Estimating Domain-based User Influence in Social Networks

Mario Cataldi, Mittal Nupur, Marie-Aude Aufaure

Estimating Domain-based User Influence in Social Networks

Mario Cataldi
École Centrale Paris
Paris, France
mario.cataldi@ecp.fr

Nupur Mittal
École Centrale Paris
Paris, France
nupur.mittal@ecp.fr

Marie-Aude Aufaure
École Centrale Paris
Paris, France
marie-aude.aufaure@ecp.fr

ABSTRACT

Social networks and microblogging systems play a fundamental role in the diffusion of information. The information, from different sources, reaches each user through multiple connections, the study of which is indispensable for the sake of understanding the dynamics of its evolution and expansion. In this paper, we propose a system which enables to delve in the spread of information over a network along with the changes in the user relationships with respect to the domain of discussion. To cope up with the goal, considering Twitter as a case study, we analyse the tweets as the starting point or as the generators of the information which later flows through subsequent retweets. Furthermore, we integrate a N-Gram model classification approach for categorizing, under various domains, the information shared within the social network under consideration. We finally leverage this formalization to propose a domain-based model which aims to estimate the influence of a user, on a community, in the domain under consideration. In conclusion, using a sample of the Twitter network we then present a set of case studies and real case scenarios that show the validity of the proposed approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Algorithms

Keywords

Social Influence, Twitter, Classification, Authoritativeness

1. INTRODUCTION

With the advent of social networks and microblogging sites, there has been a substantial increase in the amount of real-time information streaming. Within these environments, users are the generators of information spanning a wide range of domains like daily chats, politics, sports, celebrity gossips, etc.

Thus, considering the importance of this information diffusion, there is an emerging need to study and propose novel models for analysing this massive information exchange. With this intent, a fundamental challenge is to trace the spread of the information (also termed as contamination in literature) in the network along with the cause of its spread. This requires to detect the start of the contagion and modeling the paths of its propagation in the network in order to understand the cause of its popularity. For this, the contagion can be formalized as a flow of information, from one node to many other, that follows some rule explicited by some (structural or virtual) edges existing among the nodes. Having a complete view over the information flow over the network requires to study the nature of the exchanges of information among the users and to understand the influence of each one on the entire community.

Following these considerations, the goal of the paper is to provide an unsupervised technique for estimating influence of users of social network, on a community, in a particular domain, where the influence is intended as the capacity of a user to make the posting activity of the others similar to his/her own one[1]. The high-level assumption is that the structural relationships expressed by the explicit connections among the users are not useful for this goal because they only state the existence of some social relationship between them and they do not provide any detail about the real information exchange among them.

Following this intuition, instead of only taking into account the structural connections among users (for example, the following/follower relationship in Twitter), we aim at discovering the nature of their relationship based on their exchange of information about a particular domain. Finally, by using this domain-based information exchange model, we aim at estimating the influence of each user on the community, in the considered domain.

With this goal, we considered as case study the Twitter environment and we start off with a statistical analysis of the corpus with the objective of categorizing all the considered tweets with respect to a pre-determined set of domains of discussion. Subsequently, for each domain, we consider only the tweets (and retweets) positively associated to it, and formalize what we defined as *domain information ex-*

SAC'13 March 18-22, 2013, Coimbra, Portugal.
Copyright 2013 ACM 978-1-4503-1656-9/13/03 ...$10.00.

[1] The posting activity is defined as the set of text documents posted by the user.

change graph; then, within this graph, we study how each information spread over this network and estimate, with a unique value, the influence of each user on the community about the considered domain. At the end of the process, each user will have multiples influence values based on the number of topics of discussion he/she tackled along his/her activity life.

The remainder of the paper is organized as follows: Section 2 surveys related research, focusing on the works that aimed at studying the information spread in social network environments. Subsequently, after describing our motivations and assumptions, Section 3 presents our novel technique for estimating influence of users in social network environments, which tries to understand how the social connections change when the topic of discussion drifts from one argument to another. Empirical evidences collected through experiments are finally presented in Section 4. Section 5 draws the work to final remarks and conclusions, discussing limitations and future research directions.

2. RELATED WORK

In the last decade, with the explosion of the available data in social network environments, the problem of the estimation and the analysis of information diffusion has received considerable attention. Among all, many works tried to study the information propagation [1, 11], the structure and the evolution of the information [12], the dynamics of massive viral communication [14] and the link prediction in social network environments [16].

In fact, with the impressive number of user-generated content, it is fundamental to understand the dissemination, diffusion and spreading of information; processes like the topic evolution [4, 7], news and opinions spread [10], sampling methods for hidden populations and/or communities, have been demonstrated as strictly correlated with this problem.

In early work, before the explosion of social networks, many scientists tried to infer a transmission network between bloggers through various cascade model [11]. Within the same environment, [1] proposed a hybrid approach which uses diffusion trees to evaluate dependencies among bloggers.

With the advent of social networks, many authors slightly moved their attention to them in order to study information diffusion in these very dynamic environments; among all, the authors in [21] took into account user connections in Facebook, [3] analysed the friendship relationships in Second Life, [2] considered the messages in the communication network of Yahoo!, while [6, 19] took into account DBLP for evaluating collaborations among scientific authors.

Many works analysed also micro-blogging services for studying the problem of information diffusion (also due to the fact the users of this network seem more focused on information rather than other social aspect); [13] compared different measures of influence based on parameters like number of retweets, number of followers, etc., while [5] compared different influence estimation strategies discovering that the most followed users did not necessarily score highest on the other measures. The authors in [22] also revealed that the presence of reciprocity can be explained by phenomenon of homophily. Based on this finding, they proposed a system called TwitterRank to measure the influence of users by considering both the topical similarity between users and the link structure (and, again, revealing that the final ranking depends on the influence measure).

The structure of the users network has been also considered for estimating the dependency and/or influence among the users through probabilistic models [8]. Network structure learning has also been used for anonymising the networks in which massive/viral diffusions are happening and for estimating probabilistic relational models [9].

3. DOMAIN-BASED INFLUENCE ESTIMATION IN SOCIAL NETWORKS

In this section we present our three-steps approach for estimating domain-based influence in social network environments. We first present our method for categorizing each content shared by the community. We then leverage this categorization to model a *domain information exchange graph* that permits to formalize the relationships among users when sharing information about a particular domain. We finally make use of these graphs to analyse how the information is spread on the network and estimate the influence of each user on the community for each considered domain. The next sections will describe in detail these steps.

3.1 Content Pre-Processing and Classification

Considering Twitter as case study, the process starts with the extraction and categorization (also called classification) of the tweets from the stream of user-generated content. As in many alternative systems, we perform a preliminary phase of stop word elimination. Then, we formalize the content expressed by the corpus T of text tweets and associate to each considered tweet $tw_j \in T$ a representative *keyword vector*, $\vec{kw_j}$, that formalizes the information extracted from it. For this, instead of representing a tweet tw_j as a classic weighted keyword vector, we aim at calculating all the possible n-grams (portions of n characters) of each term in tw_j and preserving this information within its vector $\vec{kw_j}$. In fact, considering the brevity of the Twitter messages (up to 140 characters), we aim at building a system tolerant of small textual errors, abbreviations and minimal terms variations. For this, n-gram based representation systems has been proposed in text retrieval field and in a wide range of natural language processing applications because of their capacity to effectively treat text documents with errors and abbreviations [20]. In fact, since every keyword is decomposed into small portions, the errors/variations can be easily detected because they tend to affect only a limited number of the n-grams; thus, matching systems that leverage n-grams result resistant to a wide variety of small textual errors (as typos and/or abbreviations, which is often the case of micro-blogging systems).

In detail, each keyword of length k of the considered tweet tw_j is divided in bi-grams (portions of 2 consecutive letters within the string), tri-grams (3 consecutive letters), quad-grams (4 consecutive letters), etc. Therefore, we calculate the weight $w_{x,j}$ of the x-th n-gram in j-th tweet by using the augmented normalized term frequency [18]:

$$w_{x,j} = 0.5 + 0.5 \cdot \frac{tf_{x,j}}{tf_j^{max}} \tag{1}$$

where $tf_{x,j}$ is the frequency value of the $x - th$ n-gram in the $j - th$ tweet and tf_j^{max} returns the highest n-gram frequency value of the $j - th$ tweet. In fact, as in the standard `TF` formula, we try to give credit to any term (n-gram in this case) that appears in the corpus, but also adding some additional credit to terms that appear less frequently. In fact, the most relevant terms related to a domain are often specific keywords that do not appear so frequently in a big corpus of documents: the augmented normalized term frequency tries to preserve this information.

Thus, for each tweet tw_j, a tweet vector

$$k\vec{w}_j = \{w_{1,j}, w_{2,j}, ..., w_{v,j}\} \quad (2)$$

is defined, where K is the n-grams vocabulary of the corpus and $v = |K|$ is its size.

At this point, we perform the categorization step by leveraging the tweet vectors previously calculated. In detail, given a set of documents (tweets in our case) formalized as weighted vectors of n-grams, the classification system tries to compare them with respect to a set of template domain vectors in order to detect similarities among them.

The high-level idea of this approach is that every text document related to some specific domain (also called topic in the paper) invariably has a set of terms which tend to occur more frequently than others. This assumption has been formalized through the Zipf's Law [23], which simply proved that the n-th most common term in a text document occurs with a frequency inversely proportional to n. The implication of this law is that there is always a set of terms which occur more than the others in any considered topic of discussion. Thus, in other words, it implies that two text documents related to the same category should have similar terms and n-gram frequency distributions.

Following this intuition, for each considered domain, we generate what we call *n-gram distribution profile* by taking into account a training set of text documents, which act as a significant set of representative text entries about the considered domain. Then, for each set, we generate all the possible n-grams and calculate their overall frequency. The resulting n-gram distribution profile is therefore a n-gram frequency profile of the considered domain[2].

At this step, the classification system works as follows: we compute the cosine similarity measure between the tweet vectors and all the profile vectors of the considered set of categories. Then, for each tweet, the system selects the categories whose profiles have the smallest distance w.r.t. the tweet vector. Notice that, considering that each tweet could refer to multiple domains, we also aim at identifying this multiplicity by identifying when a tweet results strongly related to multiple domains. In order to cope with this, we provide a completely automatic model that works as follows:

1. for each tweet, the systems first ranks the categories in descending order of similarity value;

2. therefore, it computes the *average drop* (between consecutive entities) for all those categories that are ranked before a pre-defined threshold value;

3. the first drop which is higher than the computed average drop is called the *critical drop*.

At this point the tweet is categorized by the domains whose similarity value is above the adaptively computed critical point. With this step, we aim at associating to the documents only the categories that strongly emerged with respect to the others. In this way, we avoid to associate those categories which result only marginally related to the document (i.e., the similarity value does not significantly differ from the other categories). Please also notice that the document can also remain uncategorized if there is no category whose similarity is above the given threshold.

3.2 Studying the Relationships among Users based on the Topics They Share

A fundamental issue in the study of the spread of information in user-generated environments is the set of social relationships that exist among the users. Figuring out a level of importance of the source of a specific information contamination represents a key point towards a more precise information spread estimation model.

In Twitter, a user can follow the text stream of other users by making explicit his/her social relationship of *follower*. On the other hand, a user who is being followed by another user does not necessarily have to reciprocate the relationship by following him/her back, which makes the graph of the network directed. Moreover, a user can share the information another user posted by using the function of *retweet*.

In this paper, as already stated in the Introduction, we believe that the following/follower relationship can represent a wide-range of social relationships existing among the users and cannot help estimate the degree of information shared (and therefore the influence) among them. For this, we aim at analysing the information exchange among the users based on the real quantity of information shared within the network and not the pure static statement of their structural relationship. In fact, an Obama's follower, for example, is not obliged to read or share the content expressed by him, and therefore this follower relationship only state the original will of the user to create a connection with him (without any implication about sharing the same topics or being influenced by his content).

For this, given a considered domain, in order to measure the influence of each user on a community about a domain, we take into account the retweet graph only formed by those retweets that have been categorized by that domain (see also Section3.1). This social model enables us to define a *domain information exchange graph*, $G_d(U_d, E_d)$, where d is a given domain, U_d is the set of users that resulted have posted at least one tweet associated to d, and E_d is the set of directed edges. In this model, given two users u_i and u_j, the edge $< u_i, u_j >$ exists only if u_i has retweeted at least a tweet of u_j categorized by the domain d.

Thus, we measure a first high-level authoritativeness value of each user about a given domain d by analysing the connectivity in G_d. The idea of this step is to calculate a first value for estimating the authority of the user, in a community, on some domain. This value will permit to understand the overall visibility of the user about a domain and allows the system estimating a more precise influence value (explained in Section 3.3).

In particular, since users retweet the content that suppose to be interesting for them (i.e., expliciting some interest in the original tweet), we can assume that a user with a high number of retweets (incoming edges) represents an authori-

[2] An interesting observation is possible: the highest ranked n-grams are mostly uni-grams and simply reflect the distribution of the letters of the alphabet in the language of the document. In other words, the most frequent n-grams are most of the time correlated to the language. Thus, the most frequent n-grams for the considered domain profiles resulted to be very similar, while they start differing consistently in the lowest part of ranked n-grams list.

tative information source, into this social community, in the considered information domain. Moreover, the concept of authority can be also extended by taking into account the fact that the importance of a user is also related to the degree of importance of the persons who are retweeting him; considering for example the case of "Barack Obama" (millions of retweets along his activity), each user retweeted by him assumes more importance based on this authoritative relationship.

Based on this scenario, it is possible to model this problem as a topological-based computation of web pages authority in large hyper textual systems. In other words, for this task, it is possible to rely on a similar strategy to the well known PageRank algorithm [17] that aims at calculating the authority of each page by analysing the topological graph of the considered web entities. Following the same idea, the high-level authority of a user about a topic depends on the number and the authority of the user that retweeted his/her tweets on the same domain. Hence, given a user $u_i \in U_d$, its *authority* is computed as follow:

$$auth(u_i) = f \times \sum_{u_j \in retweeter(u_i)} \frac{auth(u_j)}{|retweeting(u_j)|} + (1-f) \tag{3}$$

where $retweeter(u_i)$ is a function that returns the set of users retweeting u_i, $retweeting(u_j)$ is a function that returns the set of user that u_j retweeted, and $f \in (0,1)$ is a dumping factor representing the probability that a "random surfer" on the graph moves from one node to another. Authority values are therefore calculated using an iterative algorithm, where, at the initial instant, each value is initialized as:

$$auth^0(u_i) = \frac{1}{|U_d|} \tag{4}$$

At each step, the algorithm recomputes the authority values as:

$$auth^t(u_i) = f \times \sum_{u_j \in retweeter(u_i)} \frac{auth^{t-1}(u_j)}{|retweeting(u_j)|} + (1-f) \tag{5}$$

The process ends when a convergence condition is satisfied. In detail, let $\mathbf{A}$ the column vector that contains all the authority values $auth(u_i)$, at $t=0$ the initial values are assumed as

$$A_i^0 = auth^0(u_i) = \frac{1}{|U_d|} \tag{6}$$

At each step, the algorithm recomputes the column vector $\mathbf{A}$ as:

$$\mathbf{A}^{t+1} = f\mathcal{M}\mathbf{A}^t + \frac{1-f}{|U_d|}\mathbf{1} \tag{7}$$

where $A^t = auth^t(u_i)$, $\mathbf{1}$ is the column vector of length U containing only 1, and $\mathcal{M}$ is the modified adjacency matrix:

$$\mathcal{M}_{ij} = \begin{cases} \frac{1}{|retweeting(u_j)|} & \text{if } j \text{ retweeted } i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Given a threshold value ε, the computation ends when the convergence condition is satisfied:

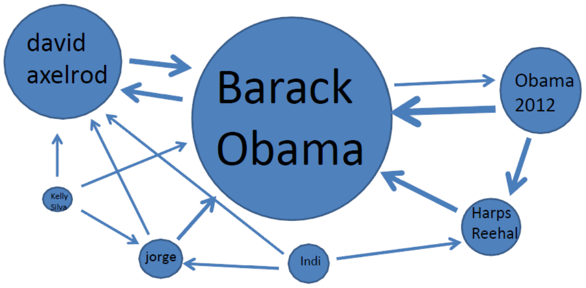

Figure 1: The authority value computation on the political domain: the size of the nodes represents their importance in the considered community.

$$\left|\mathbf{A}^{t+1} - \mathbf{A}^t\right| < \varepsilon. \tag{9}$$

In Figure 1 an example of user authority computation on the input graph, obtained by performing a graph sampling process [15] in which the "Barack Obama" vertex represents the starting point, is depicted. User authority values are visually represented by the size of the circles. In this case, "Barack Obama" is the most authoritative user. His authority is also propagated to the user "davidaxelrod" because of the strong retweet relation with "Barack Obama", even if he has a significantly lower number of retweets.

3.3 Domain-based User Influence Estimation

At this step, given a domain d, a set of tweets categorized by d and a domain information exchange graph $G_d(U_d, E_d)$ of users retweeting about the considered topic, it is possible to estimate the influence degree of each author $u_i \in U_d$ on the community expressed by U_d about the domain d, by analysing the user posting activities and their impacts within the community that resulted interested in the same domain.

In other words, for each user $u_i \in U$, we aim at analysing all the tweets $T_{i,d}$ posted by u_i and categorized by d in order to understand how much his/her content spread over the community interested in the same domain. For this, we take into account the domain information exchange graph with the goal of studying the paths each content followed over the network (i.e., who retweeted the information shared by the user) and how much authoritative were the users that re-shared the information.

More formally, given a set of users U_d interested in the domain d (i.e., they tweeted at least one tweet classified as related to d) and a set E_d of retweet relationships among them, we can estimate the influence of each user $u_i \in U_d$ on the community interested in d by analysing the graph $G_d(U_d, E_d)$ with the goal of calculating a weighted average authoritativeness of the persons that shared the information posted by u_i. For this, the influence degree $inf_{u_i,d}$ of the user u_i on the domain d can be calculated as

$$inf_{u_i,d} = \frac{\left[\sum_{tw_j \in T_{i,d}} \left(\frac{\sum_{u_x \in Ret(tw_j)} \frac{auth(u_x)}{\left|dist(u_x, start(tw_j))\right|}}{\left|Ret(tw_j)\right|} \right) \right]}{|T_{i,d}|} \tag{10}$$

where

- the function $Ret(tw_j)$ returns the set of users in U_d that retweeted the tweet tw_j (also those that retweeted tw_j through a user u_k, where $k \neq i$);

User Name	Influence	# Followers
will.i.am	0.0271	4372058
Samuel L. Jackson	0.0262	1526852
deadmau5	0.0224	1543616
Jim Gaffigan	0.0177	1292586
Aiden Grimshaw	0.0155	466957

(a) Entertainment

User Name	Influence	# Followers
Guardian news	0.0255	462584
D Wasserman Schultz	0.0198	121912
Reince Priebus	0.0191	121121
Mitt Romney	0.0094	1186236
Newt Gingrich	0.0073	1471146

(b) Politics

User Name	Influence	# Followers
Bill Simmons	0.0275	1848526
Usain St. Leo Bolt	0.0173	2178512
Boston Celtics	0.0078	762059
FOX Sports: MLB	0.0045	128283
SkySports	0.0012	855685

(c) Sports

Table 1: The most influential users, retrieved by the system, in three different topics: a) entertainment, b) politics, c) sports.

- the function $start(tw_j)$ returns the user who first posted the tweet tw_j (even before the user u_i, i.e. in this case tw_j was already a retweet);
- the function $dist(u_i, u_j)$ returns the distance (i.e., number of edges that separate u_i from u_j in $G_d(U_d, E_d)$.

Please notice that, with the function $dist(u_x, start(tw_j))$, we aim at taking into account the distance, in $G_d(U_d, E_d)$, between the person that originally posted tw_j ($start(tw_j)$) and another user u_x that shared the same information in a subsequent time frame; in fact, the higher the distance from the original information source, the less important the role of the considered person in the spread of the information. In fact, if for example CNN posts a tweet, which is retweeted, indirectly, by u_2 (i.e., u_2 retweeted the information through another user u_1, who originally retweeted the tweet of CNN), we believe that the role of u_1 cannot be considered of the same importance of the original creator of the information. In a sense, u_1 only acted as an information booster and the importance of the influence of u_1 on the community cannot be considered equal to the role of the information producer. Thus, the function $dist$ permits to take into account this parameter and weight accordingly the resulting influence.

At the end of this step, we have estimated a degree of influence for each user about each domain he expressed interested through his/her posting activity. In the next section we will illustrate real case scenario and experiments for analysing the effectiveness of the proposed influence estimation approach.

4. EVALUATION

In this section, we evaluate the results of several experiments we conducted by monitoring the Twitter community during the period included between August 20th to September 4th 2012. For this work, we analysed a connected community of $\sim$ 45k Twitter users (extracted by performing the graph sampling process [15] starting from the user "CNN"), which included more than 150k tweets with $\sim$ 350k different keywords. We then performed the classification step (Section 3.1) on three different domains (sports, entertainment and politics) by previously training our n-grams classification model using Wikipedia articles categorized accordingly.

User Name	Influence	# Followers
Victoria's Secret	0.0555	1220859
Dara O Briain	0.0375	1083658
Heidi Klum	0.0217	1019451
Glamour	0.0111	260694
Alexandra Burke	0.0045	1083658

Table 2: The most influential users, retrieved by the system, by using the following/follower relationship.

The main aim of the experimental evaluation was twofold: from one side analyse, through examples and case studies, the impact of the parameters proposed within the presented technique. On the other hand, evaluating, through real case scenarios, the effectiveness of the proposed approach in estimating domain-based influence of Twitter users.

In Tables 1(a), (b) and (c) we show the most influential Twitter users, retrieved by the system, on the three considered domains. The most evident result is that no users are shared among the lists; in other words, even in presence of widely known users (as for "Usain St. Leo Bolt", Jamaican sprinter) who tweeted about many different topics, the system was able to discriminate the users' influence and effectively understand their area of influence. In fact, when for example the twitter account of the "Guardian news" (first most influential user in the political domain) reported some information not related to the political domain, it did not result as much influential (i.e., there is not the same spread of information in terms of retweets) as when it provided fresh political news.

It is also important to notice that the pure number of followers is not directly correlated with the influence value. In fact, for example, "Mitt Romney" (nominee of the Republican Party for President of the United States in the 2012 election, and followed in Twitter by more than 1 million users) resulted less influential, within the considered community, than "Reince Priebus" (Chairman of the Republican National Committee) who has a significantly lower number of followers (Table 1(b)). The same consideration is valid for the other considered domains.

Moreover, in order to better test the assumption that guided our work, we also calculated the influence value by taking into account only the structural connections among users. In other words, we estimated the influence of the users by only considering the follower/following relationship (i.e., we construct the domain information exchange graph with the follower/following information). As it is possible to notice from Table 2 and as originally supposed, in this case, the influence is mainly reflecting the pure visibility of the users (i.e., number of followers) and it is not able to detect the capacity of the user in influencing the posting activity of the other users. Moreover, the list of most influential users, resulted mixing similar domains (fashion, music, entertainment). This result has a simple explication: each of these users has a high number of followers and they also result highly interconnected. Therefore, the resulting influence values, which is also based on PageRank-like approach, is determined by this high connectivity (and therefore is not providing any evidence about the capacity of influencing the posting activity of the others).

Finally, in order to test our system when estimating the influence of the users based on some topic of discussion, we calculated the influence values of the most influential users

User	Inf (enter.)	Inf(pol.)	Inf(sport.)
will.i.am	**0.0271**	1,5e-5	2,3e-6
Samuel L. Jackson	**0.0262**	4,2e-5	3,1e-4
deadmau5	**0.0224**	3,1e-7	5,5e-4
Jim Gaffigan	**0.0177**	8,2e-8	3,1e-6
Aiden Grimshaw	**0.0155**	5,1e-6	2,7e-5
Guardian news	6,2e-3	**0.0255**	1,4e-4
D Wasserman Schultz	8,5e-15	**0.0198**	9,9e-8
Reince Priebus	7,3e-14	**0.0191**	5,7e-13
Mitt Romney	5,2e-5	**0.0094**	5,9e-7
Newt Gingrich	4,4e-7	**0.0073**	2,8e-8
Bill Simmons	4,5e-5	2,3e-12	**0.0275**
Usain St. Leo Bolt	1,4e-6	6,4e-8	**0.0173**
Boston Celtics	1,6e-5	7,2e-9	**0.0078**
FOX Sports: MLB	9,6e-4	9,1e-5	**0.0045**
SkySports	7,5e-4	5,6e-4	**0.0012**

Table 3: The most influential users, retrieved by the system, on a) entertainment, b) politics, c) sports.

(reported in Table 1) also for the other considered topics. The results are shown in Table 3 and show the capacity of the system in effectively detecting the area of influence of the users. In fact, all the considered users resulted poorly influential outside their domain, proving the capacity of the system in estimating the domain-based influence of the users in a community.

In fact, with the proposed approach, it is possible to overtake the structural information to retrieve the degree of influence of each user within a particular domain which, as demonstrated, resulted not dependent on the pure structural relationships existing within the network (i.e., follower/following relationship).

5. CONCLUSIONS

In this paper we have presented a novel approach for automatically estimating the influence of a user on a specific domain of discussions. We introduced the high-level assumptions that guided our work and presented a novel method that first aims at categorizing the considered content and then makes use of this classification for analysing the different relationships among the users with respect to their domains of discussion. Within the proposed approach, in fact, the relationships existing within a social network environment (especially for those oriented to information) are re-interpreted with the goal of understanding how much the relationships among they are correlated with the discussed information domain. For this, we provided case studies and real case scenarios that showed how the influence values can significantly differ when we consider different domains and also highlight the effectiveness of the proposed approach when employed for detecting influential users.

6. REFERENCES

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 207–214. IEEE Computer Society, 2005.

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, Dec. 2009.

[3] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pages 325–334. ACM, 2009.

[4] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *MDMKDD '10*, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM*, 2010.

[6] L. Di Caro, M. Cataldi, and C. Schifanella. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*, pages 1–25, 2012.

[7] A. Favenza, M. Cataldi, M. L. Sapino, and A. Messina. Topic development based refinement of audio-segmented television news. In *NLDB '08*, pages 226–232, Berlin, Heidelberg, 2008. Springer-Verlag.

[8] N. Friedman. Being bayesian about network structure. In *Machine Learning*, pages 201–210, 2000.

[9] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *J. of Machine Learning Research*, 3:679–707, 2002.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501. ACM, 2004.

[11] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explor. Newsl.*, 6(2):43–52, Dec. 2004.

[12] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, Dec. 2004.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW '10*, pages 591–600. ACM, 2010.

[14] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[15] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06*, pages 631–636. ACM, 2006.

[16] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, pages 556–559. ACM, 2003.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *WWW'98*, pages 161–172, 1998.

[18] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[19] C. Schifanella, L. D. Caro, M. Cataldi, and M.-A. Aufaure. The d-index: a web environment for analyzing dependences among scientific collaborators. In *KDD*, pages 1520–1523. ACM, 2012.

[20] C. Y. Suen. n-gram statistics for natural language understanding and text processing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):164–172, Feb. 1979.

[21] E. Sun, I. Rosenn, C. Marlow, and T. M. Lento. Gesundheit! modeling contagion through facebook news feed. In *ICWSM*. The AAAI Press, 2009.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10*, pages 261–270. ACM, 2010.

[23] G. Zipf. *Human Behaviour and the Principle of Least-Effort.* Addison-Wesley, Cambridge, MA, 1949.