based on their probability distribution in a time window. Sakaki et al. (2010) deployed a functional system that detects and geo-locates earthquakes in a competitive time regarding the Japan Meteorological Agency. Their system is processed by aggregating various tweets provided by users geo-located in the earthquake area. Pohl et al. (2012) studied the efficacy of different clustering techniques : Self-Organizing Maps (SOM) and Agglomerative Clustering (AC) in order to detect sub-events related to a specific crisis event. Earle et al. (2012) implemented a simple event detector that captures any increase in the frequency of tweets containing the word "earthquake" or its equivalent in other languages.

2.3.1.3 Situational Awareness

Situation awareness "is the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley 1988). Establishing situation awareness requires three different levels of activity :

Perception : Relevant and fresh disaster-related-information shared in microblogging platforms have to be extracted instantly. By accessing this information, emergency first responders would have a global view about the different threatened and affected regions by the disaster. MacEachren et al. (2011) extracted useful information from disaster-related tweets by using a web-enabled mapping tool in order to compare, and classify tweets. Starbird & Stamberger (2010) proposed a tweet syntax including predefined keywords that facilitate information extraction such as the location and the nature of communicated emergency.

Comprehension : The disaster-related information extracted in the perception level has to be analyzed in order to acquire new knowledge. Such knowledge would be of significant help to emergency teams in order to understand what is really happening-on-the-ground. By highlighting urgently needed information, these teams would be able to intervene rapidly to manage the detected emergencies. Many research studies have been conducted for situational information comprehension in real world disaster cases. Various analysis techniques have been explored in this context. These techniques depend on the nature of the extracted data and the target of analysis. For example, natural language processing techniques are more suited to analyze and summarize the different extracted information (Rudra et al. 2015). For identifying influential users, standard ranking algorithms such as HITS and

PageRank are generally applied (Gupta et al. 2012). To categorize information according to their meanings, disaster-based ontologies and machine learning technologies can provide efficient results (Imran et al. 2013*a*).

Projection : After acquiring new knowledge, it is possible to visualize this knowledge in maps, reports or graphs. Projection can be represented by :

- Providing a complete summary of the shared disaster-related-information by on-the-ground users in order to help emergency teams to be aware of what is happening in each region affected by the disaster.
- Mapping the position of the detected prominent users requesting urgent assistance through the microblogging platform in order to be able to intervene just in time.

2.3.2 Situation Awareness during Crisis Events

Alert dissemination, event detection and situation awareness processes are all essential to ensure efficient management of crisis events. Each process has its own particularities and benefits that have to be considered during disasters. In this section, we focus on reviewing the different proposed approaches in the literature for situation awareness enhancement during crisis events. As described in the previous subsection, situation awareness process consists of retrieving the relevant and exclusive information helpful for emergency teams. Retrieving such information from the huge amount of data shared in real time during crisis events remains challenging. Tweets are expressed in various forms and languages. The same information can be expressed in different ways and by various users. Outdated information keeps spreading in microblogs even if they do not have an informative value anymore. To deal with these challenges, many situational information retrieval strategies have been proposed in the literature. We categorize these information retrieval strategies into two broad categories : *disaster-related information classification* and *disaster-related information extraction and summarization.* These two strategies are detailed on the following.

2.3.3 Disaster-related Tweets Classification

The problem of valuable information retrieval from microblogs during crisis events has been widely studied in the literature using content-based-analysis techniques (Sakaki et al. 2010, Imran et al. 2013*a*). However, there are only few works dealing with such problem using the user-centric information retrieval approach (Kumar et al. 2013). In this subsection, we detail both the *content-based classification* and the *user-based classification* approaches.

2.3.3.1 Content-based Classification

This approach consists of analyzing the content of each shared event-related information in order to judge which information have to be retained. While tweets content can be expressed in various formats (i.e. text, image, video and links), most of the literature work have evaluated these tweets according to their text content. Fewer works have explored the other possible tweet content types such as links and videos for information retrieval during crisis situations (Gupta et al. 2013).

In the following, we detail the different steps that are generally processed in the literature in order to perform tweets content analysis and classification. The choice of the pre-processing, the feature extraction and the tweets classification techniques differs according to the type of analyzed data and the data analysis goals.

Data pre-processing Most researchers and practitioners prepare microblogs content data by pre-processing it. Several pre-processing techniques have been explored for situational information retrieval. The choice of which technique to employ principally depends on the tweet content type, the targeted features that need to be extracted and the data analysis goals. Typical NLP preprocesssing operations include *tokenization*, *part-of-speech tagging* (POS), *shallow parse tagging*, *stemming and lemmatization*, etc.

Tokenization text pre-processing technique is generally adopted in order to segment tweet text into tokens and to retain only the required words to process the feature extraction step. This technique was processed by several disaster information retrieval models. Cobo et al. (2015) used such technique in order to retain only hashtags, words and user mentions. These filtered tokens were then considered in order to extract a list of features describing the frequency of each considered token. Imran et al. (2013*b*) considered tweets as a sequence of word tokens in order to algorithmically label each token as a part of their targeted disaster-related-information or not.

Part-of-Speech (POS) tagging is the process of assigning a part-of-speech to a given word using linguistic and statistical information. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Morstatter et al. (2014) used POS technique in order to extract Part-of-Speech patterns and to reduce the number of dimensions and possible noise in the disaster-related tweets dataset. The idea of patterns extraction was proposed by Munro (2011) who prove that such particular sub-word patterns improve the accuracy of relevant information identification during disasters.

Shallow parse tagging or "chunking" is the process of segmenting a text tweet into an unstructured sequence of syntactically organized text units called "chunks". These chunks describe the relations between the different words included in the tweet text such as noun phrases and verb phrases. Morstatter et al. (2014) used the shallow parse tagging in the

context of situational information retrieval in order to highlight the syntactic differences between relevant and non-relevant disaster-related tweets. This tagging technique is effective when only a brunch of words is targeted and the sub-structure of the whole tweet text is not of interest. A chunker gets directly the needed information without having to parse the full sentence words.

Stemming and lemmatization is the process of reducing inflectional forms and sometimes derivationally related forms of a word to a common base form. Various situational information retrieval systems have used stemming and lemmatization techniques (Cobo et al. 2015) (Imran et al. 2015). Such NLP techniques are generally processed together. While the lemmatization system would handle matching of synonym words including verbs and nouns like "car/automobile". Stemming would deal with grammatical differences between nouns or verbs such as "cat/cats", "mouse/mice" or "run/runs/running/ran".

There are also other NLP techniques that were applied for preprocessing tweets text such as *ARK tagger* (Owoputi et al. n.d.) and *PTB Style Tag set*. ARK tagger was especially designed for tweets text tagging (Morstatter et al. 2014, Imran et al. 2015). This tagger is able to recognize idioms like "ikr" meaning "I know, right?" and assign them the correct Part-of-Speech tag. PTB Style Tag set is a more fined grained POS. This tagger was considered by Morstatter et al. (2014) in order to compare the efficiency of this tagger with that of ARK tagger in the context of disaster-related tweets classification. Additionally, other higher level techniques can be considered such as *sentiment tags* to highlight the different parts of tweets reflecting a specific positive or negative emotion. Sentiment tags were considered by Beigi et al. (2016) in order to classify disaster-related tweets into positive, fear, anger and other emotions classes.

Feature Extraction and Selection Tweets content is typically represented as a numerical vector composed of different well-defined features. The most adopted text-content representation is the vector-space model one. Such text modeling approach characterizes the position of specific defined words or phrases in the tweet text. Each vector dimension refer to a specific term (e.g. words or phrases describing the crisis event). These terms have specific weights that can be computed using different techniques such as TF-IDF weighting. Such technique favors terms that are important however infrequent during crisis events (Baeza-Yates & Ribeiro-Neto 2011, Cobo et al. 2015).

Other efficient text-based features were considered in the literature. Herein, we split these features into three categories : *linguistic features*, *tweets specificities-related features* and *geotagging features*.

Linguistic features can be divided into two features classes *word-based features* and *POS features* (Morstatter et al. 2014). Word-based features are constructed by analyzing unigrams and bigrams frequency counts. Such unigrams and bigrams frequency can be applied

either to check the frequency of well-defined keywords characterizing the specific analyzed disaster or to check the presence of geolocation information in the tweet text. Morstatter et al. (2014) highlight eyewitness tweets by checking the existence of specific terms like the term "there" defined by Lakoff (1987) as "a mental space in which a conceptual entity is to be located". POS features are considered in order to extract the different patterns that can differentiate relevant tweets from non-relevant ones or to point out important information that have to be extracted. Morstatter et al. (2014) have proposed two notable crisis-sensitive features : *POS patterns features* in their ARK and PTB forms and *prepositional phrase patterns features* highlighting specific patterns describing crisis situation.

Tweets specificities-related features refer to the features considering principally the syntax defined by Twitter such as retweets, detected by the suffix "RT", mentions, detected by the "@" symbol, or hashtags, detected by the "#" symbol. By referring to this specific syntax, many features can be extracted such as the number of users mentioned, the number of comments related to each retweet and the number of hashtags included in the text tweet. Moreover, the metadata-related to tweets has also been considered while extracting tweets-related features like the number of likes and retweets attached to each tweet.

Geo-tagging features are typically extracted automatically from the tweet attached metadata. However, such automatic extraction is only possible if the user has enabled the option of sharing his/her current geographic position. It is also possible to extract such information from the tweet text. This approach was extensively explored in the literature. MacEachren et al. (2011) used Geonames[17] to identify geolocation information from the text and Gate[18] for tweets geo-coordinates extraction from metadata. The geolocation identification process known as *geo-tagging* is tricky. It does not necessarily consist of looking for proper nouns. There are various ambiguities that have to be considered and treated separately. For example, the name of a city "Texas" can also refer to the famous game "Texas Hold'em". There have been several interesting geo-tagging approaches dealing with such ambiguities (Morstatter et al. 2014). Sultanik & Fink (2012) proposed a probabilistic model dealing with these location mentions ambiguities using an indexed gazeteer.

There are also other various *image-based, video-based features* that can be extracted from tweet content such as colors and textures (Gupta et al. 2013). Cobo et al. (2015) have extracted 4766 features describing tweet contents using a TF-IDF vectorizer. Such high dimensional representation would increase the chance of overfitting and make the filtering or the classification process infeasible in real time. To reduce the number of features that have to be extracted and increase the classifiers performance, dimensionality reduction techniques are generally used. Cobo et al. (2015) used the Latent Dirichlet Allocation technique for the features selection task.

[17] http://www.geonames.org/
[18] http://gate.ac.uk/

Classification algorithms

Once each tweet is characterized by a specific content modeling approach, an appropriate classification algorithm is generally executed in order to differentiate between the different classes of tweets. This classification step is generally processed using the following methods :

Content-categorization Several content-categorization approaches appropriated to situational information retrieval problem have been proposed in the literature. Imran et al. (2013*a*) proposed a crisis-related-ontology for tweets-content categorization into different classes (e.g. cautions or advice, donations, causalities and damages, missing or lost people, etc.). Kiritchenko et al. (2014) explored the emotional characteristics expressed in the event's tweets in order to highlight the main shared opinions regarding the specific analyzed event. Seo et al. (2012) studied the credibility of tweets in order to extract those reflecting what is really happening on the ground. Twitinfo (Marcus et al. 2011) used disaster-related sub-events picks to identify and sort the relevant tweets by using keyword-matching techniques. The words composing each tweet are matched with a list of top-keywords extracted from the different event's detected peaks characterized by a high tweeting activity regarding the event-topic. De Longueville et al. (2009) studied both the tweets text and metadata in order to extract geographical information and identify witnesses' tweets.

Supervised content-classification Supervised classifiers have been extensively explored in the literature these last years. Sakaki et al. (2010) learned automatic text-tweets classifiers over a list of selected features extracted by referring to a set of well-defined keywords. Naive Bayes were used by ESA in order to study different tweet classification settings : identifying whether a tweet is about the disaster or not, identifying tweets related to each disaster type (e.g. earthquake, flooding, fire ...) and identifying tweets reporting a damage to infrastructure (Yin et al. 2012). The classification of the EMERSE model classifying text messages about the Haiti disaster relief was learned using SVM (Caragea et al. 2011). The AIDR model coordinating between human and machine intelligence for valuable disaster-information retrieval used random forests (Imran et al. 2014).

Unsupervised content-classification Clustering methods were essentially explored in the context of event-related information analysis either for filtering the already collected information or for grouping them according to their similarity degree. Such approach was adopted by CrisisTracker (Rogstadius et al. 2013) which extracts automatically tweets containing well-defined keywords. CrisisTracker clusters these tweets into stories according to their lexical similarity. Using such strategy, this system reduces the human efforts to analyze the content of tweets relative to different dispatched stories.

2.3.3.2 User-based Classification

As content-based classification techniques are time consuming and also sensitive to tweet content format (i.e. image, text, video and links) and language, content-unaware information retrieval techniques have been explored in the context of crisis events. Such techniques associate the relevance and quality of tweets with the importance of their authors. Identifying and tracking users who are behind the required disaster-related information, would give direct access to relevant and exclusive tweets independently of their format and language. This strategy also explores some high level data content analysis techniques in order to detect information nature or to check the existence of some keywords and hashtags. For example, text included in tweets can be analyzed in order to extract specific hashtags, keywords and location information. Linguistic, syntactic or semantic features reflecting tweets' textual content are rarely explored using such techniques.

Using this information retrieval approach, features are not coupled with tweet content, but rather with the users' specificities in general. These features characterizing microblog users typically refer to the user activity in the microblog or/and the user connectivity according to the microblogs structure. Referring to these features, users are represented either by a single feature vector-based representation or by a graph-based representation.

Herein, we detail the few works that have explored this user-centric classification approach within the field of situational information retrieval during crisis events. Using this classification strategy, various categories of key microblog users (e.g. influential users, eyewitnessers, journalists, domain experts, etc.) have been targeted during crisis events.

Hemant et al. (2014) define key users that have to be tracked during the disaster as influential users having a central position in the microblogging platform network. These users were identified by using the PageRank algorithms ranking the different microblog users interacting during the disaster based on their position in the network graph. This graph is represented by different nodes and edges referring respectively to microblog users and their various interactions (retweets, comments and mentions). Gupta et al. (2012) define microblog key users as information sources who are central in Twitter communities. In order to identify such users, they started by identifying the different users communities. Such communities are detected using a spectral clustering technique to cluster each user based on new proposed similarity metrics. These similarity metrics consider the content, link and metadata similarities characterizing the different users. Once the different communities are identified, top central users are detected using the degree centrality measure. According to the results obtained through this identification approach, 81% of the detected users are used to share the same opinion of the entire community.

De Choudhury et al. (2012) studied various categories of microblogs key users (i.e. organizations, journalists/media bloggers and ordinary individuals) during disasters. Users

belonging to such categories are classified using standard machine learning techniques. Each user was represented by a single feature vector. Network/structural features, activity features, interaction features, named entities and topic distribution features were considered for user representation. According to the classification results, it has been observed that while organizations tend to share more content links, ordinary users use to express their personal experience and opinion regarding the event.

Starbird et al. (2011) define key users who need to be tracked as on-the-ground Twitterers. To differentiate between twiterrers who are on the ground and those who are not, these authors have explored a variety of flat profile user features and recommendation features. While flat features describe user's profile metadata, recommendation features reflect how the other microblog users interact regarding the user's posts. For classifying the different microblog users based on these features, an SVM model was generated. Using this identification strategy, 68% of true on-the-ground users were identified. Such study has highlighted the fact that richer features have to be considered in order to increase the model accuracy. Similarly, Kumar et al. (2013) define key microblog users in crisis events as eyewitnesses who are geo-located in the disaster area. They categorized the different Twiterrers interested in the disaster into different classes according to both their topical and location scores. Through this categorization, they targeted users having high topical and location scores.

2.3.4 Disaster-related Information Extraction and Summarization

Unlike content-based classification techniques which evaluate the relevance of the whole tweet content, information extraction and summarization techniques analyze information nuggets included in each tweet. Such techniques automate the analysis process of tweets content in order to generate a structured report, providing an overview of the different news shared regarding a specific event or topic. In this subsection, we briefly review how *information extraction* and tweets *content summarization* techniques have been used in the context of crisis events.

2.3.4.1 Information Extraction

The task of Information Extraction (IE) consists of automatically extracting structured data from unstructured or semi-structured data forms. The extraction of structured data from tweets is a challenging task. Unlike web documents or blogs, tweets are always short due to the 140-character length limit. Such limitation encourages microblog users to express their messages by using various abbreviations, symbols and misspellings or/and by neglecting the grammatical rules. One of the typical tasks in information extraction is named entity

extraction which consists of recognizing entities included in the tweet text. Let us assume the following tweet as an example "The death toll in an earthquake in south-west China is now at least 32, with 467 injuries". There is different information that can be extracted from tweet text content such as (disaster-type=earthquake, location=south-west China, number-of-injures=467, time=current time). These extracted entities can be easily integrated with external information or/and filtered or/and associated with other entities.

The entity recognition problem is generally solved by two phases, the chopping of unstructured texts phase and the labeling of extracted parts phase. The parts of resulted information pieces of the first phase are commonly expressed using one of these two forms : tokens and word chunks. Such forms are extracted using Natural Language pre-processing techniques. In the labeling phase, a pre-trained model is processed in order to identify the labels of each extracted piece of information from unstructured texts. Nevertheless, the labels of adjacent pieces of information have generally different relations between them. These relationships between two particular pieces of information can be used to determine the label of the next analyzed piece of information. Consequently, different probabilistic models were proposed to capture the relations between the labels of adjacent pieces, such as Hidden Markov Models, Maximum entropy Markov models, and Conditional Random Fields (CRFs) (Imran et al. 2015). Imran et al. (2013*b*) applied CRFs for tweets information extraction. The information extraction process is conducted in two steps. They started by classifying the disaster-related tweets into the following categories "infrastructure damage", "people", "donations", and "caution and advice" based on a rich set of features. These features are extracted by analyzing word unigrams, bigrams, Part-of-Speech (POS) tags and other tags. Once the tweets are classified, class-relevant information are extracted. For example, by analyzing tweets classified in the category of "People", information nuggets relative to missing or lost people found or the number of missing people or the identities of found users can be easily extracted. Starbird & Stamberger (2010) proposed a micro-syntax easing the information extraction process during disasters. This syntax answers the following questions "who, what, and where" by using well defined hashtags adapted to emergency situations. These hashtags are designed to indicate the different details contained in a tweet. For example, locations are detected by referring to specific hashtags #city [name of the city] #location [place] #addy [street name]. However, such syntax was not widely adopted by Twitter users.

2.3.4.2 Summarization

Summarization is also conducted in order to deal with the rapid rate of posted disaster-related tweets during emergency situations by providing a text report summarizing the relevant information that have to be retained. Commonly, a summary is generated by considering only disaster-related tweets content without taking into account additional

information. Most of the relevant information that need to be reported are associated with well-defined keywords specific to each domain context.

In the context of crisis events, the temporal dimension of tweets reveals itself as the most important dimension. Summarizing texts based on outdated information would disorient emergency teams. Summarization systems have to be able to distinguish between outdated information and the new ones. The resulted reports have to be frequently updated through processing incremental text summarization which is challenging. Dang & Owczarzak. (2012) proposed an incremental text summarization system. This system processes the summarization task by referring only to the new or old set of posted disaster-related tweets that was not yet read by the emergency teams. Rudra et al. (2015) combined the information extraction task with the summarization task. Their approach is processed in two steps : Extracting the situational information from the different disaster-related tweets and summarizing information by considering the time-varying actionable information such as the number of injuries.

Research works presented in the TREC temporal summarization initiative attempts to summarize information related to events, by generating updates relative to crisis events immediately after their occurrence (Aslam et al. 2014). The proposed evaluation metrics through this track take into consideration the different summarization characteristics. Time-sensitive versions of recall and precision (i.e. the Expected Latency Gain and the Latency Comprehensiveness) have been proposed to evaluate the freshness of the reported information and the uniqueness of each information. Tan et al. (2015) proposed a summarization and filtering algorithm consisting of frequently updating the resulted relevant sentences by pushing the new detected information that have to be updated. Abbes et al. (2015) proposed three different temporal summarization approaches. The first one is based on named entity recognition based method, the second one refers to a rank fusion based method and the third one relies on novelty and redundancy based approach. Their named entity recognition technique has given the best results.

According to the reported experimental results of TREC 2015 Aslam et al. (2014), it is noted that none of these systems have succeeded to achieve high results in terms of precision and novelty of the update coverage.

2.3.5 Discussion

As presented in the previous subsections, various information retrieval techniques have been explored in order to harvest microblogs information contributing to situation awareness. Table 2 summarizes the different techniques that we have described in the previous subsections. These techniques are split into two main categories : information retrieval techniques

based on information-content classification and those based on information providers' classification. The classification dimensions explored for each category are briefly described with some application examples. We discuss in the following the advantages and drawbacks of these techniques :

- Information-content classification strategy consists of analyzing tweets content for situational information retrieval. Many classification dimensions have been explored to separate between relevant and irrelevant situational information (e.g. Location, time, credibility, etc.). Few features characterizing images and videos contents have been investigated in the context of crisis events as it is time consuming to analyze such kind of data. Most of the proposed dimensions have mainly focused on characterizing textual information included in tweets by neglecting the other content types. With the emergence of free live stream applications, microbloggers are used to share more videos presenting their live experiences than sharing text information. By neglecting such type of information content, a significant portion of indispensable information for crisis events management would be hidden.

- Information providers' classification strategy consists of identifying the prominent microblog users who are susceptible to provide the targeted relevant information. Such strategy is insensitive to tweets content type. Once a user is identified as prominent, he will be tracked in real time and all his/her shared information would be categorized as relevant independently of their content. While this strategy seems to be more suitable for situational information retrieval in the context of crisis events, few systems have explored this strategy for this task. The few proposed systems have focused on three main dimensions : the user's location, the user's connectivity graph and the user's interaction graph. These explored criteria are not very effective to be able to differentiate between prominent users and non-prominent ones. User tweets location is generally extracted using geotagging pre-processing technique which mainly explores tweet text data. While geotagging could perform well when applied to some tweets including some textual location indications, such pre-processing technique does not ensure the detection of the true user location. Textual location indications are either not included or hard to detect due to some ambiguities. Thus, mainly relying on the location dimension is not enough sufficient to differentiate between true witnesses and those who are not. On the other side, user's connectivity and interaction graph criteria have not provided promising identification results (De Choudhury et al. 2012). The considered dimensions are known to be sensitive to popular users who are well-connected in the microblogging network.

Table 2.3: State-of-the-art situational information retrieval techniques : advantages and drawbacks. Examples of targeted situational information by each existing technique are also specified.

Information retrieval techniques	Description/Examples	Advantages	Drawbacks
IR based on Information Content Classification	The relevance and freshness of each information is evaluated according to its content		
Classification Dimensions :			
Time	Considered for filtering the disaster-related information, detecting emergent keywords and updating the disaster reports over time(Aslam et al. 2014). **Examples :** Temporal summarization (Tan et al. 2015, Dang & Owczarzak. 2012), Disaster-related information filtering (Munro 2011), event detection (Yin et al. 2015)	-In depth analysis of the content of each shared information. -Extraction of various topical, syntactic, semantic and linguistic features that can reveal the quality of disaster-related information. -Extraction of further information regarding the location from which the information was provided. -Detection of the lexical ambiguities. -Rumors are more detectable using content-based analysis techniques. -Extraction of a large set of features that can be effective to the defined classification goals.	-Sensitive to the information content type (image, text, videos and links). -Sensitive to the content text language (e.g. english, french, Spanish, Arabic, etc.). -Neglect users' specificities. -Image and video analysis techniques are time consuming. -Text-analysis techniques would neglect any non-textual content.
Location	Extracted from text using natural language processing techniques and gazetteers or/and directly from the provided metadata. **Examples :** Eyewitnesses tweets detection (Morstatter et al. 2014), disaster-related tweets mapping (MacEachren et al. 2011), extracting information reported nearby the affected region (De Longueville et al. 2009).		
Information provided	Analyzed according to the type of the tweet content (i.e. text, image, videos and links) for the categorization of the disaster-related information. **Examples :** Infrastructure damage, people, donations, and caution and advice (Imran et al. 2013*b*), outdated and fresh information (Tan et al. 2015), positive and negative value (Beigi et al. 2016) informative and personal (Imran et al. 2013*a*)		
Credibility	Evaluated by referring to official organizations information or by analyzing the information source credibility. **Examples :** Fake images (Gupta et al. 2013), rumors detection (Seo et al. 2012), content and users credibility (Gupta et al. 2014)		

IR based on Information Providers Classification	The relevance and freshness of each information is evaluated according to the importance of their authors during the analyzed crisis event.		
Classification Dimensions :			
Location	Extracted from the user profile, user shared tweets or/and tweets metadata. Geotagging techniques are generally explored for the location extraction from text tweets. **Examples :** On-the ground Twitterers (Starbird et al. 2011), eyewitnesses (Kumar et al. 2013)	-Insensitive to information content type. -Insensitive to the texting language and abbreviations. -Exploration of various users features computationally feasible in real time. -Real-time access to the relevant information. -Evaluation of users activity evolution over time. -Analysis of the impact of the users shared information on the other users behavior. -No need to analyze all the disaster-related information, tweets relative to users who proved their prominence would be automatically retained as relevant.	-Evaluating users according to their social position in the network makes the prominent users identification task sensitive to well-connected users. -Identification approaches based on location prediction are efficient only if there are enough location indications regarding users who are really geolocated in the disaster area.
User's connectivity graph	Constructed according to the different friendship relations of users interacting about the crisis event. **Examples :** Popular and influential users (Purohit et al. 2014), users central in a specific community (Gupta et al. 2012)		
User's interaction graph	Designed according the user's interactions regarding the information shared by the other users. **Examples :** Influential users (Purohit et al. 2014)		
User's activity	Extracted by analyzing the type and nature of the information shared by each user during the event. **Examples :** Organizations, journalists/media bloggers and ordinary individuals (De Choudhury et al. 2012)		

According to these comparisons, we can conclude that while information-content classification techniques are efficient to classify disaster-related information expressed in a text format, such techniques remain sensitive to other tweet content formats and languages. On the other aside, information providers' classification strategy is insensitive to such ambiguities. However, such technique has not been efficiently explored in the context of crisis events. The explored classification dimensions are pretty basic. Richer dimensions are required in order to explore further the efficiency of such strategy for real-time information retrieval in the context of crisis events. Through this thesis, we explore existing classification dimensions that have proved their efficiency for microblog users classification in a general context for prominent microblog users identification in the context of crisis events. We also explore new dimensions in order to point out our targeted users particularities and ease the identification process.

2.4 Identifying Key Users in Microblogs

As stated in the previous subsection, there are few key users identification techniques that have been explored in the context of crisis events. In this subsection, we aim to list the main key users' identification approaches proposed in the literature in a general context and discuss their suitability to be applied in the context of crisis events. We divide these approaches into two broad categories *graph-based* approaches and *vector-based* ones. In the following, we detail (1) the different targeted key users categories that have gained the interest of researchers (2) the main specificties that can be considered while adopting graph-based classification approaches (3) the different features that can be explored while representing and classifying microblog users through a vector-based approach (4) the advantages and drawbacks of each classification approach and their degree of adaptability in the context of crisis events.

2.4.1 Targeted Key Users in Microblogs

Microblog key users identification problem has been widely discussed in the literature. These key users are known under various names having different definitions (e.g. influencers, domain experts, prominent users,etc.). These definitions differ according to the targeted research goals and domain. In the following, we list the different categories of microblog users that have been targeted in the literature.

Popular users. are defined as microblog users who are well-connected in the social network. User popularity does not necessarily rely on microblog user direct relationships (i.e. his/her followers and followees connections). It is mainly measured by considering the social relationship and the interactive relationship of microblog users related to the evaluated user.

The more popular the user is, the wider his/her visibility is. Cha et al. (2010) has raised the popularity measuring issues mentioned by Adi Avnit work[19] known by the term "The Million Follower Fallacy". Avnit has discussed various aspects leading to conclude that the number of user followers does not reflect the user popularity. Measuring popularity according to the followers number as conducted by Kwak et al. (2010) would highlight false popular users. Ordinal users seeking popularity would use spam and advertising techniques to attend a virtual popularity. The most popular followers increasing techniques are the followers recruiting technique and the "follow me, I will follow you back" principle. To study the user popularity measuring problem, Cha et al. (2010) proposed three measures, *indegree*, *mentions* and *retweets*, capturing different popular users categories. These categories vary according to the audience engaged by each user. According to their study, they proved experimentally that focusing on mentions and retweets measures would reveal popular users having higher audience than those resulted by the indegree measure. For dealing with the "follow me, I will follow you back" principle, Cha et al. (2010) applied the FlowRank measure. This measure involves the ratio between the number of a user followers and the number of other people the user follows.

Influencers or influentials. are users who can easily propagate a given information widely in a short period of time by producing large diffusion cascades (Silva et al. 2013). Influencers have generally particular characteristics like credibility, popularity, expertise or authority which make them a known reference in particular domains. Such users are generally targeted by advertisers in order to increase the diffusion of their new products demos. While the identification of popular users is mainly based on the network social graph, the identification of influent users relies on many factors taking into account both the microblogging platform structure and specificities (i.e. mentions, retweet, following activities) (Romero et al. 2011). To detect influencers, microbogs are designed as an interaction graph where nodes and edges represent respectively the users and the different interaction activity between them (Weng et al. 2010). These interactions refer to retweeting, mentioning, commenting or following activities. The identification of such users from the interaction or/and social graph is generally processed using centrality measures. Such measures evaluate each user according to its position in the graph. The most famous measures for this task are betweeness measures and eigenvector centrality measures variants such as PageRank and HITS. By adapting these measures to the context of influencers identification, many new ranking algorithms have emerged. Weng et al. (2010) proposed a topic-sensitive PageRank algorithm, TwitterRank, for user influence measuring. This algorithm takes into account both the topical similarity between users and the link structure while ranking users. Silva et al. (2013) introduced the ProfileRank model designed under the assumption that relevant content is created and propagated by influencers. This model measures user influence based

[19] https://hbr.org/2010/05/influence-and-twitter.html

on random walks over a generated user-content bipartite graph. Most of the proposed influencer ranking algorithms have mainly been inspired by the link analysis algorithms HITS and PageRank.

Domain experts. refer to microblog users having the expertise to significantly contribute in microblogs by sharing relevant and exclusive information regarding a specific topic or domain. Such users are typically not widely followed in the social network compared to popular and influencers users. They are mostly followed by thousands users who have the same interests like them or/and who are interested on their domain of expertise. In order to benefit from a direct access to trustful information sources about each specific topic, these experts have to be identified *a priori*. Information shared by the identified experts is typically used as a reference to judge the credibility of content produced by ordinary users (Wagner et al. 2012). The problem of these users identification in microblogs is generally resolved by analyzing Twitter lists or users interactions regarding a specific topic or domain of expertise. Xianlei et al. (2014) proposed a Gradient Boosted Decision Tree to identify domain experts in Sina Microblog over state-of-the-art and new linguistic features characterizing the activities and the content produced by each microblog user. This user activity-based identification approach has been significantly outperformed by Twitter lists-based identification approaches. By analyzing the various microblog user data (i.e. tweets, retweets, biography, lists and social connections), Wagner et al. (2012) proved that referring to user lists as a main criteria for user expertise evaluation outperforms tweets and retweets related features which were extensively explored in the literature. Such findings have been also confirmed by Ghosh et al. (2012) who proposed a new domain experts identification system –called Cognos– which mines the different information included in user lists. Cognos identification results outperformed standard identification systems focusing on user topical activities analysis (Xianlei et al. 2014, Pal & Counts 2011).

Topical authorities. refer to users sharing relevant information regarding a specific topic. These users are not necessarily identified *a priori*. They can be identified in real time according to the trending analyzed topics. Such users differ from popular, influencers and domain experts users. They can refer to ordinary microblog users who are interested in a specific trending topic or event like world cup event or I-phone launching topic. Identifying this category of users is trickier than it appears at first. Link analysis techniques are not suited for such task as they are time consuming and sensitive to popular and influencers users (Pal & Counts 2011). In this category, popular and influencers users, like CNN and BBC, reporting outdated information that have been already spread in the network have to be discarded. To the best of our knowledge, there is a single notable work which explored the problem of real-time topical authorities identification in microblogs (Pal & Counts 2011). Pal and Counts (Pal & Counts 2011) proposed a set of 15 features characterizing microblog users activity and connection in the network. These features include both topical

and nodal features which are computationally feasible in real time. Through this features-based user characterization, they classify and rank the different users using unsupervised classification and ranking techniques.

On-the-ground users. refer to microblog users who are geolocated in a specific area at a specific period of time. Such users are generally targeted by emergency management systems in order to gain a direct access to information shared from the affected or threatened area. While there are various works which have addressed tweets location prediction (Han et al. 2014, Mahmud et al. 2014), few works have addressed the problem of on-the-ground users identification. Starbird et al. (2011) explored a variety of flat profile user features and recommendation features for identifying these users. While flat features describe user's profile metadata, recommendation ones reflect how the other microblog users interact regarding the user's posts. For classifying the different microblog users based on these features, an SVM model was generated. Using this identification strategy, 68% of true on-ground users were identified. Similarly, Kumar et al. (2013) identify these users by categorizing the different Twiterrers interested in the disaster into four categories that are extracted according to both users' topical and location scores. Through this categorization, they targeted the category of users having both high topical and location scores.

2.4.2 Graph-based Microblog Users Classification

In order to identify key users –especially influencers, popular and domain experts users–, several works have explored the graph-based users representation and ranking approaches (Silva et al. 2013, Weng et al. 2010). In the following, we present at first the different graph-based user representation approaches. We then detail the graph analysis techniques implemented to detect such users from each appropriate graph-based representation.

2.4.2.1 Graph-based User Representation

Graph-based representation has long been utilized for information retrieval and ranking in web engines. However, using such representation for microblog users' classification has a much shorter history. The amount of work along this direction has exploded with the emergence of online social networks. Such representation was mainly exploited for friends or/and experts recommendation, predicting friendship links between users and influential users detection. In this subsection, we briefly survey related work conducted for key users' identification from microblogging platforms. We highlight the main graph-based representation approaches proposed for this purpose : *Followers graph*, *interaction graph*, *topic-sensitive graph* and *user-content graph* representations.

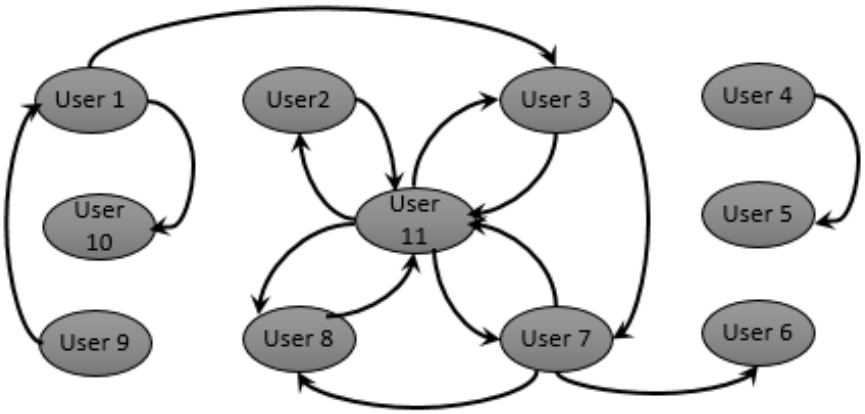

FIGURE 2.2: Followers graph-based user representation.

Followers Graph. A directed graph $G(V, E)$ is formed with the twitterers and the "following" relationships among them. V is the vertex set, which contains all the twitterers. E is the edge set. There is an edge between two twitterers if there is "following" relationship between them, and the edge is directed from follower to friend(Weng et al. 2010).

Figure 2.2 illustrates an example of followers graph connecting 11 twiterrers according to their following relationships. This followers graph representation is also known as social relation graph. This graph-based user representation was widely explored in the literature for popular and influencers detection. The different proposed influence (Weng et al. 2010) and popularity (Kwak et al. 2010) measures based on this representation will be described in the next section. These measures use various transition approaches in order to weight the different edges relating two users.

Interaction graph. is a directed graph in which nodes are individual user accounts. An edge is drawn from node A to B when either B retweets a message from A, or A mentions B in a tweet, with the weight of the edge representing the number of occurrences of the associated event (Ratkiewicz et al. 2011).

Microbloggers can also be linked according to their exchanged interactions in the network. These interactions can refer to the user retweeting activity regarding the other user original tweets, mentioning activity addressed to the user or/and replying activity regarding a particular tweets content. Such activities could be considered either separately or together while representing the user interaction graph (Conover et al. 2011, Cha et al. 2010). This category of graphs was widely explored for popular, influencers and domain experts' identification in microblogs (Ratkiewicz et al. 2011, Subbian & Melville 2011).

Topic-sensitive graph. is a bipartite graph $G(U, C, F, E)$, where U is the user set, C is the content set, and E and F are sets of edges that associate users to content and the other way around, respectively. For each user $u \in U$ and piece of content $c \in C$, there is a directed

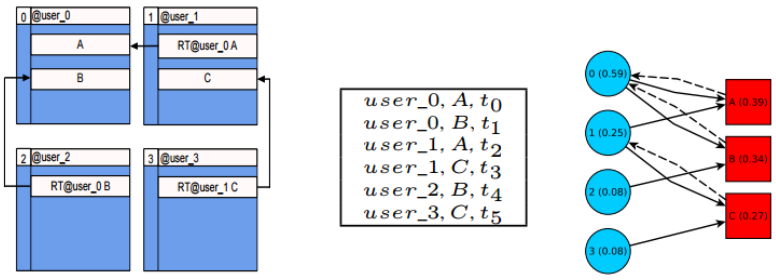

FIGURE 2.3: Topic-sensitive graph representation. Circles and squares refer respectively to users and content (Silva et al. 2013).

edge $(u, c) \in E$ if the user u has created or propagated the content c and a directed edge $(c, u) \in F$ if u created c (Silva et al. 2013).

An example of topic-sensitive graph is represented in Figure 2.3. This graph representation approach was introduced by Silva et al. (2013) in order to detect influential users. This representation is proposed under the assumption that a microblog user u is influential to v if u creates content which is relevant to u. Analyzing both the user influence and the content relevance for influential identification leaded to promising results outperforming standard followers or interactions graph-based user classification (Silva et al. 2013).

2.4.2.2 Graph Analysis Techniques for Key Users Identification

Identifying key microblog users in large scale networks remains a big challenge. Microblog networks connect a huge number of users providing millions of contents daily. To address these identification challenges, known node centrality measures and diffusion-based processes, which have proved their efficiency in the context of standard complex networks, have been explored for this task. Inspired by these existing measures, recent researches proposed extended measures adapted to the context of key users identification in microblogs. In the subsections, we describe the different analysis techniques explored in the literature for key users' identification.

Graph Analysis using Centrality Measures

Key users identification problem in microblogs is generally casted into a problem of central nodes identification in complex networks. Most research work proposed to deal with microblogs network problem using standard centrality measures. In the following, we list most of the standard and new proposed centrality measures that have been explored to analyze the microblog network structure for key users' identification.

Indegree centrality. is defined as the number of ties incident upon a node. That is, it is the sum of each row in the adjacency matrix representing the network (Borgatti 2005).

This measure was widely explored in the context of key users identification in microblogs, especially for popular and influent users detection (e.g. twitterholic[20], wefollow[21]) (Cha et al. 2010). However, according to the various comparison studies conducted for influential and popular users detection. The performance of this measure typically registers lower precision results than those obtained by the more advanced measures such as FollowRank and eigenvector centralities described in the following (Cha et al. 2010, Weng et al. 2010).

Betweenness centrality. is defined as the number of times that a node i needs a node k (whose centrality is being measured) in order to reach a node j via the shortest path (Borgatti 2005). If g_{ij} is the number of geodesic paths from i to j, and g_{ikj} is the number of these geodesics that pass through node k, then the betweenness centrality of node k could be computed as follows :

$$Bt = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}, i \neq j \neq k \tag{2.1}$$

This measure was mainly explored in the context of influential users identification. Wu et al. (2012) used betweenness centrality as a proxy of attractive and potentially influential users in order to make recommendations of future retweet and future mentioned users. The efficiency of this measure for influential identification was also compared by eigenvector centrality measures which remain the most adapted ones for this task.

Eigenvector centrality. is defined as the principal eigenvector of the adjacency matrix defining the network. The idea is that even if a node influences just one other node, who subsequently influences many other nodes (who themselves influence still more others), then the first node in that chain is highly influential (Borgatti 2005). The variants of eigenvector centrality measures such as HITS and PageRank were widely explored in the context of key users identification in microblogs. These measures have been categorized as the most suited measures for the task of influential and domain experts' identification (Weng et al. 2010, Cappelletti & Sastry 2012).

FlowRank. is defined as the ratio between the number of one's followers and the number of his/her friends (Weng et al. 2010). This measure was mainly used for popular users' identification. According to literature studies evaluating the efficiency of centrality measures, identification models based on this measure outperform those measuring popularity based on the user indegree centrality (Cha et al. 2010).

TunkRank[22]. is an extension of PageRank, calculating users' influence recursively by taking into account retweeting probability and the distribution of attentions. User influence is

[20] http://twitterholic.com/
[21] http://wefollow.com/
[22] http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

measured using the following equation :

$$Influence(X) = \sum_{Y \in Followers(X)} \frac{(1 + p * Influence(Y)}{||Following(Y)||} \tag{2.2}$$

Where p is the constant probability that X followers retweet a tweet.

These presented centrality measures only capture some aspects of the user position in the network. In order to take advantage from various centrality measures, few works proposed to combine all these individual measures to use them as inputs to a classification or clustering algorithm (Kayes et al. 2012, Subbian & Melville 2011). Key users identification models can either be trained *a priori* in order to be able to predict new key users based on their centrality scores or processed in real time using unsupervised rank aggregation algorithms. Subbian & Melville (2011) identified influential users using a supervised Kemeny ranking algorithm driven from social choice theory. Through this rank aggregation approach, it has been proved that combining aspects of different centrality measures ensures more effective results for influential users identification.

Graph Analysis using Clustering and Classification Algorithms The proposed graph analysis techniques based on centrality measures have mainly explored the user social position regarding the network structure. Although these measures have proved their efficiency for central and popular users' detection in classic networks, such measures still not adapted enough for key users identification in microblogs. Microblogs are richer than simple follower and following link relating users. Microblogs have many specificities (e.g. topics, interactions, etc.) that have to be integrated in order to enrich the user representation in the social graph. As stated in the previous subsection, a microblog network can be also represented through user interaction or topical-sensitive graphs. In order to explore the rich information characterizing users in microblogs, many works have adapted the standard centrality measures –mainly the diffusion-based ones– such as PageRank and HITS algorithms based on the new proposed user representation graphs (Silva et al. 2013, Romero et al. 2011).

TwitterRank (Weng et al. 2010). Inspired by the PageRank measure, Weng et al. (2010) proposed a topic-sensitive ranking measure TwitterRank measuring the microblog user influence based on the social graph representation. TwitterRank computes each user topical influence score by performing a topic-sensitive surf between user nodes. This random surfer visits each node based a transition matrix P_t computed for each topic t.

TwitterRank Transition Matrix. Given a topic t, each element of matrix P_t, i.e. the transition probability of the random surfer from follower s_i to friend s_j , is defined as :

$$P_t(i,j) = \frac{|T_j|}{\sum_{a:s_i \text{ follows } s_a} |T_a|} * sim_t(i,j) \tag{2.3}$$

Where $|T_j|$ is number of tweets published by s_j , and $\sum a : s_i$ follows $s_a|T_a|$ sums up the number of tweets published by all of s_i's friends. $sim_t(i, j)$ is the similarity score between s_i and s_j in topic t. TwitterRank vector for each topic t is thus computed as follows.

$$\vec{TR}_t = \gamma P_t * \vec{TR}_t + (1 - \gamma)E_t \tag{2.4}$$

Where E_t is the teleportation vector defined. γ is a parameter between 0 and 1 to control the probability of teleportation

IP-influence (Romero et al. 2011). Is an extension of the HITS algorithm. While the HITS algorithm computes the authority score for each page and the hub score for links relating webpage, the IP-algorithm considers the authority and the hub score to measure the user passivity and influence based on a weighted interaction graph. A user's influence score depends on both the number of users influenced as well as those who remain passive. A user's passivity score depends on the influence of users who have seen the user tweet content and have not been influenced. These scores are simultaneously computed by considering other properties of the network such as the acceptance and rejection rates.

Acceptance rate. *This value represents the amount of influence that user j accepted from user i normalized by the total influence accepted by j from all users in the network.*

$$u_{ij} = \frac{w_{ij}}{\sum_{k:(j,k)\in E}(w_{kj})} \tag{2.5}$$

Rejection rate. *Since the value* $1 - w_{ij}$ *is amount of influence that user i rejected from j, then the value* $v - ji$ *represents the influence that user i rejected from user j normalized by the total influence rejected from j by all users in the network.*

$$v_{ji} = \frac{1 - w_{ij}}{\sum_{k:(j,k)\in E}(1 - w_{jk})} \tag{2.6}$$

ProfileRank (Silva et al. 2013). is an extension of the PageRank algorithm. This algorithm follows the random surfer idea adopted by PageRank algorithms in order to measure the influence of each user and tweet represented in topic-sensitive graph. Starting from a random user, this surfer keeps on clicking on the user tweet and retweets at random. By clicking on user retweeted information, the surfer would be redirected to the user profile of the original author who has produced the retweeted information. The relevance of each particular tweet represented by a square, is measured according to the relative frequency that the random surfer clicks on this tweet or its attached retweet. The user's influence is measured according to the frequency that the random surfer clicks on the user's profile.

2.4.3 Vector-based Microblog Users Classification

While graph-based microblog users' classification techniques have led to good results, such techniques remain sensitive to popular and well-connected users in the network. Processing these techniques in real time is still complex due to the huge amount of user connections that have to be analyzed. To address this problem, few works proposed a vector-based classification approach consisting of classifying and modeling users mostly according to their behavior. User's behavior was generally represented by various features computationally feasible in real time. Such features were proposed by considering each user social connections, behavior and topical interest. By representing microblog users through feature vectors, supervised or/and unsupervised machine learning algorithms for identification key users can be applied. In the following, we describe at first the different explored features in the literature for microblog users modeling. Then, we list the proposed classification and ranking techniques for key users' identification.

2.4.3.1 Microblog Users Features

In order to differentiate between key users and ordinary ones, each user particularities have to be highlighted. Various features characterizing microblog users specificities have been studied in the literature. In the following we focus on describing the main features proposed in the context of key users identification problem. We split these features into five categories : *user activities features*, *topical features*, *profile features*, *network structure features* and *Twitter lists features*.

User activities features "How you tweet?". These features characterize the user tweeting behavior in microblogs. They are typically represented by statistics of the on-topical productivity of each user according to the nature of his/her shared tweets. The different nature of tweets (i.e. original tweets, retweets and mentions) are considered to reflect the user behavior tendencies. These features can be used in their raw form (e.g. number of user original tweets, number of user received mentions,etc.) or engineered form. Engineered features are generally computed by combining the different raw features into a more descriptive form reflecting the real behavior of users. A large set of user activities features were proposed by Pal & Counts (2011) and Xianlei et al. (2014). These features are described in detail in Chapter 4 where a comparative study is conducted for the evaluation of the state-of the art features effectiveness. In the following, we describe two engineered features proposed by Pal & Counts (2011) for topical authorities' detection :

Retweet impact. *indicates the impact of the content generated by the author. This definition of RT3 ensures that we dampen the impact for an author who has few overzealous users retweeting her content a lot of times* (Pal & Counts 2011).

$$\text{Retweet impact} = RT2 * log(RT3) \tag{2.7}$$

Where $RT2$ and $RT3$ refer respectively to the number of unique tweets retweeted by other users and the number of unique users who retweeted author's tweets.

Signal Strength. *indicates how strong is author's topical signal, such that for a true authority this value should approach 1* (Pal & Counts 2011).

$$\text{Signal strength} = \frac{OT1}{OT1 + RT1} \tag{2.8}$$

Where $OT1$ and $RT1$ refer respectively to the number user's original tweets and the number of retweets of other's tweets.

Topical features "what you tweet?". These features point out the user lexical usage and the main topics the user is interested in. In a general context, several techniques have been studied for tweets topic modeling like bag-of-words, TF-IDF and Latent Drichlet Allocation (LDA) (Wang et al. 2012, Xianlei et al. 2014). These techniques were explored in various tasks such as news detection, friends' recommendation and sentiment analysis (Mehrotra et al. 2013, Hong & Davison 2010*a*). While these standard text mining tools have proved their efficiency for topic modeling of long standard documents, the performance of such techniques would erodes if they are applied for short documents like tweets (Luo et al. 2012). Thus, many aggregation strategies have been considered in order to combine the different tweets responding to a well-defined criteria into a single document. In the context of key users identification, a user stream tweets could be aggregated in order to construct a single document covering all user's tweets (Sasaki et al. 2014). Identifying topics of interest from these aggregated tweets is generally conducted using LDA. Each microblog user represented by the collection of his/her produced tweets is associated with a multinomial distribution over a defined set of topics.

Hong & Davison (2010*b*) have discussed several topic modeling schemes that can be explored for this task. Xianlei et al. (2014) adopted a similar strategy introduced by Hong & Davison (2010*b*) study for modeling each microblog user by a topic vector characterizing his/her topical interest. They have experienced the LDA model in the Chinese microblog Sina[23], similar to Twitter, for domain experts' identification. Their model characterized the user topical interests by highlighting the nature of their provided messages (i.e. original, reply and conversation microblog).

Profile features "who you are?". Convert the user interface information –like name, age, location, number of user tweets from the creation of his/her account, and short summary

[23] http://weibo.com/

of interests– into a numeric values. The profile information is generally shared publicly and is accessible using open Twitter APIs. Features characterizing this information provide a digital overview of the user interests– except the location feature–. These feature were rarely explored in the context of key users' identification. Pennacchiotti & Popescu (2011) extracted user profile features in order to classify twitter users in a general context. Xianlei et al. (2014) explored these features in order to identify domain experts in Sina microblog.

Network structure features "who is seeing your tweets?". describe the user social connections in the microblog by analyzing the microblog structure. These features do not principally characterize the statistics of the user followers and followees, they can also refer to the statistics of users follower and followees meeting certain criteria. These criteria could be related to the users interest similarities (Pal & Counts 2011) or to their followers and followees interaction regarding his/her posts (Pennacchiotti & Popescu 2011). In the following, we list one of the structure engineering feature proposed by Pal & Counts (2011) for topical authorities' identification :

Network score. *consider the raw number of topically active users around the author (Pal & Counts 2011).*

$$\text{Network score} = log(G1 + 1) - log(G2 + 1) \tag{2.9}$$

Where $G1$ and $G2$ refer to the number of topically active followers and the number of topically active followees respectively.

List features "who you are regarding communities perception?". characterize the description of the different lists to which the user belongs. These features have been rarely explored in the literature. To the best of our knowledge, there is only one work which has explored the use of these lists for key users identification and more precisely for domain experts identification. This Twitter-lists-based model proposed by Ghosh et al. (2012) outperformed the most efficient baselines that have referred to various features analyzing tweets contents, activities and the network structure. The strategy of analyzing Twitter lists consists of extracting frequent topics (words) which describe the domains of interest of each user. The assumption behind this strategy is that a user assigned by many other users in various lists covering the same topic, is very likely to be expert on this particular topic.

2.4.3.2 User Activities Classification Techniques

In order to classify and rank the different users represented by a vector of features, supervised and unsupervised machine learning algorithms are generally explored. Supervised

algorithms learn to differentiate between feature vectors characterizing target and non-targeted users using training data. Unsupervised techniques learn to cluster the different vectors according to their similarities. In the following, we detail the supervised and unsupervised techniques explored in the literature for microblog users' categorization.

Supervised techniques. supervised machine learning algorithms were rarely explored in the context of key users' identification. Using these algorithms, the problem of key users' identification is cast into a binary classification problem. To the best of our knowledge, there have been only one work which has explored supervised classification techniques for key user's identification based on the vector-based user characterization approach (Xianlei et al. 2014). Xianlei et al. (2014) classify domain experts using the Gradient Boosted Decision Trees (GBDT) based on profile, tweeting activities, topical and network structure raw features. The resulted model has been compared to other baselines learned using SVM. According to their obtained results, they showed that GBDT outperforms SVM in terms of both run time and efficiency.

Unsupervised techniques. Clustering and similarity measuring approaches were also processed for key users identification in microblogs. Pal & Counts (2011) processed a Gaussian Mixture Model to cluster users into two clusters based on the vector-based user characterization approach. User vectors are composed of a set of rich raw and engineered features characterizing user social position and behavior in the microblog. The clustering approach was conducted in order to eliminate most of the non-topical authorities' users. Users selected in the retained cluster are then ranked using a Gaussian ranking algorithm. Ghosh et al. (2012) explored the Twitter list-based user representation for topical experts ranking. The topic vector (t_i, f_i) representing each user, where f_i is the frequency of occurrence of topic t_i, is compared with other users topical vectors given a specific query. This comparison is conducted by computing a topical similarity score using the cosine similarity on TF-IDF based representation. The final user similarity score is obtained by multiplying this topical similarity score by the logarithm of the number of Lists referring to the expertise of this user. By sorting to the similarity score of each user given a query (topic), topical experts are identified. The identification model proposed by (Ghosh et al. 2012) has yielded better results than the model proposed by (Pal & Counts 2011) for domain experts identification. Domain experts are generally assigned to various lists targeting topics related to their domain of expertise.

2.4.4 Discussion

As described in this section, microblog key users identification is generally processed by graph-based user classification technique or features-based one. Table 2.4 summarizes the classification dimensions explored by each technique. Examples referring to key microblog

user's detection systems presented in the literature are also cited according to their explored classification dimensions. In the following, we discuss these different techniques and their special strengths and weaknesses for detecting each key microblog users' category.

- Graph-based classification. consists of analyzing the graph representation in order to detect the targeted key users' category. The users graph representation could either reflect the users position regarding the network structure (i.e. followers and followees connections) or the topical interaction introduced by- or/and intended to-each user. Aside from considering the "user to user" relations, the graph representation can be modeled in order to connect the different users regarding their shared information content. This "user to content" graph is known as the topical-sensitive graph. These different graph representation techniques were mainly analyzed using the standard centrality measures for popular, influencers and domain experts' detection. While such graph-based user representation and classification techniques have yielded promising results in the context of influencers and domain experts identification, they are still unsuitable for crisis situations requiring a real-time identification process. Moreover, mainly referring to the network structure and user interactions information makes the model sensitive to users having a central social position or/and a high activity in the network.

- Features-based classification. consists of representing microblog users using a vector-based representation. Through this representation, users are evaluated according to their characteristics extracted using different features computationally feasible in real time. Several features have been explored for this purpose. As presented previously, we split these features into four categories : profile features, user activities features, social features and Twitter lists features. These features have been mainly explored for the detection of topical authorities and domain experts. Considering all these features simultaneously while representing the user behavior may either erode or improve the detection results. The effectiveness of these features depends on the targeted users specificities. However, most of the state-of-the-art works have selected the user modeling features without any study evaluating their efficiency in the targeted context. Such step is important in such cases as the identification model performance is directly associated with the user modeling approach efficiency. The better the user modeling approach is, the easier the identification of targeted users is.

Table 2.4: State-of-the-art key users identification techniques in microblogs : Advantages and drawbacks. Examples of targeted key users by each existing technique are also specified.

Key Users Identification Techniques	Description/Examples	Advantages	Drawbacks
Graph-based Classification			
User Modeling Graphs :			
Followers Graph	Users are represented according to their social connections extracted from the list of their followers and followees. **Examples :** influential users detection (Weng et al. 2010) and popular users identification (Kwak et al. 2010)	-Eases the detection of popular and central users. -Eases the identification of users communities. -Details the different direct and indirect connections between any users.	- Unsuitable for influencers, domain experts and topical authorities detection. - Sensitive to popular users or spam accounts having many relations in the network. - Computationally infeasible in real time.
Interaction graph	Users are evaluated according to their received and shared interactions with the other users in the network. These interactions include shared original tweets, retweets, mentions and comments. **Examples :** information spreaders detection (Ratkiewicz et al. 2011), influencers prediction (Subbian & Melville 2011)	-Eases the identification of information sources. -Details the information spread in the network. -Highlights influential users in the network.	-Unsuitable for domain experts and topical authorities detection. -Favors users provoking huge interactions regarding their tweets independently of their content. -Computationally infeasible in real time.
Topic-sensitive graph	Users are represented according to the information content they are pointing to. The relation between users is not modeled explicitly like in the interaction and follower graph. Only the relation between users and content is highlighted. **Examples :** Influential users detection (Silva et al. 2013)	-Eases the detection of both relevant users and relevant content. -Analyzes users according to the relevance of their content independently of their social position in the network.	-Neglects many user features that can lead directly to influent users. -Over-complicates the user influence measuring process.

Vector-based Classification			
User Modeling Features :			
User activities features	Users are characterized according to their topical activities in a specific period of time. These activities typically cover original tweets, retweets, mentions and comments activities. **Examples :** Topical authorities detection (Pal & Counts 2011), domain experts detection (Weng et al. 2010).	-Reflects the user implication regarding a specific topic. -Characterizes the different nature of interaction conducted by each user over time. -Points out many sub-metrics that can ease the differentiation between key users and those who are not.	-Tends to promote users having a high topical activity even if they are sharing outdated information. -Makes the model sensitive to popular and influential users highly active in the network but providing irrelevant or outdated information such as news outlets channels.
Profile features	Users are characterized according to their profile information generated automatically by the microblog or set manually by the user such as their location and biography. **Examples :** Domain experts identification (Xianlei et al. 2014), on-the ground users detection (Starbird et al. 2011).	-Highlights the global image of each user in the microblog. -Provides a statistical description of the user activity from the creation of his/her microblog account.	-Are not enough strong to point out the differences between key users and those who are not.
Network structure features	Users are represented according to their topical social connections in the network. Both raw or/and engineered features are generally extracted to take advantage from the user social connections information. **Examples :** Topical authorities detection (Pal & Counts 2011).	-Reflects the user connections using metrics computationally feasible in real time. -Highlights the user social position in the network regarding the specific analyzed topic.	-Makes the identification model sensitive to popular users having many connections.
List features	Users are characterized according to the Twitter lists they belong to. The short description included in these lists are explored in order to point out the topical expertise of each user. **Examples :** Cognos domain experts identification model (Ghosh et al. 2012)	-Ensures the detection of domain experts in real time. -Characterizes the different topical expertise of each user. -Insensitive to user social connections and popularity.	-Unsuitable to detect ordinary key users relative to a new trending crisis event. These features are more adapted for domain experts' detection.

According to these comparisons, we can note that there are various key users detection approaches that have not yet been explored in the context of crisis events. The graph-based and vector-based user classification techniques have been cursorily considered in the context of crisis events. By studying the advantages and drawbacks of the existing identification techniques presented in the literature, we perceived that there are various identification dimensions that can be explored in the context of crisis events.

Graph-based ranking technique proposed in the context of crisis events can be enriched by exploring the different user interactions and topical interests specificties. However, as mentioned in the previous section such technique remains unsuitable for a real-time detection process. It is mainly adapted for an *a priori* detection of specific key users, who remain important regarding a specific topic or domain over time, such as domain experts. However, for the detection of on-the ground or prominent ordinary users during crisis events, the identification process has to be ensured in real time. Vector-based key users identification techniques proposed in a general context cover various features that seem to be adapted to the context of crisis events. However, the current proposed combination and integration form of these features for both users representation and classification hide many important factors that could lead to a better identification of prominent users in the context of crisis events. This thesis proposes a rich key users identification approach characterizing and evaluating users based on a set of new features adapted to the crisis events context. Such approach explores new dimensions that are not covered neither by the vector-based users identification techniques nor the graph-based ones.

Chapter 3

MASIR: A Modular Multi-agent System for Information Extraction and Retrieval from Microblogs

Contents

3.1 Introduction

Popular microblogging platforms, like Twitter, are always crowded during major events. The number of shared tweets regarding an emerging crisis event can easily reach tens of thousands in few minutes. While this large number of shared tweets can provide valuable information for situation awareness during crisis events, it also makes the access and retrieval of such information challenging. Twitter APIs provide access to a limited amount of data (i.e. around 1% of data can be extracted through these APIs). Accessing to this real world microblogs data has become a constant hindrance for both researchers and organizations. Scientists typically need to collect historic data for learning and testing new research models which are able to point out the wealth of information behind these microblogs. Organizations need to access in real time the required information in order to have an overview of what is happening during major events. However, such needs remain unmet. How could we access the various historic data shared in microblogs? How could we retrieve the required relevant information in real time for any given topic or event?

To the best of our knowledge, there are no available open Twitter data or crawling systems providing or enabling access to real-time or historic Twitter data suitable to the problem of key microblog users identification. Data that has been explored for such identification problem typically covers either social connections characterizing the evaluated users relationships or tweets shared regarding a specific analyzed topic. Information describing the behavior of users complying with certain criteria is generally neglected. Such neglected information can point out the behavioral patterns specific to the targeted key users.

In this chapter, we propose a modular Multi-Agent System for Information extraction and Retrieval (MASIR). This system responds to research data needs in the context of key users identification in microblogs during crisis evens. MASIR collects not only most of tweets shared regarding a specific topic or event but also most of the characteristics and activities of such information providers. It also supports key users identification and tracking in real time during specific events. MASIR is based on a distributed architecture integrating various agents with different roles. These agents can be adapted for both historic and real-time Twitter data crawling. In Section 3.3, we describe how MASIR historic data extraction module ensures an intensive data extraction. In Section 3.4, we detail the different functions integrated in the MASIR key users tracking module in order to gain a real-time access to the relevant information shared in microblogs. We evaluate the efficacy and efficiency of the two presented modules of MASIR during crisis events in Section 3.5.

3.2 Research Questions

In this research, we propose a modular-agent system, named MASIR, for extracting historic information regarding interested users in a specific event on one hand and tracking the selected key users in real time on the other hand. To the best of our knowledge, MASIR is the first system providing an extraction architecture which enables the identification and tracking of key microblog users in real time using public Twitter APIs. This system integrates various collaborative agents with different roles and goals. This research explores multi-agents flexibility to answer the following questions:

1. How to build a scalable architecture supporting the identification and tracking process of key microblog users in real time?
2. How to explore standard microblog APIs in order to be able to access both real-time and historic data?
3. How to manage the extraction and tracking modules in order to avoid IP banning and boost the limits imposed by Twitter?

3.3 MASIR for Boosting Historic Data-Access

In order to collect most of information shared during specific events, we propose *a historic data extraction module*. The main purpose of this model is to extract historic information shared by users who have interacted at least one time regarding the event. By following this extraction strategy, we aim to extract most of data that can help researchers to test and explore new key users identification approaches. Such collection strategy ensures the collection of social graph data relative to event-related information providers on one hand and information providers produced content on the second hand. In the following subsections, we describe how MASIR processes the collection of this required data and how it deals with the imposed restrictions of Twitter APIs. The MASIR architecture for historic data extraction is introduced in subsection 3.3.1. The role of the different agents integrated in this architecture is described in subsection 3.3.2.

3.3.1 MASIR Crawling Principle

MASIR collects the shared information relative to a specific event or topic by crawling only the profiles of users who have shared at least one tweet regarding the analyzed topic or event. The MASIR extraction module is executed in two steps. The first step consists of identifying any microblog users interacting regarding the event. This first identification

step is processed by extracting the identity of any user providing at least one information regarding the event. The second step consists of crawling the profiles of users identified in the first step. Any information shared by the identified users has to be extracted even if it is not about the analyzed event. The idea behind storing all users shared information consists of having a complete view of the user behavior during specific events.

Figure 3.1 describes the decentralized structure of the parallel historic crawling module integrated in MASIR. This crawling module is composed of 3 different kinds of agents (i.e. Stream Retrieval Agent, Historic Listener Agents Manager and Historic Listener Agent) designed to execute well-defined related tasks. The crawling process starts when the keywords and/or hashtags describing the targeted event were specified to the Stream Retrieval Agent (SRA). Using these parameters, SRA searches for the list of new users sharing real-time information about the event. Once we declare the event end, SRA sends the list of users who have interacted regarding the event to the Historic Listener Agents Manager (HLAM). HLAM assigns a Historic Listener Agent (HLA) to each identified user by SRA in order to extract and store his/her profile information (e.g. biography, followers, followees, etc.) and all his/her shared tweets from the beginning of the event until its end. The collected information is then stored in the Historic and Social Information Base (DB2). HLAs are processed in parallel in different containers and hosts as described in Figure 3.1. The role and specificities of these agents are described in detail in the following sub-sections.

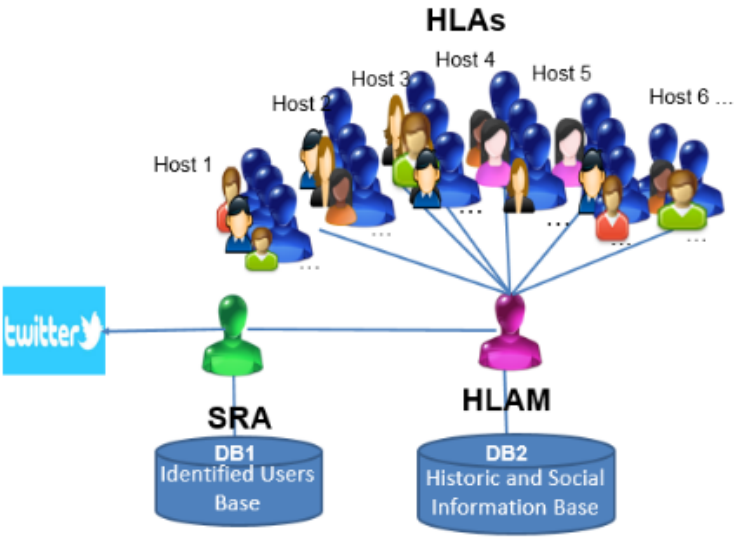

FIGURE 3.1: MASIR historic data extraction module

3.3.2 MASIR Crawling Agents Role

As described previously, the different agents integrated in the historic data extraction module are complementary. Each agent is indispensable to efficiently complete the crawling task. The parallel and distributed crawling architecture maintains the continuity of the crawling process even if one of the crawler agents has been stopped or blocked by Twitter. In the following, we describe the main characteristics of these agents and their roles.

3.3.2.1 The Stream Retrieval Agent (SRA)

SRA retrieves the tweets published in real time about the specific analyzed event and extracts the identities of users who are sharing it by following these monitoring operations:

1. *Streaming search:* SRA remains connected to Twitter during the event in order to search in real time for new tweets using the assigned hashtags or keywords characterizing the targeted event.
2. *Users' identification:* SRA extracts the identity of users sharing on-topic tweets.
3. *Users' filter:* SRA dynamically applies new filters by making reference to the Identified User Base (DB1) in order to force the streaming search to retrieve only tweets shared by new users.
4. *Users' storage:* SRA stores in DB1 the identifier of any new detected user posting information related to the event.
5. *List of users sending:* SRA has to send the list of detected users by the end of the event to HLAM.

3.3.2.2 The Historic Listener Agents Manager (HLAM)

HLAM manages the extraction process of the social and historic information from the identified users profiles. It controls multiple HLA agents which are in charge of the historic and social information extraction from the profile of each identified user. HLAM may undergo different transitions according to the following processed operations:

1. *Users assignment:* When HLAM receives the list of users sent by SRA, it adds the new identified users in a waiting list. It then assigns each user to one of the available HLAs by respecting the FIFO (First In First Out) principle.

2. *Information reception:* This operation is processed after the reception of a message from a HLA precising that the historic extraction process was successfully accomplished. HLAM then stores the returned information collected by this HLA in DB2.

3. *HLA status change:* Once HLAM has received all the extracted information from a HLA, it sets this HLA status to "free" in order to be able to assign it to other users.

3.3.2.3 The Historic Listener Agents (HLAs)

HLAs have to extract historic information shared by each assigned user. Once a HLA has finished the extraction of the needed information belonging to a specific user, it sends a message to HLAM to store the collected information in DB2. Then, the HLAM will change this HLA status to "free". Each HLA has to be able to process the following operations:

1. *Receiving a user's identity:* When HLA status is set to "free", HLA could be assigned to a unique user recognized by his/her unique identifier.

2. *Historic information extraction:* HLA extracts all the historic information shared by the assigned user from the beginning of the analyzed event until its end.

3. *Social information extraction:* HLA extracts the followers and followees list associated to the assigned users.

4. *Extracted information Sending:* HLA sends all collected information to HLAM in order to store it in DB2.

3.4 MASIR for Real-time Tracking of Key Microblog Users

In order to gain real-time access to the relevant and valuable information shared during specific events, we integrate a real-time tracking module complementary to the MASIR data extraction module. This module is in charge of analyzing users historic data extracted by the MASIR data extraction module in order to identify and track the most prominent microblog users. The idea behind the identification and tracking of key users consists of gaining a direct real-time access to the relevant and exclusive information. Based on the already extracted historic information, any identification approach can be processed for key microblog users detection. In this section, we propose a straightforward identification approach to test our tracking model efficiency. More efficient identification approaches are experimented in the next chapters. The proposed tracking module is designed to ensure real-time access to any information shared by top key microblog users. This module thus integrates adapted functions coping with Twitter APIs limitations for real-time tracking.

The architecture of this additional retrieval module and the roles of its integrated agents managing the key user's selection and tracking processes are described in subsections 3.4.1 and 3.4.2 respectively.

3.4.1 MASIR Tracking Principle

MASIR integrates a tracking module consisting of selecting and tracking the most key microblog users in real time during events. While the historic data extraction module ensures the extraction of the historic of any user interested in the analyzed event independently of their prominence, the tracking module analyzes this extracted information in order to select and track only top key users information which are shared in real time. This module ensures real-time access to any information shared by the selected key users who are susceptible to share the required valuable information. Twitter limits users streaming tracking from 3 to 15 users. The architecture of this module boosts the number of tracked users in real time during real world events cases. The parallel and distributed tracking architecture of MASIR encounters the Twitter limits and ensures a parallel processing of the extraction, the analysis and the crawling processes. This process parallelization guarantees real-time identification access to the different information shared by most key users.

Figure 3.2 describes how the key users tracking process is managed in order to ensure real-time access to the relevant information shared during events. The tracking process module is mainly composed of 3 different kinds of agents (i.e. Key Users Detector, The Stream Listeners' Agents Generator and Stream Listener Agents) communicating with the historic data extraction module through the HLAM agent. In order to ensure the real time function of these two compliant modules, the first module is managed to extract the historic of users detected by SRA over time. SRA has to send continuously the detected users list to HLAM after each 30 seconds. This real-time crawling is processed in order to ensure the analysis of the user historic over time during the analyzed event. The resulted data is continuously analyzed by the Key Users Detector (KUD) agent which is in charge of the identification of key microblog users. The detected key users are then tracked in real time by the Stream Listener Agents (SLAs). SLAs are generated and manged by the Stream Listeners' Agents Generator (SLAG). These different agents are described in detail in the following sub-sections.

3.4.2 MASIR Tracking Agents Role

The agents integrated in the crawling and tracking modules are both indispensable for the identification and tracking of key microblog users. In the following, we describe the role of the different agents integrated in the identification and tracking module. This module is

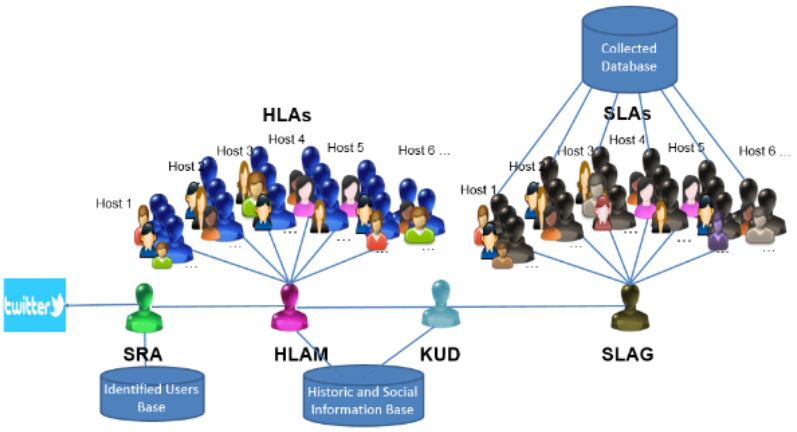

FIGURE 3.2: A decentralized multi-agent system for real-time information retrieval from Twitter

activated once the agent KUD has received a signal from HLAM mentioning the update of some users information. We detail below the main roles of these agents starting from the agent ensuring the selection of key users until the agent ensuring the users tracking in real time.

3.4.2.1 The Key Users Detector (KUD)

KUD acts as the intermediary between the historic data extraction process and the streaming data tracking process. This agent detects key users with reference to the data collected during the historic data extraction process. The identification of these key users optimizes the tracking process by assigning the limited number of parallel SLAs only to microblog users who have proved their prominence. The detection of such users can be processed using any identification technique. For the evaluation purpose of the MASIR proposed architecture, we propose a straightforward key users detection approach. This approach consists of detecting key microblog users by calculating and updating periodically the Prominence Score (PS) of the already watched users. This identification approach is processed by the agent KUD. KUD estimates the final prominence score of each user according to his/her geo-location and social positions on one hand and the recency of his/her first provided event-related information on the other hand. PS is computed using the following ranking model :

$$PS(u) = w_1 * RS(u) + w_2 * GPS(u) + SPS(u) \tag{3.1}$$

Where w_1 and w_2 reflect the importance of RS and GPS and are set to 0.38 and 0.02 respectively. The sum of the weights' values (from w_1 to w_6) need to be 1. The w_3 to w_6 weights attached to the SPS score formula are described in Equation 3.4. All these weights were estimated *a priori* through a user study evaluating the active Twitter users in the

South Korea ferry disaster. This study was conducted by a group of volunteers who have evaluated the Twitter users according to the relevance and recency of their information about the disaster. These volunteers have rated these users from 1 to 10 according to their prominence. These rates were retained for fitting a linear regression model composed of the different predictor scores (i.e. RS, GPS, SPS predictors) proposed to evaluate the prominence of each user. The weights evaluating each predictor were normalized to form the sum 1 for all the weights.

The Recency Score (RS) indicates the recency of the user's first shared event-related-topic information. To compute this score, the time of share of this first on-topic tweet (t_{on}) is compared with the time of the event occurrence (t_{event}). The difference in time between t_{on} and t_{event} is measured in minutes.

$$RS(u) = \frac{1}{t_{on} - t_{event} + 1} \tag{3.2}$$

The Geo-location Position Score (GPS) indicates the inclusion rate of the geo-location (i.e. longitude, latitude) specified by the user in the event area. The event area is represented by a polygon or a set of polygons (Pe) that may include many distant zones. For each user u, we extract from his/her different historic tweets collected by HLAs the set of his/her geo-locations (Cu). For example, if all the geolocations specified by the user are included in the event area, his/her GPS score will be set to 1.

$$GPS(u) = \frac{Cu \cap Pe}{Cu \cup Pe} \tag{3.3}$$

The Social Position Score (SPS) indicates how much the user's followers (F) and followees (Fe) are interested in the analyzed event. The higher the RS score of the evaluated user's *on-topic followers* (OnF) and *followees* ($OnFe$) is, the more important the user's social position is. As well-connected users such as CNN and BBC would have a large number of OnF and $OnFe$ due to their celebrity, we adjust their on-topic social connections statistics by the total number of their *followers* (F) and *followees* (Fe). Through this adjustment, the SPS score would be insensitive to well connected users. SPS is computed as follows using the social information already extracted by HLA and stored in DB2 :

$$SPS(u) = w_3 * \frac{\sum_{i=1}^{OnF} RS(i)}{log(OnF+1)} + w_4 * \frac{OnF}{log(F)} + w_5 * \frac{\sum_{i=1}^{OnFe} RS(i)}{log(OnFe+1)} + w_6 * \frac{OnFe}{log(Fe)} \tag{3.4}$$

Where $w_3 = 0.21$, $w_4 = 0.1$, $w_5 = 0.23$ and $w_6 = 0.04$ are the weights reflecting the importance of the different predictors comprised in the SPS score of each user.

3.4.2.2 The Stream Listeners' Agents Generator (SLAG)

SLAG manages the tracking process of the identified key microblog users during the event. It starts the agents' generation and management process when it receives the list of selected key users by KUD. SLAG generates one SLA for each user in the list. These SLAs are generated in different hosts in order to avoid the risk of IP banning by Twitter. The following operations are processed by SLAG :

1. *Receiving detected users :* SLAG receives periodically an updated list of key users that have to be tracked in real time.
2. *Killing existing SLAs :* After receiving the updated list, SLAG kills SLAs which are tracking users who are not mentioned in the new list. By killing these SLAs, SLAG will release the place in some hosts in order to be able to track the new key users.
3. *Generating a new SLA :* Once there is free hosts that can be assigned for new agents, SLAG generates new SLAs in order to track the new detected key users.

3.4.2.3 The Streaming Listener Agents (SLAs)

SLAs differ in various points with HLAs. While HLA stops the historic extraction process once all information shared by the assigned user have been extracted, SLA has to keep listening to a user profile continuously. It need to be connected all the time in order to track any new update shared by the assigned user. SLAs are dynamically generated by the SLAG. Each SLA is in charge of tracking the assigned user profile in real time. SLAs store in real time any new detected information shared by its assigned user in the Retrieved Information Base (DB3).

3.5 Experiments and Evaluation

In order to evaluate the efficiency and the efficacy of the proposed crawling and tracking modules, we implement and test these MASIR modules using Java Agent DEvelopment framework (JADE). Using this framework, each agent is created in a running instance named container. As illustrated in Figure 5.3, MASIR agents are executed in various containers distributed in different hosts connected via a Virtual Private Network (VPN). For the purpose of evaluation and testing, 5 hosts are used in these experiments. The MASIR crawling and tracking modules are based on the two public Twitter APIs; the Search API for the historic information extraction process and the Streaming API for the real-time tracking of key microblog users.

3.5.1 Experimental Set-up

As the Streaming and Search APIs limit the number of simultaneously crawled profiles to around 5 per host and per Twitter developer account, MASIR has encountered this limit by distributing the listener agents in different hosts and using different Twitter accounts. SLAs and HLAs are manged in various hosts on one hand and are processed using different developer Twitter accounts on the other hand. This agents' distribution aims not only to avoid IP banning when the authorized crawling limit rate is reached, but also to boost the number of tracked and crawled profiles. The 5 hosts used for these experiments each incorporates a main container. These containers enable manager agents (i.e. SRA, SLAG, HLAM) to communicate together and to manage the different listener agents according to the number and capacity of the available hosts. Through implementing this architecture using 5 hosts and 7 developer accounts, HLAM is expected to manage up to 175 HLAs (35/host). Similarly for SLAG, 175 SLAs (35/host) are expected to be simultaneously processed. Extra Twitter developer accounts have been also considered in order to be used if one of the already active accounts is banned by Twitter.

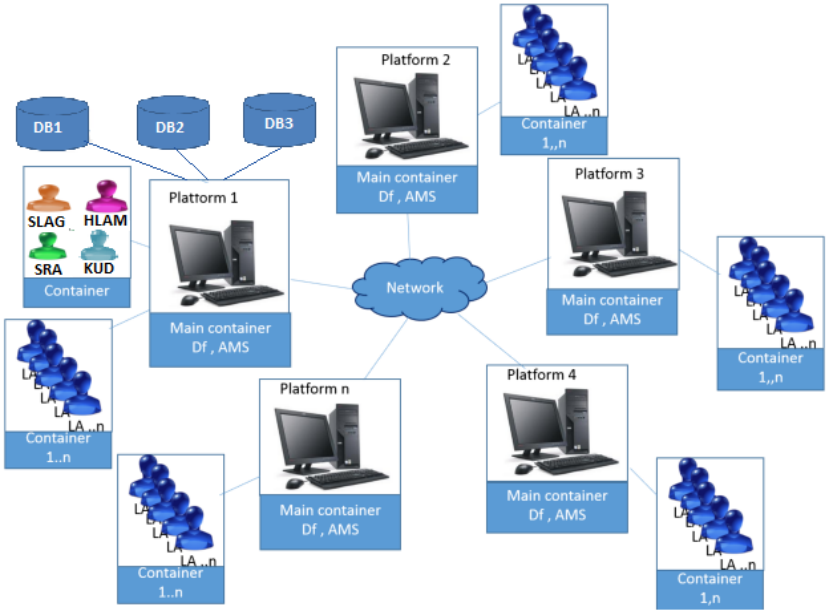

FIGURE 3.3: MASIR implementation environment

3.5.2 MASIR Efficiency for Historic Data Collection

MASIR Collected Data

In order to collect research data that can be explored by researchers on the field of key microblog users identification, we have run the MASIR historic data extraction module during two different flooding events : *Herault flooding* and *Alpes-Maritimes flooding*. The

identification process of users interacting regarding the analyzed event is managed by SRA. This process has been launched after a while from the announcement of each disaster by referring to the keywords listed in Table 3.1. At the end of each event, the users crawling process managed by HLAM is processed to collect data relative to any user who has shared at least one event-related information. Followers and followees relationships specific to each evaluated user are extracted and recorded in DB2. Similarly, the historic registered activities shared by each user during the flooding duration are stored in the same database. By processing this MASIR extraction steps, we have obtained the two following collections :

Collection 1. Herault DB2 : contains the different tweets and social connections relative to users who have shared at least one event-related tweet regarding the Herault flooding event. This event has occurred in the south-east of France from 29^{th} to 30^{th} September 2014. $3,338$ users who have interacted regarding the event have been identified by SRA. The $44,330$ tweets shared by these users during the event were extracted and stored in this collection.

Collection 2. Alpes-Maritimes DB2 : covers the different tweets and followees and followers relationships belonging to $21,364$ users who have shared at least one event-related tweet regarding the Alpes-Maritimes flooding event. This event has occurred in the south of France from the 3^{rd} to 7^{th} October 2015. The $152,402$ tweets shared by these users from the beginning of the event until its end are included in this collection.

TABLE 3.1: Setted keywords for Herault and Alpes-Maritimes floodings events' tweets extraction using MASIR, the number of resulted detected users interested in each event and the two events duration.

	AlpesMarDB	HeraultDB
Keywords	AlpesMaritimes, Orage, Alpes-Maritimes, Intenpéries, Orages, Antibes, Nice, Nice06, Cannes, Inondations	Herault, Hérault, Crue, Crues, Orage, Orages, Intempéries, Flooding, Montpelier, Alert, RedAlert
Number of users	3,338	21,364
Event duration	2 days	4 days

Table 3.2 reports the statistical details of the collected tweets at each phase of each flooding event. $P1$, $P2$ and $P3$ refer to the standard disaster phases *Preparedness*, *Response* and *Recovery* phases respectively. According to these statistics, we observe that the number of extracted information differ according to the scale of the disaster. During the Herault flooding event, few users have been interested in the event as the damages caused by this disaster were not huge. However, during the Alpes-Maritimes flooding serious financial and human damages have been registered which explains the extent of this event. The number

of tweets shared regarding such events varies according to the level of threat characterizing the targeted event.

TABLE 3.2: Number of the different natures of tweets recorded in the two datasets **Alpes-Maritimes DB2** and **Herault DB2** at each phase. **#OnT** and **#OffT** refer to the number of flooding-related (On) original tweets and the off-topical ones respectively. **#OnRT** and **#OffRT** refer to the number of on- and off-topical retweets shared by the different users. **#OnM** and **#OffM** refer to the number of on- and off-topical mentions respectively.

Event Phases		#OnT	#OnRT	#OnM	#OffT	#OffRT	#OffM
Collection 1	P1	513	329	36	9,102	4,333	2,165
	P2	3,357	2,480	202	5,823	2,904	1,427
	P3	2,229	1,260	208	4,586	2,293	1,083
Collection 2	P1	155	91	32	1,506	788	434
	P2	6,692	4,046	300	5,840	3,547	1,064
	P3	2,2343	1,3579	1,960	51,596	28,736	9,693

Listening Process Evaluation

In order to evaluate the historic extraction process of MASIR crawling module, we compare the extracted tweets from each user profile by HLAs with those displayed in the user profile web interface. To conduct this comparison study, we randomly selected 25 users from "Herault DB1" and "Alpes-Maritimes DB1". These two databases refer to the identified users bases during the Herault and Alpes-Maritimes flooding events respectively. We then compared the number of tweets extracted automatically from the selected 25 users' profiles using HLAs with the true number of tweets computed by referring to the users' profiles interface in Twitter. The results of our comparison are described in Figure 3.4. The percentage of tweets extracted from users profiles varies between 80% and 100% for the Herault flooding dataset and between 61% and 100% for the Alpes-Maritimes flooding dataset. The low extraction percentages are due to the short disconnections of HLAs when the tweets extraction limit rate is reached. The MASIR crawling module can be reprocessed once the first extraction process is accomplished and the duration of 7 days is not yet elapsed in order to check if there is some missing data that can be recovered. This recovery step was processed during only the Herault flooding event as it is characterized by a short duration compared to the Alpes-Maritime floodings duration.

We also observe that MASIR has registered more attractive results through Herault flooding collection rather than the Alpes-Maritimes one. This can be explained by the fact that Alpes-Maritime flooding has gained wider interest and has lasted longer than the Herault flooding event which over-complicates the data verification and collection process. For checking the completeness of the Alpes-Maritimes collected data, MASIR has not enough time to perform this process. Only tweets shared during the past 7 days are provided through Twitter APIs. Thus, MASIR has to process the collection within 7 days starting from the beginning of the event. The collection of users data process followed by a recovery of lost data process could take more than 7 days especially for events of long duration.

Data collection for long duration events need to be processed using various hosts to boost and speed further the collection and recovery processes.

Overall, we conclude that the obtained results are promising for the two extracted collections knowing that only 5 hosts and 15 Twitter developers accounts were used by MASIR. Missing data can be avoided and recovered by using further distributed hosts and Twitter developers accounts. Such module can not be used for long duration events lasting more than 7 days.

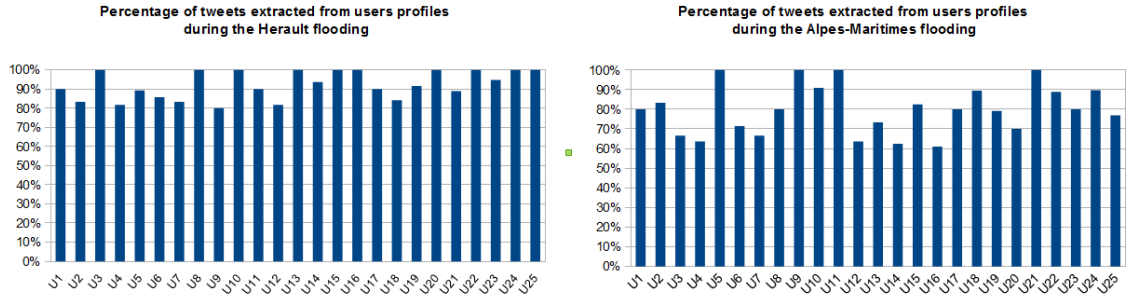

FIGURE 3.4: The percentage of extracted tweets from users profile by MASIR during the two flooding events : Herault and Alpes-Maritimes

3.5.3 MASIR Evaluation for Tracking Key Users

In order to evaluate the efficiency and efficacy of MASIR detection and tracking module during real-world cases, MASIR was launched after 10 minutes from the official announcement of the Herault flooding event. It has collected 44,330 historic tweets and 22,136 fresh ones shared respectively by 3,338 users managed by HLAM and 604 users managed by SLAG. 175 users were simultaneously tracked in real time by SLAs. MASIR has thus coped with the limits imposed by Twitter APIs by tracking an important number of users in real time.

In the following, we conduct a thorough evaluation of the key users tracking process integrated in MASIR. For the purpose of this evaluation, we compared the list of users tracked by MASIR with the ground-truth selected key users. The construction of this ground-truth was conducted by manually evaluating the prominence of users in Collection 1. We describe in the following how we have built Collection 1 ground-truth and how we have evaluated the MASIR tracking process.

Building a Ground-truth for Collection 1

In order to evaluate the quality of users tracked in real time by MASIR, we conducted a user study consisting of labeling each user included in Collection 1 according to the relevance

and freshness of user's tweets. As described previously, this collection contains all the tweets of users interested in the disaster event.

Each user in this collection was evaluated by a group of volunteers in order to build this collection ground-truth. This group was composed of 10 voluntary participants from our laboratory. All these participants are familiar with Twitter and fluent in french which is the official language used in the affected Herault region. These participants were asked to rate each user in this collection from 1 to 10 according to the relevance and freshness of their tweets. To ease this user evaluation process, we gave each participant a detailed report listing in a chronological order most of the important flooding news with their time of first announcement. These news information were extracted from *vosgesmatin*[1] news website. Once all users are rated, we sort these users according to their obtained scores and we retain the top rated 175 users in order to check if they were tracked by MASIR.

MASIR Tracking Process Evaluation

In order to evaluate the MASIR tracking process efficiency, we calculate the true key microblog users that have been listened over time by MASIR. Table 3.3 presents the total number of key and non key users identified by MASIR during the two days of the disaster and the number of the true key users tracked at each period of time with reference to the ground-truth results.

According to these results, an important number of ground-truth key users were identified by SRA and tracked by SLAs from the first day of the disaster. We also note that the precision of our detection process was improved at the end of the second day by tracking 46% of the ground-truth key users continuously.

TABLE 3.3: The evaluation results of the identified and tracked microblog users by MASIR over time with reference to the validated ground truth top 175 key users list

		Identified users by SRA	Ground-truth key users	True key users listened by SLAs
1^{st} day	12 am-00 pm	1,254	157	67
	00 pm-12 am	1,264	157	67
2^{nd} day	12 am-00 pm	2,433	173	57
	00 pm-12 am	3,143	175	81

[1] http://www.vosgesmatin.fr/actualite/2014/09/30/intemperies-l-herault-reste-en-alerte-rouge-de-nouvelles-pluies-possibles

MASIR Detection Process Evaluation

We compared the vector-based key users detection approach integrated in MASIR with three graph-based baseline algorithms : *eigenvector centrality*, *PageRank* and *HITS* algorithms. This experiment aims not only to point out the efficiency of our key users detection approach, but also to prove that the resulted MASIR collections stored in DB2 are suitable for both vector-based and graph-based key users identification approaches. The graph-based measures selected for this comparison are typically used for the detection of such key users. In these experiments, users' graph was designed by taking into account the users followers and followees relationships.

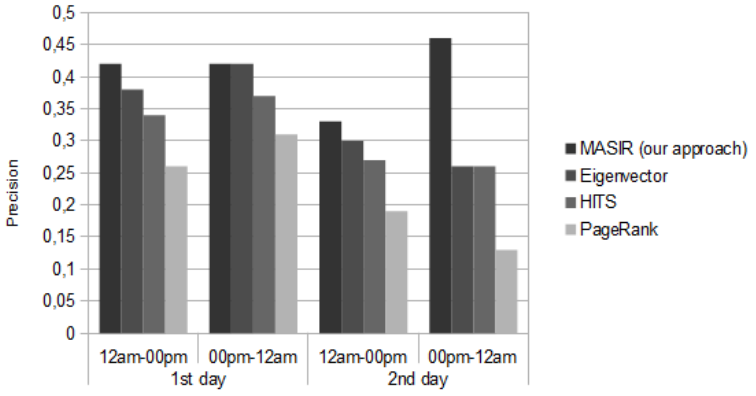

FIGURE 3.5: Comparing the performance of the vector-based key users identification approach integrated in *MASIR* with respect to state-of-the-art graph-based approaches *Eigenvector, Hits* and *PageRank* during the Herault flooding

To evaluate the quality of results returned by each baseline in each period of time, we measure the precision of the returned key users by each algorithm. The obtained results are shown in Figure 3.5.

Compared to the time consuming centrality measures, our model gains a significant increase in performance at the different stages of the event. We also note that the performance of the graph-based measures decreases over time as they are sensitive to well-connected users.

According to these obtained results, MASIR outperforms the identification models based on centrality measures. MASIR detected most of ground-truth key users at an early stage of the event. Based on the multi-agent parallel processing architecture, MASIR has proved its ability to detect and track targeted key users in real time. Even if MASIR has not detected a large number of key users, the obtained results remain promising as they show the capacity of such a system to process both the key users detection and tracking processes in real-world scenarios. Any graph-based or vector-based detection approach can be integrated in our modular MASIR architecture if time requirements are not strict.

3.6 Conclusion

This chapter highlights the capacity of multi-agent systems for both extracting rich tweets collections, identifying, and tracking key microblog users in real time. The historic extraction module integrated in MASIR ensures access to valuable collections. These collections are suitable to be used for key microblog users identification models learning, testing and evaluation. MASIR uses various collaborative agents enabling a real-time detection of key users who tend to share valuable information. This first research effort to deal with the detection and tracking of key users in real-world crisis events cases has achieved promising results. The different agents executed in parallel ensured a real-time analysis and tracking of the needed data. The integration and distribution of these agents in different hosts have coped with current Twitter APIs limits. MASIR was able to track 175 microblog users in parallel using only 5 hosts and 30 Twitter developers accounts.

While the detection approach integrated in MASIR has outperformed the identification results obtained by standard centrality measures, this approach can be enhanced by exploring more richer features. Various features can be extracted by referring to the user related information collected by MASIR. In the next chapter, we aim to evaluate both the existing state-of-the-art features and other new proposed features for key users identification during crisis events.

Chapter 4

Microblog User's Features Categories Effectiveness for Prominent Users Identification

Contents

4.1 Introduction

Microblogging platforms, especially Twitter, provide various information that can be explored for microblog users characterization. Twitter shared information, commonly known as tweets, are generally expressed in various languages and formats. Tweets can be expressed using short texts, images, links or/and videos. Transforming this variety of unstructured content into a structured format remains complex. Each content needs to be processed separately according to its type. This variety of tweets content over-complicates the microblog users characterization process.

Given such complexity, users information shared content is generally neglected while modeling microblog users. Most of prior works have modeled users in terms of their behavior and social connections in the microblogging platforms (Pal & Counts 2011, Xianlei et al. 2014). Microblog users behavior and social position were generally projected either in social graphs or in feature vectors .

As discussed in Sections 2.3 and 2.4 of Chapter 2 and proved in Chapter 3, vector-based user characterization approach is more suited to our key microblog users identification problem. The effectiveness of this characterization approach mainly lies in the effectiveness of the extracted and selected features for users modeling. These features have to be meaningful in order to point out the particularities of key microblog users. Various raw feature have been explored in the literature for the identification of different categories of key users like topical authorities and domain experts. These raw features measure quantitatively the different activities and social relationships characterizing user behavior. Such features are generally selected without any prior study which evaluates their effectiveness in the specific key users identification context (Pal & Counts 2011).

In this chapter, we focus on evaluating the effectiveness of both state-of-the-art and our new topical proposed features categories for the identification of prominent microblog users in the context of crisis events. Through this evaluation study, we aim to select the most descriptive categories of features pointing out the main differences between prominent and non-prominent users in crisis events context. The purpose of this evaluation is to select the most effective raw features categories which could be explored to derive better discriminative engineered features. As defined previously, *prominent microblog users* in the context of crisis events are microblog users who are susceptible to share relevant and exclusive information regarding the event. This category of users does not necessarily refer to users geolocated on the crisis event area or/and to users who are experts in the domain of crisis. These users may refer to ordinary users geolocated far from the crisis event area, however, transmitting exclusive news regarding their friends or family who are geolocated there.

The rest of this chapter is organized as follows. Section 4.2 details the research questions addressed in this chapter. The role of features in the identification of prominent microblog users is described in Section 4.3. The transformation of unstructured microblog user specificities into structured features is presented in Section 4.4. In Section 4.5, the different state-of-the-art and new proposed features categories explored for user characterization are listed. The experiments are discussed in Section 4.6. Finally, we conclude this chapter and discuss the obtained results in Section 4.7 and 4.8.

4.2 Research Questions

In this research, we list the different raw features explored in the literature for key users identification. We also propose additional topical features that could be effective for this task. We evaluate the effectiveness of these feature categories by experimenting the effect of each category on the identification results using real world crisis event data. This feature categories study helps us to answer the following research questions :

1. Which raw feature categories best reflect the prominent microblog users behavior and particularities during crisis events?
2. What are the feature categories that could be neglected while representing microblog users in this problem context?
3. How effective can identification algorithms be while considering all the feature categories?

The answers of these research questions help us to identify the categories of features that we have to focus on for prominent users identification in the context of crisis events. The selection of these categories paves a way for proposing further engineered features derived from these categories and hence describing better microblog user's behavior and interactions.

4.3 Features Role in Microblog Users Categorization

Features play a central role in both microblog key users modeling and identification. Identification models based on either classification, clustering or/and ranking algorithms would fail to identify the targeted users if the selected features are difficult to learn. Selected features might not have any correspondence with the real-world targeted user behavior or specificties and thus would over-complicate the identification task. Most of the identification models follows these main steps :

1. **Select Data :** Collect data belonging to each user.
2. **Pre-process Data :** Format, clean and sample this data according to their specificities and type.
3. **Transform Data (Feature Extraction) :** Extract suitable features for user characterization with reference to the preprocessed raw data.
4. **Model Data :** Learn and test models, identify and learn key users behavior patterns, rank key users.

Various features can be extracted for microblog users characterization. However, the effectiveness of the features extraction step relies on their discriminative power in prominent users identification. Extracting and modeling users using various features which are not relevant to the analyzed problem would erode the performance of the identification model. Features have to be evaluated in terms of their effectiveness regarding both the modeling and the identification problem. As shown in Figure 4.1, users belong to different user categories (e.g. experts, journalists, celebrities, ordinal users...). Each category of users has its specific behavior and characteristics. Key users identification models are generally designed to target at least one user category information. In order to be able to identify the targeted user categories, features highlighting the specificities of these categories have to be extracted. For example, in the case of targeting spammers, extracting features reflecting the credibility (i.e. trust features as shown in Figure 4.1) of each evaluated user would make the detection more accurate. However, the effectiveness of such features would not be the same in the case of targeting other user categories like influencers. Each extracted feature has to be relevant to the specific key users identification problem by pointing out the particularities of the targeted users. There are various methods to identify features which fit the best to the modeling and identification context :

- **Brainstorming Method :** The extracted features using this method are defined by observing the targeted and non-targeted users data, pointing out the particularities of the targeted ones, listing the existing features used for other problems, studying these features and selecting those which may be suitable to characterize the observed specificities.
- **Learning and Analyzing Raw Data Method :** The extracted features are constructed either automatically using features learning algorithms (e.g. auto-encoders and restricted Boltzmann machines) or manually by observing data or using a mixtures of the two techniques.
- **Features Selection Algorithms based Method :** Features are selected by evaluating the effectiveness of various set of features using feature selection algorithms.

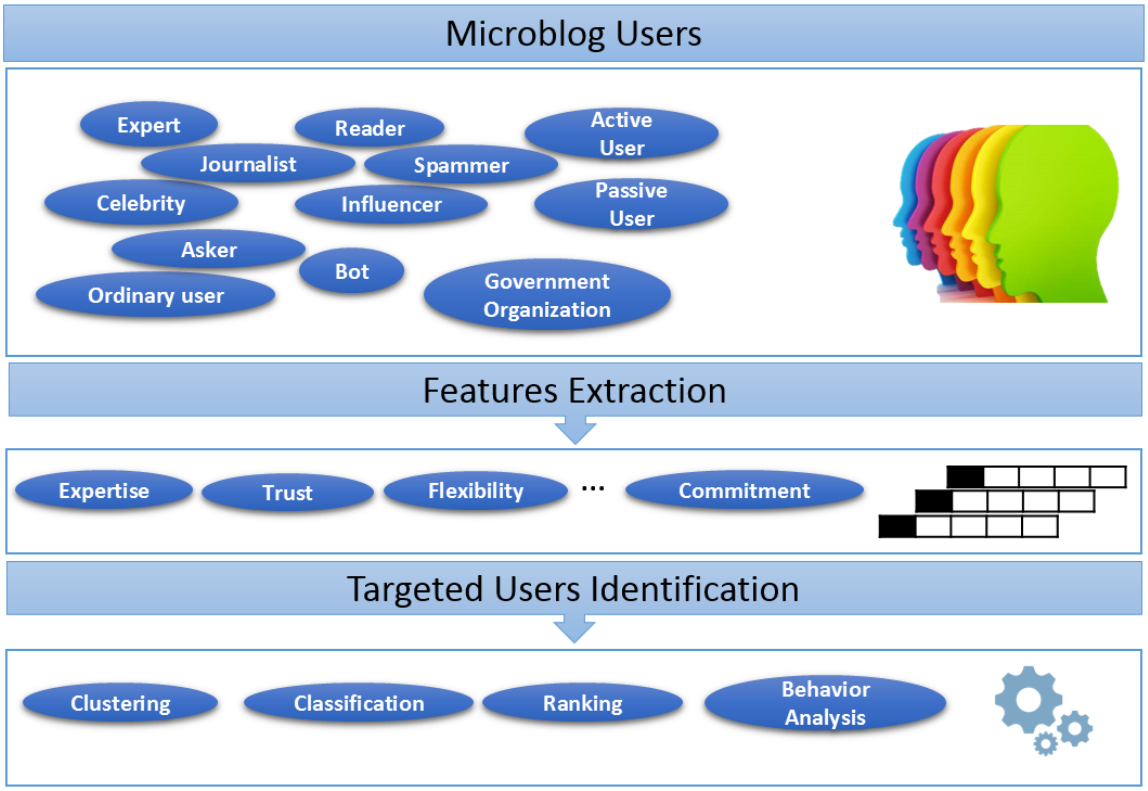

FIGURE 4.1: Features role for key users identification in microblogs.

These algorithms generally fall in three categories, algorithms-based on filter methods, algorithms-based on wrapper methods and algorithms-based on embedded methods.

- **Models based Method :** User features are selected by evaluating the identification model performance on unseen data incrementally using at each step a different set of features.

4.4 Mapping Microbog Users Specificities into Features

In order to identify the most effective features that can be suited to our identification problem, we aim to extract and study existing features and also newly proposed ones which characterize microblog users. These features have to be adapted to our problem by being computationally feasible in real time and relevant to our identification problem context.

Following the brainstorming method for feature extraction defined in the previous section, we extract a set of features characterizing the different specificities of microblogs. Figure 4.2 describes the main microblogs specificities that could be valuable to characterize microblog users in crisis context. Such specificities are generally described and modeled using

a graph-based representation. However, as mentioned in Section 2.4 of Chapter 2, identifying targeted users through graph-based user modeling is time consuming and sensitive to celebrities. These limitations make such representation unsuitable to our identification problem context. However, these graphs generally cover the main specificties characterizing users. By exploring the different specificities of microblog user activities represented in such graphs as described in Figure 4.2, we extract the maximum of metrics that can be computationally feasible in real time. These metrics known as *features* are extracted by studying different possible relations between users and Twitter specificities. The following relations are considered :

- **The relation between the evaluated user and his/her shared content** is represented in the user graph by different edges describing the nature of the user shared content (i.e. tweets, retweets, replies, received or/and sent mentions). To benefit from the wealth of these relations, we mapped them into various topical features characterizing the user attachment to the analyzed topic.
- **The relation between the evaluated user and the content shared by others** is characterized by various edges reflecting the effect of the content shared by others on the user behavior (i.e. retweet, like or/and reply). These edges are also transformed into topical features characterizing user's interactions regarding the topical content of other users.
- **The relation between the evaluated user and the other users** characterizes the social relationship between the user and his/her followers and followees. We map these kinds of relations into social features characterizing the user topical followership relations.
- **The relation between the evaluated user and his/her profile metadata** reflects the user's main information. This relation is mapped into profile features describing the user activity and interest in a general context.
- **The relation between the evaluated user and his/her content metadata** describes additional information regarding the shared user content (e.g. time, location, number of likes, etc..). This relation is mapped into spatial features on one hand and topical user features on the other hand.

This mapping process results in four raw feature categories (i.e. user social features, user profile features, topical user activity features and user spatial features) which are computationally feasible in real time. These features cover the main characteristics describing Twitter users.

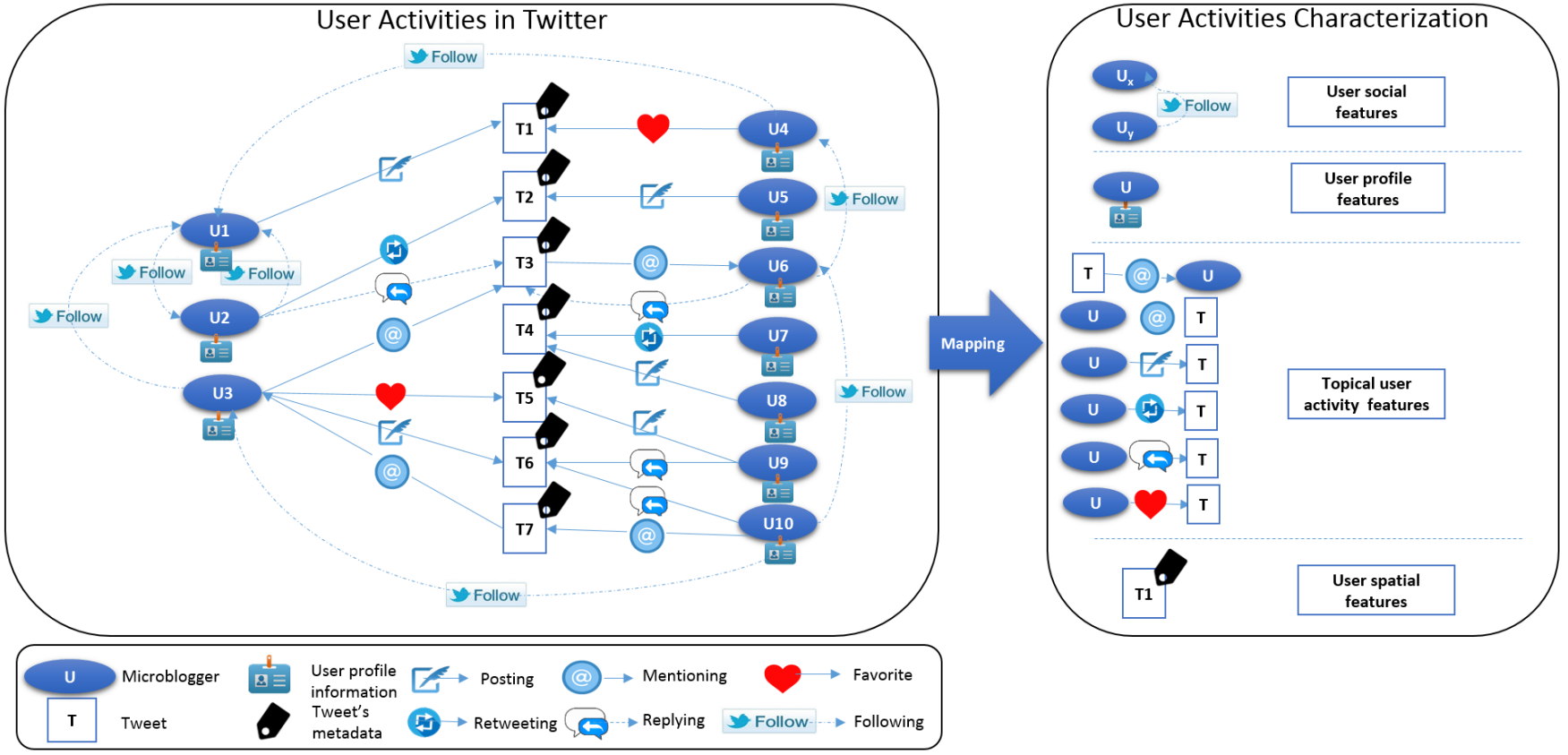

FIGURE 4.2: Mapping microblog user activities into different categories of features.

4.5 Microblog User Features Categories

In order to efficiently identify the targeted prominent users in the context of crisis events, we study a large set of state-of-the-art features and some new features that can be suitable for microblog user modeling in the context of crisis events (Bizid et al. 2015e,b). The studied features mainly reflect the behavior and the activity of each evaluated user regarding both the analyzed event topic and the other topics. As described previously, feature categories can be split into four broad categories : *profile features (PrF)*, *topical features (OfAF+OnAF)*, *spatial features (SpF)* and *social network structure features (SnF)*. In this study, topical features are categorized in two distinct categories : on-topical feature category and off-topical one. The rest of this section describes these different categories in detail.

4.5.1 Profile Features

Profile Features *(PrF)* characterize the user profile description in the microbogging platform. This description (e.g. location, domains of interest...) is either registered by the profile owner himself or automatically generated by the microblogging service in order to report the user activeness rate in the platform. The generated information are computed according to the registered historical activities belonging to the evaluated user (e.g. Number of collected favorites, Number of followers...). Table 4.1 presents the set of user profile features selected for this study. These analyzed features are easily extractable from any user profile using Twitter APIs.

Table 4.1: Extracted Profile Features *(PrF)* for having a global view of microblog users specificities.

Name	**Features**
P1	Certified user (Xianlei et al. 2014)
P2	Enabled geolocation (Bizid et al. 2015e)
P3	Protected (Xianlei et al. 2014)
P4	Number of produced tweets (Xianlei et al. 2014)
P5	Number of collected favorites (Bizid et al. 2015e)
P6	Creation date of the Twitter account (Bizid et al. 2015e)
P7	Number of followers (Xianlei et al. 2014)
P8	Number of followees (Xianlei et al. 2014)

PrF provide a digital representation of the user identity, activity and influence. Such broad description could be valuable to identify prominent users susceptible to share relevant and exclusive information during a given crisis event. By examining these features, we note that P2 and P1 features could be enough descriptive for prominent microblog users in the context of crisis events. P2 could give some valuable information regarding the user geographical

zone of interest and location. P1 could be a strong indicator to evaluate the veracity and credibility of the information shared by each user. P7 and P8 features which are generally explored for celebrities and domain experts detection could also be valuable to identify the targeted users in this thesis context. P4 and P5 which refer to the user activeness in the network are studied in order to evaluate if daily active users in the microblogging platform would be prominent during unexpected disasters or not.

4.5.2 User Activity Features

Various features reflecting user activity have been proposed in the literature (Pal & Counts 2011). However, all these designed features are explored for user on-topical activity characterization while neglecting their off-topical ones. In the context of the identification of prominent microblog users during crisis events, we aim to explore both the user's on-topic tweets related to the disaster and the off-topic ones. The rationale behind thus proposed strategy is to highlight users interested only by the analyzed crisis event and neglect those toggling between several topics such as news outlets. Users interested in several topics would generally share outdated information which were already spread in microblogs.

Thus, we divide the different user activities features extracted from the user timeline during the disaster into two categories : *On-topic Activities Features* (*OnAF*) and *Off-topic ones* (*OffAF*). These features are measured respectively according to the on-topic and off-topic activities belonging to each user :

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 crisis event under consideration.

Off-topic : an off-topic activity refers to any activity that was not recorded as an on-topic one.

Additionally, we assume that tweets referring to the disaster and including at least one keyword reflecting non-serious or non-valuable contents (e.g. advertising or joke words and symbols such as sale, rent, pub, lol and so on), will be automatically recorded as an off-topic one. Thus, users who share non valuable contents would be penalized.

TABLE 4.2: On-topic User Activities Features (***OnAF***) and Off-topic User Activities Features ***(OffAF)*** extracted according to the user related tweeting activities. The "*" symbol refers to the new proposed features that we propose for user topical activities characterization.

Id	Features	On	Off*
	Original tweets		
T1	Number of original tweets (Pal & Counts 2011, Xianlei et al. 2014, Bizid et al. 2015e)	+	+
T2	Number of links shared (Java et al. 2006, Bizid et al. 2015e)	+	+
T3	Number of keyword and hashtags (Pal & Counts 2011, Bizid et al. 2015e)	+	-
T4	Number of collected likes for user original tweets* (Bizid et al. 2015e)	+	+
	Retweets		
T5	Number of retweets of other's tweets (Boyd et al. 2010, Xianlei et al. 2014, Bizid et al. 2015e)	+	+
T6	Number of unique users retweeted by the evaluated user* (Bizid et al. 2015e)	+	+
T7	Number of retweets of the evaluated user's tweets (Bizid et al. 2015e)	+	+
T8	Number of unique users who retweeted the evaluated user's tweets (Boyd et al. 2010, Bizid et al. 2015e)	+	+
	Mentions		
T9	Number of mentions of other users by the evaluated user (Pal & Counts 2011, Bizid et al. 2015e)	+	+
T10	Number of unique users mentioned by the evaluated user (Pal & Counts 2011, Bizid et al. 2015e)	+	+
T11	Number of mentions by others of the evaluated user (Honey & Herring 2009, Pal & Counts 2011, Bizid et al. 2015e)	+	+
T12	Number of unique users mentioning the evaluated user (Honey & Herring 2009, Pal & Counts 2011, Bizid et al. 2015e)	+	+

Both on- and off-topic features will be studied in the context of crisis events. Table 4.2 presents state-of-the-art features characterizing user activities and our new proposed features marked by the "*" symbol.

In the following, we describe our new proposed features :

- Number of collected likes for the evaluated user original tweets (*T4*) : represents the sum of the collected likes by each user (i.e a small heart icon attached to each original tweet indicating how many users have liked the shared information). Such metric reflects how many users' tweets are of interest regarding the specific topic ($T4_{on}$) and the other topics ($T4_{off}$).
- The number of unique users having tweets retweeted by the evaluated user (*T6*) : indicates how many users who are actively communicating about the crisis event topic ($T6_{on}$) or the other off-topics ($T6_{off}$) have attracted the attention of the evaluated user. Prominent users in a specific topic could retweet tweets produced by different users in order to provide a wide range of relevant tweets produced by different sources.

We separately study these on- and off-topic user activities feature categories. Through this study, we aim to estimate the effectiveness of each category for prominent users identification in the context of crisis events.

4.5.3 Spatial Features

Spatial Features (SpF) characterize microblog users according to their assigned location and geolocation regarding the threatened crisis event zone. Such features may be essential to determine who are the users geolocated in the crisis event zone. On-the-ground users could play the role of sensors by providing fresh information in real time. We thus evaluate the effectiveness of the following spatial features, described in Table 4.3 :

TABLE 4.3: Spatial Features (***SPF***) characterizing the geographic position of microblog users regarding the analyzed crisis event.

Name	**Features**
S1	Spatial co-location* (Bizid et al. 2015b)
S2	Spatial co-geolocation* (Bizid et al. 2015b)

S1 indicates if the user's location has been stricken by the crisis event or not. This feature is measured by computing the matching rate between the set of unique locations Lu specified by each user and the set of unique locations included in the crisis event zone Ld. The extracted locations are drawn from the user's profile.

$$S1 = \frac{Lu \cap Ld}{Lu \cup Ld} \tag{4.1}$$

S2 measures the inclusion rate of the geo-coordinates related to the user shared tweets in the territory threatened by the crisis event. The crisis event area is represented by a polygon or a set of polygons Pg that may include many distant zones. This feature takes into account only specific geographic coordinates Cu.

$$S2 = \frac{Cu \cap Pg}{Cu \cup Pg} \tag{4.2}$$

4.5.4 Network Structure Features

Network structure features (SnF) are extracted from the user followers and followees lists. Based on these lists, we count the number of user followers and followees who have shared at least one event-related information. These features have been widely explored in the context of influential microblog users identification (Romero et al. 2011). However, such features are generally criticized and judged as sensitive to well-connected and popular microblog users (Pal & Counts 2011). To avoid this problem, we have proposed two additional network structure features *NS3* and *NS4*. These features adjust the number of on-topic followers and followees with the total number of both off- and on-topic followers (Bizid et al. 2015e). Table 5.6 presents the network structure features studied in this Chapter.

TABLE 4.4: Network Structure Features (***SnF***) characterizing the social position of microblog users.

Name	**Features**
NS1	Number of user's topical followers (Bizid et al. 2015b, Pal & Counts 2011)
NS2	Number of user's topical followees(Bizid et al. 2015b, Pal & Counts 2011)
NS3*	Number of user's topical followers adjusted by the total number of his/her followers* (Bizid et al. 2015b,e)
NS4*	Number of user's topical followees adjusted by the total number of his/her followees* (Bizid et al. 2015b,e)

4.6 Selection of Feature Categories

In order to select the best subset of features categories that can be suited to our identification problem context, we employ a selection approach following the same principle as the forward greedy wrapping one. This approach consists of learning an identification model using different feature categories subsets and measuring each category subset effect on the model performance. The selection process starts with evaluating the identification

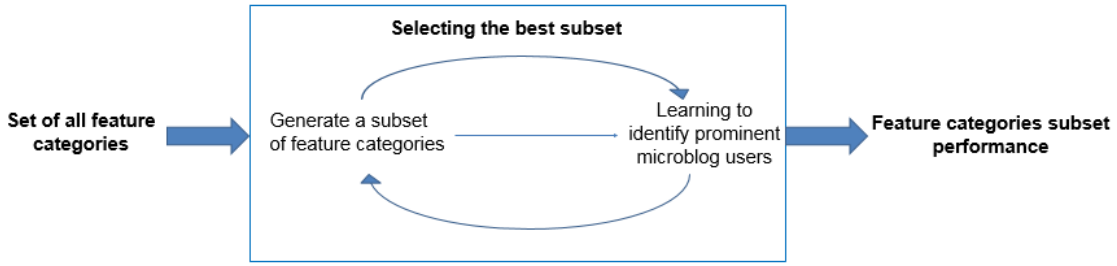

FIGURE 4.3: Feature categories evaluation process using the forward greedy wrapping method.

model performance using features included in one single category and then incrementally adds other feature categories. Only feature categories which lead to better performance are retained at each step. As described in Figure 4.3, this selection process is incrementally executed until no further improvement can be achieved. We use both ANN and SVM machine learning algorithms to test the identification performance of the different subsets.

After examining which categories of features increase the performance of the identification model, only effective features categories that have been approved by the two machine learning algorithms are retained.

4.7 Experiments and Results

4.7.1 Dataset Definition and Labeling

To conduct experimental performance evaluation on real data, we use the Herault database collected using our modular multi-agent system MASIR. This database was described in depth in Chapter 3. In the previous described user study conducted regarding this database as detailed in Section 3.5.2 of Chapter 3, we asked participants to attribute a prominence rank according to each user tweets relevance and freshness. Herein, we conduct another different user study aiming to label each user according to his/her prominence during the whole analyzed flooding event independently of its phases.

Through this study, we have asked three volunteers to manually classify the tracked users in $C1$ (prominent users class) or $C2$ (non-prominent users class) according to the relevance and freshness of their tweets during the whole period of the event. The complete list of news sorted in a chronological order was provided to each participant. Two of the selected participants were asked separately to label each user according to his/her prominence. The resulted labeling results of the two participants are then evaluated by the third one who has to break the detected disagreements in terms of evaluated users' labels. This third

participant has to decide whether the user labeled as $C1$ by one participant and $C2$ by the other one deserves to be labeled as C1 or not.

According to this user study, 90 users were labeled in $C1$ and $3,248$ in $C2$. Using these labels, we can measure the performance of the prominent users identification models in the context of crisis events and thus evaluate the effectiveness of each user feature categories.

4.7.2 Experimental Set-up and Evaluation Metrics

For experimental set-up, we use two different learning algorithms for studying the effectiveness of each category using ; Support Vector Machine (SVM) (Osuna et al. 1997) and Artificial Neural Networks (ANN) (Zhang 2000). Based on these algorithms, we tested the main combinations of feature categories in order to find the most effective one in the context of prominent users identification during crisis events.

According to the obtained user study results, the number of prominent users is greatly larger than the number of non-prominent ones. This data unbalance complicates the classification process. In order to deal with this problem, we set a larger weight to the class C1 of prominent users ($W_1 = 10$) than the class C2 of non prominent users ($W_2 = 1$). These parameters were set experimentally in the training phase of SVM.
On the other side, as there are no parameters to tune the class weights using ANN, we have duplicated the dataset of prominent users 30 times in order to balance the two datasets of prominent and non prominent users in the training phase of ANN.

For test and training purposes, we randomly sampled 60% of both prominent and non-prominent labeled users datasets as training data to learn the classification and ranking models based on different feature categories, and the remaining 40% as test data to evaluate the efficiency of the learned model.

Table 4.5: Training and test datasets description

	Training Dataset (60%)	Test Dataset (40%)
Number of Prominent users	54	36
Number of Non-prominent users	1945	1297

Through the different experiments conducted in the following, we use standard precision, recall and F1-score (i.e. F-measure) evaluation metrics.

$$\text{Precision (Prec.)} = \frac{\#\text{Correctly classified prominent users}}{\#\text{Users classified as prominent users}}$$

$$\text{Recall (Rec.)} = \frac{\#\text{Correctly classified prominent users}}{\#\text{Ground truth prominent users}}$$

$$\text{F1-score (F1)} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

4.7.3 Evaluation of Feature Categories Effectiveness

In order to select the most representative feature categories for prominent microblog users identification in the context of crisis events, we evaluate the effectiveness of each category of features separately. Table 7.2 reports the experimental results evaluating the effectiveness of each features category using two different learning algorithms.

TABLE 4.6: Effectiveness of each feature category for prominent users identification in terms of Precision, Recall and F1-score evaluation metrics.

Feature Category	#Features	**SVM**			**ANN**		
		Prec.	Recall	F1	Prec.	Rec.	F1
AF_{On}*	**12**	**0.43**	**0.86**	**0.57**	**0.29**	**0.80**	**0.42**
AF_{Off}	11	0	0	0	0.04	0.33	0.07
PrF	8	0	0	0	0.01	0.33	0.03
SnF	4	0.05	0.02	0.03	0.09	0.61	0.15
SpF	2	0	0	0	0	0	0

According to the identification results recorded by both the learned SVM and ANN models, the category of features characterizing the on-topical user activity in microblogs (AF_{On}) is the most representative category for prominent users modeling in the context of crisis events. The remaining categories have yielded poor results. However, these categories may yield improvement in terms of precision and recall if they are combined with other categories. Therefore, we study the effectiveness of these categories with associating them with the selected feature category AF_{On}. Table 7.4 reports the results of the different evaluated feature categories pairs for learning prominent microblog users identification models based on both ANN and SVM.

TABLE 4.7: Effectiveness of each pair of feature categories (AF_{On}, An additional Feature Category) for prominent users identification in terms of Precision, Recall and F1-score evaluation metrics.

Feature Categories	#Features	**SVM**			**ANN**		
		Prec.	Recall	F1	Prec.	Rec.	F1
AF_{On}+AF_{Off}*	**23**	**0.47**	**0.75**	**0.58**	**0.43**	**0.80**	**0.56**
$AF_{On} + PrAF$	20	0.42	0.86	0.56	0.36	0.86	0.51
$AF_{On} + SnF$	16	0.40	0.86	0.55	0.24	0.66	0.35
$AF_{On} + SpF$	14	0.43	0.86	0.57	0.39	0.88	0.54

According to the reported results by the ANN and SVM identification models, we observe that the combination of the two categories of features AF_{On} and AF_{Off} improves the identification results. However, the other feature categories combinations negatively affect the initial identification results obtained in the previous iteration. Thus, we only retain

the (AF_{On}, AF_{Off}) combination for the next iteration. The results obtained based on this retained combination with an additional feature categories are reported in Table 7.5.

TABLE 4.8: Effectiveness of 3 combined feature categories (AF_{On}, AF_{Off}, An additional Feature Category) for prominent users identification in terms of Precision, Recall and F1-score evaluation metrics.

Feature Categories	#Features	**SVM**			**ANN**		
		Prec.	Recall	F1	Prec.	Rec.	F1
AF_{On}+AF_{Off}+SpF	**25**	**0.48**	**0.75**	**0.60**	**0.41**	**0.80**	**0.54**
AF_{On}+AF_{Off}+PrF	31	0.43	0.72	0.54	0.32	0.75	0.45
AF_{On}+AF_{Off}+SnF	27	0.45	0.75	0.56	0.36	0.77	0.50

According to these results, we observe that there is no significant enhancement when adding a third category of features to OnAF and OfAF. Only the spatial category of features slightly improves the identification results in the case of using the SVM model. We also note that the learned ANN model based on these same categories, decreases the identification performance compared to the previous resulted ANN learned based on AF_{On} and AF_{Off} categories. These results show that two different models learned using the same user representation could lead to different results. Features need to be discriminative enough in order to be able to identify prominent users using any machine learning algorithm. In this case, spatial features can not be retained as relevant features for prominent users modeling and identification.

4.8 Discussion

The obtained results in this study have led us to validate the effectiveness of both the on-topical and off-topical activities features categories for the identification of prominent microblog users in the context of crisis events. On- and off-topic features are extremely useful in disaster management scenarios where prominent users mainly focus on sharing disaster-related information. Thus, using off-topic activity features, users toggling between different topics will be penalized. In addition, referring to the on-topical activities features, users focusing potentially on the unexpected disaster will be promoted. Such a property has shown that users faced by a disaster would mainly share on-topical information and neglect the other topics-related information. Moreover, we have shown that users geolocated in the disaster area can not be systematically detected using spatial features. Such features are not strong enough to make the identification of prominent users easier. Thi can be explained by the fact that users rarely share their geolocation via microblogiing platforms. As discussed in Chapter2, only 1% of user tweets are attached to geolocation-coordinates. Such features have been slightly useful using SVM and have eroded the identification results using the ANN learning algorithm. An open access to Twitter data would be necessary to confirm further these findings. However, such access could not easily be afforded.

4.9 Conclusion

In this chapter, we analyzed the effectiveness of different state-of-the-art and new proposed feature categories for prominent users identification during crisis events. We tested different combinations that may lead to an efficient classification model. The different experiments were conducted using two different learning algorithms ANN and SVM. We found that on- and off-topic user activities feature categories are the most relevant for users behavior modeling in the context of crisis events. Moreover, we showed that a similar user characterization can lead to different identification results using different classification algorithms. The SVM algorithm learned using AF_{On}, AF_{Off} and SpF features have provided better results than the ANN algorithm. The selected learning algorithm for prominent users identification has to be adapted to the chosen user representation approach.

In next steps, we aim to analyze the effectiveness of each feature characterizing prominent users independently of their category using a different feature selection algorithm. Moreover, we wish to propose additional engineered features derived from these selected categories of raw features.

Chapter 5

Features Engineering for Prominent Users Identification during Crisis Events

Contents

5.1 Introduction

Microblogging platforms offer services of convenient access to- and sharing of- exclusive information on any topic. To evaluate the freshness and relevance of this shared topical information, most of researchers have focused on analyzing this information content and more precisely its textual content. However, as explained in Chapter 2, information content formats are not restricted to text. Images, links and videos formats are also extensively used in microblogs. Textual content is generally associated with this variety of non-textual formats in the form of tags referring to well-defined hashtags or/and keywords specific to the targeted topic. These defined hashtags and keywords ensure a wider visibility of the shared information in the microblogging platform. However, they do not reflect in any case the attached content relevance and freshness. Thus, evaluating the quality of the shared topical information by analyzing mainly textual content, restricted in some keywords or/and hashtags or in short expressive phrases, is not sufficient. Additionally, information retrieval techniques based on analyzing each information content according to its format are not feasible in real time and thus unsuitable to be applied during crisis events.

Having the aforementioned particularities of microblogs in mind, associating the relevance and the quality of tweets content with the user's prominence strategy remains the most adapted strategy to the context of crisis events (Wagner et al. 2012, Liao et al. 2012). However research works following this strategy have mainly focused on modeling microblog users quantitatively according to their activity on the specific analyzed topic or event. The on-topical raw features studied in the previous chapter are generally considered for the modeling purpose. Through such modeling approach, prominent microblog users are generally detected following this principle : "the more the user is active regarding the analyzed topic the more he/she is prominent and thus his/her shared information are relevant and exclusive".

While on-topical features-based modeling approaches have succeeded to achieve promising results for influencers (Romero et al. 2011, Chen et al. 2009), domain experts (Xianlei et al. 2014, Bozzon et al. 2013) and topical authorities (Pal & Counts 2011) identification in microblogs, such techniques are still unsuited for prominent microblog users detection in the context of crisis events. These techniques are sensitive to users who are extremely active in sharing outdated information regarding the analyzed event. Let us assume the example of news outlets channels, these channels accounts usually share various information regarding different topics. They are usually active regarding major events including crisis ones. However, such accounts are not necessarily considered as prominent to track. They generally report outdated information already shared in the microblogging platforms. As can be seen from the earlier chapter, considering both on- and off-topic user activities

while modeling microblog users is recommended in order to improve the effectiveness of the identification models.

However, explored topical features, as presented previously in their raw form, are very straightforward and do not effectively reveal the quality of content shared by users. These features have many correlations among them that could be explored. Most of on-topical raw features have their corresponding off-topical ones. Raw features reflecting the same type of user activities from different angles of view can also be combined. These raw features should be designed in an optimal conceptual form that could better represent the targeted users in terms of their topical interactions. Proposing derived efficient engineered features from the already selected raw ones would ease and speed the learning of predictive models.

In this chapter, we propose a set of engineered features derived from the selected effective raw features evaluated in the previous chapter. Unlike state-of-the-art engineered features for microblog users representation, our proposed features characterize each user by considering both his/her on- and off-topic activities during the analyzed crisis event. These features are designed in order to ensure the promotion of users mainly focusing on the event under consideration, and the penalization of those who are toggling among several topics. We represent microblog users by a vector of engineered features. Based on this vector-based user representation, we learn to differentiate between the topical activity of prominent microblog users and non-prominent ones based on a SVM machine learning algorithm.

The rest of this chapter is organized as follows : Section 5.3 highlights the importance of considering both on- and off-topical user activities for user modeling. Section 5.4 presents the set of our proposed features for microblog users modeling. Section 5.5 describes the classification and ranking approach employed to identify prominent users. Section 5.6 presents the experiments and results obtained by our model. Section 5.7 concludes this chapter and discusses future steps.

5.2 Research Questions

The main purpose of this research is to explore the possible combinations of the selected raw features categories for microblog users modeling in the context of crisis events. We propose a new list of engineered features that are well suited to the problem of prominent users identification. These features have to promote users focusing on the analyzed crisis event and penalize those toggling between several topics. By exploring the raw selected features, we seek answers to the following questions :

1. How to qualify users activities features in the context of crisis events? How do we make the identification model less sensitive to celebrities sharing various relevant however outdated information regarding the event?
2. Is it more effective to combine raw features instead of considering them separately?
3. Adjusting users on-topical features with the off-ones strategy : does it enhance the identification model efficiency in the context of crisis events?

The answers to these research questions help us to explore the relation between the different extracted raw features representing users activities during the studied event. Finding real time processable techniques for pointing out the quality of user activities would lead us to neglect the complexity of users generated content and focus on user behavior patterns. Including the proposed engineered features in effective modeling format would ease the understanding of the targeted user behavior specificities. Such features would thus speed up the prominent users prediction process in real time.

5.3 Focus on User Topical Activities

We showed in the previous chapter the effectiveness of both on- and off-topic user activities features categories for microblog user representation in the context of crisis events. To the best of our knowledge, the off-topical features have never been considered in the literature for the purpose of key users identification. Microblog users are typically characterized and evaluated regarding their on-topic activities. Such characterization does not reflect neither realistically nor accurately the real user behavior regarding the specific analyzed event or topic. User off-topic activities shared during the analyzed event period have to be considered in order to reflect the real attachment of each user to the analyzed topic.

Let us assume the example of three users having the same recorded values regarding their on-topic activities features as represented in Table 5.1. By referring only to the on-topic activities of each user, the three different users would be modeled similarly. Such representation will over-complicate the identification task. The prediction model would not be able to differentiate between prominent and non-prominent users in such case. By focusing on these users off-topic activities, many outstanding differences can be stated. While User 1 on-topic and off-topic activities features values are equal, the recorded on-topic activities of User 2 are remarkably lower than the off-ones. On the other side, on-topic activities statistics of User 3 are similar to those of users 2 and 1. However, this user can not be characterized similarly as 2 and 1 as his/her off-topic activities statistics are lower compared to them. Neglecting such notable differences between these users would promote microblog users who have to be penalized regarding their over-interest in the other topics.

By considering both on- and off-topic user activities while modeling each user, we can have a clear overview of the user attachment to the specific analyzed topic or event. Such representation could also highlight the behavior of particular users who are generally used to intervene in crisis events. In the following, we describe the standard behavior of some of these users categories in terms of their on-topic and off-topic activities :

- *News outlet channels.* These users are extremely active in microblogs. They usually toggle between several topics rather than focusing on a single one. Information shared through these accounts is typically relevant however not exclusive enough.
- *Passengers.* These users are also known as sympathizers. They share or/and report little information regarding the event by expressing their solidarity with people affected by the event. Such solidarity messages are generally recorded as on-topic which makes it difficult for the identification model to distinguish such users and to classify them as non-prominent.
- *Locals.* These users would share various on-topic activities regarding the analyzed event and would neglect any other information regarding the other topics. During crisis events, users geolocated in the threatened areas are generally in panic and they are interested only in what is happening around them.

By learning on- and off-topic activities of both prominent and non-prominent users separately, the identification model will be able to distinguish the behavior particularities of each user category. However, the proposed on- and off-topic features in their current form are not expressive enough and cannot efficiently highlight the balance between on-topic activities and the off-ones. The real prominence of users would be better revealed if the user on-topic activities are adjusted accurately with respect to the off-topic ones.

5.4 Qualifying the Quantified User Activities

The selected raw feature categories, presented in the earlier chapter, have proved their effectiveness in the context of prominent users identification during crisis events. However, the current form of these features mainly point out the quantity of produced and shared information by each user independently of their quality and freshness.

These raw features have many correlations between them that can be explored to highlight the quality of user on-topic activities. For example, $R1$ and $R2$ raw features describe both the retweeting activity of the user regarding the other users' produced tweets. Similarly, $R1_{on}$ and $R1_{off}$ referring to the user on- and off-topic retweeting activeness respectively characterize the same type of user activity.

TABLE 5.1: Statistics of the activities of example three users. These users have the same statistics in terms of on-topic activities and different statistics in terms of off-topic ones.

	User 1		User 2		User 3	
Features	On	Off	On	Off	On	Off
T1 : Number of original tweets	3	3	3	6	3	1
T2 : Number of links shared	2	2	2	4	2	0
T3 : Number of keyword and hashtags	10	10	10	20	10	2
T4 : Number of favorites of original tweets	14	14	14	28	14	2
R1 : Number of retweets of other's tweets	2	2	2	4	2	1
R2 : Number of unique users retweeted by the user	1	1	1	2	1	1
R3 : Number of retweets of author's tweets	20	20	20	40	20	5
R4 : Number of unique users who retweeted author's tweets	13	13	13	26	13	5
M1 : Number of mentions of other users by the author	2	2	2	4	2	0
M2 : Number of unique users mentioned by the author	1	1	1	2	1	0
M3 : Number of mentions by others of the author	4	4	4	8	4	0
M4 : Number of unique users mentioning the author	2	2	2	4	2	0

In order to take advantage of these different raw selected features, we manually design new engineered features derived from the raw ones. These features aim at firstly to better reflect the on-topic user activities for describing the user on-topic behavior. We then design the features in a way to point out the quality of user on-topic activities by evaluating their rate of interest in the analyzed topic or event.

Inspired by the features presented by Pal & Counts (2011), we propose a new list of engineered features reflecting the user information quality according to his/her topical behavior. We present each user by a vector of features. The features included in this vector are designed by aggregating the user raw features of same nature on one hand and by adjusting his/her on-topic activities by the off-topic ones on the other hand. Our proposed engineered features are described in the following :

Topical Strength (F1) : estimates the value (or worthiness) of the evaluated user's on-topic tweets with respect to the off-topic ones. This feature promotes users that have collected more likes regarding their on-topic tweets than off-topic ones.

$$F1 = \frac{T4_{on}}{T4_{off} + 1} \tag{5.1}$$

Topical Attachment (F2) : indicates the involvement rate of the user regarding the analyzed topic by referring to the number of his/her original on-topic tweets adjusted by the off-topic ones. The more a user produces on-topic tweets compared to off-topic ones, the higher his/her Topical Attachment score would be.

$$F2 = \frac{T1_{on} + T2_{on}}{T1_{off} + T2_{off} + 1} \tag{5.2}$$

Retweeting Rate (F3) : measures the impact of the original tweets shared by the other users on the evaluated user topical activities. This measure is adjusted by the retweeting activity of the evaluated user regarding others' off-topic original tweets.

$$F3 = R1_{on} * log(R2_{on} + 1) - R1_{off} * log(R2_{off} + 1) \tag{5.3}$$

Retweeted Rate (F4) : calculates the impact of the topical original tweets produced by the evaluated user in the other microblog users. This feature is adjusted by the user's influence rate on the other off-topics.

$$F4 = R3_{on} * log(R4_{on} + 1) - R3_{off} * log(R4_{off} + 1) \tag{5.4}$$

Incoming Mention Rate (F5) : 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.

$$F5 = M3_{on} * log(M4_{on} + 1) - M3_{off} * log(M4_{off} + 1) \tag{5.5}$$

Outcoming Mention Rate (F6) : promotes users producing many on-topic mentions intended to several users on one hand and penalizes users addressing more off-topic mentions than on-topic ones on the other hand.

$$F6 = M1_{on} * log(M2_{on} + 1) - M1_{off} * log(M2_{off} + 1) \tag{5.6}$$

Centrality Degree (F7) : adjusts the number of on-topic followers and followees of each user with the number of his/her off-topic relations. This feature promotes users connected to more on-topic users than off-topic ones.

$$F7 = log(\frac{G1_{on} + 1}{G1_{off} + 2}) - log(\frac{G2_{on} + 1}{G2_{off} + 2}) \tag{5.7}$$

These hand-crafted features combine the different selected on- and off-topic user raw features that have proved their efficiency in the context of crisis events. The resulted engineering features offer a better representation of users by pointing out both the quantity of their topic activities and their quality by considering their off-topic activities.

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

$$x_i = (F1_i, F2_i, F3_i, F4_i, F5_i, F6_i, F7_i, T5_i) \tag{5.8}$$

5.5 Classification and Ranking of Prominent Users

To identify prominent users 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 goal behind this classification step is to reject most of the non-prominent users and retain the prominent ones. Such classification process would significantly reduce the number of users that have to be ranked. Through the ranking step, we would mainly focus on identifying the top prominent users regarding the specific analyzed event.

5.5.1 Prominent Users Classification using an SVM-trained Model

For our supervised classification problem, we have chosen support vector machines (SVM) as they are theoretically well-founded among machine learning techniques (Vapnik 1995, Boser et al. 1992). This machine learning model generally ensures a good empirical performance in a wide variety of pattern recognition and data mining applications. Our problem is a two-class problem, we want to discriminate prominent users in a specific event versus all other users. SVM separates the two classes of users by constructing a maximal linear hyperplane that has the largest distance to the nearest training-data point of any class. Generally, the larger the margin between the parallel constructed hyperplanes the lower the generalization error of the classifier will be. Data points U representing each user are expressed as follows :

$$U = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \tag{5.9}$$

Where x_i is a 8-dimensional vector of features representing each user in the training set, y_i denotes the class to which each user i belongs and is either 1 for prominent users or -1 for non-prominent. The SVM classification function $F(x)$ takes the following form :

$$F(x) = w \times x - b \tag{5.10}$$

Where w is the weight vector and b is the bias, which is computed by SVM in the training process to construct the classification model.

To correctly classify each user in U, $F(x)$ must return positive values for prominent users and negative values for the non prominent ones.

$$\begin{aligned} w \times x_i - b > 0 \text{ if } y_i = 1 \\ w \times x_i - b < 0 \text{ if } y_i = -1 \end{aligned} \tag{5.11}$$

If there exists a function F that correctly separates the users in the training set, then F has to maximize the margin zone in order to minimize misclassification errors. The hyperplanes bounding the margin are represented as :

$$\begin{aligned} w \times x_i - b = 1 \text{ , and} \\ w \times x_i - b = -1 \end{aligned} \tag{5.12}$$

To measure the distance between the hyperplane to a vector x_i is formulated as :

$$\frac{|F(x_i)|}{||w||} \tag{5.13}$$

Where $\frac{1}{||w||}$ is the margin value.
Hence, to build our classification model, we need to minimize w by solving the following

optimization problem :

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & Q(w) = \frac{1}{2}||w||^2 \\ \text{subject to} \quad & y_i(w \times x_i - b) \geq 1 \ \forall (x_i, y_i) \in U \end{aligned} \tag{5.14}$$

Once the minimizer w is obtained, the induced SVM classifier is given as :

$$SVM(x) = sgn((w \times x_i - b) \tag{5.15}$$

Different types of SVM nonlinear kernels have also been considered to select the best function for prominent users identification based on the defined features. These kernels classify microblog users based on nonlinear boundaries learned *a priori*. The linear kernel has been experimentally selected as the most efficient kernel in our case.

5.5.2 Ranking Prominent Users using an SVM-trained Model

As the classification model is built using linear separators, it is possible to use the learned parameters which resulted from the training phase directly to rank each user in the test set. Assume that T represents m data points which have been classified by our model as prominent users during the classification phase of test set. T is composed of 8-dimensional feature vectors of m users $\{\tilde{x_i}\}$:

$$T = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m\} \tag{5.16}$$

In order to rank these users, we extract the learned values of $\tilde{w}$ and $\tilde{b}$ resulting from the training phase, and we compute the score obtained for each user x_i using :

$$R(x_i) = \tilde{w} \times x - \tilde{b} \tag{5.17}$$

The score $R(x_i)$ is then used in order to attribute a rank for a user, such that for users i and j : if $R(x_i) > R(x_j)$, this means that i is more prominent than j.

5.6 Experiments and Evaluation

In order to evaluate the importance of qualifying user on-topic activities, we conduct in a first step an in-depth study for analyzing the distribution of the topical activity of prominent and non-prominent users regarding their social connections. Topical activities belonging to prominent and non-prominent users who have been interested in the Herault flooding

dataset were used for this study. We aim to prove that a high user's on-topic activeness does not necessarily imply the prominence of the user. In subsections 5.6.3 and 5.6.2, we conduct some experiments evaluating the identification performance of our proposed model learned by considering on-topic user activities adjusted by the off-ones compared to other state-of-the-art baselines.

5.6.1 Studying Microblog Users Topical Activities during Herault Floods

State-of-the-art key users identification systems are mostly criticized for being sensitive to popular microblogs users and celebrities. As explained previously, this sensitivity makes such systems unsuitable for prominent users identification in the context of crisis events. Through this study, we first explore microblog users' networking dimension (i.e. number of followers) in order to understand the involvement rate of popular and ordinary microblog users in crisis events. We then study the topical activeness of the different microblog users according to both their popularity in the microblogging platform and their prominence during the Herault floods. Through this study, we aim to point out the topical behavior specificities distinguishing prominent users.

In order to conduct this study, we categorize the detected active microblog users during the Herault floods event into 4 categories :

- Category 1 (Cat1). refers to users having less than 1,000 followers.
- Category 2 (Cat2). includes users having more than 1,000 followers and less than 10,000.
- Category 3 (Cat3). refers to users having a followers number between 10,000 and 100,000.
- Category 4 (Cat4). includes users having more than 100,000 followers.

Evaluating Microblog User Prominence per Category

This evaluation is conducted using the Herault floods dataset and its ground truth described in Chapter 3. We evaluate both prominent and non-prominent microblog users included in this dataset according to their popularity and prominence. Figure 5.1 and Table 5.2 report both the number of selected prominent microblog users and the total number of active users interacting during the analyzed event per category. According to the reported results, we can observe that ordinary microblog users having less than 1000 followers are the most interested in the analyzed event. Such results were expected. During such events, ordinary users are the most susceptible to share the required information by the emergency

teams. On the other side, 30% of prominent users having more than 1000 followers have been judged as prominent and the other 70% refer to users having less than 1000 followers. Highly connected users belonging to categories 4 and 3 which mostly refer to celebrities and news outlet accounts were mainly judged as non-prominent as reported in Table 5.2.

Overall, we can conclude that prominent users in the context of crisis events mostly refer to ordinary users. User popularity does not necessarily imply his/her prominence. Such conclusions confirm and support the results found in the previous chapter where the user social network features have been experimentally classified as irrelevant for prominent users identification during crisis events. We also note that prominent users in the context of crisis events do not principally refer to domain experts. Most of the selected prominent users according to our conducted ground truth have less than 1000 followers. Domain experts generally have an important number of followers interested in their expertise domain. Thus, the identification process of such users in the context of crisis events has to be distinguished from the context of other identification systems targeting domain experts or/and influencers. Prominent users targeted in our context can refer either to ordinary microblog users or popular and domain expert microblog users.

Evaluating Prominent and Non-Prominent Microblog Users' Topical Activities per Category

We study herein both microblog users prominence and topical activeness during Herault floods per category. Through this study, we aim to point out the topical characteristics that can differentiate prominent and non-prominent users per user category. In other words, we study if there is any correlation between on- and off-topic raw features that penalize non-prominent users sharing many relevant however outdated on-topic information.

Tables 5.3, 5.4 and 5.5 report both on-topic and off-topic activities belonging respectively to all active microblog users, prominent ground-truth users and the non-prominenent ones. These activities are measured by computing the averages of resulted raw features of each user category.

According to these reported statistics, we observe that most of the evaluated users have been more active regarding the other topics than the Herault floods-related topic. Moreover, we note that the recorded averages of users on-topic features increase as we go from category 1 to 4. Popular users have registered the highest on-topical activity. However, such high activity does not necessarily indicate that they are prominent. As stated previously, few users from this category has been retained as prominent. In order to understand why such users are discarded, we study the topical activity of retained prominent users and the non-retained ones per category as presented in Table 5.4 and 5.5. According to the reported results, we can observe that the average of on-topic features related to prominent users is

TABLE 5.2: Prominent and non-prominent users statistics per category

	Cat.1	Cat.2	Cat.3	Cat.4
Prominent Users	62	23	2	3
Non-prominent Users	2,596	570	57	19
All active Users	2,663	594	59	22

significantly higher than their off-topic ones. However, the averages of on-topic features registered by non-prominent users are lower than the off-topic ones.

Overall, we conclude that there is a high correlation between on- and off-topic raw features that have to be considered in order to distinguish between prominent and non-prominent users. The recorded statistics and findings support our assumptions that users toggling between several topics have to be discarded on one hand, and that the relation between on- and off-topic features have to be considered to measure users topical attachment. Moreover, we show that a high on-topical activity does not automatically reflect the prominence of the user, such activities have to be adjusted by the off-ones.

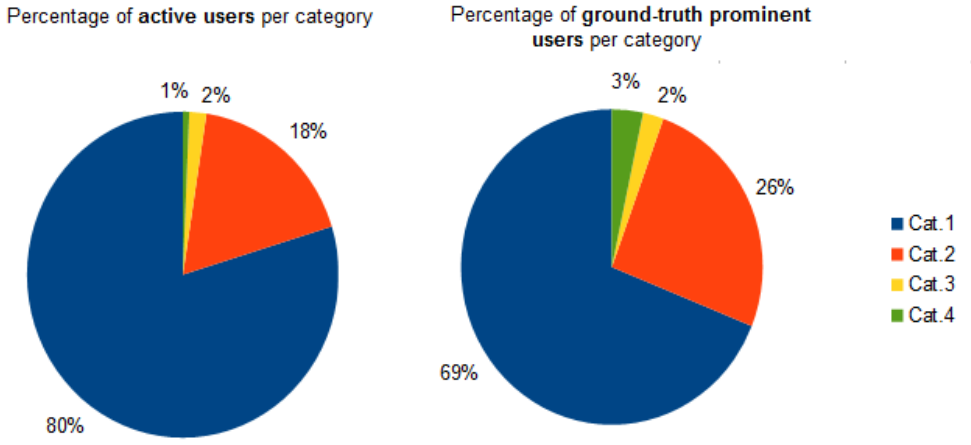

FIGURE 5.1: Prominent and non-prominent microblog users distribution per category. These statistics are computed by referring to the MASIR extracted data during the Herault flooding event. Active users refer to users who have shared at least one event-related tweet.

5.6.2 Performance of the Automatic Users Classifier Model

In this subsection, we evaluate the performance of our proposed identification model based on new designed engineered features exploring the correlations between on- and off- user activities features.

In order to train and test our proposed model, we divided the Herault dataset into training and test sets using two different partitions as described in Table 5.6. We also applied the

TABLE 5.3: Recorded on-topic and off-topic raw features averages for the **3338 users active** regarding the event. Features averages are computed per user category.

		T1	T2	T4	R1	R2	R3	R4	M1	M2	M3	M4
Cat.1	on	0,39	0,24	0,14	1,22	1,06	0,40	0,40	0,16	0,09	0,03	0,02
	off	1,50	0,86	0,63	2,93	2,27	0,23	0,23	1,37	0,80	0,13	0,04
Cat.2	on	0,76	0,53	0,62	1,32	1,1	2,79	2,79	0,21	0,13	0,19	0,15
	off	2,36	1,58	2,91	3,55	2,9	0,5	0,5	1,96	1,71	0,3	0,18
Cat.3	on	1,08	0,98	2,27	0,47	0,39	6,53	6,53	0,41	0,15	1,17	1,05
	off	2,88	2,66	17,36	2,29	1,95	5,90	5,90	1,27	1,14	1,95	1,05
Cat.4	on	1,64	1,32	7,77	1,09	0,73	28,95	28,95	0,23	0,14	3,64	3,23
	off	2,18	1,64	17,27	0,86	0,59	17,18	17,18	1,14	2,91	7,09	5,14

TABLE 5.4: Recorded on-topic and off-topic raw features averages for the **90 prominent users** selected using our user categorization study. Features averages are computed per user category.

		T1	T2	T4	R1	R2	R3	R4	M1	M2	M3	M4
Cat.1	on	4,56	3,32	2,18	6,6	4,27	10,02	10,02	1,26	0,68	0,53	0,47
	off	0,56	0,45	1	0,34	0,32	5,42	5,42	0,16	0,39	0,31	0,26
Cat.2	on	6,39	3,7	5,3	9,3	6	34,65	34,65	1,22	0,7	2,43	1,74
	off	0,13	0,09	0,17	1,52	1,39	0,26	0,26	0,04	0	0,52	0,35
Cat.3	on	10	10	5	0	0	72	72	0	0	19,5	17,5
	off	0	0	0	0	0	46	46	0	0	18,5	13,5
Cat.4	on	4,67	3	17,67	0,67	0,33	89,33	89,33	1	0,67	10,67	9
	off	0	0	0	0	0	30,33	30,33	0	0	13,67	9

TABLE 5.5: Recorded on-topic and off-topic raw features averages for the **3248 non-prominent users** that have to be rejected as resulted through our user categorization study. Features averages are computed per user category.

		T1	T2	T4	R1	R2	R3	R4	M1	M2	M3	M4
Cat.1	on	0,29	0,17	0,09	1,09	0,98	0,17	0,17	0,13	0,07	0,01	0,01
	off	1,52	0,87	0,62	2,99	2,31	0,1	0,1	1,4	0,81	0,12	0,04
Cat.2	on	0,54	0,4	0,43	1	0,91	1,51	1,51	0,16	0,11	0,1	0,09
	off	2,45	1,64	3,02	3,63	2,96	0,51	0,51	2,04	1,78	0,29	0,17
Cat.3	on	0,77	0,67	2,18	0,49	0,4	4,23	4,23	0,42	0,16	0,53	0,47
	off	2,98	2,75	17,96	2,37	2,02	4,49	4,49	1,32	1,18	1,37	0,61
Cat.4	on	0,77	0,67	2,18	0,49	0,4	4,23	4,23	0,42	0,16	0,53	0,47
	off	2,98	2,75	17,96	2,37	2,02	4,49	4,49	1,32	1,18	1,37	0,61

principle of 3-fold cross validation for both partitions 1 and 2 in order to avoid any bias in experiments. The resulted models are trained with the libSVM library under Matlab.

TABLE 5.6: The different partitions of data used in the training and test phases. C1 and C2 refer to prominent microblog users class and non-prominent microblog users class respectively.

	Partition 1		Partition 2	
	Training_1 (60%)	Test_1 (40%)	Training_1 (80%)	Test_1 (20%)
C1	54	36	72	18
C2	1945	1,297	2,593	649

We compared our proposed classification model with several baselines and state-of-the-art methods as described in the following :

Our model : our model learns both the user on- and off-topic behavior characterized by a list of new engineered features (see Section 5.4). These features are designed by considering the importance of adjusting on-topic user activities by the off-ones.

Baseline 1 : this model also uses engineered features similar to those described in Section 5.4. However, these features neglect the on-topic adjustment principle introduced in this chapter. Through this baseline, we aim to prove the role of the on-topic features adjustment to enhance the identification model performance.

Baseline 2 : this model uses all the on-topic raw features presented in Table 4.1 of Chapter 4. Based on this baseline, we aim to prove that on-topic raw features are not rich enough when they are considered separately without exploiting the possible correlation between them.

Baseline 3 : this model uses the engineered features proposed by Pal & Counts (2011). Through, this baseline, we aim to prove that our proposed adjusted features are more effective than those proposed by Pal & Counts (2011) for both identifying and ranking prominent microblog users.

Baseline 4 : 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. Through this model, we aim prove that our vector-based ranking model is more efficient than the graph-based models.

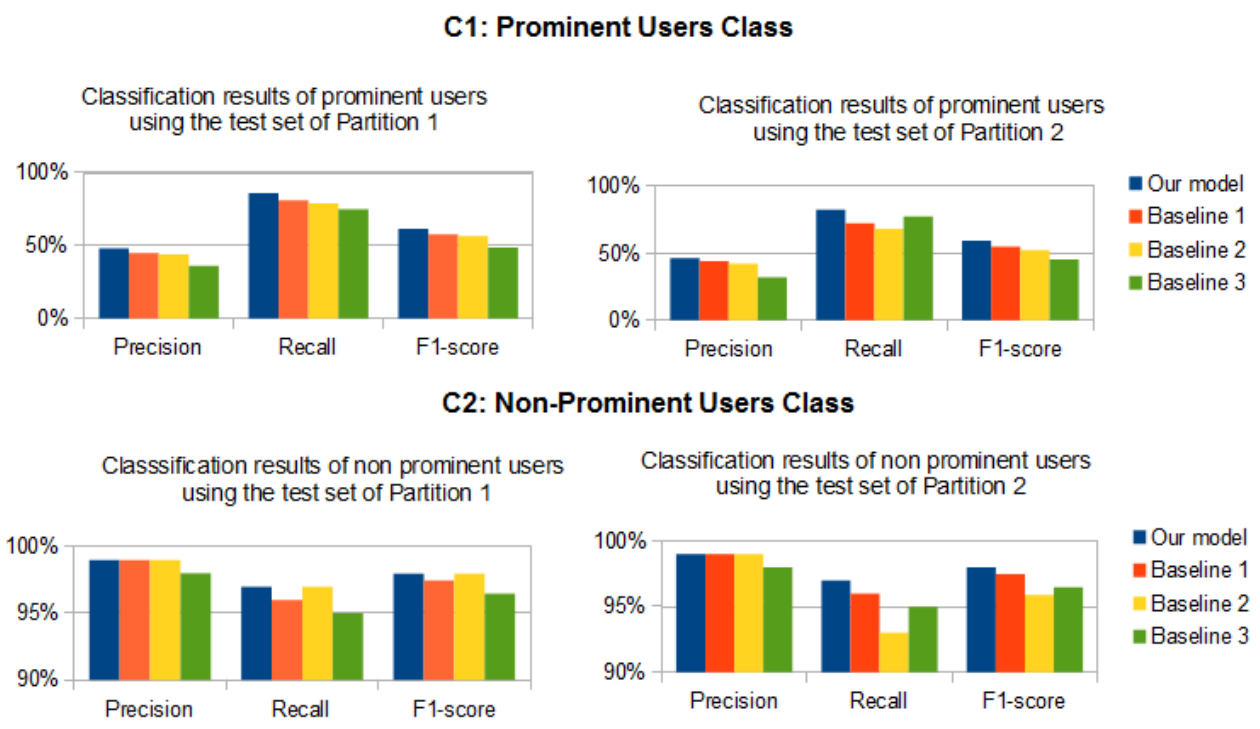

FIGURE 5.2: Comparing the classification performance of our proposed model for the classes **C1** and **C2**, referring respectively to detected prominent microblog users and non-prominent microblog users, with the baselines 1, 2 and 3. These models are evaluated in terms of Recall, Precision and F1-score.

Figure 5.2 shows the precision and recall results of our classification model for prominent users identification compared to all the other baselines. We note that the obtained results by our classification model are significantly higher than those obtained by 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 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 other baseline methods which use only on-topic features to represent user importance. Hence, we demonstrate that adjusting on-topical metrics with off-topical ones improves the classification results.

5.6.3 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 for top prominent users detection. We thus compare our model with the baselines 3 and 4.

We have ranked the set of users identified in class C1 using the different ranking baseline models. Then, we picked out the top 15 users detected by each baseline. The precision of the ratings accorded by each baseline is 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 5.3.

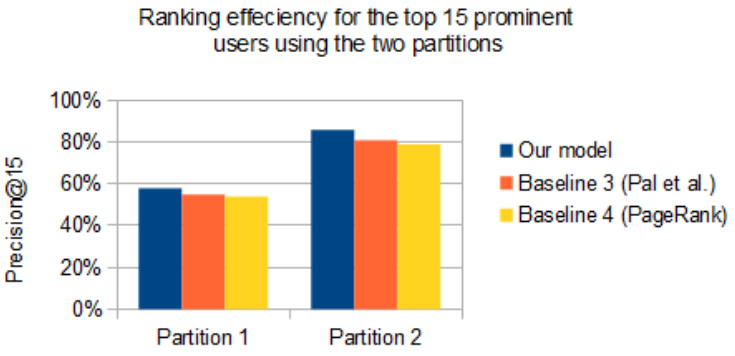

FIGURE 5.3: A comparison between our ranking model and state-of-the-art ranking baselines performance. The prominent users ranking results are computed in terms of Precision@15 measure.

According to these results, we observe that our model achieves the highest precision compared to the other baseline models, with a precision of 86% in Partition 2. Therefore, our designed high level features outperform the evaluated state-of-the-art features for both the identification of prominent users and the detection of the top ones. Moreover, we note that the graph-based model represented by baseline 4 achieves the worst results compared to the models constructed using vector-based classification and ranking models. Such results proves that our model is the most adapted model for prominent users identification in the context of crisis events.

5.7 Conclusion

This chapter has presented a classification and ranking model to identify prominent users in a specific topic or event. This model is constructed by learning the on-topic prominent and non-prominent users activities adjusted by the off-ones. 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 the other models that are based only on on-topic features. Our engineered features exploring the various correlation between the on- and off-topic raw features categories have outperformed the different baselines based on standard features.

Furthermore, we have shown through this chapter the importance of adjusting the on-topic raw features by the off-ones for making the model insensitive to popular users sharing various relevant however outdated on-topic information. We have also noted that prominent users in the context of crisis events do not necessarily refer to popular users referring to either influencers or/and domain experts, such users mainly refer to ordinary users implicated voluntarily or involuntarily in the event. Through the analysis of prominent and non-prominent users activities, we observed that considering only the statistics of users on-topic activities is not enough sufficient to distinguish prominent users.

Despite the challenges related to the nature of our real data, we have shown how the used supervised learning algorithm (SVM) can still be effective using appropriate features. Our model outperforms the graph-based ranking models for the identification of prominent users. While our proposed model based on new engineered features achieved good results, such vector-based representation does not efficiently reflect the real user behavior specificities over time. Users interacting from the beginning of the event are represented similarly to those interacting at its end. Future design work is needed to propose a more descriptive representation characterizing the user behavior differences over time.

Chapter 6

Learning Temporal Sequences of Features for Prominent Users Prediction

Contents

6.1 Introduction

The performance of prominent users identification systems is directly associated with the effectiveness of the adopted microblog user modeling approach. The chosen microblog user modeling approach has to reflect the main behavioral differences that can make prominent users detectable among the large number of non-prominent ones. Microblog users behavior is typically not static, it undergoes various changes over time according to the user activeness and interests. Modeling effectively user behavior evolution over time can help to facilitate the identification of prominent users in the context of crisis events.

General purpose existing approaches dealing with prominent users behavior detection have neglected the temporal dimension of users activities. This would give a misleading image of users behavior in real scenarios. Most of these proposed approaches mainly focused on characterizing users statistically (i.e. using straightforward mathematical formula such as the sum of a user shared tweets) regarding their different shared activities independently of their temporal distribution. Using such standard statistical user characterization approach, users interacting at an early stage of the event would be represented similarly as those who have become active only at its end. Modeling the temporal distribution of user activities would highlight many hidden patterns reflecting prominent users behavior in the context of crisis events.

This work is thus designed to alleviate this shortcoming. We thus present the following contributions : (1) a novel representation of microblog user behavior as a temporal sequence of features that characterize both the on- and off-topic user activities, (2) a probabilistic model for the prediction of prominent microblog users during crisis events. The prediction model learns to differentiate between prominent and non-prominent users behavior using ergodic Mixture of Gaussians Hidden Markov Models (MoG-HMM).

The rest of this chapter is organized as follows. Section 6.3 describes the temporal dimension importance for user behavior modeling and classification. Section 6.4 details how the temporal dimension has to be integrated for user behavior modeling. Section 6.5 discusses our proposed modeling approach. Section 6.6 describes the MoG-HMM learning model for the classification and raking of prominent microblog users. Experimental evaluation is presented in Section 6.7. Finally, we conclude with directions for future work in Section 6.8.

6.2 Research Questions

To the best of our knowledge, the temporal dimension has never been explored for microblog users behavior modeling in the context of key microblog users identification. In this research, we aim to integrate this dimension in both prominent microblog users modeling and identification processes. Both prominent and non-prominent user behavior temporal patterns need to be learned to ease the identification process. This research helps us to answer several questions :

1. How can we model the temporal distribution of users topical activities? How to differentiate between users interacting at an early stage of the event and those who have become active at its end?
2. How to learn prominent and non-prominent users behaviors patterns over time? How to predict prominent microblog users over time during a real-world case?
3. Can we improve prominent users identification performance by considering the temporal distribution of users activities?

6.3 Focus on User Activities Temporal Distribution

The temporal dimension characterizing user activities distribution over time is generally neglected while representing microblog user behavior during a specific period of time. Such dimension is of particular importance in the context of crisis events. It can point out many behavior specificties characterizing microblog users who are the source of the required information. Focusing only on user on- and off-topic information statistics would judge and represent users who have been active in the beginning of the event the same way as those who have become active at its end.

Let us assume the example of the following four microblog users having the same number of on- and off-topic activities. As described in Figure 6.1, while the different user activities statistics are equal for all users, their temporal distribution is completely different :

User 1 on- and off-topic tweets are shared in a balanced way from the beginning of the event until its end.

User 2 on-topic tweets are distributed in a balanced way during the two third period of time of the event. This user off-topic activities have been mainly shared at the end of the event.

User 3 on- and off-topic activities have been mostly concentrated in a short period of time of the event. During the remaining event periods, this user was inactive.

User 4 related off-topic tweets were shared at the beginning of the event. This user has focused his/her attention regarding the event only before a while of its end.

By analyzing the topical activities distribution of these users, many behavioral differences can be pointed out. Users permanently toggling between on- and off-topic information differ from those who have focused on on-topic information for a long period of time.

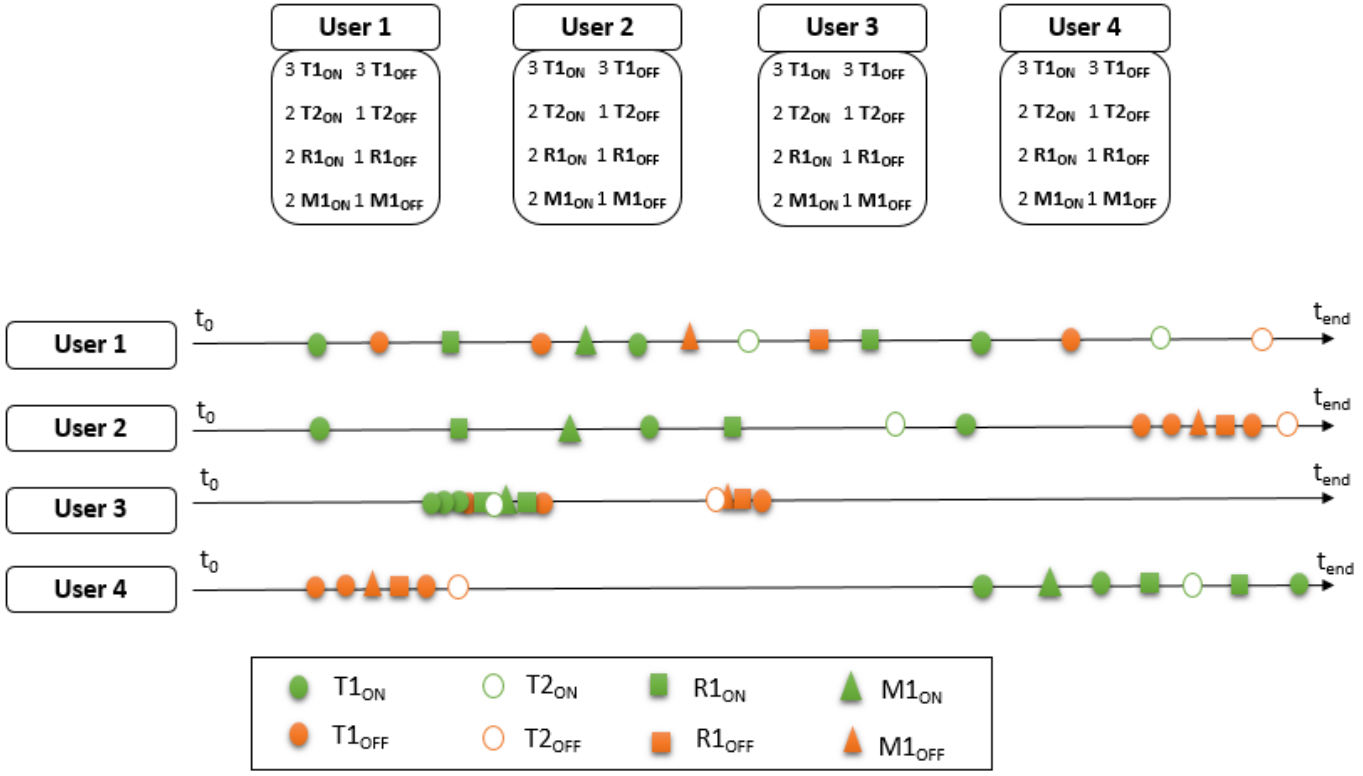

FIGURE 6.1: Mapping the temporal distribution of four microblog users topical activities during a specific event. The characteristics of these users are similar in terms of topical activities statistics but different in terms of their topical behavior over time. The different activities are described using the raw topical features defined in Chapter 4.

Considering such temporal distribution differences while modeling microblog user behavior would highlight the specific behavioral patterns characterizing prominent users. Describing users only regarding the statistics of their topical activities would make the identification model sensitive to users sharing outdated information.

6.4 Temporal Dimension Integration

In order to predict prominent users according to their realistic behavior, we integrate the temporal dimension for user behavior modeling and analysis. The integration of this dimension is processed as described in the following approaches :

1. *Time-sensitive user behavior modeling approach.* consists of representing users so that to reflect their temporal behavior during an event. Each user u has to be represented by a temporal sequence of m vectors $R_u = (V_u^1, V_u^2, ..., V_u^m)$ instead of a single one 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.

2. *Time-sensitive prominent user behavior analysis approach.* consists of learning to classify users' temporal sequences of features in prominent users class c_1 or non-prominent users class c_2 . We thus train two probabilistic models H_{c_1} and H_{c_2} by training the 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.

We detail how these steps are performed to build and test our time-sensitive prediction model in the next sections.

6.5 User Behavior Modeling as Temporal Sequences

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.

The temporal dimension of user activities shared during the event is considered while modeling user behavior. Each user has to be represented by a temporal sequence of feature vectors rather than a single one. These feature vectors are computed based on the engineered features proposed in Chapter 5. The time-line of each event phase is divided into

equispaced intervals at m time-stamps t_1,t_2,t_3,...,t_m from the beginning of the event until its end. During each interval, users activities are characterized by a set of features rather than a single one 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 Chapter 5– describe the user behavior regarding an event (on-topic activity) and also regarding other topics (off-topic activity) during each time interval. Figure 6.2 illustrates – in its upper part – such a user representation.

As described in this Figure, at each time-stamp t_i, we represent each microblog user u by a feature vector $V^{t_i}(u)$ characterizing his/her behavior from the time-stamp t_{i-1} to t_i. For each user u and each time stamp t_m, we compute the engineered features by taking into account both the on-and off-topic user activities. Once these features are computed, each user u can be represented by the following feature vector at each time stamp t_i.

$$V_{P_j}^{t_i}(u) = (F_1^{(t_i)}, F_2^{(t_i)}, ..., F_7^{(t_i)}) \tag{6.1}$$

Then, the resulted vector is added in the sequel of the temporal sequence $R^{t_{i-1}}(u)$ composed of the previously calculated vectors from the beginning of that phase.

$$R^{t_i}(u) = (V^{t_1}(u), V^{t_2}(u), ..., V^{t_{i-1}}(u), V^{t_i}(u)) \tag{6.2}$$

The set of concatenated feature vectors computed at all the time stamps represent the temporal sequence of the user behavior. Segmenting the sequence of user activity into time-series feature vectors offers a rich and personalized user representation. Users sharing the same quantity of information with different temporal distribution would not be similarly represented. Our user modeling approach offers a more comprehensive vision of users behavior by taking into account the evolution of user activity over time. It provides a detailed user representation closer to his/her real behavior in microblogs. Such representation eases the identification of users behavior regularities, similarities and dissimilarities at each phase.

6.6 Learning to Classify User Temporal Sequences

In order to classify the time-series of feature vectors V describing each microblog 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 6.2 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.

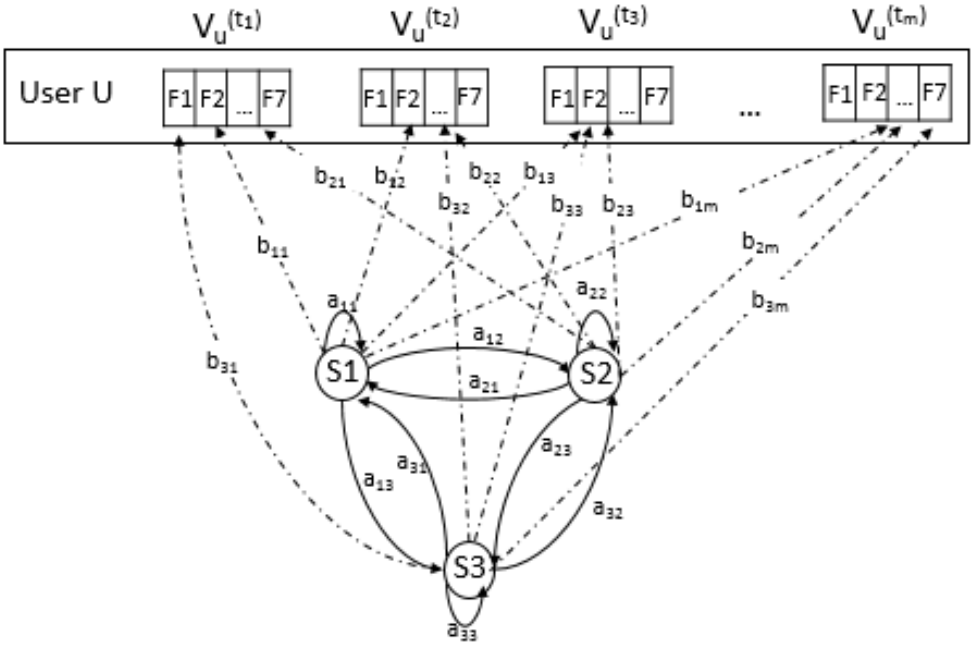

FIGURE 6.2: A 3-state ergodic HMM example for time-series user activities representation during a specific event.

To learn the parameters for our MoG-HMM ergodic models, we use the Baum-Welch algorithm (Dempster et al. 1977). This algorithm is based on the EM algorithm to search for the maximum probability of the HMM models parameters that better fit the observed temporal users sequences in the training data.

$$H = \arg\max_{H} P(V_{training}|H) \tag{6.3}$$

A MoG-HMM model H is described by the quadruplet $H = \{S, \pi, A, B\}$.

where

1. $S = S_1, S_2, S_3, ..., S_k$ 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 $(X_t \in S)_{1 \leq t \leq m}$.

2. π denotes the initial probability of the different states.

3. A is the state transition probability matrix to change from state S_i to S_j, $A = a_{ij}$ where $a_{ij} = P(X_{t+1} = S_j | X_t = S_i)_{1 \leq i,j \leq k}$.

4. 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 $b_i(t) = P(V^t|X_t) = (S_i)_{1 \leq i \leq k}$.

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 6.4.

$$b_i(V^t) = \sum_{k=1}^{M} c_{ik} \mathcal{N}[V^t, \mu_{ik}, W_{ik}] \tag{6.4}$$

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. These PDFs can be constructed using a single univariate Gaussian. In this case there is no mixture weight ($M = 1$), PDF is characterized by the mean μ and the covariance σ.

$$b_i(V_u^t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{\frac{(V_t - \mu)}{2\sigma^2}} \tag{6.5}$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are computed from the observed users characterized by m feature vectors V^t which all have the same state S_i associated.

$$\widehat{\mu} = \bar{x} = \frac{1}{m} \sum_{t=1}^{m} V_t \tag{6.6}$$

$$\widehat{\sigma}^2 = \frac{1}{m} \sum_{t=1}^{m} (V_t - \bar{x})^2 \tag{6.7}$$

In order to find better parameters for our MoG-HMM ergodic models, we use the Baum-Welch algorithm (Dempster et al. 1977). This algorithm is based on the EM algorithm to search for the maximum probability of the HMM models parameters that better fits the observed temporal users sequences in the training data.

$$H = \arg\max_{H} P(V_{training}|H) \tag{6.8}$$

Once the models parameters H_{c1} and H_{c2} are set through training, we can compute the probabilities $P(V_u|H_{c1})$ and $P(V_u|H_{c2})$ of any microblog user to belong to each class given the two learned models. These probabilities are obtained using the forward-backward algorithm (Baum & Eagon 1967). 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.

6.7 Experimental Evaluation

To conduct experimental evaluation on real data, we used the collected Herault database described in Section 3.5.2 of Chapter 3 and its corresponding ground-truth described in Section 4.7.1 of Chapter 4. We recall that according to this ground-truth, 90 users were classified in the prominent users class $C1$ and the remaining 3242 ones in the non-prominent users class $C2$. In the next sub-sections, we describe the results of the different conducted experiments for our prominent user prediction model evaluation. Our prediction model was learned with the HMM toolbox under Matlab.

6.7.1 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} prediction models, and the remaining 40% as test data. Features characterizing user behavior were sequentially extracted at each time interval of 90 minutes from the beginning of the event until its end. Thus, at the end of the event each user would be 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 key users identification, we use *Precision, Recall* and *F1-score* measures, described in Section 4.7.2 of Chapter 4, to evaluate the performance of our prediction model for classifying users. We also use the *Precision@K* measure to evaluate our model efficiency in terms of ranking.

$$\text{Precision@K} = \frac{\#\text{detected true prominent users in top K}}{K}$$

where : K =number of ground-truth prominent users (i.e. $K = 90$ by referring to the full dataset describing the Herault Floods)

We learn new $H_{c1}(t_i)$ and $H_{c2}(t_i)$ prediction models after each 90 minutes starting from the beginning of the event. Overall, 32 H_{c1} and 32 H_{c2} models were learned. Each model characterizes prominent or non-prominent users behavior from the beginning of the event until each time-stamp t_i. The training data input is composed of a temporal sequence of feature vectors characterizing the user behavior. The features composing each vector are sequentially computed at each interval of 90 minutes from the beginning of the event. For example, after 6 hours from the announcement of the event, each user would be represented by a sequence of 4 vectors.

TABLE 6.1: Prominent users identification performance for different N_S and N_G in terms of Precision@K measure.

N_s/N_G	1	2	3	4	5	6
1	49%	71%	66%	64%	61%	50%
2	61%	68%	68%	49%	5%	5%
3	44%	**74%**	66%	5%	5%	5%
4	49%	0%	0%	64%	5%	5%
5	49%	0%	0%	5%	61%	5%
6	44%	0%	5%	0%	0%	5%

In order to estimate the values of parameters for the representation of microblog users behavior through H_{c1} and H_{c2} prediction 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 6) over time with the training dataset. The experimental results relative to H_{c1} and H_{c2} prediction models are shown in Table 6.1. Models learned using the parameters values giving the best Precision@K results are retained to test their performance using the test dataset. Table 6.1 reports the obtained Precision@K results using different parameters values at the end of the Herault event. According to these experiments, the parameters values $N_S = 3$ states and $N_G = 2$, yield the best result of 74%.

6.7.2 Importance of Time-series Representation

To demonstrate the effectiveness of our temporal sequence representation approach for the classification and ranking 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 is represented by considering longer periods of time when extracting a new feature vector. Figure 6.3 shows how the sequence length variation affects the performance of our model. The classification and ranking performance of our model using different time granularities for users modeling are measured using the F1-score and the Precison@K respectively. According to the obtained 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 by considering multiple time stamps.

6.7.3 Our Prediction Model Efficiency Comparison

In order to evaluate the effectiveness of our proposed HMM temporal sequence classification and ranking model, we conduct several experiments comparing the performance of the following baselines :

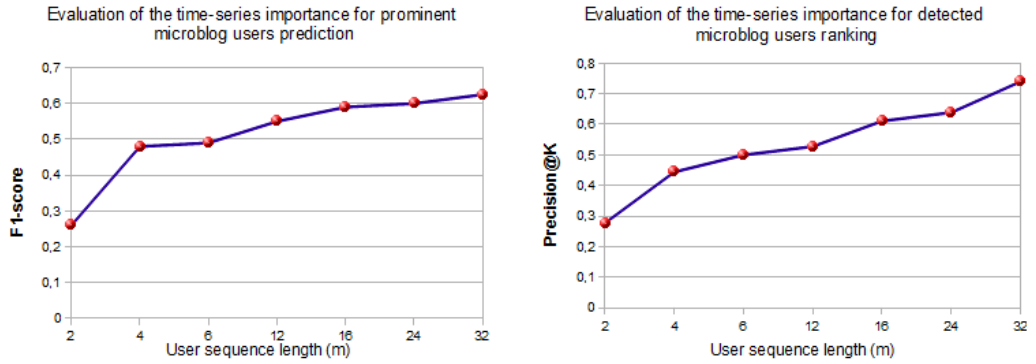

FIGURE 6.3: The effect of time-series granularity variation on our time-sensitive model performance.

our model : refers to our described microblog users prediction models. This model represents users by a sequence of feature vectors. Prediction models are built by learning the user behavior at different time stamps of the event.

Baseline 1 (Pal) : refers to the topical authorities identification model proposed by Pal & Counts (2011). This model represents users using a single features vector composed of on-topical engineered features. It uses unsupervised machine learning algorithms for both clustering and raking microblog users.

Baseline 2 (SVM) : refers to the identification model proposed in the previous chapter. This model represents users using a single vector user representation. It is learned using the supervised machine learning algorithm SVM.

Figures 6.6 and 6.7 report the prediction results obtained by the different baselines in terms of Precision, Recall, F1-score and Precision@K. The reported results in Figure 6.6 show that our model significantly outperforms the other baseline models classifying prominent microblog users independently of the temporal distribution of their topical activity. Our model has yielded promising prediction results, where it has identified a significant number of prominent users with a high precision from one third of the event duration. We also observe that Pal's model outperforms our model in terms of Recall. However, this model has registered low results in terms of Precision as an important number of non-prominent users were misclassified. Such high Recall results could be advantageous only on the case where most of true prominent users are high ranked or a small number of non-prominent users have been misclassified which is not the case for Pal.

The ranking performance comparison of the different baseline models is reported in Figure 6.7. According to these results, our model has also recorded promising results. Most of detected prominent users were highly ranked starting from the one third of the event duration. We also observe that the Pal's model ranking results are extremely lower than

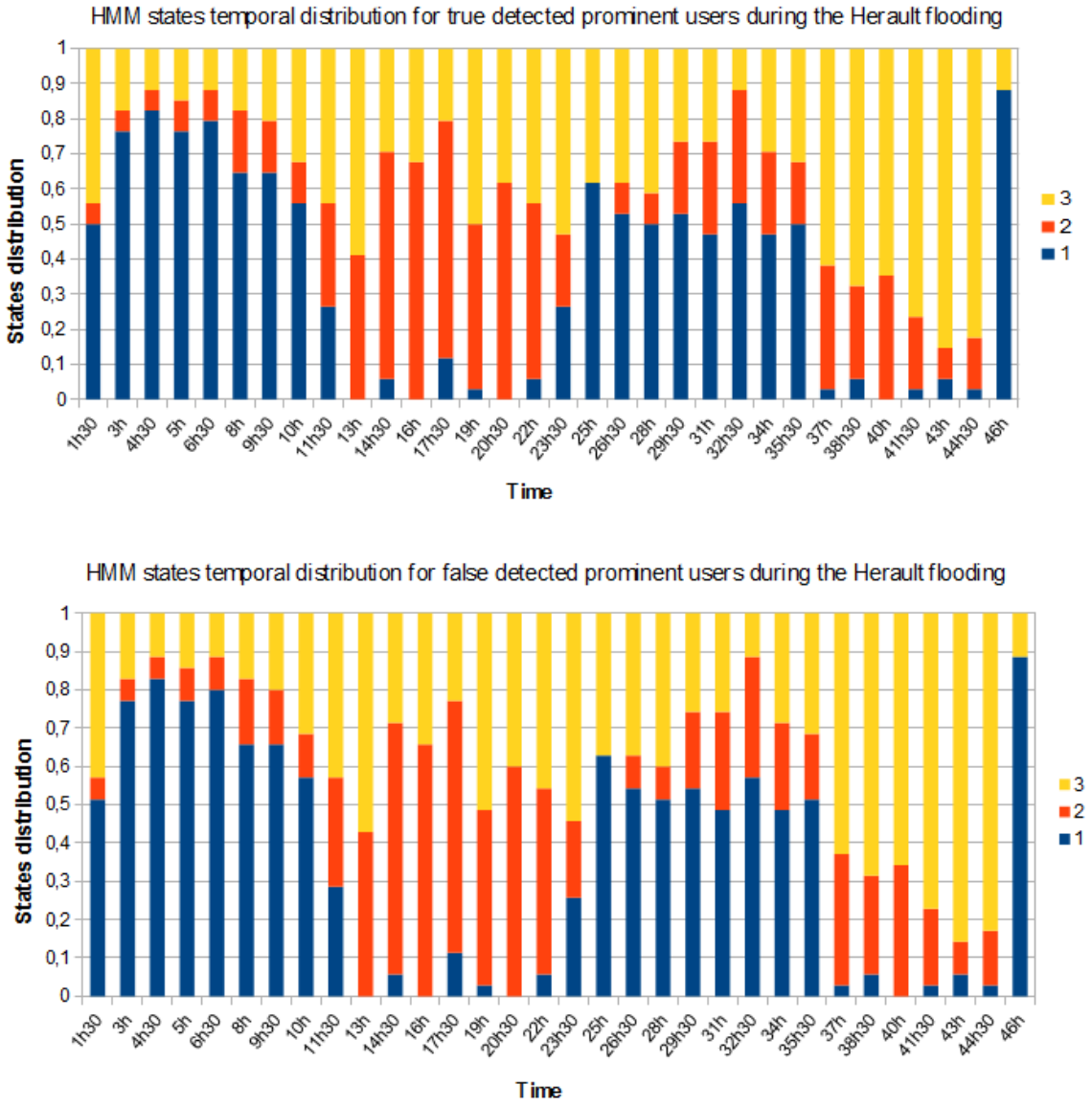

FIGURE 6.4: Temporal distribution of the **true and false detected prominent microblog users** by our learned ergodic HMM model. States 1,2 and 3 are the states set by the model to learn users behavior similarities and dissimilarities. The number of states was set experimentally as described in Section 6.7.1.

the other baseline. Such results confirm the in-adaptability of time-insensitive models for prominent users identification in the context of crisis events.

Through these experiments, we prove the importance of characterizing the temporal distribution of user activities over time for prominent users identification in the context of crisis events. We also show that our prediction model is able to predict most of prominent at an acceptable stage of the event.

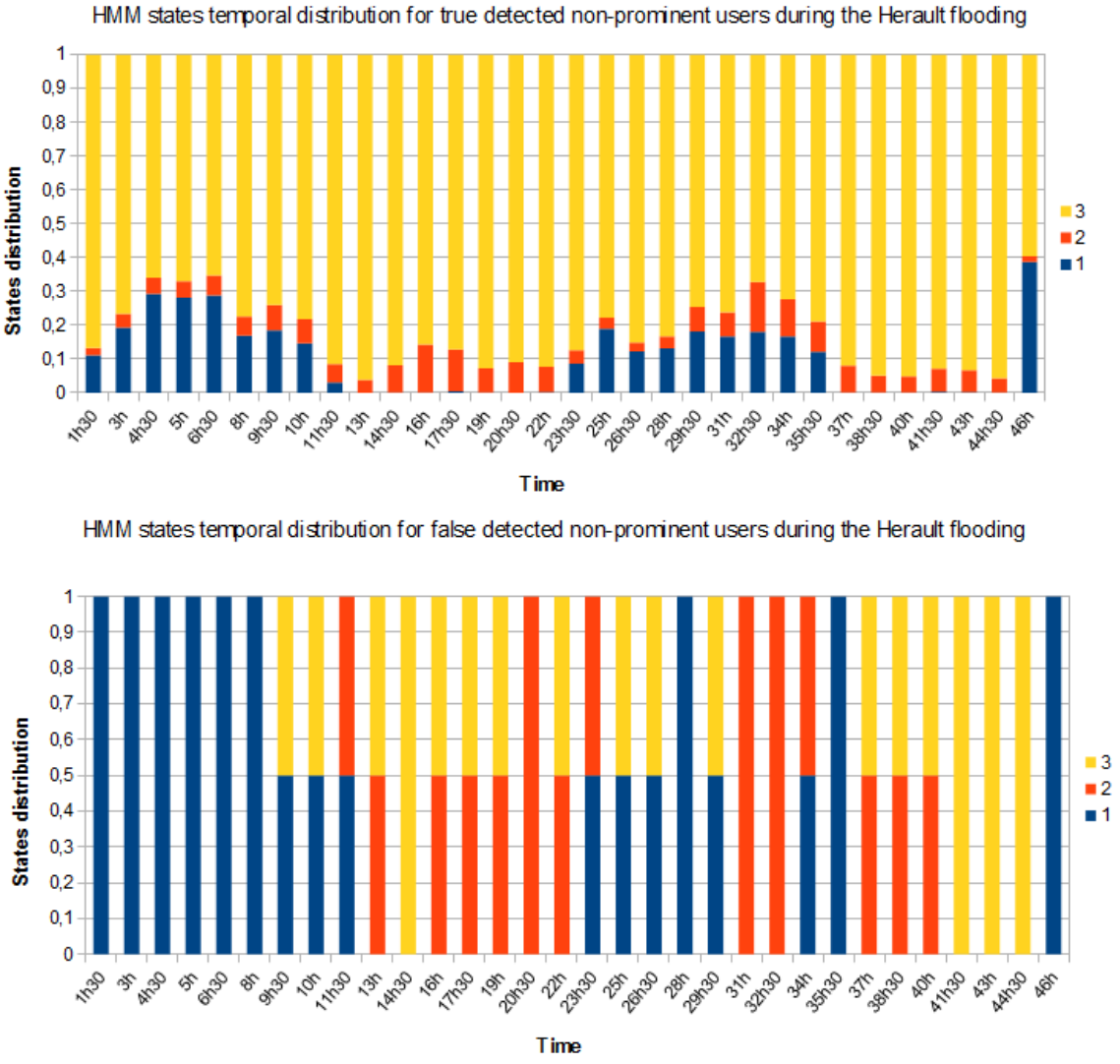

FIGURE 6.5: Temporal distribution of the **true and false detected non-prominent microblog users** by our learned ergodic HMM model. States 1,2 and 3 are the states set by the model to learn users behavior similarities and dissimilarities. The number of states was set experimentally as described in Section 6.7.1.

6.7.4 Importance of User Behavior States Learning

In order to evaluate the importance of learning users behavior states evolution over time for prominent users behavior detection, we analyze the distribution of user states –reflecting the user activeness– predicted by our model. These states are predicted using the Viterbi algorithm decoding user behavior during the Herault floods by a sequence of discrete states. Figures 6.4 and 6.5 report the behavior states distribution of the evaluated users classified as prominent and those classified as non-prominent respectively during the Herault floods.

Comparing the states distribution of true prominent and non-prominent users, we observe a

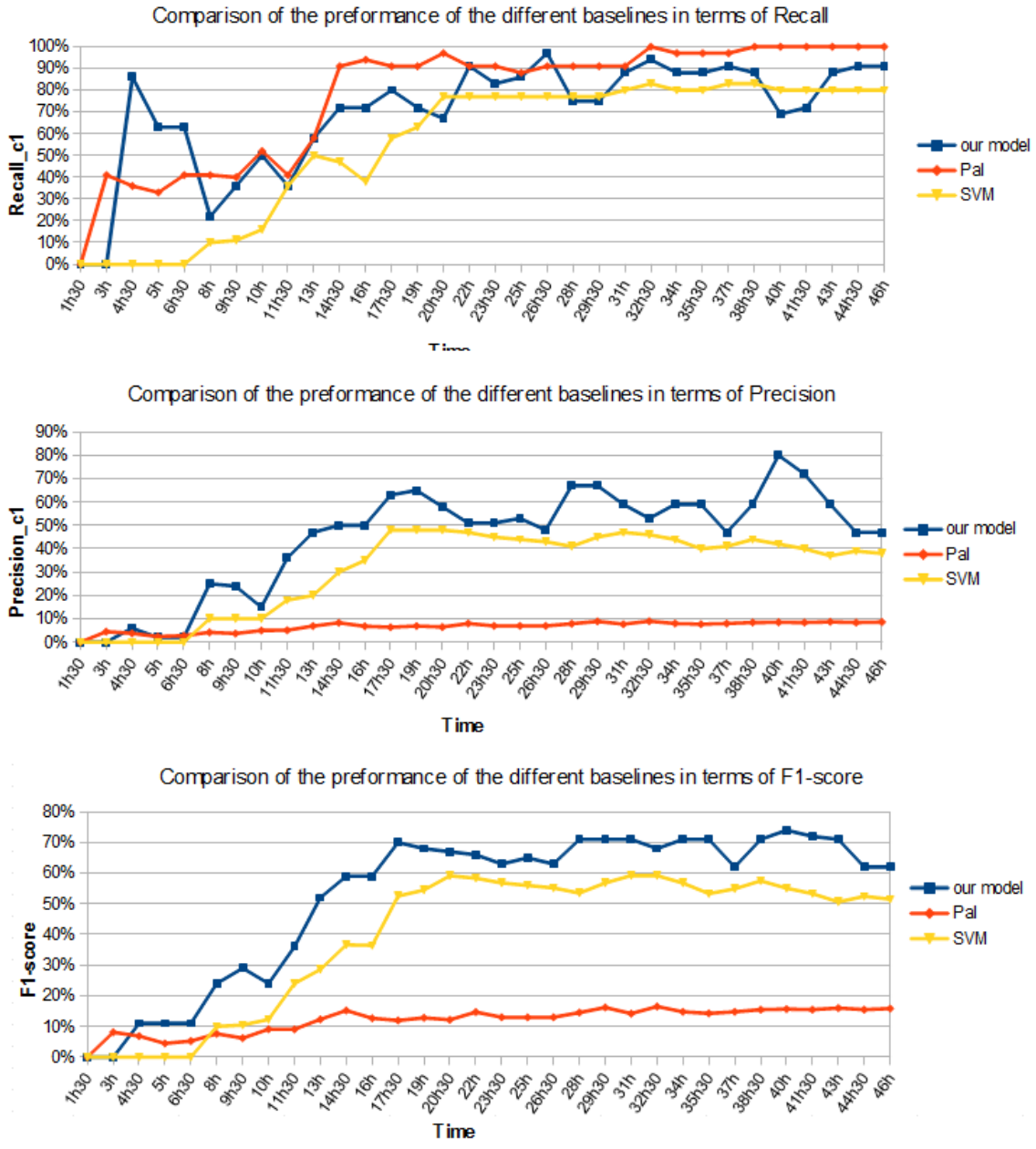

FIGURE 6.6: Comparing our time-sensitive model performance for prominent users prediction with different state-of-the-art baselines. The different prediction models are evaluated in terms of Recall, Precision and F1-score.

large difference between the behavior of the two categories of users. The detected prominent users are characterized by a high presence of user activeness states 1 and 2 especially in the first two third period of the event. However, referring to non-prominent users behavior, these same states 1 and 2 are dominated by the state 3. We also observe that the state 2 is distributed in a balanced way over time for the non-prominent users compared to the prominent ones. This explains the temporal behavioral representation importance to

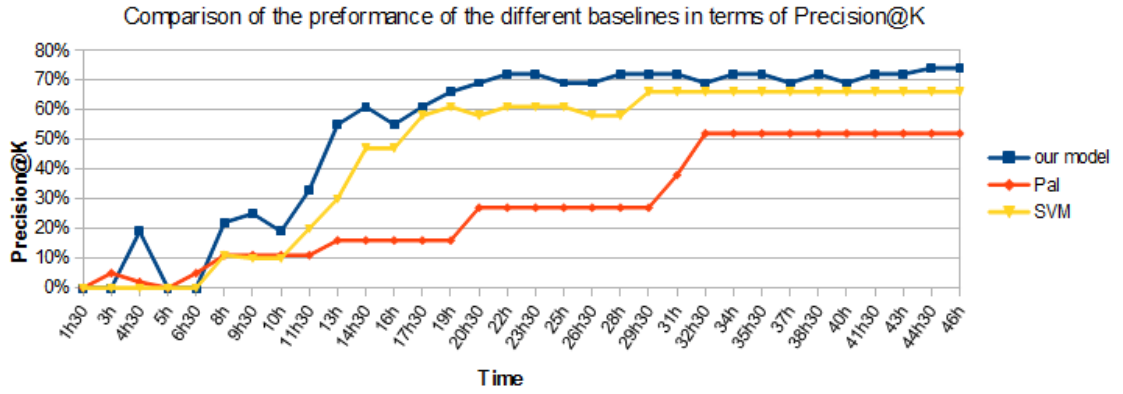

FIGURE 6.7: Comparing our time-sensitive model performance for prominent users ranking with different state-of-the-art baselines. The different ranking models are evaluated in terms of Precision@K.

differentiate between prominent and non-prominent users. Modeling such differences would enhance the precision of the prominent users identification model as proved through the experiments presented previously.

We also show that the states distribution of the false detected prominent users is almost similar to the true prominent users behavior. The misclassification of this small set of non-prominent users does not significantly affect the model performance as most of prominent users were detected. Only two prominent users behavior have been classified as non-prominent. It is possible that some non-prominent users have a similar behavior to the non-prominent ones. In these cases, our ranking model would assign a high rank the true prominent users regarding the false prominent ones by taking into account the different transitions of the different users from state to state.

6.8 Conclusion

This chapter has presented a novel microblog users modeling approach characterizing users according to their topical behavior over time. Based on this modeling approach, we learn a MoG-HMM model for prominent users prediction during crisis event. Users are characterized by a temporal sequence composed of feature vectors recorded in different periods of time during the events. These features characterize the on-topic and off-topic users activities at each time interval.

Through the conducted experiments, we have proven the importance of characterizing the temporal distribution of both the user on- and off-activities over time. We have also shown that our prediction model is processable in real time world cases. We have also found that larger sequence length characterizing detailed users activities over time works significantly better than smaller ones. Thus, user behavior is better characterized by considering various

time stamps. In addition, we have noted that our learned MoG-HMM model outperforms traditional machine learning SVM models learned based on time-insensitive user modeling approach in terms of both classification and ranking. This performance can be explained by the fact that MoG-HMM models detects better the particularities of prominent and non-prominent users behavior as they are more adapted for sequences of feature vectors learning.

While our time-sensitive model has identified most of prominent users and has outperformed state-of-the-art models, it still lacks adaptability to the particularities of crisis events cases. Few prominent users have been predicted at an early stage of the event. Emergency first responders need to access to most of the exclusive information from the beginning of the event. User behavior need to be characterized more efficiently by considering the evolution of the event over time. Such characterization is needed to highlight most of prominent users behavior while still maintaining the real time processing condition.

Chapter 7

Phase-Aware Microblog Prominent Users Modeling and Identification

Contents

7.1 Introduction

User activities are generally distributed differently during crisis events. Such distribution differences are not produced by chance. They are related to the event evolution over time. Crisis events are never static. Considering an analyzed event as a single fragment delimited by a start and an end point would hide the different co-relations between the event and the user behavior change. The evolution of crisis events is generally characterized by phases. Each event phase has its own characteristics that can differently impact users behavior.

By neglecting this event evolution impact on user behavior, microblog users would be uniformly characterized and evaluated from the beginning of the event until its end. However, realistically, the behavior of users differs according to the event evolution. Assume the case of a storm, the behavior of users during the first orange alert announcing a possible weather worsening would not be the same during the occurrence of the storm or during the recovery phase. Users act differently according to each event phase. Moreover, measuring users prominence according to their behavior during the whole event period of time independently of its phases would penalize prominent users who were active in only one – however important – event phase. Additionally, there is no need to track users who were prominent only during a specific event phase of the whole event. Users have to be evaluated according to their behavior at each single event phase.

In this chapter, we alleviate these shortcomings by proposing a phase-aware user modeling approach to highlight prominent users' behavior particularities over the different phases of crisis events. This approach takes into account new characterization aspects considering : (1) User on- and off-topical activity through the extraction and selection of the most relevant set of features reflecting the user behavior evolution over crisis events phases, (2) User activity distribution over time during the different event phases, (3) User prominence evolution over the different event phases and (4) User behavior representation according to each event phase context. Based on this phase-aware user representation, we propose a real-time prominent users prediction model identifying prominent users over time according to their represented behavior over event phases. This probabilistic phase-aware model is learned *a priori* using prior similar crisis events data.

The rest of this chapter describes the integration of these ideas in the context of prominent users identification during crisis events. In Section 7.3, we describe our phase-aware user behavior modeling approach. In Section 7.4, we detail the features extraction and selection step. Our temporal phase-aware probabilistic model for the classification and ranking of microblog user behavior is detailed in Section 7.5. The evaluation set-up is described in Section 7.6. Experimental evaluation is presented in Section 7.7. Finally, we present the discussion and conclusions along with directions for future work in Sections 7.8 and 7.9.

7.2 Research Questions

To the best of our knowledge, events specificties have never been considered while characterizing or identifying specific categories of microblog users. In this research, we take into account the possible correlations between the analyzed event particularities and the microblog users behavior evolution over time. We aim to learn both prominent and non-prominent user behavior evolution from phase to phase and select the most appropriate approach for users modeling per phase. This research helps us to answer several questions :

1. How to predict most of prominent microblog users at any time of the event? How to provide a real-time identification model processable in real world cases?
2. How can we model user behavior evolution over time by considering the analyzed event evolution over time? How to differentiate between users who are prominent only on a specific important phase and those who are non-prominent during it?
3. Which are the best features that have to be considered to characterize users behavior at each particular phase? Is there any behavioral change of users from phase to phase? Does the event evolution over time have an impact on users behavior?
4. How to insure a fair evaluation of the different users at each phase? Is it more rational to evaluate users' prominence by considering only their activities at each particular event phase independently of the other prior ones or not?

7.3 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 modeling approach that alleviates the stated shortcomings in Section 7.1. The different aspects considered in our microblog users behavior modeling approach are described in this section. We outline at first how crisis events characteristics are considered while representing the user behavior and prominence. We then detail how the user behavior change is reflected in our per-phase user modeling approach. Figure 7.1 summarizes the different modeling characteristics that have been considered in order to model each microblog user over time.

7.3.1 Crisis Events Evolution and their Impact on Microblogs Users' Behavior and Prominence

Like users have their own specificities, events and even topics have their own criteria that have to be considered while modeling user behavior. This subsection describes how our novel user behavior characterization approach takes into account the impact of the event evolution on both the user behavior and prominence over time.

7.3.1.1 Crisis Events Particularities

Crisis events are generally represented as a sequence E of d different successive "*phases*" describing the event evolution over time. The characteristics and level of importance of such events change at each particular phase.

$$E = (P_1, P_2, ..., P_d) \tag{7.1}$$

These phases are defined *a priori* by the domain experts according to the analyzed event context. In this thesis, we categorize crisis events phases into three main phases (Perez-Lugo 2004). Such categorization is widely adopted in disaster management systems in order to coordinate between the different organizations that have to intervene at each phase. The boundaries of each phase are defined in real time by referring to official organizations possessing expertise in this purpose (e.g. meteorological organizations in the case of flooding). In the following, we detail the specificities of these three different phases :

Phase 1 Preparedness (pre-warning) : is the phase announcing a possible risk that may arise on the next hours or minutes. During this phase, the risk has to be analyzed in order to predict its spatio-temporal evolution. According to the prediction reports, emergency first responders prepare the evacuation plans to be ready to deal with any possible menace.

Phase 2 Response (warning) : is the most delicate phase as it covers the period of the event occurrence. During this phase, emergency responders have the responsibility of detecting the affected areas, localizing people requiring an imminent intervention, reassuring and guiding people geo-located in the threatened and stricken area. Efficient actions taken at this phase can save an important number of lives and reduce damages.

Phase 3 Recovery : refers to the period of time following the crisis event occurrence. During this phase, official crisis events management organizations have to inventory the caused damages and make the required recovery actions to regain the usual level of functioning. The duration of such phase could be either short or long depending on the damages caused in the second phase. During such phase official governmental organizations have to

respond to personal and community needs. They also need to identify reconstruction and rehabilitation measures that have to be considered in the future.

In the following subsections, we detail further the specificities of these phases and their impact on microblog users behavior.

7.3.1.2 Event Phases Impact on Users' Prominence

As each crisis event phase has its particularities, we associate microblog users' prominence with each phase rather than with the whole event. Prominent users differ according to the analyzed phase. During the first phase where the risk is not yet confirmed, expert meteorologists are involved to analyze and communicate any news. Once the risk is confirmed and the red alert is raised, the response phase has to be managed. Emergency first responders such as police officers, fire-fighters, paramedics and emergency medical technicians intervene in order to address the immediate threats. When the situation becomes under control, emergency first responders retire in order to give way to experts who are charged to recover the disaster consequences.

Similarly, in microblogs, not all users are interested in a crisis event from its beginning to its end. For example, prominent users in the first phase will not necessarily remain prominent in the second or the third one. Through our microblog users characterization approach, we model each microblog user by a sequence of d representations reflecting his/her behavior at each phase.

$$R(u) = (R_{P_1}, R_{P_2}, ..., R_{P_d}) \tag{7.2}$$

By characterizing user behavior separately at each phase, we can evaluate microblog users prominence at each phase independently of the other prior ones. Users detected as prominent in a particular phase would be tracked only during it unless they prove their prominence in the next ones. In this way, we avoid to track users who were prominent just in a particular phase during the whole event. We would also insure a fair evaluation among microblog users at each phase. The high or low feature values characterizing users behavior in a particular phase will be neglected in the next phases. Each user features values are reset to zero at the beginning of each new phase. Only their shared activities in the current analyzed phase are considered.

Therefore, in order to insure a fair prominence evaluation for all users at each phase, we classify over time each microblog user who has shared at least one-event-related-information in d classes according to his/her prominence at each phase. We define two classes for each event phase j: $C_1^{P_j}$ and $C_2^{P_j}$ refer respectively to prominent and non-prominent microblog users during the phase j. Each user characterized by a sequence of vectors R_{P_j} is classified

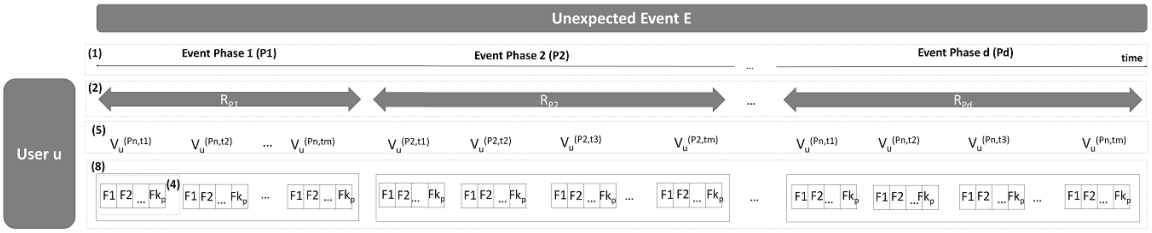

FIGURE 7.1: User behavior representation during an event E. E is divided into d phases. At each phase, the user is represented by a sequence of vectors R_{Pj}. These vectors are composed of a list of feature reflecting the user behavior specificities at each phase. Such user representation is modeled over time according to each crisis event phase particularities. Boxes (1), (2), (4) and (8) refer to the equations presented in Section 3.

in one of the two defined classes according to his/her behavior during that phase j. The classification and prediction model appropriated for each phase will be described in Section 7.5.

Our user characterization is based on the principle that user prominence is associated to his/her activeness at each event phase and not the whole event. As illustrated in Figure 7.1, each user is represented according to his/her activity at each phase independently of the other ones $\{R_{P_1}\}$,$\{R_{P_2}\}$,...,$\{R_{P_d}\}$. Each characterized user has to be classified in one of these corresponding classes $\left\{C_1^{P_1}, C_2^{P_1}\right\}$,$\left\{C_1^{P_2}, C_2^{P_2}\right\}$,..,$\left\{C_1^{P_d}, C_2^{P_d}\right\}$.

7.3.1.3 Event Phases Impact on Users' Behavior

In the context of crisis events, user behavior differs in the first and third phase from the second phase. Prominent microblog users in the second phase are generally in panic and would mainly concentrate on expressing what they are seeing and experiencing regarding the event. However, they will act somehow like ordinary days during the other phases.

To cover these users' behaviors changes according to each event phase, we model each user differently at each phase by using different features. We select the best k representative features reflecting users' behavior at each phase j.

$$F^{P_j} = (F_1^{P_j}, F_2^{P_j}, ..., F_k^{P_j}) \tag{7.3}$$

The sequence of feature vectors characterizing each user R_{P_j} is only composed by the selected features F^{P_j} reflecting users behavior during that phase P_j. These features are selected from a large set of raw and engineered features characterizing user activity in microblogs using a multi-variate feature selection algorithm Corona (Yang et al. 2005*b*) (See Section 7.4). Using this strategy, we represent users behavior differently according to the analyzed phase by using appropriate features selected *a priori*. This selection is conducted by learning the behavior of users during the different phases of similar events.

The experimented and appropriate features selected for each phase will be described in depth in the next Section.

7.3.2 User Behavior Modeling as Temporal Phase-aware Sequences

As proved in the previous chapter, considering the temporal distribution of the user's shared activities highlights new behavioral patterns differences easing the prominent users identification process. This temporal distribution is also considered in our phase-aware user representation. In the previous proposed user modeling approach, we represented each user as a single sequence of vectors characterizing his/her behavior during the whole duration of the event. Through our new proposed phase-aware user modeling approach, each user is represented by d-sequences of vectors characterizing his/her topical activities distribution at each single phase. These feature vectors are computed based on the selected features reflecting the user behavior at that specific phase. The time-line of each event phase is divided into equispaced intervals at m time-stamps t_1,t_2,t_3,...,t_m from the beginning of the phase P_j until its end. At each time-stamp t_i, we represent each microblog user u by a feature vector $V_{P_j}^{t_i}$ characterizing his/her behavior from the time-stamp t_{i-1} to t_i.

$$V_{P_j}^{t_i}(u) = (F_1^{P_j,(t_i)}, F_2^{P_j,(t_i)}, ..., F_k^{P_j,(t_i)}) \tag{7.4}$$

Then, the resulted vector is added in the sequel of the temporal sequence $R_{P_j}^{t_{i-1}}$composed of the prior calculated vectors from the beginning of that phase.

$$R_{P_j}^{t_i}(u) = (V_{P_j}^{t_1}(u), V_{P_j}^{t_2}(u), ..., V_{P_j}^{t_{i-1}}(u), V_{P_j}^{t_i}(u)) \tag{7.5}$$

Segmenting the sequence of user activity at each phase into time-series feature vectors offers a rich and personalized user representation. This user modeling approach would offer a complete vision of users behavior by taking into account the evolution of both users and events over time. This eases the identification of users behavior regularities, similarities and dissimilarities at each phase.

7.4 Extraction and Selection of Microblogs Users' Features

In order to efficiently model the user behavior particularities at each phase, we evaluate the effectiveness of a large set X of the raw and engineered features described respectively in Chapters 4 and 5. Through this evaluation, we need to select the features subset X_s that best reflects the real user behavior at each event phase. Through this selection process, microblog users behavior would be effectively reflected and the computational cost of features would be reduced. As explained in box A in Figure 7.3, both the X features extraction

and X_s features selection steps are processed *off-line*. Using prior crisis event datasets, the best representative features of users behavior have to be selected according to each defined phase. The resulted selected features are then used for user behavior representation in previous and real-time crisis events. In the following, we briefly describe these *off-line* features extraction and selection processes.

At each phase, we extract and compute the defined set of raw and engineered features for each microblog user u and each time-stamp t_m during a particular event phase P_j. Once both raw and engineered features are computed at each time-stamp during each event phase, we represent each user u by an initial feature vector $\tilde{V}_{P_j}$ characterizing his/her activity at each time-stamp t_i during each phase P_j. Each feature vector is composed of the complete features set X (i.e. 30 raw features and 14 engineered features composed of adjusted and non-adjusted on-topical features).

$$\tilde{V}_{P_j}^{t_i}(u) = (T1_{on}^{P_j,(t_i)}, T1_{off}^{P_j,(t_i)}, ..., EF7^{P_j,(t_i)}) \tag{7.6}$$

By assembling the different feature vectors computed at each time-stamp t_i during P_j, we model each user with an initial temporal sequence of vectors $\tilde{R}_{P_j}$ describing his/her behavior at that phase.

$$\tilde{R}_{P_j}^{t_i}(u) = (\tilde{V}_{P_j}^{t_1}(u), \tilde{V}_{P_j}^{t_2}(u), \tilde{V}_{P_j}^{t_3}(u), ..., \tilde{V}_{P_j}^{t_i}(u)) \tag{7.7}$$

Once each user has been characterized, we select the best representative features set $X_s^{P_j}$ for each phase P_j. Through this process, we can reduce the dimensionality of each feature vector $\tilde{V}_{P_j}^{t_i}(u)$ and obtain an optimal user characterization $R_{P_j}(u) = \tilde{R}_{P_j}^{*}(u)$ at each phase by eliminating redundant and irrelevant features.

We use the Corona (Yang et al. 2005*b*) supervised feature subset selection technique appropriated for the Temporal Sequence of Feature Vectors (TSFV) user representation as a feature selection algorithm. Corona maintains the correlation between the different feature vectors $\tilde{V}_{P_j}^{t_1}(u)$ computed at different time-stamps t_i corresponding to the same event phase P_j. Each TSFV represented by $\tilde{R}_{P_j}(u)$ is treated as a whole. Using this algorithm, we select *off-line* the top relevant features at each event phase.

We first compute the correlation coefficient matrix of each TSFV using Equation 7.8. This correlation matrix represents the relationship between each two feature vectors included in the TSFV at each phase according to the used training data. Assume that a and b refer respectively to the feature vector $\tilde{V}_{P_j}^{t_i}(u)$ characterizing the user behavior at time-stamp t_i and the feature vector $\tilde{V}_{P_j}^{t_{(i+1)}}(u)$ of the same user at time-stamp t_{i+1}. The dimension of

those vectors is 44. This number corresponds to the initial number of features.

$$corr(a,b) = \frac{\sum_{k=1}^{44}(a_k - (\bar{a}))(b_k - (\bar{b}))}{(43)\sigma_a\sigma_b} \tag{7.8}$$

Where $(\bar{a})$ and $(\bar{b})$ are respectively the averages of the feature vectors computed at time-stamp t_i and time-stamp t_{i+1}; σ_a and σ_b are the standard deviations of a and b.

Each resulted correlation coefficient matrix is then vectorized. Using these vectors, we subsequently train a SVM model to obtain the weights relative to each feature included in the training stage. We then aggregate the resulted weights in order to have one weight value relative to each feature. Based on these aggregated values, we select the worst feature using a greedy approach consisting of identifying the feature whose maximun weight is the minimum compared to all the other features weight. Subsequently, we remove the selected worst feature.

This whole process is then repeated until the k best features that reflect users behavior at each phase P_j are obtained. The selected features are then used to represent each microblog user at that phase.

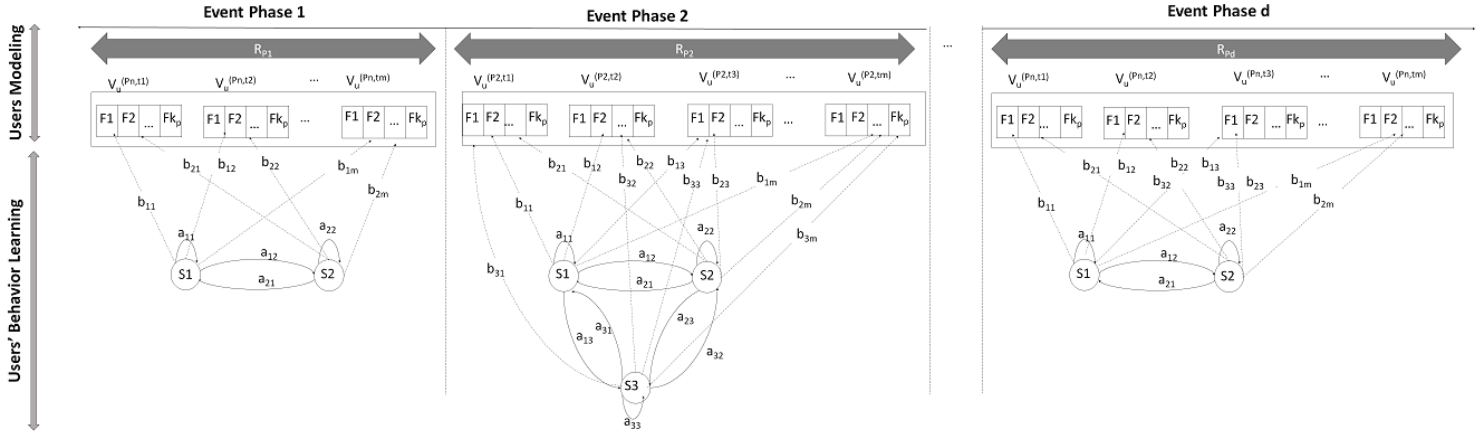

FIGURE 7.2: The different ergodic MoG-HMM models trained for prominent users detection at each event phase. A MoG-HMM is learned for each timestamp relative to each event phase. These models are constructed by learning the different users behavior over time at each phase. Each model relative to each phase is learned separately from the other ones in order to be able to distinguish between prominent and non-prominent users behavior at that phase.

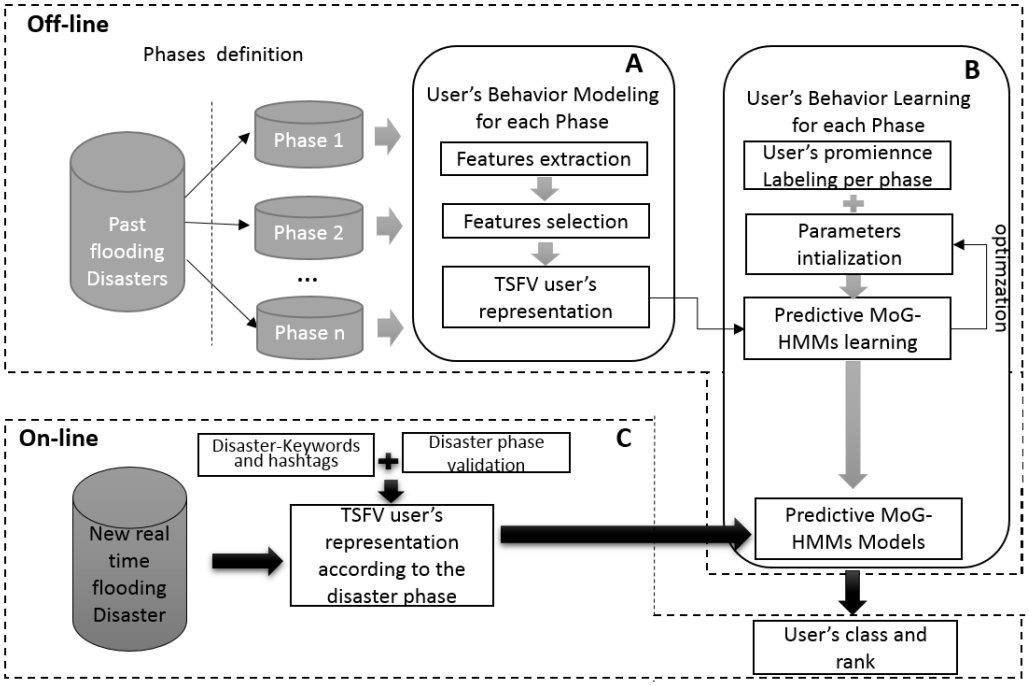

FIGURE 7.3: The prediction model process during crisis events. This model is learned **off-line** by referring to prior crisis event events having the same nature. Once the model is learned, the prediction process of prominent users can be executed **on-line** in real time during crisis events having the same nature. Such model receives as input a TSFV representing the user behavior according to each phase and gives as output the user class. Microblogs users classed as prominent are then retained and ranked.

7.5 Phase-aware MoG-HMMs for Users' Prominence Prediction and Ranking

In this Section, we describe our phase-aware probabilistic model for prominent microblog users prediction in real time during crisis events. Figure 7.3 describes how this model is learned *off-line* and how it works *on-line* during real-time crisis events. The phase-aware model is built by learning the different behaviors of prominent and non-prominent microblog users during each phase belonging to prior crisis events having the same nature. Once the model has learned to differentiate between prominent and non-promient users behavior over time, it can be applied in real time during similar crisis events.

During real time crisis events, each microblog user behavior is represented by the TSFV user representation corresponding to each phase. This representation is processed once the keywords and hashtags relative to the crisis event have been defined and the current crisis event phase has been identified by domain experts. User features are then automatically extracted according to the analyzed crisis event phase. In the following, we further detail

the learning step described in the *Box B* of Figure 7.3 and the real-time prediction model process represented in *Box C*.

7.5.1 Learning the Phase-aware Prominent Users Identification Model

In order to be able to evaluate the prominence of each new microblog user interacting during the analyzed crisis event phase, we aim to learn *a priori* phase-aware models for each crisis event category and test the resulted model. These learned models have to classify over time each microblog user behavior characterized by the TSFV $R_{P_j}^{t_i}$ in either class $C_1^{P_j}$ or $C_2^{P_j}$ referring respectively to whether the user is prominent or not in phase P_j. Learning such binary classification models is critical in crisis events context, where training data from the positive class $C_1^{P_j}$ are inherently rare, and are costly to analyze. In fact, although there is a huge amount of crisis event-related-information shared in microblogs during the different crisis events phases, the number of real prominent users who provide valuable information is small. Thus, collecting samples describing prominent microblog users' behavior during crisis events for the model learning remains difficult as described in Chapter 5.

Taking into account the stated training data limits, we address the phase-aware prominent user behavior identification problem using the generative classification MoG-HMM. Indeed, both theoretical and empirical studies pointed out that while discriminative models achieve lower asymptotic classification error, generative methods tend to be superior when training data are limited (Deng & Jaitly 2015).

Such generative classification MoG-HMM model was already adopted in the previous chapter in order to learn topical users behavior during the whole event. In the context of crisis events phases considered in this chapter, HMM models are separately learned according to the user behavior at each specific event phase. Thus, we train separate ergodic MoG-HMM models, $H_{P_j}^{C_1}$ and $H_{P_j}^{C_2}$, for each class at each time-stamp during each event phase as described in Figure 7.2.

7.5.2 Real-time Users' Prominence Prediction and Ranking

Once the HMM-MoG models $H_{P_j}^{C_1}$ and $H_{P_j}^{C_2}$ are learned using the training dataset, each microblog user can be classified into one of the analyzed event phase classes by computing the following probabilities $P(R_{Pj}^{t_i}(u)|H_{P_j}^{C_1})$ and $P(R_{Pj}^{t_i}(u)|H_{P_j}^{C_2})$. The TSFV corresponding to each user is automatically extracted by referring to the keywords and hashtags describing the analyzed crisis event phase.

Once the user behavior is modeled from the beginning of the analyzed phase until the process time t_i, the probabilities can be computed given the two learned models using

the forward-backward algorithm (Baum & Eagon 1967). If the returned probability by the model $H_{P_j}^{C_1}$ is greater than $P(R_{Pj}^{t_i}(u)|H_{P_j}^{C_2})$, then the evaluated user is classified as prominent and has to be tracked until the end of the phase P_j.

In order to rank the selected prominent users, we sort the likelihood $P(R_{Pj}^{t_i}(u)|H_{P_j}^{C_2})$ of the different microblog users sequences regarding the model $H_{P_j}^{C_2}$. The smaller this probability is, the bigger the prominence of that user is. Our rationale behind ranking users by referring to their likelihood regarding $MoG - HMM_{C_2}$ rather than $MoG - HMM_{C_1}$ consists of targeting the model which tends to be the most precise. $MoG - HMM_{C_2}$ is generally learned using a large number of samples covering most of the non-prominent users behaviors. We thus refer to its resulted likelihood for prominent users ranking as it tends to be more precise than the one resulted by $MoG - HMM_{C_1}$ trained using limited data.

7.6 Experiments and Evaluation

7.6.1 Datasets Description

For experimental and evaluation purpose, we use the collected data belonging to the two disaster datasets relative to the two different flooding events : *"Alpes-Maritimes floods"* and *"Herault floods"*. The extraction process of these two datasets was described in Chapter 3. As these two events fall in the same category of natural disasters, we use the first dataset for our model training and the second one to test the learned model performance for the identification of prominent users in real-time in similar flooding cases. We conduct also some experiments to test the efficiency of our prediction model while using the first dataset for testing and the second one for training.

During the training and testing of our models, we consider each disaster as a sequence of three phases : $P1$, $P2$ and $P3$ referring to the standard disaster phases *Preparedness*, *Response* and *Recovery* phases respectively. The different phases boundaries were set by referring to the official meteorological organizations of the regions threatened and affected by the disaster. Such organizations use to precise the level of alert and the disaster evolution state during natural disasters. Table 5.6 shows statistics of the collected tweets at each phase relative to each dataset. The first dataset *"Alpes-MaritimesDB"* is used to build our user behavior characterization and prediction model. This dataset refers to the floods that have occurred in the Alpes-Maritimes area between the 3^{rd} and 7^{th} October 2015. $152,402$ tweets shared by $21,364$ users were collected during this event. The different disaster phases $P1$, $P2$ and $P3$ have lasted respectively 3.5, 18.5 and 72 hours according to the information provided by the meteorological vigilance center of Provence-Alpes-Cote d'Azur.

The second dataset *"HeraultDB"* is used in order to test the model learned using the first dataset. *"HeraultDB"* refers to the floods that have occurred from 29^{th} to 30^{th} September 2014 in the Herault area. This dataset consists of $44,330$ on- and off-topic tweets shared by $3,338$ users during the whole event. The duration of each phase $P1$, $P2$ and $P3$ was set respectively to 15, 17, 15 hours according to the phases boundaries reported by the French inter-regional meteorological center of Aix-en-Provence.

7.6.2 Datasets Labeling

To create the ground-truth of our two collected datasets, we conducted a subjective user study for manually labeling each user at each phase Pj as $C1^{Pj}$ "prominent" or $C2^{Pj}$ "non-prominent". Three participants were selected for this purpose. These participants have known the two flooding disasters' areas and followed the different news and evolution of these two disasters in both online social media and news outlet channels. They were also required to be familiar in the concept of tweets and fluent with the languages used by microblog users interested by the analyzed disasters.

Two of these participants were separately asked to label manually all users according to the relevance and exclusiveness of their shared disaster tweets at each phase. To check the exclusivity of user tweets, these participants possessed a report listing in a chronological order most of the important disaster news with their time of first announcement. These news information were extracted from *20minutes*[1] news website. Once, all users included in the two different datasets were labeled separately by the two first participants, the third one is asked to break the labels' disagreement between the two participants by deciding which label has to be retained. The final study results of the two datasets labeling are described in Tables 7.1 and 7.2.

A second study is conducted for ranking the already validated prominent users. We have asked the same participants to attribute a score on a scale from 4 to 10 to each user labeled as prominent. Each score has to reflect the relevance and freshness of each prominent user tweets during each phase revealing his/her prominence. The average of scores set for the different prominent users is then calculated. We sort prominent users relative to each phase according to their prominence score.

7.6.3 Evaluation Set-up

Following the off-line steps described in Figure 7.3, we start by selecting the appropriate features for user modeling at each phase. We use the *"Alpes-MaritimesDB"* dataset for the

[1] http://www.20minutes.fr/nice/1701427-20151004-direct-intemperies-alpes-maritimes-bilan-alourdit-12-morts

TABLE 7.1: Results of the subjective user study for the two datasets ground-truth building for each phase.

Event Phases		#Prominent users	#Non-prominent users
AlpesMarDB	P1	20	21,344
	P2	99	21,265
	P3	157	21,207
HeraultDB	P1	35	3,303
	P2	87	3,215
	P3	67	3,271

TABLE 7.2: Common ($\cap$) and distinct ($\cup$) prominent users in the different phases of each dataset.

Prominent users sets	AlpesMarDB	HeraultDB
$\left\{C_1^{P_1} \cap C_1^{P_2}\right\}$	12	21
$\left\{C_1^{P_2} \cap C_1^{P_3}\right\}$	31	20
$\left\{C_1^{P_1} \cup C_1^{P_2} \cup C_1^{P_3}\right\}$	233	148

different model learning steps. We represent at first each microblog user in this dataset by a temporal sequence of vectors composed of the different features described in Section 4. These features are extracted, at each 30 minutes from the beginning of each phase until its end for user behavior representation. Once these users are represented, we process the Corona algorithm separately for each phase. The resulted top features selected by Corona for each phase are then considered for user representation during the model learning and testing steps.

To learn the different models $H(c1)_{P_j}^{t(i)}$ and $H(c2)_{P_j}^{t(i)}$ for predicting user prominence over time at each phase, we use the *"Alpes-MaritimesDB"* dataset considered as our training dataset. We learn a new $H(c1)_{P_j}^{t(i)}$ and $H(c1)_{P_j}^{t(i)}$ after each 30 minutes starting from one hour of the beginning of each phase. For example, in the 1^{st} phase, we learn 14 H_{c1} models corresponding to each time-stamp t_i. The training data input is composed of a temporal sequence of feature vectors characterizing the user behavior. Features composing each vector are computed sequentially at each interval of 30 minutes from the beginning of each phase. For example, after 3 hours from the beginning of the 1^{st} phase, each user would be represented by a sequence of 6 vectors.

As the *"Alpes-MaritimesDB"* 1^{st} phase is short, we extended it in our experiments by repeating the same characterized user behavior recorded with real data to cover the same duration of the 1^{st} phase relative to the test dataset.

In order to choose the optimal parameters for $H(c1)_{P_j}$ and $H(c2)_{P_j}$ models relative to each phase, the models performance is evaluated with tuning the values relative to the number of states N_S (1 – 4) and the number of multivariate Gaussian N_G (1 – 4) using the training dataset. For each phase model, we select the parameters giving the best $Precision@K$

(this measure is described on the following) at the end of each phase. Table 7.3 reports the $Precision@K$ registered at the end of phase P_2 of the Alpes-Maritimes floods while training the models $H_{c1}^{P_2}$ and $H_{c2}^{P_2}$ using different parameters. In the second phase of Alpes-Maritmes floods, $K = 99$. As shown in this table, the best $Precision@99$ result recorded at the end of the second phase was performed using these parameters $N_S = 2$ states and $N_G = 1$. Thus, the different temporal models $MoG - HMM_{C1}$ and $MoG - HMM_{C2}$ learned with these parameters during the phase P2 will be retained as our final models. These resulted models are then applied during new flooding events.

TABLE 7.3: Prominent users identification performance for different N_S and N_G in terms of $Precision@99$ during the second phase using the training dataset.

N_S/N_G	1	2	3	4
1	0.88	0.92	0.26	0.26
2	**0.94**	0.27	0.1	0.1
3	0.27	0.1	0.1	0.1
4	0.27	0.1	0.1	0.1

Once the different models are learned, we test their performance for prominent microblog users detection at each phase using a new dataset relative to a new flooding event. To conduct our experiments, we use the *HeraultDB* dataset for testing. We model each microblog user in this dataset over time from the beginning of the disaster until its end using our proposed phase-aware temporal user characterization approach. Features selected in the training phase are extracted at each 30 minutes for user representation according to each analyzed phase. Phases boundaries are considered while modeling each microblog user. Each microblog user behavior represented by a TSFV is evaluated over the floods time-line using our learned models for prominent users prediction. The ground-truth labels relative to the *HeraultDB* dataset are used to check the relevance of the prediction results of our learned models.

To evaluate the performance of our learned models, we use standard evaluation metrics : recall, precision and ranking measures such as Recall@10 and Precision@K.

$$\text{CommonPromP1P2} = \frac{\#\text{detected Common Prominent Users in P1 and P2}}{\#\text{True Common Prominent Users in P1 and P2}}$$

$$\text{CommonPromP2P3} = \frac{\#\text{detected Common Prominent Users in P2 and P3}}{\#\text{True Common Prominent Users in P2 and P3}}$$

$$\text{Recall@10} = \frac{\#\text{detected top10 ground-truth prominent users}}{10}$$

$$\text{Precision@K} = \frac{\#\text{detected true prominent users in top K}}{\text{K}}$$

where : K =number of ground-truth prominent users (i.e. $K = 35$ in P1, $K = 87$ in P2 and $K = 67$ in P3 by referring to the test dataset *HeraultDB*)

7.7 Experimental Results

To experimentally validate our real-time prominent microblog users identification model during crisis events, we compare its performance with several baselines. These baselines were especially implemented to evaluate our proposed model. We describe below these different baselines :

Ours : This refers to our proposed model which represents each user by a sequence of feature vectors characterizing the user behavior from the beginning of each analyzed phase independently of the other ones as described in Sections 7.3 and 7.4. It uses an additional Boolean feature Bf indicating the user prominence in the previous phase.

***Pal* :** This refers to the system built by Pal & Counts (2011). This system represents microblog users uniformly during the whole event. Microblogs users are represented by a single feature vector composed of 15 features. *Pal* classifies and ranks users according to their behavior from the beginning of the event without considering event phases. Through this state-of-the-art model, we aim to evaluate the performance of our phase-aware model considering only the Bf feature as the only indication of the user activity during previous event phases.

***Pal** :** This baseline uses the same specificities of *Pal* model presented above. However, this model considers the different event phases while representing user activities. The user temporal representation is not considered in this model. Through this baseline, we aim to prove that our phase-aware modeling approach can improve the prediction results of *Pal*.

***b1* :** This baseline uses the same specificities of our model, but, it does not consider the Boolean feature Bf. Through this baseline, we want to evaluate the contribution of the Boolean feature Bf on enhancing the prediction results over time.

***b2* :** This baseline follows the same user representation and classification principles used in our model. However, it is learned at each phase by referring to all the prominent microblog users $\left\{C_1^{P_1} \cup C_1^{P_2} \cup C_1^{P_3}\right\}$ independently of their phase of prominence. Through this baseline, we aim to validate our assumption considering that user prominence has to be associated to each phase rather than the whole event.

***b3* :** This model has the same specificities as our model. However, it characterizes users

uniformly during the whole event. It uses Corona to select the relevant features that better reflect users' behavior during the whole event and not during each particular phase. Through this baseline, we evaluate the efficiency of our per-phase user modeling approach.

7.7.1 Efficacy of the Real-time Prominence Prediction Model

Through the conducted experiments in this subsection, we evaluate the real-time prediction efficiency of our phase-aware model. More precisely, we evaluate the impact of considering the Bf feature, as the only indication of the user activity in the previous phase, on the model efficiency. We also evaluate the importance of considering the event phases for prominent users prediction by comparing the two baselines Pal and Pal*.

We test the performance of the different learned ergodic MoG-HMM models relative to each phase during the Herault floods. Our prediction results are compared with those obtained by the state-of-the-art clustering and ranking system *Pal* and the baselines *Pal** and *b1*. While *Pal* has prior knowledge about all the users activities from the beginning of the event, it is not the case for *Pal** and *b1*. These two models do not have any prior knowledge about the user prominence or activity in the previous phases. Figure 7.4 shows the prediction results obtained by these models at each time-stamp relative to each phase in terms of $Recall_{C1}^{Pj}$ and $Precison@K$. Additionally, Table 7.4 reports more detailed results of this comparison study in terms of $Recall@10$, $Recall_{C2}^{Pj}$, $CommonPromP1P2$ and $CommonPromP2P3$ at the beginning, one-third, half and the end of each phase. $CommonPromP1P2$ and $CommonPromP2P3$ refer to the detection rate of the common prominent users at $\left\{C_1^{P_1} \cap C_1^{P_2}\right\}$ and $\left\{C_1^{P_2} \cap C_1^{P_3}\right\}$ respectively.

According to the classification results reported by the $Recall_{C1}^{Pj}$ and $Recall_{C2}^{Pj}$ measures, our model detects most of the prominent users at an early stage of each phase and eliminates a large number of the non-prominent ones. We also note that the model $Precision_{C1}^{Pj}$ is low for class $C1$. However, in our case, this measure is not really important as the number of prominent and non-prominent users are extremely unbalanced. Thus, even if the precision results registered by our model are low and do not exceed the 16%, our model is still promising as it detects most of the prominent users and rejects a large set of non-prominent ones. In realistic information retrieval cases, such as search engines, the returned results relative to a specific query do not contain only relevant answers, it can return both relevant and irrelevant ones. However, what matters the most is the rank of these returned relevant results. Similarly in our case, if our model returns most of the prominent users and ranks them in the top lists then our identification model would be efficient. In order to incorporate this scenario, we evaluate the ranking of the detected prominent users in the top K list where K refers to the number of ground-truth prominent users at each phase

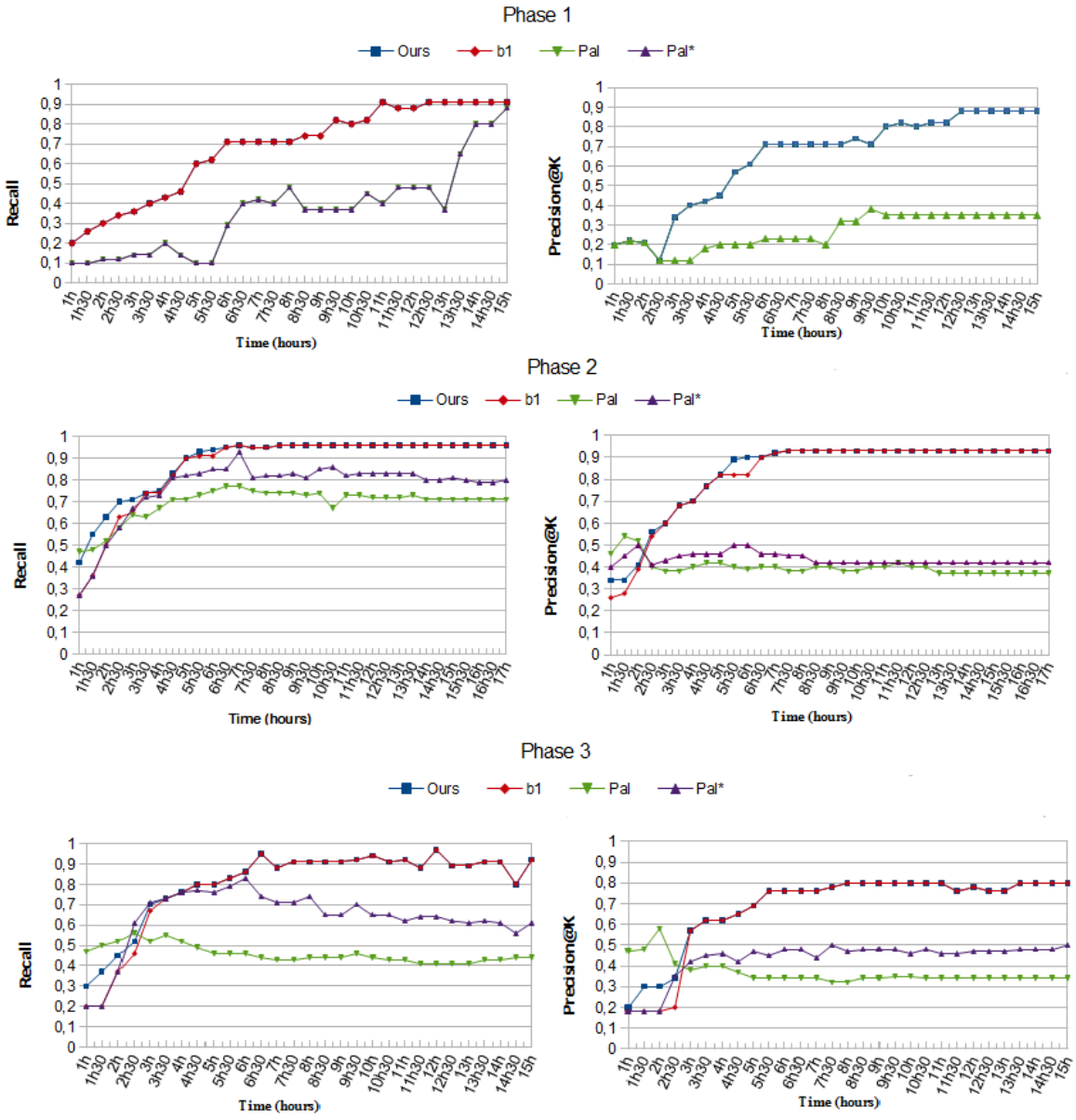

FIGURE 7.4: Comparing the prediction results of **ours** model with **Pal***, **Pal** and **b1** baselines in terms of $Recall_{C1_{Pj}}$ and $Precision@K$ during each phase. *Alpes-MaritimesDB* is used for training and *HeraultDB* for testing. At the first phase, *b1* and *Ours* are identical. Similarly, **Pal*** and **Pal** are similar in P1 as there are no prior phases. The different results were registered while testing the model at different timestamps during the Herault floods.

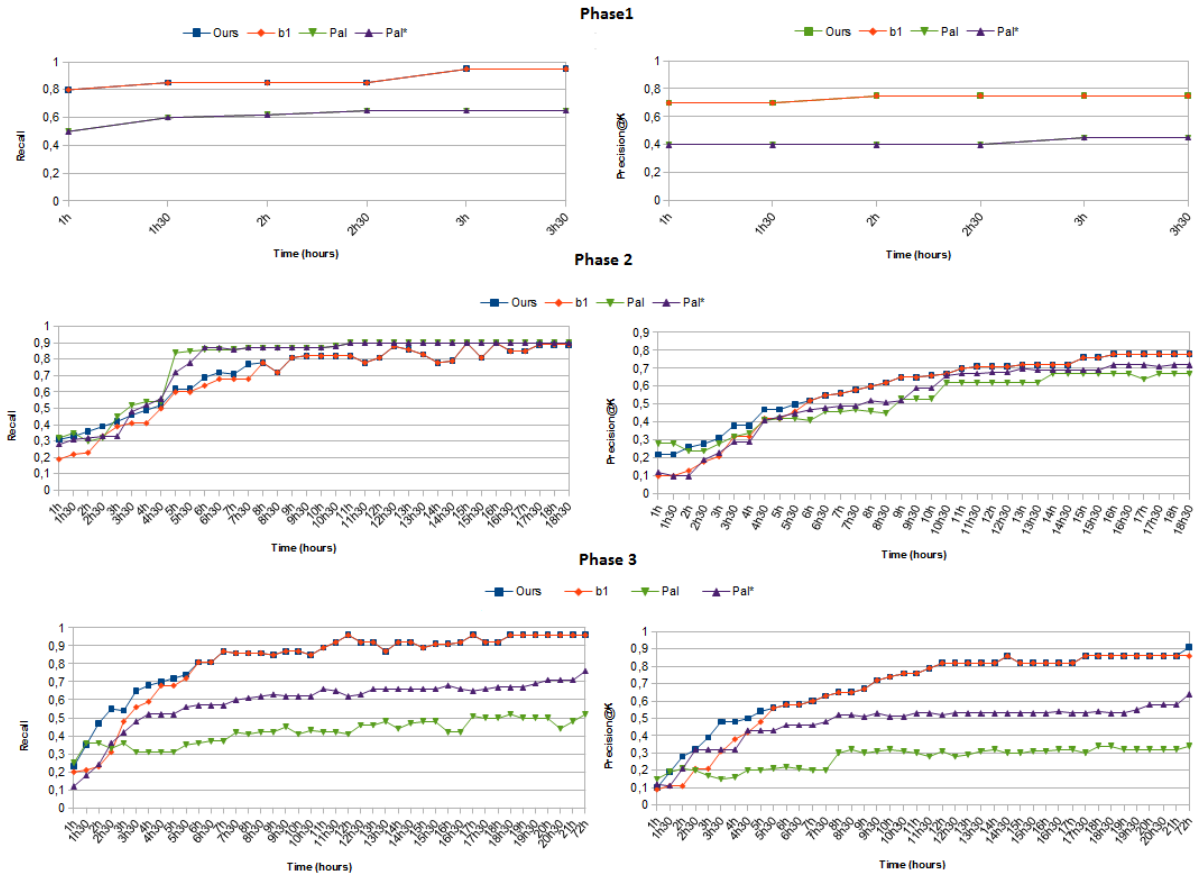

FIGURE 7.5: Prediction results comparison of **ours**, **Pal***, **Pal** and **b1** baselines in terms of $Recall_{C1_{Pj}}$ and $Precision@K$ during each phase. *HeraultDB* is used for training and *AlpesMaritimesDB* for testing. At the first phase, *b1* and *Ours* are identical. Similarly, **Pal*** and **Pal** are similar in P1 as there are no prior phases. The different results were registered while testing the model at different timestamps during the Herault floods.

(i.e. $K = 35$ in P1, $K = 87$ in P2 and $K = 67$ in P3). According to the obtained results, most of the prominent users were detected and top ranked at an early stage of the disaster as indicated by the $Precision@K$ curves in Figure 7.4.

We also note that our model detects all the top 10 prominent users (i.e. 100% Recall@10) after a few hours of each phase. Comparing these results with *Pal*, *Pal** and *b1* baselines, our model performs the best in terms of classification and ranking. Using the Bf feature, we succeed to identify more prominent users at the beginning of the event compared to the *b1* model. This feature helps to identify common prominent users between the current and the previous phase as indicated by the reported $CommonPromP1P2$ and $CommonPromP2P3$ results. We also observe that *Pal* slightly outperforms our model and *Pal** at the beginning of $P2$ and $P3$ as it detects most of the prominent users that were already detected in the previous phases by considering their tweeting activity from the beginning of the event. However, the performance of the baseline *Pal* erodes further with time as it is not able to

detect the new prominent users relative to the current phase. Moreover, we note that *Pal** which does not consider any information about user activity in prior phases outperformed *Pal* results after few hours. This validates our assumption. Using the phase-unaware model *Pal*, the new prominent users will not be favored with respect to the prior ones.

In order to prove the efficacy of our model for prominent users prediction independently of the duration of the crisis events phases, we train our model this time using *HeraultDB* and we test it using *Aples-MaritimesDB*. The prediction results of the obtained models are illustrated in Figure 7.5. According to these results, we observe that our model has detected most of prominent users even during the first phase which is characterized by a short duration of 3.5 hours. The obtained experimental results also confirm the comparison findings pointed through comparing the different models learned using *Aples-MaritimesDB* and tested using *HeraultDB*.

We conclude that the phase-unaware baseline *Pal* considering all the users' activities in the previous phases leads to better results in the first hours of each phase than our phase-aware model. However such recorded prior activities would erode the model performance after few hours (*Pal* vs. *Pal**). The obtained results also demonstrate the positive impact of the Bf feature which improves the detection results during the first hours of each new phase (*Ours* vs. *b1*). Such feature promotes users who were previously detected as prominent without biasing the real user activity at the analyzed phase. We also note that our phase-aware prediction model considering both the user behavior and event evolution over time outperforms *Pal* and *Pal** models characterizing users quantitatively and uniformly during the whole event.

7.7.2 Phase-aware vs Phase-unaware Models

Through this experiment, we aim to validate our assumption considering that the user prominence and behavior have to be associated with each event phase rather than the whole event. Thus, we compare our model with the phase-unaware baseline *b2*, and the phase-unaware-model *Pal* with *Pal**. Both *Pal* and *b2* consider that user prominence has to be associated according to their prominence during the whole event. In this experiment, we evaluate the different baselines' performance to identify prominent users at the end of each phase. Figure 7.6 reports the prominent users' identification results of each baseline.

TABLE 7.4: Prediction performance comparison of the baselines : **ours**, **Pal** and **b1** at the **beginning**, **one-third**, **half** and at the **end** of each event phase. The different baselines are compared mainly in terms of $Recall_{C1}$, $Recall@10$, $Precision@K$, $CommonPromP1P2$ and $CommonPromP2P3$. These measures reflect the baselines performance at the different stages of each disaster phase.

	Ours				b1				Pal				Pal*			
	Phase 1															
Evaluation Metrics	2h	6h	8h	15h	2h	6h	8h	15h	2h	6h	8h	15h	2h	6h	8h	15h
$Recall_{C1}$	30%	71%	71%	91%	30%	71%	71%	91%	12%	12%	20%	88%	12%	12%	20%	88%
$Precision_{C1}$	10%	9%	8%	9%	10%	9%	8%	9%	5%	5%	6%	11%	5%	5%	6%	11%
$Recall_{C2}$	93%	92%	92%	92%	93%	92%	92%	92%	97%	97%	96%	92%	97%	97%	96%	92%
$Precision_{C2}$	97%	97%	97%	99%	97%	97%	97%	99%	97%	97%	98%	98%	97%	97%	98%	98%
$Recall@10$	40%	100%	100%	100%	40%	100%	100%	100%	30%	30%	50%	100%	30%	30%	50%	100%
$Precision@K$	21%	71%	71%	88%	21%	71%	71%	88%	21%	23%	20%	35%	21%	23%	20%	35%
	Phase 2															
	2h	7h	9h	17h	2h	7h	9h	17h	2h	7h	9h	17h	2h	7h	9h	17h
$Recall_{C1}$	63%	96%	96%	95%	50%	96%	95%	94%	52%	77%	74%	71%	58%	93%	83%	81%
$Precision_{C1}$	11%	12%	10%	12%	20%	15%	10%	15%	17%	16%	17%	16%	26%	17%	20%	19%
$Recall_{C2}$	88%	81%	78%	92%	92%	85%	85%	85%	93%	89%	90%	90%	95%	91%	91%	91%
$Precision_{C2}$	98%	99%	99%	99%	98%	99%	99%	99%	95%	96%	96%	96%	96%	96%	96%	96%
$Recall@10$	90%	100%	100%	100%	60%	100%	100%	100%	90%	100%	100%	100%	80%	100%	100%	100%
$Precision@K$	41%	93%	93%	93%	39%	93%	93%	93%	52%	41%	41%	37%	50%	52%	46%	42%
$CommonPromP1P2$	90%	100%	100%	100%	68%	100%	100%	100%	86%	95%	95%	90%	63%	95%	95%	95%
	Phase 3															
	2h	6h	8h	15h	2h	6h	8h	15h	2h	6h	8h	15h	2h	6h	8h	15h
$Recall_{C1}$	45%	86%	97%	95%	37%	86%	91%	91%	52%	46%	44%	44%	50%	85%	82%	81%
$Precision_{C1}$	12%	16%	14%	12%	14%	16%	14%	14%	13%	11%	11%	10%	17%	15%	18%	19%
$Recall_{C2}$	88%	91%	87%	92%	93%	91%	87%	88%	92%	92%	92%	92%	94%	90%	93%	94%
$Precision_{C2}$	97%	99%	99%	99%	98%	99%	99%	99%	96%	96%	96%	96%	94%	90%	97%	97%
$Recall@10$	50%	90%	100%	100%	40%	90%	100%	100%	60%	60%	50%	50%	70%	70%	70%	70%
$Precision@K$	30%	76%	80%	80%	18%	76%	80%	80%	58%	34%	32%	34%	48%	52%	47%	50%
$CommonPromP2P3$	66%	90%	80%	100%	57%	90%	95%	90%	95%	90%	95%	95%	71%	90%	90%	90%

According to the obtained results, we observe that phase-aware-models (*Pal* vs. *Pal**) (*Ours* vs. *b2*) perform better than phase-unaware-models. The classification results of *Pal* and *Pal** models are the same at P1, as it is the first phase and all the users' features are already set to zero for the two models. *Pal** performs better than *Pal* in the next phases. The phase-unaware-users' representation of *Pal* promotes users who were prominent in the prior phases over the new prominent ones.

Comparing our phase-aware model with the phase-unaware model *b2*, we observe that *b2* registers low results at P1 and good results close to ours at P2 and P3 in terms of $Recall_{C1}^{Pj}$. These results can be explained by the fact that learning identification models by referring to all prominent users independently of their phases of prominence tends to bias the learning of the classification and ranking model.

Overall, we conclude that the consideration of event phases for representing user activity during the event leads to a better prominent users detection. Evaluating and representing microblog users according to their prominence at each phase would guarantee the construction of a more efficient prediction model (As demonstrated by the comparison of *Ours* vs. *b2*) and insure a fair evaluation for all users at any time of the event (As demonstrated by the comparison of *Pal** vs. *Pal*).

7.7.3 Phase-based User Characterization Evaluation

Through this experiment, we aim to prove the importance of modeling users behavior differently according to the specificities of each phase. Our model is compared with the *b3* model which characterizes users uniformly using the same features during the different phases. Figure 7.7 reports the results of this experiment.

By referring to the different evaluation metrics, our approach performs better than *b3* for both the classification and ranking of prominent users in the different phases. *b3* failed to identify the prominent users in P1. Modeling users uniformly during the whole event leads to good results only in phases characterized by high activity of prominent users such as P2 and P3. However, it would fail to identify the prominent ones during phases recording a low activity regarding the event topic such as P1.

Characterizing users' behavior differently at each phase would highlight the relevant users behavior characteristics for each phase. As demonstrated in this experiment, such characterization improves the identification results.

In order to understand more the users behavior differences at these different phases, we analyzed in Table 7.5 the nature of features selected by Corona at each phase in the preprocessing step. According to the obtained results, we observe that the number of selected

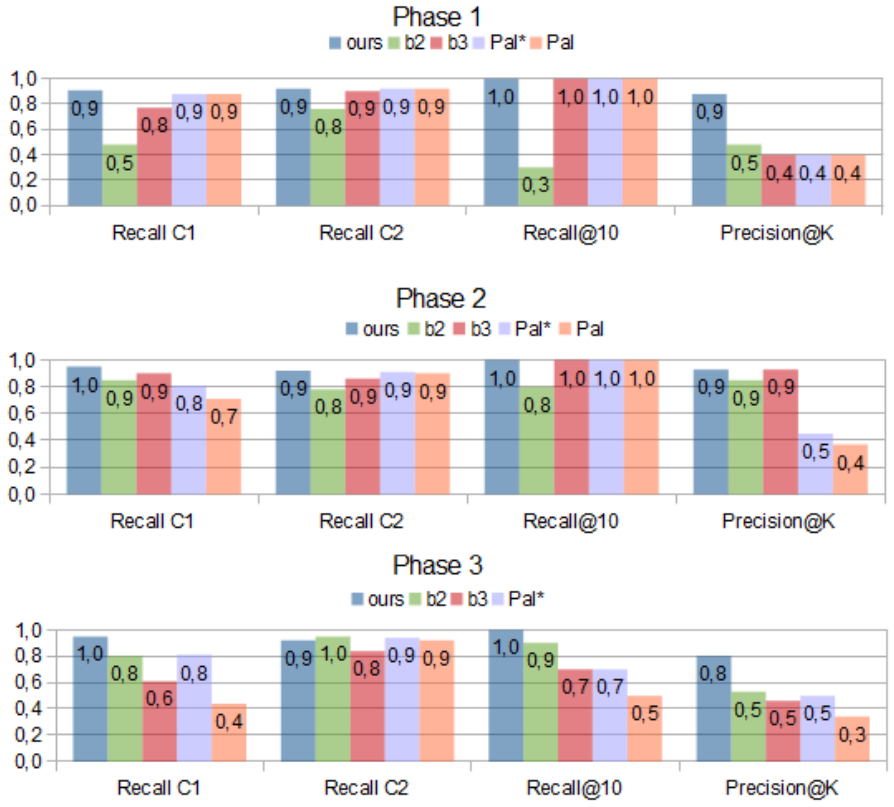

FIGURE 7.6: Comparison of the classification and ranking results of our model (**ours**), and the other phase-unaware baselines **Pal**, **b2** and **b3** at the end of the different phases **P1, P2** and **P3**. By comparing **Ours** Vs. **b2**, we aim to prove the importance of associating the user prominence with each phase independently of the other ones. Through **Ours** Vs. **b2**, we evaluate our proposed approach consisting of representing the user behavior differently at each phase. Through **Pal** Vs. **Pal*** comparison, we test the performance of our modeling approach consisting of evaluating users by only considering their shared activities at the analyzed phase.

on-topic features is close to the number of off-topic ones in P2, unlike the other phases P1 and P3. This can be explained by the fact that the behavior of prominent microblog users during P1 and P3 would be similar to their behavior in regular days as the danger are either not yet confirmed or removed. In such situations, users would share relevant information regarding the disaster but keep also tweeting about other topics. Thus, there is no need to penalize them regarding their off-topic behavior.

However, during the 2^{nd} phase, prominent microblog users who are generally concerned by the disaster would be in panic and would frequently share updates describing what they are seeing, hearing and experiencing. They would focus mainly on sharing the disaster event news. Thus, it is more rational to consider the off-topic features (i.e. On-engineered features adjusted by the off-topic ones or off-topic raw-features) in order to penalize those toggling between different topics and who are not necessarily concerned by the disaster. Using this strategy, the identification model would rank users active only regarding the

TABLE 7.5: Selected feature categories statistics recorded by different feature selection algorithms. **On** and **Off** refer respectively to on- and off-topic raw and engineered (**Eng**) features.

	Phase 1				Phase 2				Phase 3			
	Raw vs Eng		On vs Off		Raw vs Eng		On vs Off		Raw vs Eng		On vs Off	
Feature delection algorithm	Raw	Eng	On	Off	Raw	Eng	On	Off	Raw	Eng	On	Off
Corona	**6**	**3**	**6**	**3**	**5**	**4**	**5**	**4**	**6**	**3**	**7**	**2**
Clever	4	5	5	4	4	6	3	6	6	3	7	2
ReliefF	4	5	8	1	4	6	5	4	4	5	7	2
Average	0.52	0.48	0.7	0.3	0.48	0.59	0.48	0.52	0.59	0.41	0.78	0.22

disaster higher than those who are extremely active in several topics (e.g. news outlet channels users).

Through this experiment, we have shown the importance of selecting the most appropriated features for each event phase. This phase-based feature selection highlights the behavioral differences between prominent and non-prominent users, and hence improves the precision and the efficiency of the detection model.

7.7.4 Adequacy of the Feature Selection Algorithm

Through the previous experiments, we have shown the importance of the feature selection process per phase. In this experiment, we aim to prove the appropriateness of our chosen feature selection algorithm to our user modeling approach. Thus, we compare our adopted algorithm Corona with the following two feature selection algorithms :

Clever (Yang et al. 2005*a*) belongs to the family of unsupervised feature subset selection methods for multivariate time-series based on principal component analysis.

ReliefF (Robnik-Šikonja & Kononenko 2003) is a supervised feature selection algorithm which selects relevant features which works only on vectorized data. To apply this technique we vectorized each time-series sequence representing each user by summing the values of the same features recorded at each time-stamp.

As in our model, the number of selected features k is set to 9 for both Clever and ReliefF. Table 7.5 describes the statistics of the different categories of the selected features by each algorithm. According to these statistics, we observe that the number of selected raw, engineered, on- and off-topic (except P1 for On and Off) by the different algorithms is nearly the same for the different phases. For the phases P1 and P2, we note that there is a low number of Off-topic features considered compared by the number of On-topic ones.

As the selected features by the different algorithms are not necessarily the same even if they belong to the same category, we trained our model using the selected features by each feature selection algorithm in order to evaluate their effectiveness. Table 7.6 describes

TABLE 7.6: Performance comparison of different feature selection algorithms for the detection of prominent users at each phase in terms of $Recall_{C1}^{Pj}$ and $Precision@K$.

	Phase 1		Phase 2		Phase 3	
Feature selection algorithm	Recall C1	Precision@K	Recall C1	Precision@K	Recall C1	Precision@K
Corona	**0.91**	**0.89**	**0.95**	**0.95**	**0.95**	**0.59**
Clever	0.42	0.7	0.94	0.43	0.74	0.51
ReliefF	0.31	0.6	0.82	0.5	0.62	0.48

the obtained results by the different models in terms of $Recall_{C1}^{Pj}$ and $Precision@K$ for the selected prominent users class $C1$ at the end of each phase. We observe that features selected by Corona give the best results.

Through these experiments, we note that vectorizing the time-series representation without taking into account the different correlations of data hides the real importance of each feature. Thus, the choice of an appropriate feature selection algorithm has to take into consideration the nature of user representation.

7.7.5 Temporal User Sequence Representation Analysis

Through this experiment, we aim to demonstrate the importance of detailing the temporal distribution of user activities while modeling microblog users behavior at each phase. Thus, we evaluate the performance of our temporal user characterization approach by increasing the intervals of time m of our model from 30 minutes to 9 hours while extracting time series feature vectors representing each user. For example, after 4 hours from the beginning of a particular phase P_j, each microblog user would be represented by a sequence of 8 vectors of features if $m = 30$ min and a sequence of 2 feature vectors if $m = 2$ hours. Figure 7.7 shows the obtained prediction results in terms of $Recall_{C1}$ and $Precision@k$ at the one-third, half, two-third and the end of each disaster phase while tuning the temporal sequences' interval m.

According to the obtained results, we note that representing microblog users behavior into short sequences of vectors erodes the model ranking and classification results. By setting m to $9h$, the identification results at the end of each phase become close to those obtained by *Pal**. This explains the large differences between our temporal model results and those recorded by the phase-aware baseline *Pal**. Detailing the temporal distribution of user activity would lead to higher identification results.

According to this experiment, we find that extracting and representing users activities features at shorter periods of time works significantly better than longer ones. Detailing the temporal distribution of users activities highlights hidden user behavior patterns. Such

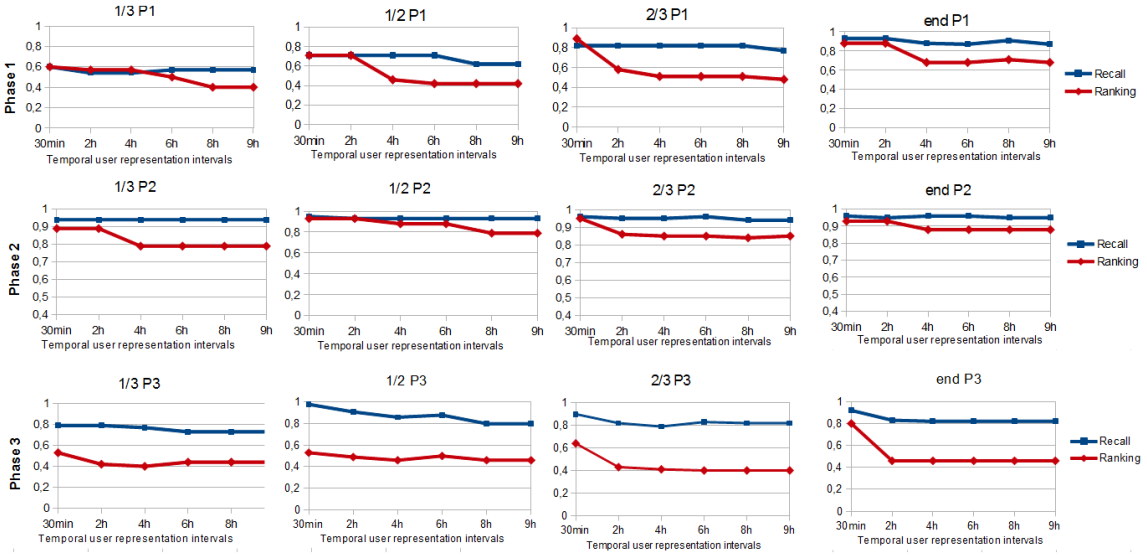

FIGURE 7.7: Prediction results comparison at **one-third**, **half**, **two-third** and the **end** of each disaster phase using different temporal sequence intervals for user representation. The prediction results are evaluated in terms of $Recall_{C1}$ and $Precision@k$. The temporal user representation intervals are increased at different stage of each disaster phase. The temporal distribution of users activities is represented using both short intervals (e.g. 30 minutes) detailing users behavior over time and long intervals representing the overall users behavior by a single or two series of features vectors. 30 minutes corresponds to the user representation interval set while learning and testing our model.

patterns would point out the behavior differences between prominent and non-prominent users.

7.8 Discussion

In this chapter, we have presented a phase-aware probabilistic model for real-time prominent microblog users identification during crisis events. The level of importance of these events changes over time. This evolution has to be considered while modeling users behavior and evaluating users prominence over time.

Prominent users may change at each new phase. As demonstrated by our ground-truth study, only few prominent users have been prominent from the first phase until the last one. We have also noted that only 2% of all the users who have interacted regarding the disaster were labeled as prominent. Such statistics were expected. During crisis events, many microblog users share or/and report event-related information. However, few of them would share the exclusive and relevant information required by emergency teams. Our experiments also show that while learning the identification models relative to each phase, only users who were prominent at the targeted phase have to be considered (*b2* vs. *Ours*). Users behavior learning by considering all the prominent users independently of their phase

of prominence would bias the results as the model will not be able to differentiate the true prominent users relative to each phase.

We also show that considering the event phases while representing the prominent users behavior leads to better results. As reported in Figure 7.6 through the comparison of *Pal* vs. *Pal** results, modeling users by referring only to their recorded activities in the current analyzed phase independently of the other prior ones improves the identification results. In phases two and three, *Pal** have detected most of the prominent users relative to each phase. However, the Recall results recorded by *Pal* are lower from phase to phase. As assumed in this chapter, considering the activities of users in prior phases would promote users who were active previously and penalize those who began to prove their prominence regarding the event.

Our strategy to model users behavior change according to the event evolution has also proved its effectiveness. We see that the models representing users uniformly using the same features from the first phase until the last one (*b3* models) would not highlight the real behavior of users at each particular phase. *b3* has registered lower results than our models at each phase. Users behavior changes during the event according to the event evolution. As reported in Table 7.5 which details the statistics of the selected features at each phase using different algorithms, around 48% of the selected features were on-topical and 52% were off-topical in the second phase. This can be argued by the fact that real prominent users during this phase are generally in panic, so, they tend to focus their attention only on what is happening during the disaster by sharing only on-topical information. Thus, by considering fairly both on- and off-topical metrics in such phase, the identification model will be able to distinguish between microblog users who are toggling between several topics and those active only regarding the disaster. We also note that during the first and last phase, the off-topical features were not extensively considered by the selection algorithms. This can be explained by the fact that in such phases there is no potential danger thus even prominent users tend to be active regarding on- and off-topics. In such cases, the off-topical features do not make prominent and non-prominent users more distinguishable.

We also show that our temporal sequence representation approach characterizing the user activity details at different timestamps of each event-phase has proved its importance. In Figure 7.7, we have shown that our model performance tends to decrease if we consider longer intervals between the different timestamps. The more we detail the user activities differences by considering several timestamps, the better the identification results are. Highlighting the temporal distribution of user activity can point out the hidden patterns reflecting the prominence of each user according to his/her behavior over time during each phase.

Lastly but most importantly, we demonstrate that our model can identify prominent users in real time at an early stage of each event phase. For example, 63% of prominent users

were detected after two hours from the beginning of the most important phase which is the second one. Even with learning our classification models *a priori* using similar events data, as described in Figure 7.4, our model outperforms the state-of-the-art phase-aware and unaware models (*Pal* and *Pal**) which are using unsupervised algorithms for classifying and ranking microblog users. We have also shown that with considering the user prominence in prior phases –using the Bf feature– , we can detect more prominent users at the first hours of each event phase. As reported in Table 7.4 by referring to the *CommonProm* measure results of *b1* and *Ours*, we succeed to detect the common prominent users relative to the prior and the current phases from the first hours. However, we note that the *Pal* models outperform our model on the first hours of each phase.

7.9 Conclusion

This chapter has proposed a phase-aware prediction model for detecting prominent microblog users in real time at each event phase. This model is based on a novel user modeling approach taking into account both the user behavior and the event evolution over time. Using this approach, microblog users are characterized differently in the beginning of each event phase using the best relevant features that can characterize their behavior according to the analyzed event phase particularities. Users are judged according to their prominence in an analyzed event phase independently of their activity prior to that phase.

Through the conducted experiments, we have shown that our prediction model significantly outperforms state-of-the-art models by detecting most of the prominent users at an early stage of each phase. We also proved that associating user prominence with event phases insures a fair evaluation for all users at each phase. We thus demonstrated that characterizing users using different features at each event phase improves the detection results and helps to highlight the user behavior differences according to each event phase specificities. Finally, we have shown that the choice of the feature selection algorithm has to be in harmony with the chosen user characterization.

For future work, we aim to automate the phases definition process by detecting phases boundaries in real time. This detection process will be explored in two ways by automatically extracting and analyzing the official organizations tweets in real time; and by detecting the emerging keywords describing each phase. We also aim to automate the detection of the different hashtags and keywords describing the targeted crisis event. The identified prominent users tweets over time will be explored in order to detect new relevant keywords and hashtags. This detection process will continuously enrich the list of referenced keywords and hashtags. We also would like to test our model in different crisis events types through collecting new datasets.

Chapter 8

Conclusions

This dissertation has made a number of contributions towards the goal of prominent microblog users identification in the context of crisis events. We define these targeted *prominent microblog users* as users who are susceptible to share relevant and exclusive information during a specific analyzed crisis event. These users do not typically refer to domain experts, they may refer to ordinary microblog users geolocated in the event area or having many relations with users who are geolocated there. Key contributions proposed for the identification of such users are: the use of multi-agent systems for microblog users tracking in real time, the use of new engineered features taking into account both the user on- and off-topical activities for user activity characterization, the proposition of a new phase-aware and time-sensitive user modeling approach and the use of MoG-HMM machine learning algorithm for a real-time identification of prominent microblog users. Let us summarize these contributions and discuss their implications as well as their limitations in more detail.

The MASIR multi-agent system based architecture presented in this dissertation integrates two complementary modules: a data extraction module and a data analysis and tracking module. The main purpose of this system is mainly the extraction of the needed Twitter data for the experimentation of new key users identification systems. The flexibility and modularity of multi-agent systems has also led us to integrate further complementary modules enabling both the identification and tracking of key users in real time cases based on Twitter APIs. The coordination between these modules is insured by the manager agents managing the extraction and listening processes. Such managers generate various listener agents in different hosts in order to be able to deal with the limits on both the volume of extracted data and the number of tracked users.

Based on this architecture, we extracted two data collections describing two different flooding events. Compared to the available standard collections, these extracted collections are exploitable by any key users identification system: graph-based ones or vector-based ones. The main particularity of these databases is the coverage of any users activities shared during the targeted event. Such particularity allows the exploration of new identification methods of key microblog users by mainly focusing on microblog users behavior. This architecture has also shown its efficacy to integrate any identification process and to insure the tracking of the detected key users in real time. The conducted experiments have

provided promising results. MASIR has coped with the limits imposed by Twitter APIs and has tracked around 175 users in real time using only 5 hosts. However, there is still some data lost due to the disconnections laps invoked by Twitter. This lost data can only be easily recovered for event of short duration. For long-term events, various hosts and Twitter developers connections are required to deal with this point.

In order to analyze the evaluated users behavior for the identification of prominent users, we focused in a first stage on the definition of an adequate modeling approach highlighting the different behaviors of these users. This approach has to reflect the particularities of the prominent and non-prominent users in the context of crisis events. The proposed user modeling approach presented in this dissertation considered three new dimensions that have not been explored in prior research work. These dimensions refer to: user topical activities dimension, users activities temporal dimension and event phases dimension.

Modeling microblog users according to their shared topical activities, more precisely their on- and off-topic activities, has been revealed to be more efficient than representing users only regarding their on-topic ones. Our proposed user modeling approach eases the identification of prominent users by focusing on evaluating the quality of the users activities rather than their quantity. User topical activities dimensions are reflected via new engineered features penalizing users having a higher off-topic activity regarding the on-topical one. As demonstrated in this dissertation, users toggling between several topics generally refer to popular microblog users such as news outlet channels CNN and BBC. Such users generally share various relevant but outdated on-topic information. They typically report what was already shared in the microblogging platform. Our engineered features point out these stated users particularities. As proved experimentally, these proposed features outperform on-topic based features defined in prior research works. Moreover, learning microblog users behavior based on user vector-based representation composed of these features insures better results than the state-of-the-art identification graph-based models considering mainly users relationships in the network.

While this proposed features vector representation results in good identification performance, such representation remains sensitive to active users sharing outdated information. Users active from the beginning of the event would be represented similarly as those who have become active by the end of the event. To deal with such ambiguities, we proposed a new efficient strategy to model the temporal distribution of users activities. This strategy consists of representing users by a sequence of feature vectors rather than a single one. Each vector has to represent both user on- and off- activities at a specific period of time rather than the whole event period of time. The use of such user characterization approach highlights the different temporal behavior specificities distinguishing prominent users from the other non-prominent ones. The experiments conducted to evaluate such representation show that a more detailed temporal distribution of user activities yields better identification

results. Such time-series vectors representation points out the behavioral patterns specific to prominent users. It has outperformed standard time-insensitive identification models in terms of prominent users prediction over time.

The identification model trained based on this user modeling approach has provided efficient identification results. However, most of prominent users predicted by this model has been detected after one day from the beginning of the event. Such prediction results are unconvincing. Prominent users need to be detected at an early stage of the event in order to be able to access in real time the needed valuable information shared over the analyzed event phases. To deal with this cold start prediction performance, we have considered an additional dimension while characterizing and evaluating microblog users.

This dimension refers to the event phases characterizing the evolution of the event over time. We have characterized each evaluated user by d sequences of feature vectors representing his/her behavior at each particular event phase. As shown by the conducted experiments in this dissertation, this phases-aware user characterization has many advantages compared to the standard characterization methods. First, it deals with the uniform characterization of users during the whole event. Only specific features that best reflect the user behavior at each particular phase are considered. Second, users are evaluated fairly at each phase independently of the other ones. Third, users prominence is associated with each phase independently of the other ones. By considering such user modeling approach, we have succeeded to predict most of prominent users at an early stage of each event phase.

This dissertation has also explored the use of machine learning techniques to improve the performance of prominent users detection. The idea consists of learning the behavior of prominent and non-prominent microblog users, based on the proposed user modeling approach. The different users behavior is learned by referring to past events databases. Through this learning process, we build new identification and prediction models that can be exploited during future similar events.

We have experimented two machine learning algorithms SVM and ANN for vector-based user representation and MOG-HMM for time-series vectors-based user representation. The phase-aware MOG-HMMs models built by learning microblog users behavior represented by the time-series representation provided the best prediction results. Such models separately learn the behavior of prominent microblog users and non-prominent ones over time per phase. They point out the behavioral patterns appropriate to each category of users. These models use the forward backward algorithm to compute the probability of each evaluated user to belong to the prominent or non-prominent class. Such generative models rank and predict prominent microblog users by measuring the similarity between the new evaluated user behavior and the learned prominent users behavior in past events situations. The user behavior is encrypted into various states defining his/her level of activeness. Such

models are suited to our problem nature where prominent users are rare regarding to the non-prominent ones for both learning and testing.

Overall, these contributions are major advancements in the research of information retrieval during crisis events. To the best of our knowledge, this is the first dissertation focusing on identifying prominent microblog users to gain a direct access to relevant and exclusive information shared during crisis events. The hope is that, such contributions provide the basis for the development of efficient information retrieval systems classifying, predicting, ranking and tracking prominent microblog users for the benefit of end users.

Future Work

There are many directions to proceed in the work presented in this dissertation.
In terms of the MASIR architecture, it can be enriched by speeding up the analysis process of users activities using big data analysis tools such as Hadoop[1] or Spark[2]. Another possible improvement is to integrate an additional tracking module extracting identified prominent users information through crawling their web interface. Such module would deal with the lost of data caused by Twitter APIs accounts frequent disconnections.

Moreover, as event-related keywords, hashtags and phases are defined by humans in order to launch our model process, it would be more convenient to automate this definition step. Event-related keywords can be automatically set and updated by analyzing the trending keywords shared by the already identified prominent microblog users. On the other side, phases can be defined and updated by analyzing the updated information shared by official accounts providing event-phases related information.

One possible improvement is to classify and rank the information content provided by the tracked prominent microblog users. Such process can be made by standard information content analysis techniques or by deep learning algorithms which are now applied for any type of information content. Such information categorization would help decision makers to access the most important information in real time independently of the prominence of their providers.

More databases of different crisis events natures could be collected for building more robust models, and also for performance evaluations. Additional features characterizing the activity of the evaluated microbog users prior the event (e.g. visited places, domain of interests, activeness, etc) could also be explored. By considering such features, we could study the prominent users behavior evolution prior- and post-event.

It is also possible to make our proposed user behavior modeling approach more dynamic by automatically detecting the user behavior state change over time. Users could be characterized by temporal sequences having different lengths. The length of the temporal sequence

[1] http://hadoop.apache.org/
[2] http://spark.apache.org/

would depend from the user behavior states change timestamps. Such dynamic user modeling approach could speed the identification process as the length of the temporal sequences characterizing each user would be optimized.

In terms of applications, our prominent microblog user identification system could be adapted for different other contexts: identifying experts in question and answer platforms, identifying "bot" accounts according to their behavior, predicting users behavior during the launch of a new product, etc.

Bibliography

Abbes, R., Moulahi, B., Chellal, A., Pinel-Sauvagnat, K., Hernandez, N., Boughanem, M., Tamine, L. & Yahia, S. B. (2015), IRIT at TREC temporal summarization 2015, *in* 'Proceedings of the 24th Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015'.

Abdulrahman, R., Neagu, D. & Holton, D. R. W. (2011), Multi agent system for historical information retrieval from online social networks, *in* 'Proceedings of the 5th KES International Conference on Agent and Multi-agent Systems: Technologies and Applications', KES-AMSTA '11, Springer-Verlag, Berlin, Heidelberg, pp. 54–63.

Ashktorab, Z., Brown, C., Nandi, M. & Culotta, A. (2014), Tweedr: Mining twitter to inform disaster response, *in* 'Proceedings of the 11th International ISCRAM Conference', ISCRAM '14, ACM.

Aslam, J. A., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., McCreadie, R. & Sakai, T. (2014), TREC 2014 temporal summarization track overview, *in* 'Proceedings of the 23rd Text REtrieval Conference', TREC '14.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011), *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books)*, 2 edn, Addison-Wesley Professional.

Baum, L. E. & Eagon, J. A. (1967), 'An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology', *Bulletin of the American Mathematical Society* **73**(3), 360–363.

Beigi, G., Hu, X., Maciejewski, R. & Liu, H. (2016), *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, Springer International Publishing, Cham, chapter An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief, pp. 313–340.

Bizid, I., Boursier, P., Morcos, J. & Faiz, S. (2015a), Masir : A multi-agent system for real-time information retrieval from microblogs during unexpected events, *in* 'Proceedings of the 9th KES Conference on Agent and Multi-Agent Systems Technologies and Applications', KES-AMSTA '15, pp. 1–8.

Bizid, I., Boursier, P., Morcos, J. & Faiz, S. (2015b), A classification model for the identification of prominent microblogs users during a disaster, *in* 'Proceedings of the 12th International Conference on Information Systems for Crisis Response and Management', ISCRAM '15.

Bizid, I., Nayef, N., Boursier, P., Faiz, S. & Doucet, A. (2015c), Identification of microblogs prominent users during events by learning temporal sequences of features, *in* 'Proceedings of the 24th ACM International on Conference on Information and Knowledge Management', CIKM '15, pp. 1715–1718.

Bizid, I., Nayef, N., Boursier, P., Faiz, S. & Doucet, A. (2016d), 'Modeling the behavior of microblog users for prominent users detection over crisis events phases', *ACM Transactions on Information Systems (TOIS)* p. [under review].

Bizid, I., Nayef, N., Boursier, P., Faiz, S. & Morcos, J. (2015e), Prominent users detection during specific events by learning on- and off-topic features of user activities, *in* 'Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining', ASONAM '15, pp. 500–503.

Bizid, I., Nayef, N., Naoui, O., Boursier, P. & Faiz, S. (2015f), A comparative study of microblogs features effectiveness for the identification of prominent microblog users during unexpected disaster, *in* 'Proceedings of the 2nd International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries', ISCRAM-med '15, pp. 15–26.

Borgatti, S. P. (2005), 'Centrality and network flow', *Social Networks* **27**(1), 55 – 71.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, *in* 'Proceedings of the 5th annual workshop on Computational learning theory', COLT '92, pp. 144–152.

Bošnjak, M., Oliveira, E., Martins, J., Mendes Rodrigues, E. & Sarmento, L. (2012), Twitterecho: A distributed focused crawler to support open research with twitter data, *in* 'Proceedings of the 21st International Conference on World Wide Web', WWW '12 Companion, ACM, New York, NY, USA, pp. 1233–1240.

Boyd, D., Golder, S. & Lotan, G. (2010), Tweet, tweet, retweet: Conversational aspects of retweeting on twitter, *in* 'Proceedings of the 43rd Hawaii International Conference on System Sciences', HICSS '10, pp. 1–10.

Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M. & Vesci, G. (2013), Choosing the right crowd: Expert finding in social networks, *in* 'Proceedings of the 16th International Conference on Extending Database Technology', EDBT '13, pp. 637–648.

Canali, C., Colajanni, M. & Lancellotti, R. (2011), Data acquisition in social networks: Issues and proposals, *in* 'Proceedings of the International Workshop on Services and Open Sources', SOS '11.

Cappelletti, R. & Sastry, N. (2012), Iarank: Ranking users on twitter in near real-time, based on their information amplification potential, *in* 'Proceedings of the 2012 International Conference on Social Informatics', SOCIALINFORMATICS '12, pp. 70–77.

Caragea, C., Mcneese, N., Jaiswal, A., Traylor, G., woo Kim, H., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J. & Yen, J. (2011), Classifying text messages for the haiti earthquake, *in* 'In Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management', ISCRAM '11.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K. (2010), Measuring user influence in twitter: The million follower fallacy, *in* 'Proceedings of 4th International AAAI Conference on Weblogs and Social Media', ICWSM' 10.

Chau, D. H., Pandit, S., Wang, S. & Faloutsos, C. (2007), Parallel crawling for online social networks, *in* 'Proceedings of the 16th International Conference on World Wide Web', WWW '07, ACM, New York, NY, USA, pp. 1283–1284.

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

Cleverdon, C. (1997), Readings in information retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, chapter The Cranfield Tests on Index Language Devices, pp. 47–59.

Cobo, A., Parra, D. & Navón, J. (2015), Identifying relevant messages in a twitter-based citizen channel for natural disaster situations, *in* 'Proceedings of the 24th International Conference on World Wide Web', WWW '15 Companion, ACM, New York, NY, USA, pp. 1189–1194.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A. & Menczer, F. (2011), Political polarization on twitter, *in* 'Proceedings 5th International AAAI Conference on Weblogs and Social Media (ICWSM)', ICWSM '11.

Dang, H. T. & Owczarzak., K. (2012), Overview of the tac 2008 update summarization task, *in* 'Proceedings of of CSCW', CSCW '12, Unfolding the event landscape on Twitter: Classification and exploration of user categories.

De Choudhury, M., Diakopoulos, N. & Naaman, M. (2012), Unfolding the event landscape on twitter: Classification and exploration of user categories, *in* 'Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work', CSCW '12, ACM, New York, NY, USA, pp. 241–244.

De Longueville, B., Smith, R. S. & Luraschi, G. (2009), "omg, from here, i can see the flames!": A use case of mining location based social networks to acquire spatio-temporal

data on forest fires, *in* 'Proceedings of the 2009 International Workshop on Location Based Social Networks', LBSN '09, ACM, New York, NY, USA, pp. 73–80.

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

Deng, L. & Jaitly, N. (2014), The 2014 quadrennial homeland security review, Technical report.
URL: *https://www.dhs.gov/sites/default/files/publications/2014-qhsr-final-508.pdf*

Deng, L. & Jaitly, N. (2015), Deep discriminative and generative models for pattern recognition, Technical Report MSR-TR-2015-59.
URL: *http://research.microsoft.com/apps/pubs/default.aspx?id=251677*

Earle, P., Bowden, D. & Guy, M. (2012), 'Twitter earthquake detection: earthquake monitoring in a social world', *Annals of Geophysics* **54**(6).

Endsley, M. R. (1988), 'Design and evaluation for situation awareness enhancement', **32**(2), 97–101.

Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N. & Gummadi, K. (2012), Cognos: Crowdsourcing search for topic experts in microblogs, *in* 'Proceedings of the 35th International ACM SIGIR conference on research and development in Information Retrieval', SIGIR '12, pp. 575–590.

Gupta, A., Joshi, A. & Kumaraguru, P. (2012), Identifying and characterizing user communities on twitter during crisis events, *in* 'Proceedings of the 2012 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media', DUBMMSM '12, ACM, New York, NY, USA, pp. 23–26.

Gupta, A., Kumaraguru, P., Castillo, C. & Meier, P. (2014), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, Springer International Publishing, Cham, chapter TweetCred: Real-Time Credibility Assessment of Content on Twitter.

Gupta, A., Lamba, H., Kumaraguru, P. & Joshi, A. (2013), Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy, *in* 'Proceedings of the 22nd International Conference on World Wide Web', WWW '13 Companion, ACM, New York, NY, USA, pp. 729–736.

Han, B., Cook, P. & Baldwin, T. (2014), 'Text-based twitter user geolocation prediction', *Journal of Artificial Intelligence Research* **49**(1), 451–500.

Hemant, P., Shreyansh, B., Andrew, H., Valerie, S., Amit, S. & John, F. (2014), With whom to coordinate, why and how in ad-hoc social media communities during crisis response,

in 'Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management', ISRAM '14, University Park, PA.

Honey, C. & Herring, S. (2009), Beyond microblogging: Conversation and collaboration via twitter, *in* 'Proceedings of the 42nd Hawaii International Conference on System Sciences', HICSS '09, pp. 1–10.

Hong, L. & Davison, B. D. (2010*a*), Empirical study of topic modeling in twitter, *in* 'Proceedings of the 1st Workshop on Social Media Analytics', SOMA '10, ACM, New York, NY, USA, pp. 80–88.

Hong, L. & Davison, B. D. (2010*b*), Empirical study of topic modeling in twitter, *in* 'Proceedings of the 1st Workshop on Social Media Analytics', SOMA' 10, ACM, New York, NY, USA, pp. 80–88.

Imran, M., Castillo, C., Diaz, F. & Vieweg, S. (2015), 'Processing social media messages in mass emergency: A survey', *ACM Computing Surveys* **47**(4), 67:1–67:38.

Imran, M., Castillo, C., Lucas, J., Meier, P. & Vieweg, S. (2014), Aidr: Artificial intelligence for disaster response, *in* 'Proceedings of the 23rd International Conference on World Wide Web', WWW '14 Companion, ACM, New York, NY, USA, pp. 159–162.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. & Meier, P. (2013*a*), Practical extraction of disaster-relevant information from social media, *in* 'Proc. of Workshop on Social Media Data for Disaster Management', WWW'13 Companion, Republic and Canton of Geneva, Switzerland, pp. 1021–1024.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. & Meier, P. (2013*b*), Practical extraction of disaster-relevant information from social media, *in* 'Proceedings of the 22nd International Conference on World Wide Web', WWW '13 Companion, ACM, New York, NY, USA, pp. 1021–1024.

Java, A., Pranam, Finin, T. & Oates, T. (2006), Modeling the spread of influence on the blogosphere, *in* 'Proceedings of the 15th International World Wide Web Conference', WWW '06.

Kayes, I., Qian, X., Skvoretz, J. & Iamnitchi, A. (2012), *Social Informatics: 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter How Influential Are You: Detecting Influential Bloggers in a Blogging Community, pp. 29–42.

Kiritchenko, S., Zhu, X. & Mohammad, S. M. (2014), 'Sentiment analysis of short informal texts', *Journal of Artificial Intelligence Research* **50**(1), 723–762.

Kumar, S., Barbier, G., Abbasi, M. A. & Liu, H. (2011), Tweettracker: An analysis tool for humanitarian and disaster relief., *in* L. A. Adamic, R. A. Baeza-Yates & S. Counts, eds, 'ICWSM', The AAAI Press.

Kumar, S., Morstatter, F., Zafarani, R. & Liu, H. (2013), Whom should i follow?: Identifying relevant users during crises, *in* 'Proceedings of the 24th ACM Conference on Hypertext and Social Media', HT '13, ACM, New York, NY, USA, pp. 139–147.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010), What is twitter, a social network or a news media?, *in* 'Proceedings of the 19th International Conference on World Wide Web', WWW '10, pp. 591–600.

Lakoff, G. (1987), *Women, fire, and dangerous things: what categories reveal about the mind*, University of Chicago Press Chicago.

Leskovec, J. & Krevl, A. (2014), 'SNAP Datasets: Stanford large network dataset collection', `http://snap.stanford.edu/data`.

Liao, Q. V., Wagner, C., Pirolli, P. & Fu, W.-T. (2012), Understanding experts' and novices' expertise judgment of twitter users, *in* 'Proceedings of the 2012 Conference on Human Factors in Computing Systems', CHI '12, New York, NY, USA, pp. 2461–2464.

Liben-Nowell, D. & Kleinberg, J. (2003), The link prediction problem for social networks, *in* 'Proceedings of the 12th International Conference on Information and Knowledge Management', CIKM '03, ACM, New York, NY, USA, pp. 556–559.

Luo, Z., Osbornez, M., Petrovic, S. & Wang, T. (2012), Improving twitter retrieval by exploiting structural information, *in* 'Proceedings of the 26th AAAI Conference on Artificial Intelligence', AAAI '12, AAAI Press, pp. 648–654.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanov, S., Savelyev, A., Blanford, J. & Mitra, P. (2011), Geo-Twitter analytics: Application in crisis management, *in* 'Proceedings of the 25th International Cartographic Conference'.

Mahmud, J., Nichols, J. & Drews, C. (2014), 'Home location identification of twitter users', *ACM Transactions on Intelligent Systems and Technology Journal* **5**(3), 47:1–47:21.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S. & Miller, R. C. (2011), Twitinfo: Aggregating and visualizing microblogs for event exploration, *in* 'Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems', CHI '11, ACM, New York, NY, USA, pp. 227–236.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I. & McCullough, D. (2012), On building a reusable twitter corpus, *in* 'Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '12, ACM, New York, NY, USA, pp. 1113–1114.

Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. (2013), Improving lda topic models for microblogs via tweet pooling and automatic labeling, *in* 'Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '13, ACM, New York, NY, USA, pp. 889–892.

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J. & Liu, H. (2014), 'Finding eyewitness tweets during crises', *CoRR* **abs/1403.1773**.

Munro, R. (2011), Subword and spatiotemporal models for identifying actionable information in haitian kreyol, *in* 'Proceedings of the 15th Conference on Computational Natural Language Learning', CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 68–77.

Olteanu, A., Castillo, C., Diaz, F. & Vieweg, S. (2014), CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises, *in* 'Proceedings of the 8th International Conference on Weblogs and Social Media', ICWSM '14, pp. 376–385.

Osuna, E., Freund, R. & Girosi, F. (1997), Support vector machines: Training and applications, Technical report, Cambridge, MA, USA.

Owoputi, O., O?Connor, B., Dyer, C., Gimpel, K., Schneider, N. & Smith, N. A. (n.d.), *in* 'Proceedings of NAACL-HLT'.

Pal, A. & Counts, S. (2011), Identifying topical authorities in microblogs, *in* 'Proceedings of the 4th ACM International Conference on Web Search and Data Mining', WSDM '11, pp. 45–54.

Pennacchiotti, M. & Popescu, A.-M. (2011), A machine learning approach to twitter user classification., *in* L. A. Adamic, R. A. Baeza-Yates & S. Counts, eds, 'Proceedings of the 5th International AAAI Conference on Weblogs and Social Media', ICWSM '11, The AAAI Press.

Perez-Lugo, M. (2004), 'Media uses in disaster situations: A new focus on the impact phase', *Sociological Inquiry* **74**(2), 210–225.

Pohl, D., Bouchachia, A. & Hellwagner, H. (2012), Automatic sub-event detection in emergency management using social media, *in* 'Proceedings of the 21st international conference companion on World Wide Web', WWW '12 Companion, New York, NY, USA, pp. 683–686.

Purohit, H., Hampton, A., Bhatt, S., Shalin, V. L., Sheth, A. P. & Flach, J. M. (2014), 'Identifying seekers and suppliers in social media communities to support crisis coordination', *Computer Supported Cooperative Work (CSCW)* **23**(4), 513–545.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A. & Menczer, F. (2011), Truthy: Mapping the spread of astroturf in microblog streams, *in* 'Proceedings of the 20th International Conference Companion on World Wide Web', WWW '11, ACM, New York, NY, USA, pp. 249–252.

Robnik-Šikonja, M. & Kononenko, I. (2003), 'Theoretical and empirical analysis of relieff and rrelieff', *Machine Learning* **53**(1-2), 23–69.

Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E. & Laredo, J. A. (2013), 'Crisistracker: Crowdsourced social media curation for disaster awareness', *IBM Journal of Research and Development* **57**(5), 4:1–4:13.

Romero, D. M., Galuba, W., Asur, S. & Huberman, B. A. (2011), Influence and passivity in social media, *in* 'Proceedings of the 25th International World Wide Web Conference', WWW '11, pp. 113–114.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P. & Ghosh, S. (2015), Extracting situational information from microblogs during disaster events: A classification-summarization approach, *in* 'Proceedings of the 24th ACM International on Conference on Information and Knowledge Management', CIKM '15, ACM, New York, NY, USA, pp. 583–592.

Sakaki, T., Okazaki, M. & Matsuo, Y. (2010), Earthquake shakes twitter users: real-time event detection by social sensors, *in* 'Proceedings of the 19th international conference on World wide web', WWW '10, New York, NY, USA, pp. 851–860.

Saroop, A. & Karnik, A. (2011), Crawlers for social networks amp; structural analysis of twitter, *in* 'Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on', pp. 1–8.

Sasaki, K., Yoshikawa, T. & Furuhashi, T. (2014), Online topic model for twitter considering dynamics of user interests and topic trends, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing', EMNLP '14, pp. 1977–1985.
URL: *http://aclweb.org/anthology/D/D14/D14-1212.pdf*

Seo, E., Mohapatra, P. & Abdelzaher, T. (2012), Identifying rumors and their sources in social networks.

Silva, A., Guimarães, S., Meira, Jr., W. & Zaki, M. (2013), Profilerank: Finding relevant content and influential users based on information diffusion, *in* 'Proceedings of the 7th Workshop on Social Network Mining and Analysis', SNAKDD '13, pp. 2–9.

Smailovic, V., Striga, D., Mamic, D.-P. & Podobnik, V. (2014), Calculating user's social influence through the smartsocial platform, *in* 'Proceedings of the 22nd International Conference on Software, Telecommunications and Computer Networks', SoftCOM' 14, pp. 383–387.

Spring, N., Peterson, L., Bavier, A. & Pai, V. (2006), 'Using planetlab for network research: Myths, realities, and best practices', *ACM SIGOPS Operating Systems Review* **40**(1), 17–24.

Starbird, K., Muzny, G. & Palen, L. (2011), Learning from the crowd: Identifying on-the-ground twitterers learning from the crowd: Collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions, *in* 'In Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management', ISCRAM '11, ACM, New York, NY, USA, pp. 23–26.

Starbird, K. & Stamberger, J. (2010), Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting, *in* 'Proceedings of the 7th international ISCRAM Conference', ISCRAM '10.

Subbian, K. & Melville, P. (2011), Supervised rank aggregation for predicting influencers in twitter, *in* 'Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)', SocialCom '11, pp. 661–665.

Sultanik, E. A. & Fink, C. (2012), Rapid geotagging and disambiguation of social media text via an indexed gazetteer, *in* 'Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management', ISCRAM '12.

Tan, L., Roegiest, A. & Clarke, C. L. A. (2015), University of waterloo at TREC 2015 microblog track, *in* 'Proceedings of the 24th Text REtrieval Conference', TREC '15.

Theodore, J. (2013), 'Esri and geofeedia expand social media with location analytics'. **URL:** *Retrieved June 13, 2013, from http://www.esri.com/esri-news/releases/13-2qtr/esri-and-geofeedia-expand-social-media-with-location-analytics*

Valkanas, G., Saravanou, A. & Gunopulos, D. (2014), *Web Information Systems Engineering – WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part II*, Springer International Publishing, Cham, chapter A Faceted Crawler for the Twitter Service, pp. 178–188.

Vapnik, V. N. (1995), *The nature of statistical learning theory*, Springer-Verlag New York, Inc.

Wagner, C., Liao, V., Pirolli, P., Nelson, L. & Strohmaier, M. (2012), It's not in their tweets: Modeling topical expertise of twitter users, *in* 'Proceedings of the 2012 Conference on Social Computing and Networkin', SocialCom '12, pp. 91–100.

Wang, A. H. (2010), Don't follow me - spam detection in twitter., *in* S. K. Katsikas & P. Samarati, eds, 'SECRYPT', SciTePress, pp. 142–151.

Wang, Y., Agichtein, E. & Benzi, M. (2012), Tm-lda: Efficient online modeling of latent topic transitions in social media, *in* 'Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '12, ACM, New York, NY, USA, pp. 123–131.

Weng, J., Lim, E.-P., Jiang, J. & He, Q. (2010), Twitterrank: Finding topic-sensitive influential twitterers, *in* 'Proceedings of the 3rd International Conference on Web Search and Data Mining', pp. 261–270.

Wu, S., Gong, L., Rand, W. & Raschid, L. (2012), Making recommendations in a microblog to improve the impact of a focal user., *in* P. Cunningham, N. J. Hurley, I. Guy & S. S. Anand, eds, 'Proceedings of the 6th ACM Conference on Recommender Systems', RecSys '12, ACM, pp. 265–268.

Xianlei, S., Chunhong, Z. & Yang, J. (2014), Finding domain experts in microblogs, *in* 'Proceedings of the 2014 International Conferenceon Web Information Systems and Technologies', WEBIST '14.

Yang, K., Yoon, H. & Shahabi, C. (2005*a*), Clever: A feature subset selection technique for multivariate time series, *in* 'Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining', PAKDD '05, pp. 516–522.

Yang, K., Yoon, H. & Shahabi, C. (2005*b*), A supervised feature subset selection technique for multivariate time series, *in* 'Proceedings of the 2005 International Conference on Fuzzy Systems and Data Mining', FSDM '05.

Ye, S., Lang, J. & Wu, F. (2010), Crawling online social graphs, *in* 'Web Conference (APWEB), 2010 12th International Asia-Pacific', pp. 236–242.

Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B. & Power, R. (2015), Using social media to enhance emergency situation awareness: Extended abstract, IJCAI1 '15.

Yin, J., Lampert, A., Cameron, M., Robinson, B. & Power, R. (2012), 'Using social media to enhance emergency situation awareness', *Intelligent Systems, IEEE* **27**(6), 52–59.

Zhang, G. P. (2000), Neural networks for classification: a survey, *in* 'and Cybernetics - Part C: Applications and Reviews'.

Prédiction des utilisateurs primordiaux des microblogs durant les situations de crise : Modélisation temporelle des comportements des utilisateurs en fonction des phases des évènements

Durant les situations de crise, telles que les catastrophes, le besoin de recherche d'information (RI) pertinentes partagées dans les microblogs en temps réel est inévitable. Cependant, le grand volume et la variété des flux d'informations partagées en temps réel dans de telles situations compliquent cette tâche. Contrairement aux approches existantes de RI basées sur l'analyse du contenu, nous proposons de nous attaquer à ce problème en nous basant sur les approches centrées utilisateurs tout en levant un certain nombre de verrous méthodologiques et technologiques inhérents : 1) à la collection des données partagées par les utilisateurs à évaluer, 2) à la modélisation de leurs comportements, 3) à l'analyse des comportements, et 4) à la prédiction et le suivi des utilisateurs primordiaux en temps réel.
Dans ce contexte, nous détaillons les approches proposées dans cette thèse afin de prédire les utilisateurs primordiaux qui sont susceptibles de partager les informations pertinentes et exclusives ciblées et de permettre aux intervenants d'urgence d'accéder aux informations requises quelque soit le format (i.e. texte, image, video, lien hypertexte) et en temps réel. Ces approches sont centrées sur trois principaux aspects. Nous avons tout d'abord étudié l'efficacité de différentes catégories de mesures issues de la littérature et proposées dans cette thèse pour représenter le comportement des utilisateurs. En nous basant sur les mesures pertinentes résultant de cette étude, nous concevons des nouvelles caractéristiques permettant de mettre en évidence la qualité des informations partagées par les utilisateurs selon leurs comportements. Le deuxième aspect consiste à proposer une approche de modélisation du comportement de chaque utilisateur en nous basant sur les critères suivants : 1) la modélisation des utilisateurs selon l'évolution de l'évènement, 2) la modélisation de l'évolution des activités des utilisateurs au fil du temps à travers une représentation sensible au temps, 3) la sélection des caractéristiques les plus discriminantes pour chaque phase de l'évènement. En se basant sur cette approche de modélisation, nous entraînons différents modèles de prédiction qui apprennent à différencier les comportements des utilisateurs primordiaux de ceux qui ne le sont pas durant les situations de crise. Les algorithmes SVM et MOG-HMMs ont été utilisés durant la phase d'apprentissage. La pertinence et l'efficacité des modèles de prédiction appris ont été validées à l'aide des données collectées par notre système multi-agents MASIR durant deux innondations qui ont eu lieu en France et des vérités terrain appropriées à ces collections.
Mots clés : Recherche d'information, modélisation du comportement des utilisateurs des microblogs, prédiction des utilisateurs primordiaux, gestion des situations de crise, système multi-agents.

Prominent Microblog Users Prediction during Crisis Events : Using Phase-aware and Temporal Modeling of Users Behavior.

During crisis events such as disasters, the need of real-time information retrieval (IR) from microblogs remains inevitable. However, the huge amount and the variety of the shared information in real time during such events over-complicate this task. Unlike existing IR approaches based on content analysis, we propose to tackle this problem by using user-centric IR approaches with solving the wide spectrum of methodoligical and technological barriers inherent to : 1) the collection of the evaluated users data, 2) the modeling of user behavior, 3) the analysis of user behavior, and 4) the prediction and tracking of prominent users in real time.
In this context, we detail the different proposed approaches in this disseration leading to the prediction of prominent users who are susceptible to share the targeted relevant and exclusive information on one hand and enabling emergency responders to have a real-time access to the required information in all formats (i.e. text, image, video, links) on the other hand. These approaches focus on three key aspects of prominent users identification. Firstly, we have studied the efficiency of state-of-the-art and new proposed raw features for characterizing user behavior during crisis events. Based on the selected features, we have designed several engineered features qualifying user activities by considering both their on-topic and off-topic shared information. Secondly, we have proposed a phase-aware user modeling approach taking into account the user behavior change according to the event evolution over time. This user modeling approach comprises the following new novel 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 characterizing users at each event phase. Thirdly, based on this proposed user modeling approach, we train various prediction models to learn to differentiate between prominent and non-prominent users behavior during crisis event. The learning task has been performed using SVM and MoG-HMMs supervised machine learning algorithms. The efficiency and efficacy of these prediction models have been validated thanks to the data collections extracted by our multi-agents system MASIR during two flooding events who have occured in France and the different ground-truths related to these collections.
Keywords: Information retrieval, microblog user behavior modeling, prominent users prediction, crisis events management, multi-agent systems.

Laboratoire Informatique, Image, Interaction
Faculté de Sciences & Technologie
Avenue Michel Crépeau