Online Social Networks and Media 5 (2018) 61–75

Contents lists available at ScienceDirect

Online Social Networks and Media

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

User-centered probabilistic models for content diffusion in the blogosphere

Cédric Lagnier[a,b,*], Eric Gaussier[a,b], François Kawala[c]

[a] University Grenoble Alpes, CNRS, Grenoble INP, LIG, France
[b] Skopai, France
[c] Société Karos, France

ARTICLE INFO

Article history:
Received 30 September 2017
Revised 23 January 2018
Accepted 23 January 2018

Keywords:
Information diffusion
Modelization
Social networks

ABSTRACT

Predicting the diffusion of information in social networks is a key problem for applications like Opinion Leader Detection, Buzz Detection or Viral Marketing. Many diffusion models are direct extensions of the *Cascade* and *Threshold* models, initially proposed for epidemiology and social studies. In such models, the diffusion process is based on the dynamics of interactions between neighbor nodes in the network (the social pressure), and largely ignores important dimensions as the content diffused and the active/passive role users tend to have in social networks. We propose here a new family of models that aims at predicting how a content diffuses in a network by making use of additional dimensions: the content diffused, user's profile and willingness to diffuse. In particular, we show how to integrate these dimensions into simple feature functions, and propose a probabilistic modeling to account for the diffusion process. These models are then illustrated and compared with other approaches on two blog datasets. The experimental results obtained on these datasets show that taking into account the content diffused is important to accurately model the diffusion process. Lastly, we study the influence maximization problem with these models and prove that it is NP-hard, prior to propose an adaptation of the greedy algorithm to approximate the optimal solution.

1. Introduction

Propagation models in content networks, i.e. social networks in which content are shared and diffused among users, aim at reproducing the diffusion of information between users. Being able to accurately model this diffusion has several practical applications, as the identification of influence hubs, the choice of initial diffusers for a maximal diffusion, or the identification of links one has to remove in order to limit the diffusion (e.g. for stopping rumors).

Most of the models proposed in the domain of information diffusion are extensions of the Independent Cascade model (IC) [1] and the Linear Threshold model (LT) [2]. The IC model is based on the following simple principle: as soon as a user (i.e. a node in the social network) n_j is infected, she has a unique chance to infect each of her direct neighbors n_i with a probability P_{ji} that depends on both n_j and n_i. The LT model considers that a node n_i of the social network (i.e. a user) is contaminated if the sum of the weights on its incoming edges are above a threshold θ_i specific to n_i, this threshold being chosen randomly in many instances of the model [3]. They nevertheless fail to take into account for two important elements:

- They ignore the content of the information diffused even though, in a given social network, two different pieces of information will not propagate in the same way;
- They tend to ignore users characteristics even though the interest of a particular user plays a major role in the diffusion process.

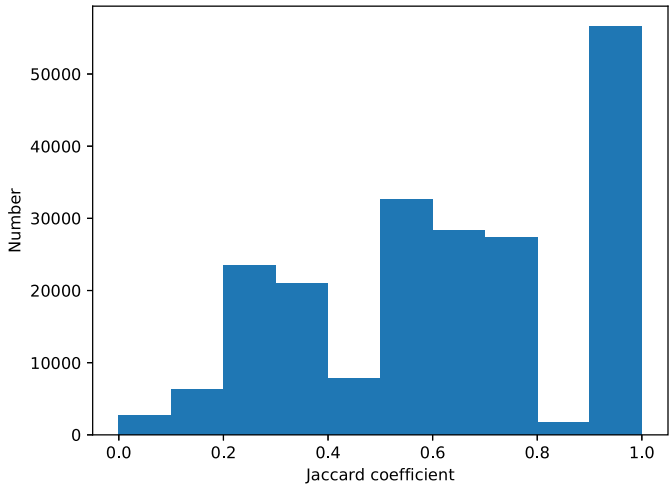
Ignoring the content being diffused entails that, in these models, different contents issued from the same user will diffuse in the same manner. In other words, in content-agnostic models, the diffusion cascades[1] originating from a given user are the same, regardless of the content being diffused.

However, this does not correspond to what is happening in real social networks. To illustrate that, we compared diffusion cascades

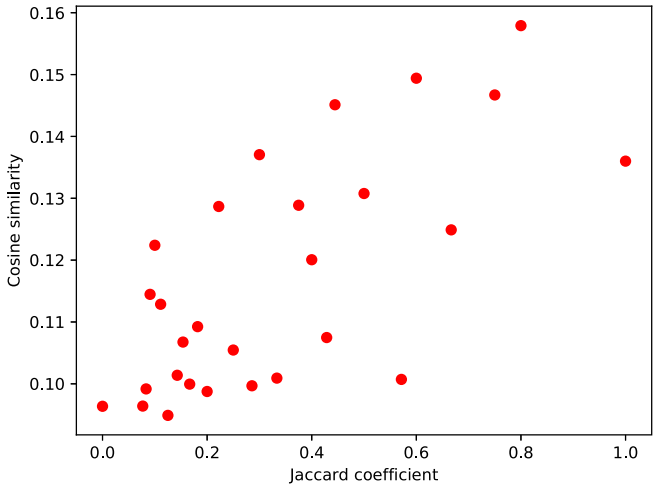
* Corresponding author at: Skopai, France.
E-mail addresses: cedric.lagnier@skopai.com (C. Lagnier), eric.gaussier@imag.fr (E. Gaussier), francois@karos.fr (F. Kawala).

https://doi.org/10.1016/j.osnem.2018.01.001
2468-6964/

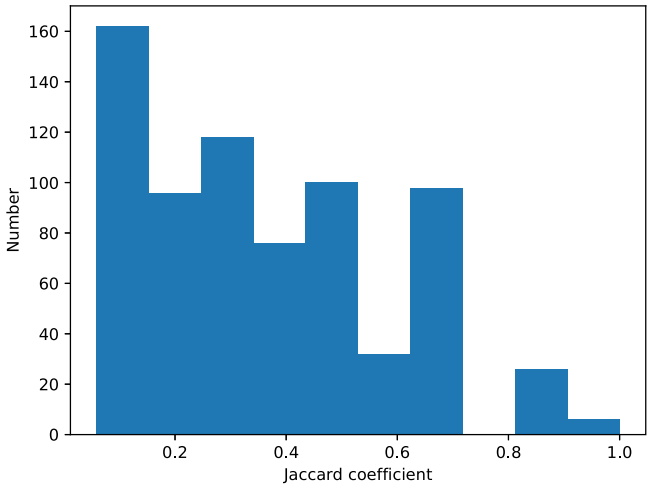
[1] A cascade corresponds to the production of a content by one user of the network, as well as the ensuing re-diffusions of this content by other users. It is thus characterized by the set of users involved in the diffusion of a content. A cascade of size n involves the initial diffuser and $n-1$ re-diffusers.



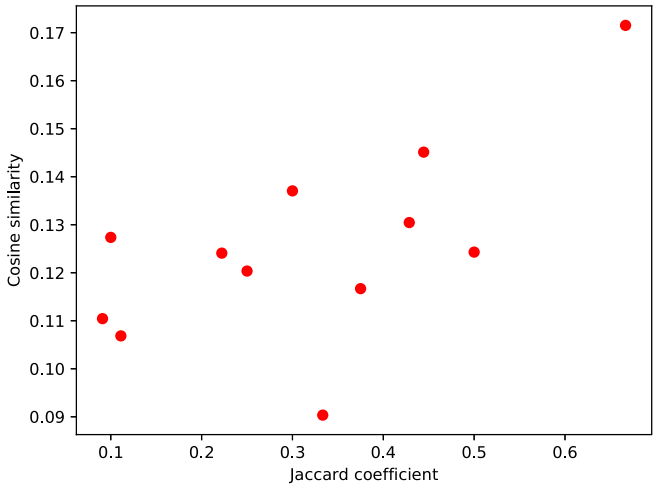
(a) Jaccard coefficient on cascade pairs ($|C| \geq 2$)



(b) Content vs. Diffusion ($|C| \geq 2$)



(c) Jaccard coefficient on cascade pairs ($|C| \geq 5$)



(d) Content vs. Diffusion ($|C| \geq 5$)

Fig. 1. Importance of content in diffusion: histograms of the different values (discretized with bins of size 0.1) taken by the Jaccard coefficient on pairs of cascades issued from the same user and involving re-diffusion (left), and relation between content similarity (computed with the cosine) and the Jaccard coefficient on cascade pairs.

issued by the same users in the ICWSM blog dataset[2] The comparison is based on the Jaccard coefficient, that measures here the proportion of users involved in two different cascades. Let A and B be two sets, the Jaccard coefficient between those two sets is defined by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|X|$ denotes the cardinal of set X. Thus, for any pair of diffusion cascades issued from the same user, one obtains a score between 0 and 1 indicating whether the users involved in the cascade are the same or not. If the score is close to 0, then the users involved in the two cascades are different; if the score is 1, then the same users are involved in the two cascades. The scores obtained can then be binned into fixed-size intervals so as to better visualize their behavior.

Fig. 1(a) displays such an histogram for the ICWSM blog dataset using only cascades of size 2 or more (*i.e.* involving the initial diffuser and at least one re-diffuser) , where the bins considered are of size 0.1, starting from 0 to 1. As one can note, this coefficient takes on very different values. If many cascade pairs get a score close to 1 (last bin on Fig. 1(a)), the majority of cascade pairs (74.5%) get a score strictly lower than 0.8. This shows that cascade pairs do not diffuse in the same way. This finding is similar to the one reported in [4] for the diffusion of hashtags related to different topics in Twitter. It has to be noted that, on average, the longer the cascades are, the smaller the Jaccard coefficient between two cascades is. This is illustrated in Fig. 1(c) that corresponds to the same experiment as before but using only cascades of size 5 or more. One can see that there are even less cascades with a high Jaccard coefficient here.

Having established that cascade pairs do not diffuse in the same way, we now consider the question whether or not the diffusion process is influenced by the content being diffused. Purely probabilistic models, as IC or LT, can generate cascade pairs that behave as in Fig. 1(a) and (c). They are however agnostic to content and would fail to accurately model cascades if they depend on the content being diffused. In order to study the relation between

[2] This dataset is fully described in Section 5.