Twitter User Profiling Model Based on Temporal Analysis of Hashtags and Social Interactions

Abir Gorrab, Ferihane Kboubi, Bénédicte Le Grand, Henda Ben Ghezala, Ali Jaffal

Twitter User Profiling Model Based on Temporal Analysis of Hashtags and Social Interactions

Abir Gorrab[1], Ferihane Kboubi[1], Ali Jaffal[2], Bénédicte Le Grand[2] and Henda Ben Ghezala[1]

[1]RIADI laboratory, National School of computer Science (ENSI)
University of Manouba, Tunisia
{Abir.Gorrab,Ferihane.Kboubi,henda.benghezala}@riadi.rnu.tn
[2]Centre de Recherche en Informatique
University Paris1 Panthéon Sorbonne, Paris, France
Ali.jaffal@malix.univ-paris1.fr
Benedicte.le-grand@univ-paris1.fr

Abstract. Social content generated by users' interactions in social networks is a knowledge source that may enhance users' profiles modeling, by providing information on their activities and interests over time. The aim of this article is to propose several original strategies for modeling profiles of social networks' users, taking into account social information and its temporal evolution. We illustrate our approach on the Twitter network. We distinguish interactive and thematic temporal profiles and we study profiles' similarities by applying various clustering algorithms, by giving a special attention to overlapping clusters. We compare the different types of profiles obtained and show how they can be relevant for the recommendation of hashtags and users to follow.

Keywords. User profile, temporality, social interactions, hashtags, similarity.

1 Introduction

With the success of social networks, the integration of social information has become strategic. In this paper, we investigate strategies to build profiles of Twitter users that exploit social information, which is heterogeneous and evolves over time. Our final goal is to use these profiles to cluster users with similar profiles in order to suggest new hashtags and users to follow. In addition, thematic content reflects users' interests and is then prominent in analyzing their preferences. The temporal aspect of social content is also used in social works in order to track the evolution of users' social behaviors. With this in mind and inspired by studies of time-sensitive social profiles, we propose a new social user profile construction strategy and analysis.

The rest of this paper is organized as follows. Section 2 reviews related work. We detail in Section 3 our proposal including temporal social interactions' and temporal hashtags' analysis, social profiles' construction and users' clustering. In Section 4, we analyze and discuss the effectiveness of our model on a dataset of tweets. Finally, Section 5 concludes the paper and introduces future work.

2 Related Work

In this section, we expose some related works on temporal information exploitation. We then review works on user profiles similarity. Temporal characteristics have been investigated in various research works, but for different purposes. In [4], authors proposed a time-aware user profile model based on social relations, by measuring freshness and importance of social users' interests. In [1], a language model document prior is proposed that uses social and temporal features to estimate documents' relevance. A wide range of researchers have focused on measuring the similarity of social user profiles. In [9], authors propose a social user profiling model and use it in an Information Retrieval system. They also propose a new process of search results' classification. Besides, diffusion kernels are exploited in [10] to calculate tags similarities, by connecting users based on social similar preferences. Furthermore, authors in [8] analyze Twitter profiles, by calculating their similarities using TF-IDF, after applying an indexation algorithm with Lucene.

What distinguishes our work from the existing approaches is the strategy of social profiles' constitution and analysis of users' clusters obtained, with a focus on overlapping clusters.

3 Proposed Model

Our methodology of social users' profiling comprises three main steps, as shown in Figure 1. The first step consists in information preprocessing, by collecting and filtering social information. The input data consists, for each user, of a set of tweets written during a time interval. We differentiate six features: the user ID, the number of tweets he wrote, the list and the number of hashtags contained in each tweet, their timestamps and the number of followers. In a second step, we build interactive and thematic social and temporal user profiles. In fact, we use our generic social user profile model proposed in [8] to instantiate and build original social and temporal profiles. The third step exploits these profiles to build users' clusters. Clustering interactive and temporal user profiles allow us to regroup similar users based on social properties like activity and popularity and study their temporal evolution.

In Section 3.1, we detail our methodology for social and temporal user profiles construction. In Section 3.2, we describe users' clustering process.

3.1 Social and Temporal Profiles' Construction

To build social profiles, we distinguished two social information types, notably social interactions which include the number of followers, the number of tweets and also the number of hashtags contained in each tweet; and also the thematic hashtags.

3.1.1 Interactive and Temporal User Profiles' Construction

The number of followers of a given user, the number of his tweets published in a given time interval and the number of hashtags contained in each tweet provide a clear vision of this user's social activity and popularity. We set a time variable t that

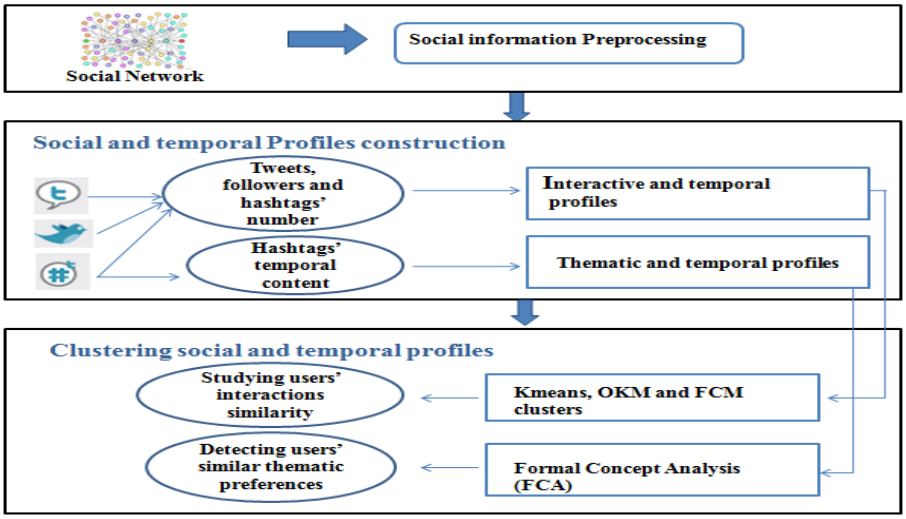

Fig.1. Our methodology for building and exploiting social and temporal user profiles

determines the duration of the social interactions that are considered, where $t \in [t_0 \dots t_{current}]$; with t_0 is the timestamp of the oldest tweet in the dataset. We then calculate the number of tweets written by the user in this time interval, and also the number of hashtags contained in his tweets. Considering the number of followers, we always consider the latest update of the followers list.

3.1.2 Thematic and Temporal User Profiles' Construction

We distinguish three types of hashtags' lists that could be used to characterize user profiles. The difference between them results from tweets' temporality and frequency.

☐**Historical profile (P_H):** we consider all the tweets of the user since the initial time.

$$P_H = \{H_i(t)\} \quad (1)$$

☐**History and frequency based profile (P_{HF}).** In this case, we use all the social content starting from t_0. We then remove unfrequent hashtags from the user profile, unless they are recent. TF measure is used here to calculate hashtags' frequency in each user model.

$$P_{HF} = \{H_i(t)\} \setminus \{H(NF_i)(t)\} \quad (2)$$

☐**Instantaneous profile (P_I),** considers only the hashtags from the most recent tweet sent by the user.

$$P_I = \{H_i(t_{current})\} \quad (3)$$

$\forall i \in [1 \dots N]; \forall t \in [t_0 \dots t_{current}]$; Where N is the total number of hashtags in the tweets sent by the user, $Hi(t)$ corresponds to the i^{th} hashtag at instant t; and $H(NF_i)(t)$ is the i^{th} unfrequent hashtag at instant t. A hashtag is considered unfrequent if its appearance frequency in the user model does not exceed a threshold $\theta \in [0 \dots 1]$.

3.2 Users' Clustering Based on Social and Temporal Profiles

To cluster users according to the similarity of their social interactions, we apply Kmeans, OKM [4] and FCM [3] algorithms to each component, notably the number of followers, tweets and hashtags in order to emphasize the significance of each social feature. To cluster users according to their thematic and temporal profiles, we construct a similarity matrix. We then apply the same clustering algorithms to the similarity matrix obtained and compare resulting clusters. Our aim is to track the level of users' belonging to the various clusters, based on their hashtags' temporal similarities. Furthermore, we apply another approach to cluster users based on their thematic profiles similarities: Formal Concept Analysis (FCA) developed in [9]. FCA builds overlapping clusters with native labels, by constructing conceptual graphs called Galois lattices. This approach takes as input a set of objects characterized by attributes called formal context. In our case study, the objects represent the ids of twitter users and the attributes are the hashtags associated to these users. From each formal context, FCA groups objects (users) into clusters according to their common attributes (hashtags). These clusters are called formal concepts.

From each clustering result obtained, we can provide various recommendations of hashtags or users to follow. We can provide a given user with common hashtags in the cluster to which he belongs, users having the same interactive properties or thematic similarities, and even users from other clusters that are very active or popular.

4 Experimental Illustration

We conducted a series of experiments on a Twitter dataset, used in [5]. From this dataset, we extracted a significant sample of tweets corresponding to 1050 users, notably 4000 tweets. We chose users with different values of social features, i.e. different numbers of tweets, followers and hashtags.

4.1 Illustration of Clustering Based on Social Interactions

In these experiments, we study the interactive profiles considered from initial time. We compared the number of profiles contained in each cluster corresponding to the three dimensions of the interactive and temporal profile, and respectively to each dimension. The results are shown in Figure 2 where Nfollows, Ntweets and Nhashtags denote respectively the number of followers, tweets and hashtags. C1, C2 and C3 are respectively cluster 1, 2 and 3. We notice that the number of users in each cluster changes according to the features that have been considered. If we consider the number of followers, the number of profiles is reduced in C1 and 2, but it is higher in C3. When compared in terms of tweets number, the number of users is the highest in C2 and less important in C1 and 3.

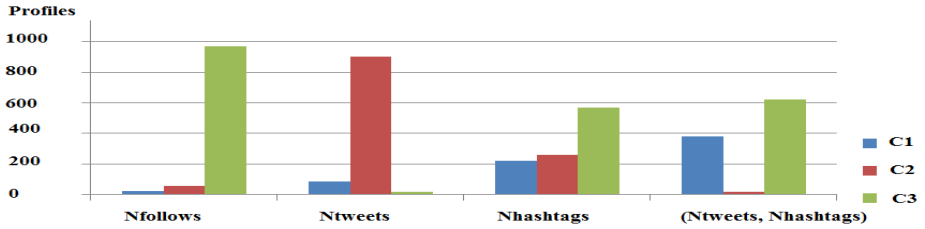

Fig.2. Comparative graph of users' numbers by cluster

For Nhashtags, the number of users varies from 220 in C1 to 260 in C2, and reaches 570 in C3. The number of profiles in each cluster gives a valuable indication of the quantitative aspect of user's social activity.

4.2 Illustration of Clusters Based on Thematic Profiles

To study the impact of thematic user profiles on clustering results, we start by choosing the adequate value of θ. We choose $\theta = 0.25$ to eliminate the most unfrequent hashtags. Therefore, we analyze tweets starting from the initial instant t_0. We analyze then the three profiles types: P_H, P_{HF} and P_I. We also eliminate from our dataset the users who have only one tweet, since they stay invariant in all the strategies.

We summarize the results of OKM clustering. For each type of user profiles, we show the list of overlapping clusters and the relative number of profiles they contain. For P_H profiles, there are three overlapping clusters that contain respectively 59, 7 and 1 profile. That means that 59 users belong simultaneously to both clusters 1 and 2, 7 users are in clusters 1, 2 and 3 and one user belongs to clusters 1, 2, 3 and 4. P_{HF} profiles are also formed of three overlapping clusters of 24, 23 and 16 users, while with the P_I profiles, 28 users belong to both clusters 1 and 2, and 10 users are in clusters 1, 2 and 3 simultaneously, with different membership degrees. We can use these clusters to recommend users, considering each type of thematic profiles. A limit of this approach is the lack of cluster's labeling. This is where the contribution of FCA appears, as it facilitates clustering results' interpretation.

4.3 Formal Concept Analysis of Temporal Thematic Profiles

From the lattice, we calculate the conceptual similarity between users and between hashtags. Users are conceptually similar if they belong to the same concepts; which means that they use the same hashtags as other users in different concepts. This similarity is defined by equation (4).

$$\text{Conceptual similarity}(ui, uj) = \frac{\text{Nb concepts containing } ui \text{ and } uj}{\text{Nb concepts containing ui or uj}} \quad (4)$$

Table 3 presents the pairs of most similar users based on the P_H profile. The pair of most similar users is (374; 231) with a conceptual similarity of 60%. These results

are relevant and can be exploited to recommend hashtags of similar users, since users who present a high similarity level are likely to be interested in similar topics.

(456 ; 514) : 40%	(456 ; 363) : 40%	(363 ; 456) : 40%	
(1242 ; 235) : 50%	(235 ; 1242) : 50%	(1008 ; 514) : 50%	(738 ; 374) : 50%
(514 ; 903) : 57%	(903 ; 514) : 57%		
(374 ; 231) : 60%	(231 ; 374) : 60%		

Table 3. Users' similarities percentage for the historical profile P_H

5 Conclusion and Future Work

In this work, we conducted a deep study to investigate the efficiency of the temporal strategy of deriving social user profiles and its impact on users' clustering. We built interactive and thematic temporal social profiles, formed respectively by users' social interactions and hashtags' content. Thematic profiles are differentiated by the history taken into account and also the frequency of hashtags in each profile. The formula of deriving profiles was proved so useful to similarities' calculation and clusters' construction. To go further, we will integrate these analysis in hashtags and users' recommender system, taking into account user's preferences and the clusters to which he belongs.

References

1. Badache, I., Boughanem, M.: Document Priors Based On Time-Sensitive Social Signals. In *ECIR* (2015).
2. Bezdek, J. C., Trivedi, M., Ehrlich, R., Full, W.: Fuzzy clustering: A new approach for geostatistical analysis. *International Journal of Systems, Measurement and Decision*, 1(2), pp. 13-24 (1981).
3. Canut, M. F., On-At, S., Péninou, A., Sèdes, F.: Time-aware Egocentric network-based User Profiling. In *ASONAM* (2015).
4. Cleuziou, G.: A generalization of k-means for overlapping clustering. *Technical report*, 54 (2007).
5. Danisch, M., Dugué, N., Perez, A.: On the importance of considering social capitalism when measuring influence on Twitter. In *Behavioral, economic, and socio-cultural computing* (2014).
6. Gorrab, A., Kboubi, F., Le Grand, B., Ghezala, H. B: Towards a dynamic and polarity-aware social user profile modeling. In *AICCSA (2016).*
7. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys*, pp. 199-206 (2010).
8. Jaffal, A., Le Grand, B.: Towards an Automatic Extraction of Smartphone Users' Contextual Behaviors. In *RCIS* (2016).
9. Nathaneal ,R., Andrews , J.: Personalized Search Engine using Social Networking Activity. In *Indian Journal of Science and Technology* (2015).
10. Wang, X., Liu, H., Fan, W.: Connecting users with similar interests via tag network inference. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1019-1024 (2011).