Undefined 1 (2011) 1–5
IOS Press

Social Influence Analysis in Microblogging Platforms – A Topic-Sensitive based Approach

Editor(s): Aba-Sah Dadzie, The University of Sheffield, UK; Milan Stankovic, Université Paris-Sorbonne, France; Matthew Rowe, Lancaster University, UK
Solicited review(s): Claudia Wagner, Joanneum Research, Austria; Guillaume Erétéo, INRIA, France; Sofia Angeletou, BBC, UK

Amparo E. Cano [a,*], Suvodeep Mazumdar [a] and Fabio Ciravegna [a]
[a] *Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom*
E-mail: {firstinitial.surname}@dcs.shef.ac.uk

Abstract.
The use of Social Media, particularly microblogging platforms such as Twitter, has proven to be an effective channel for promoting ideas to online audiences. In a world where information can bias public opinion it is essential to analyse the propagation and influence of information in large-scale networks. Recent research studying social media data to rank users by topical relevance have largely focused on the "retweet", "following" and "mention" relations. In this paper we propose the use of semantic profiles for deriving influential users based on the retweet subgraph of the Twitter graph. We introduce a variation of the PageRank algorithm for analysing users' topical and entity influence based on the topical/entity relevance of a retweet relation. Experimental results show that our approach outperforms related algorithms including HITS, InDegree and Topic-Sensitive PageRank. We also introduce VisInfluence, a visualisation platform for presenting top influential users based on a topical query need.

Keywords: social awareness streams, microblogging, social influence, semantic profiles

1. Introduction

The rise of ubiquitously available social media services has contributed to a change in how people engage with their social context and environment, and how they consume, produce and share information. The Social Web has become a space, which exhibits real-time social perceptions and interests on current events and topics. These social space consists of a set of users who leave trails in the form of discussions and social interaction patterns forming a historical dataset of social activities. Particularly microblogging platforms, such as Twitter and Facebook, have emerged as a new form of communication characterised by high social connectivity and the ability to communicate trends. These real-time-message-routing platforms gather a collection of semi-public, natural-language messages (a.k.a Social Activity Streams [23]), which are further shared via different channels including the Web, email and text messaging services[21], extending in this way their online audience.

In this work we investigate the derivation of users influence in the Twitter Graph. By influence we refer to the capacity of users to produce effects on the actions of other users. These actions include for example: i) to retweet a comment from an other user; ii) to follow other users; or iii) to tweet about a particular topic. The unprecedented, rapid rate of information propagation in Twitter streams along with its high topical diversity present new challenges for its analysis. This

*Corresponding author.

paper investigates the challenge of identifying which users become influential in a particular topic, and in a particular entity – by entity we refer to an instance of a particular type (e.g Location, People, Product)–. For example a user can be influential in the topic: "Sports News", but not reach audiences interested in the tennis player Roger Federer (which is an instance of an entity of type People).

We present a graphical model with multiple semantic edges for capturing the nature of a tweet and the way users engage with it. More specifically we explore the relationship between users, topics and entities in terms of lightweight ontologies. We introduce the Topic-Entity PageRank, as an extension of the Topic-Sensitive PageRank algorithm [11], for measuring the influence of users in Twitter. Topic-Entity PageRank measures users influence taking into account the topical and entity relevance of a retweet link.

This work improves the state-of-the-art by making the following contributions: 1) The Topic-Entity PageRank, which is an approach that leverage users' semantic profiles for deriving topical and entity ranked influence on the Twitter graph; 2) Metrics for deriving the topical and entity relevance of a retweet relationship; 3) A dynamic interface for visualising the rise and vanish of influential users by topic and entity on time.

The paper is structured as follows: section 2, present an overview of our approach. Section 3, introduces the formal model in which we analyse the Twitter graph. It also presents our approach for the semantic enrichment of tweets' content and the derivation of triples for generating semantic profiles. Section 4, presents an analysis of the type of nodes and edges contained on a Twitter graph and provides a comparison between the Web graph and the Twitter graph. Section 5, presents our approach for deriving topical and entity-based influential users on the Twitter graph. Section 6 presents the evaluation of our approach. Section 7 describes a dynamic interface for visualising influential users based on topical and entity relevance. Section 8 presents the related work including relevant work from the Semantic Web community using tweets and similar work within the field of social influence analysis. Section 9 finishes the paper with conclusion and our plans for future work.

2. Approach Overview

In this work we discover top rank influential users on the Twitter graph, given a topic or an entity. Our approach takes advantage of semantic trails (i.e semantic relationships) left as side effect of tweeting. These semantic trails include the generation of relationships such as: 1) the social relationship between a user retweeting a post and the author of the post; 2) the relationship between a user and the topic of the post he retweeted; and 3) the relationship between a user and the entities (e.g. person, products) mentioned on the content of his posts or retweets.

To extract these semantic trails for discovering influential users we utilise the approach presented in Figure 1. *First* we collect a Twitter dataset, by using the Twitter API. *Second*, in order to leverage the semantics of the Twitter relationships, we enrich the content using entity extraction and topical detection services such as Zemanta[1] and OpenCalais[2]. *Third*, in order to generate users' semantic profiles, we translated the enriched content and Twitter posts' metadata into triples using the SIOC [3], OPO [4], and AO [5] ontologies. *Forth*, we leverage these semantic profiles for calculating topical and entity relevance of retweet relationships. These metrics are used in our proposed Topic-Entity PageRank. Finally, we introduce VisInfluence, an interface for visualising topical influential users derived from the Twitter graph.

3. Modelling Twitter

In this section we present the elements of a tweet data structure. Based on these elements we describe the graph model that will be used for representing the Twitter graph. Following this model we will study the two-mode graph networks presented in section 4.2.

3.1. A Tweet Close-up

The microblogging platform Twitter allows users to publish text limited to a maximum of 140 characters. On Twitter a user has two main roles, to publish tweets (*writer*) or to subscribe to other users and read their posts (*reader*).

As a *writer* you : *1*) are followed by other users; *2*) republish (or *retweet*) other users' posts; *3*) make

[1] Zemanta, http://www.zemanta.com/

[2] OpenCalais, http://www.opencalais.com/

[3] SIOC Types, http://rdfs.org/sioc/spec/#sec-modules-types

[4] OPO, http//www.oline-presence.net

[5] Annotation Ontology, http://code.google.com/p/annotation- ontology/wiki/v2Main

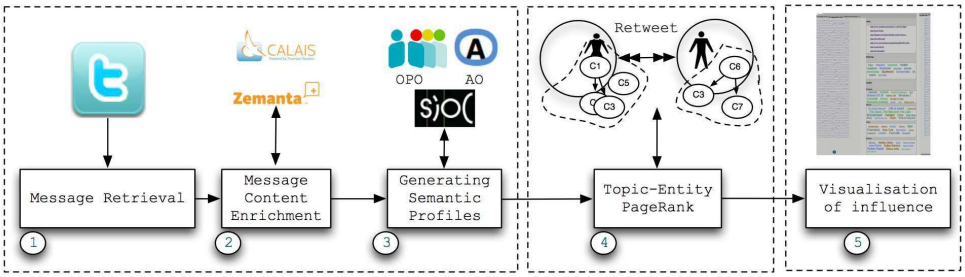


Fig. 1. Overview of the approach for 1) generating semantic profiles, 2) derving topical and entity-based influential users, and 3) visualising top ranked influential users

reference to other users within the published content (a.k.a *mentions*) by using the '@' character before the user's user name; *4*) *reply* to another tweet, replies always start with '@'username (author of the tweet you are replying to); *5*) include resources to your post (i.e. hashtags and links); and *6*) are *listed* by your followers.

As a *reader* you:*1*) *follow* other users' posts; and *2*) organise into groups (*lists*) the users you follow.

The users who follow your tweets are referred to as your *followers* and the set of users you follow are referred to as your *friends*. Lists help users in organising their *friends* into groups, these groups can be topical like for example "Semantic Web" or categorical e.g. "celebrities". In this way lists act as a user's personal interest taxonomy.

The following subsections describes the formalisation of the Twitter Graph model that will be used in the rest of the paper.

3.2. The Full Twitter Graph Model

Following the Tweetonomy model suggested by Wagner and Strohmaier[29], we describe a Twitter stream as a sequence of tuples S, so that:

$\mathtt{S} := (\mathtt{U}_{\mathtt{q1}}, \mathtt{M}_{q2}, \mathtt{R}_{q3}, \mathtt{T}, ft)$, where

- $\mathtt{U}$,$\mathtt{M}$,$\mathtt{R}$ are finite sets whose elements are called users, messages and resources.
- Each of these sets are qualified by $\mathtt{q1}$,$\mathtt{q2}$, and $\mathtt{q3}$ respectively (explained below).
- $\mathtt{T}$ is the ternary relation $\mathtt{T} \subseteq \mathtt{U} \times \mathtt{M} \times \mathtt{R}$ representing a hypergraph with ternary edges. The hypergraph of a Twitter stream $\mathtt{T}$ is defined as a tripartite graph $\mathtt{H}(\mathtt{T}) = \langle \mathtt{V}, \mathtt{E} \rangle$ where the vertices are $\mathtt{V} = \mathtt{U} \cup \mathtt{M} \cup \mathtt{R}$, and the edges are: $\mathtt{E} = \{\{u, m, r\} \,|\, (u, m, r) \in T\}$. Each edge represents the fact that a given user associates a certain message with a certain resource.
- f_t is a function that assigns a temporal marker to each ternary edge.

In this study we focus on user-centric Twitter streams (the data set is described in section 4.3), using the following qualifiers:

- The way a user can be related to a message is represented by the qualifier $\mathtt{q1}$. For this analysis we consider the authorship relationship, and we differentiate it into two types: $\mathtt{U}_{oa}$ (the original message's author) and the $\mathtt{U}_{rt}$ (author retweeting a message that is not his own).
- The qualifier $\mathtt{q2}$ represents the types of messages. For this analysis we consider two types: $\mathtt{M}_d$ (direct message) and $\mathtt{M}_r$ (re-tweeted message).
- The qualifier $\mathtt{q3}$ for resources considers: $\mathtt{R}_k$ (keywords), $\mathtt{R}_h$ (hashtags), $\mathtt{R}_{li}$ (URLs), $\mathtt{R}_x$ (a typed entity (e.g location, people), and $\mathtt{R}_t$(topic derived from the message's content).

A user stream aggregation is defined as a tuple: $\mathtt{S_a}(\mathtt{U}') = (\mathtt{U}, \mathtt{M}, \mathtt{R}, \mathtt{Y}', ft)$, where

$$\mathtt{Y}' = \{(\mathtt{u}, \mathtt{m}, \mathtt{r}) \mid \mathtt{u} \in \mathtt{U}' \vee \exists \mathtt{u}' \in \mathtt{U}', \tilde{\mathtt{m}} \in \mathtt{M}, \mathtt{r} \in \mathtt{R} : (\mathtt{u}', \tilde{\mathtt{m}}, \mathtt{r}) \in \mathtt{Y}\} \tag{1}$$

and $\mathtt{U}' \subseteq \mathtt{U}$ and $\mathtt{Y}' \subseteq \mathtt{Y}$. $\mathtt{S_a}(\mathtt{U}')$, consists of all messages related with a user $\mathtt{u}' \in \mathtt{U}'$ and all the resources and users related with these messages.

3.3. Semantic Enrichment of Tweets

Given a message from a user stream aggregation $\mathtt{S_a}(\mathtt{U}')$, we perform a lightweight message enrichment by using Zemanta [6], and OpenCalais [7]. These services

[6]Zemanta, http://www.zemanta.com/
[7]OpenCalais, http://www.opencalais.com/

perform entity-extraction on the input message identifying resources which can be qualified as $\mathtt{R}_x$, where x can be for example: organisation-entities ($\mathtt{R}_o$), people-entities ($\mathtt{R}_p$) and, location-entities ($\mathtt{R}_l$). The OpenCalais service also provide a topical categorisation of the message ($\mathtt{R}_t$). Consider the example in Figure 2, where the extracted entities and the topical categorisation for a Twitter message are shown.

The qualified entities allows to build a rich RDF graph representing semantic relationships among the content and authors of content produced in a datastream.

The semantic representation of the message in Figure 2 is shown in Figure 3. Tweets are represent as instances of sioc:MicroblogPost from the SIOC Types ontology [8] (3(a)). User related information is expressed using the Online presence ontology (OPO) [9].

The relationship of a tweet with a typed entity is expressed using sioc:AnnotationSet from the SIOC Types and the Annotation ontology [10]. For linking the post with a typed entity we define instances of sioc:AnnotationSet (e.g. the location annotation set, the people annotation). The annotation sets act as containers of items of a particular type. Figure 3(b), presents an extract of the location annotation set, which contains the item representing the resource Palo_Alto,California, this item annotates the tweet document through the aof:annotatesDocument property (Figure 3(c)).

The semantic profile of a user is the graph resulting from aggregating those tweets related to this user. The advantage of modelling the tweet following the structure we present in RDF in Figure 3, is that partial representation of the user can be extracted. For example a user can be profiled by a topic, which would provide a graphical representation of the those tweets related to this user and to this topic.

This RDF representation facilitates the retrieval of tweets and users' related information, by traversing the posts' surrounding information (e.g. topics, and entities).

In the next section we define the two mode network graphs in which we will analyse the Twitter Graph. We also present a comparison between the type of edges and nodes found on the Twitter Graph and those found on the Web Graph.

[8] SIOC Types, http://rdfs.org/sioc/spec/#sec-modules-types

[9] OPO, http//www.oline-presence.net

[10] Annotation Ontology, http://code.google.com/p/annotation-ontology/wiki/v2Main

4. Analysis of the Twitter Graph

This section highlights differences and analogies between the Twitter Graph and the Web graph[6][17]. This section also defines the inlinks and outlinks distributions that can be derived from the Twitter graph. These distributions will be further used in section 5 where the Topic-Entity PageRank will be presented.

4.1. Twitter Graph and Web Graph Highlights

The Web graph model [6][17] consists of one type of node which is a page, and one type of edge linking a pair of pages, which is a hyperlink. It is represented as a directed graph model using an $n \times n$ matrix W where n represents the number of pages on the Web. Each element of the matrix (W_{ij}) represents the weight in which page i links to page j. A common way of weighting each elements it to consider the outlinks (hyperlinks within the page pointing to external pages). In this case, M_{ij} is equal to $\frac{1}{c_j}$ where c_j is the number of outgoing links going from page i to page j.

Similarly the Twitter graph can be considered as a directed three-mode graph consisting of three main types of nodes, which are users, messages (posts) [11], and resources. Table 1 presents the different types of edges linking one node type to another.

On the the Web graph, a link from page *x* to page *y* indicates that the author of page *x* confers some importance to page *y*. In the case of the Twitter Graph, there are three main types of link endorsements between user *a* and user *b* through a tweet; which are: the *mention*; *following*; and the *retweet* links.

A *mention* link can signify an endorsement of quality from a user *a* to a user *b*. User *a* will mention user *b* if he either expects user *b* to be interested in his post or to promote with his friends (followers) a relation (e.g. topical) with user *b*.

Welch et al [31] differentiate a *following* link and a *retweet* link based on the roles of users *a* and *b*, as *readers* or *writers*. In their work they state that a *following* link signifies that user *a*, in the role of *reader*, is interested in user *b*, in the role of *writer*.

In a similar way, a *retweet* link acts as an endorsment of quality. In this case, a user *a* will retweet a post from user *b* if he has topical interest in the post as a *writer* or if he expects his *readers* to be interested in this post.

[11] In the paper we will use post, message and tweet interchangeably

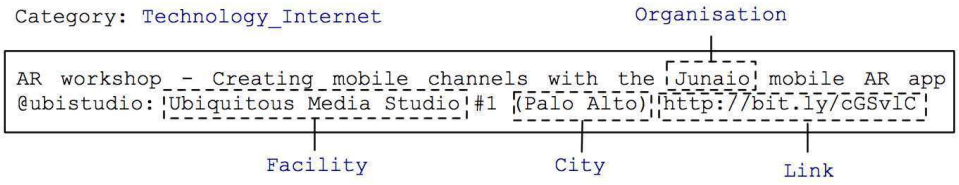


Fig. 2. Message Enriched with Zemanta and OpenCalais services. These services return entity labels as well as topical categorisation of the message

```
a @prefix sioc: <http://rdfs.org/sioc/ns#>
  @prefix sioct: <http://rdfs.org/sioc/types#>
  <http://twitter.com/userA/statuses/29298111037833216> rdf:type sioct:MicroblogPost;
       sioc:content "AR workshop - Creating mobile channels with the Junaio mobile AR
                     app @ubistudio: #1 (Palo Alto) http://bit.ly/cGSvlC";
       sioc:topic <http://my.example.org/topic/TechnologyInternet>;
       ...
       sioc:about <http://my.example.org/resource/Palo_Alto,_California>.
```

```
b @prefix sioct:<http://rdfs.org/sioc/types#>
  @prefix ao:<http://purl.org/ao/core/#>
  <http://my.example.org/set/location> rdf:type ao:AnnotationSet;
       rdf:type sioct:AnnotationSet;
       ao:item <http://my.example.org/resource/Palo_Alto,_California>;
       ...
       ao:item <http://my.example.org/resource/Pine_City,_New_York>.
```

```
c @prefix sioct:<http://rdfs.org/sioc/types#>
  @prefix ao:<http://purl.org/ao/core/#>
  @prefix aof:<>
  <http://my.example.org/resource/Palo_Alto,_California> rdf:type ao:Item;
       rdf:type ao:Item;
       rdf:type sioc:Item;
       rdf:type ao:Annotation;
       ao:hasTopic <http://dbpedia.org/resource/Palo_Alto,_California>;
       aof:annotatesDocument <http://twitter.com/userA/statuses/29298111037833216>;
       .
       aof:annotatesDocument <http://twitter.com/userB/statuses/29298111037833217>.
```

Fig. 3. RDF/Turtle extract of a tweet after the semantic enrichment, user related metadata is not shown in this extract.

Table 1
Twitter Graph Nodes and Egdes

Node	**User**	**Message**	**Resource**
User	Follow, List, Friend	Original Author, Retweet Author	Cites
Message	Mention,	Retweet, Reply	Contain
Resource	Cited by	Contained in	Related

Previous work[31] has highlighted that the propagation of topical influence through *following* links is problematic since traversing a single following link dramatically reduces the probability of topical relevance. In the same work it has been shown that there is a significant difference in precision between using the follow and the retweet links for topical influence propagation, favouring the latter.

Following these results, in this paper we will focus on the retweet type links, leaving the *mention* links study as future work. The following subsection presents a simplified version of the Twitter Graph, which will be used in further sections.

4.2. Twitter Graph Edges

Since measures for analysing three-mode graphs is still an area of research, several authors have proposed to study this type of graphs by taking 3 two-mode networks [20][30]. In this case: *1*) the user-message network UM; *2*) the resource-user network RU; and *3*) the resource-message network RM;

In this analysis we weight each of these two-mode networks following a term frequency inverse document frequency (tf-idf) weighting function. The edges we will analyse in this work include the retweet, topical, and entity edges.

4.2.1. Retweet Edges

We define the user-user *retweet* links graph (G_R) as a qualified two-mode network; representing a retweet relationship between an author of an original post (U_{oa}) and a user retweeting this post (U_{rt}) through the messages matrix (M). The *retweet* graph G_R is expressed as:$G_R = (U_{oa}M)(U_{rt}M)^T$.

The *retweet* links graph, G_R contains an edge between user *a* and user *b* if *a* has retweeted *a* post from *b*.

4.2.2. Topical Edges

We define the topic-user subgraph (G_T) as a qualified two mode network representing the relationship hold between a topic and a user through the messages this users posts. The topic graph G_T is expressed as: $G_T = (R_tM)(UM)^T$. By qualifying the U network by type of user we can generate the:

1. Original authors' topic graph

$$G_{T_{oa}} = (R_tM)(U_{oa}M)^T \tag{2}$$

   which contains an edge between a topic *t* and an author user *a* if *a* has posted a message related to topic *t*.
2. Retweeting authors' topic graph

$$G_{T_{rt}} = (R_tM)(U_{rt}M)^T \tag{3}$$

   which contains an edge between a topic *t* and a retweeter user *a* if *a* has retweeted a message related to topic *t*.

The topic-user graph represents the topical interest of a user.

4.2.3. Entity Edges

We define the qualified entity-user sbgraph (G_{Ex}), as a qualified two mode network representing the relationship hold between an entity and a user through a message. It is derived as : $G_{Ex} = (R_xM)(UM)^T$, where x represents the entity type. The entity-user graph G_E contains an edge between entity *e* and user *a* if *a* has posted a message related to entity *e*.

4.3. Twitter Dataset

Over a period of five days we captured Twitter public statuses returned via the Twitter streaming API [12].

We obtained a set of over 2.2 million tweets from which 393,700 where retweets relationships.

The distributions of tweets per user (Figure 4a), followers per user (Figure 4b) and retweets per user (Figure 5a) follow a power law function.

We examine the correlation between the counts of retweets and the counts of followers, for all the retweeter users set. Figure 5b, presents this correlation, which shows that the more a user retweets the more followers she has, and vice versa.

For the retweeter users, we generated a subset of 100,000 retweets written in English (using the Apache Tika language classifier [13]). In this subset there are a total of 32,626 unique users, from which 1,417 are retweeters and 31,299 are original authors. For this retweeter subset, the median number of statuses is 1452 tweets. These users have a median of 568 followers, and a median of 288 friends.

The retweet network extracted from this dataset consists of 32,626 users (nodes) and 50739 retweet relations (edges), with a diameter of 181 and a density of 4.76e-05. Table 2 presents (ID) Indegree, and (OD) outdegree properties of this graph.

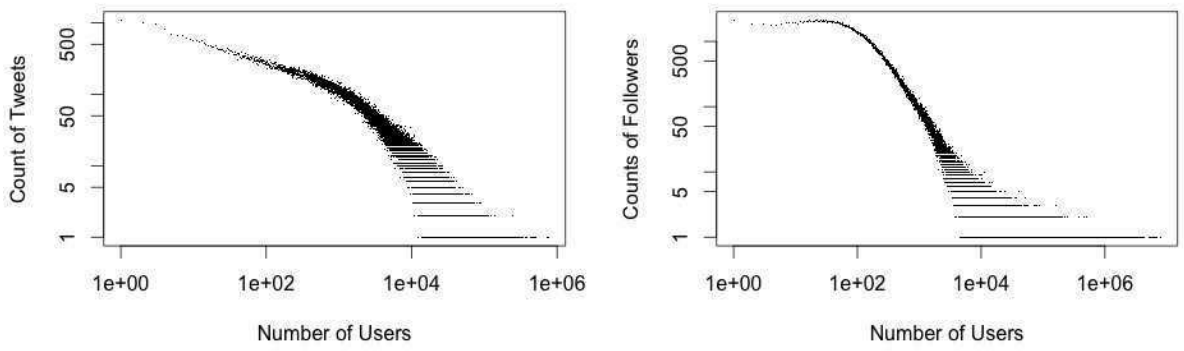
Table 2
Retweet Network ID and OD Properties

Property	Median	Mean	Max
ID	1	1.55	367
OD	0	1.55	746

In this subset we performed the semantic enrichment and RDF conversion described in section 3.3. We obtained 33 different types of entities including for example: SportEvents, Technology, Person, City, Movie, and Political Event, among others. There were 18 different types of Topics returned by the OpenCalais service, including for example: business_finance, entertainment_culture, politics, and technology_internet.

The following subsection presents our approach for deriving influential users given a context described by a topic and/or an entity.

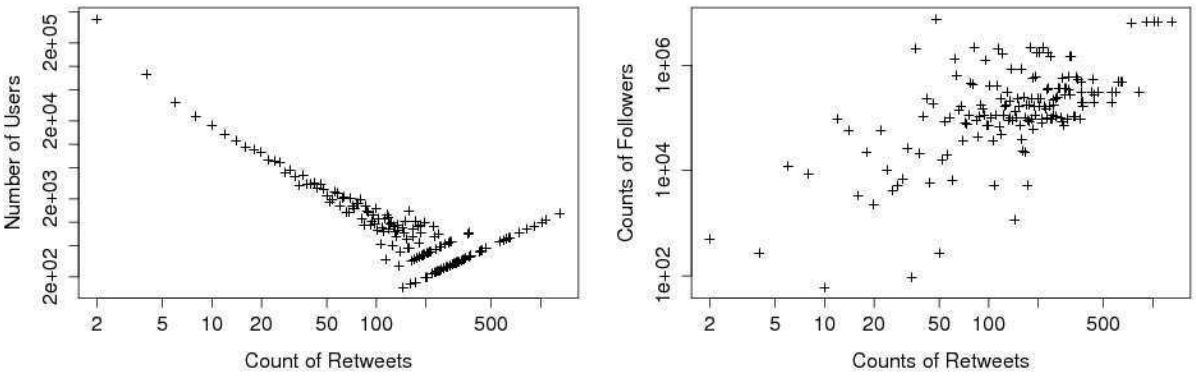
[12] Twitter API, https://dev.twitter.com/docs/streaming-api
[13] Apache Tika, http://tika.apache.org/



(a) Distribution of tweets per user

(b) Followers per User

Fig. 4. Distribution of Tweets per User and Followers per user.



(a) Retweets per User

(b) Count of Retweets vs. Count of Followers

Fig. 5. Distribution of Retweets per User and Retweets vs. Followers.

5. Topic-Entity Influence Measure

Previous studies have investigated user influence in the Twitter graph, in analogy to the study of an "authority" web page in the Web Graph [18] [32].

The following subsection presents a review of PageRank, and Topic-Sensitive Page Rank, and how they can be applied to the Twitter graph, before introducing our approach in subsection 5.2.

5.1. Review of PageRank

According to the PageRank algorithm [5] [24] applied to the Web graph, if a page u has a link to page v, then the author of page u is implicitly giving some importance to page v. In this case, the problem definition reduces to finding how much importance a page u confers to its outlinks. In this subsection we present a review of the PageRank algorithm, applied on the Twitter Graph.

In the case of the Twitter graph we are interested to find how much importance a user a confers to user b by either following b or retweeting posts from b.

Consider the problem of deriving the importance of a user on the Twitter Graph, based on the retweets relationships. Let $Rank(p)$, be the importance of user p, and let N_a be the outdegree of user a (number of

users from whom user a retweets). The link (a, b) gives $\frac{Rank(a)}{N_a}$ units of rank to user b. In this way, we can derive a rank vector $Rank^*$ over all of the users on the Twitter graph. If N is the number of users, then we assign an initial value of $\frac{1}{N}$ to all users. Let P_b be the set of users retweeting a post from b, then in each iteration the rank is propagated by computing:

$$\forall_b Rank_{i+1}(b) = \sum_{a \in P_b} \frac{Rank_i(a)}{N_a} \tag{4}$$

This computation ends when the $\vec{Rank}$ vector stabilizes to within a threshold. When this threshold is reached, then the final $Rank^*$ is the PageRank vector over the users on the Twitter Graph. This computation can be derived as well by considering the matrix representation of the the retweet direct graph G_R. Let M be such matrix, and let the matrix entry m_{ij} have the value $\frac{1}{N_j}$ if there is a retweet link from user j to user i, and the value 0 otherwise. Multiplying iteratively the $M \times \vec{Rank}$ yields to the the dominant eigenvector representing $\vec{Rank}^*$.

Studies on the convergence of the PageRank alogithm suggest the use of a damping factor $1 - \alpha$ to the rank propagation [22]. The damping factor guarantees the convergence to a unique rank vector. The PageRank algorithm with the damping factor is expressed as:

$$\begin{aligned} \vec{Rank} &= M' \times \vec{Rank} \\ &= (1 - \alpha)\, M \times \vec{Rank} + \alpha \vec{p} \end{aligned} \tag{5}$$

$$\tag{6}$$

where $\vec{p} = [\frac{1}{N}]_{N \times 1}$ represents the personalisation vector [4].

5.1.1. The Topic-Sensitive PageRank Algorithm

The Topic-Sensitive PageRank algorithm [11] suggested to biased the pageRank computation to increase the effect of a particular set of pages belonging to a category by using a non-uniform personalisation vector $\vec{p}$. In the case of the Twitter graph, rather than considering a set of pages we bias the vector $\vec{p}$ by considering sets of users belonging to a topical category.

The Topic-Sensitive PageRank considers as many damping vectors as topic you want to model. Let T_j be the set of users belonging to a topic t_j, when computing the PageRank for topic t_j, rather than using the uniform damping vector $\vec{p} = [\frac{1}{N}]_{N \times 1}$, we use a non-uniform vector $\vec{p} = \vec{v_j}$ where

$$v_{ij} = \begin{cases} \frac{1}{|T_j|} & \text{if } i \in T_j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The following subsection presents our approach to measure Twitter users influence by proposing a novel topic-based metric in which to bias the personalisation vector.

5.2. Topic-Entity PageRank

Our intuition is that users' writing "interest" can provide a better insight of the influence of a user based on topical relevance. The topical influence of a Twitter user depends on the topical diversity contained on his posts and the number of times other users retweet his posts. The more *retweets* a post generates the larger the topical audience it reaches. The same applies for the case of an entity.

Let RT be a set of triples of the form (i, k, j), representing the retweet relationship between user i and user j through post k. We define the topical relevance of a the retweet relation (i, j) to a given topic t as follows:

Definition 1. *Given the a retweet relationship (i, j), the topical relevance of this pair of users to a topic t is defined as:*

$$TR_t(i, j) = \frac{|K_t ij|}{|Kij|} \tag{8}$$

where $|K_t ij|$, represents the number of messages of topic t written by user j and retweeted by user i; and $|Kij|$ represents the number of messages from user j retweeted by user i.

This definition captures the notion that the more a user i retweets posts from user j about topic t, the higher the relevance of user j to this topic. Generally this leads to a higher influence on i, corresponding to a higher rank conferred from user i to user j.

The topical relevance of a retweet pair can be computed using the retweet graph (subsection 4.2.1); the original authors' topic graph $G_{T_{oa}}$ and the retweeting authors' topic graph $G_{T_{rt}}$ (subsection 4.2.2).

Taking into account the topical relevance of topic t given all the retweets relationships on the Twitter graph, we propose the algorithm presented in 1 for deriving a topical influence damping vector.

Algorithm 1 Calculating the Topical Influence Damping Vector

Let t be a topic from the set of topics T.
Let d be the total number of users posting tweets related to topic t.
Let RT be a set of retweet triples of the form (i, k, j), where i is the user retweeting post k produced by user j.
Let v_t be the damping vector for topic t
for all triples in RT **do**
    if user i hasTopic t **then**
        $v_t ij = TR_t(i, j) \times \frac{1}{d}$
    else
        $v_t ij = 0$
    end if If the Let $|K_i|$ be the number of messages related to topic t retweeted by user i
end for

Following the same intuition we can derive an entity influence of a Twitter user based on the entity relevance of his posts and on the number of times other users retweet his posts. However in the case of the entity approach we are interested on following the influence of an entity instance rather an entity type. For example, we would like to know which Twitter user is influential on information relating Barack Obama, which is an instance of an entity of type *Person*; or which user is influential on information relating IPhone, which is an instance of an entity of type *Technology*.

Given RT, the set of triples of the form (i, k, j), representing the retweet relationship between user i and user j through post k. We define the entity relevance of the retweet relation (i, j) to a given entity-instance y as follows:

Definition 2. *Given the a retweet relationship (i, j), the entity relevance of this pair of users to a entity-instance y is defined as:*

$$ER_y(i, j) = \frac{|K_y ij|}{|Kij|} \tag{9}$$

where $|K_y ij|$, represents the number of messages related to the entity instance y written by user j and retweeted by user i; and $|Kij|$ represents the number of messages from user j retweeted by user i.

Similar to the topical relevance, the entity relevance definition captures the notion that the more a user i retweets posts from user j related to the entity-instance y, the higher the relevance of user j to this instance.

The entity relevance of a retweet pair can be computed using the retweet graph (subsection 4.2.1); the original authors' entity graph $G_{E_{oa}}$ and the retweeting authors' entity graph $G_{E_{rt}}$ (subsection 4.2.3), qualifying them to the entity instance y.

We propose the algorithm presented in 2 for deriving a topical influence damping vector.

Algorithm 2 Calculating the Entity-Instance Influence Damping Vector

Let y be an instance from the set of entities of type x.
Let d be the total number of users posting tweets related to instance y.
Let RT be a set of retweet triples of the form (i, k, j), where i is the user retweeting post k produced by user j.
Let v_y be the damping vector for the entity instance y
for all triples in RT **do**
    if user i hasEntity y **then**
        $v_y ij = ER_y(i, j) \times \frac{1}{d}$
    else
        $v_y ij = 0$
    end if If the Let $|K_i|$ be the number of messages related to the entity instance y retweeted by user i
end for

The following section presents the evaluation of the proposed algorithms 1 and 2.

6. Evaluation

This section presents the results of applying the topic-entity PageRank algorithms to our dataset. The problem definition consists of identifying influential users from the Retweet Graph for a given topic.

In order to compare the performance of our approach, we used the following methodologies for deriving topic based influence in graphs: i) the Hyperlink-Induced Topic Search (HITS) algorithm[15]; ii) the In-degree (ID) algorithm; and iii) the Topic-Sensitive Page Rank (TSPR) algorithm [11] , which are described in subsection 6.1.

In order to differentiate our approach with others we denote it by **TPR**. For this comparison we used a testbed of 18 topics for which influential users were derived using these algorithms. Each of these ap-

proaches gives as a result, a list of users which are ranked according to how influential they are in a given topic.

We first compared these algorithms in terms of the correlation between the influential-users ranked lists results obtained for each topic (see Subsection 6.2). This correlation measures the ranking agreement among the resulting lists of the algorithms. After analysing this correlation, we performed an evaluation of this algorithms based on a recommendation task described in Subsection 6.3.

6.1. Comparison with Existing Algorithms

We compared the results obtained for the topical-based approach with the proposed algorithms against the following related algorithms:

- **Hyperlink-Induced Topic Search (HITS)** [15] (a.k.a. hubs and authorities), which applied to the retweet graph measures the "authoritativeness" (or influence) of a user as the degree to which this user is retweeted by important "hub" users. Being the "hubness" of a user the degree to which a user links to other important authorities.
- **In-degree (ID)**, when applied to the retweet graph, measures the influence of a user based on the degree in which this user has been retweeted. Currently this is the metric used by third party services (e.g. wefollow.com, twitterholic.om) for calculating a user's influence) applied to the following graph.
- **Topic-Sensitive PageRank (TSPR)** [11], which when applied to the retweet graph, calculates the influence of a user on a particular topic t by altering the damping vector in the PageRank algorithm . As discussed in subsection 5.1.1 this alteration only involves the number of users related to the topic t. For our comparison we also calculate the damping vector for an entity y in analogy to a topic t.

6.2. Correlation

After computing the rank lists generated through the TPR approach and the related algorithms, we measured the correlation among them using the Kendall's τ [14]. This metric calculates the pairwise disagreement between two lists. The τ distance ranges from -1 to 1. If two list are in the same order then $\tau = 1$; while if one list if the reverse of the other then $\tau = 0$. The closer τ is to 1 the closer the agreement between the lists.

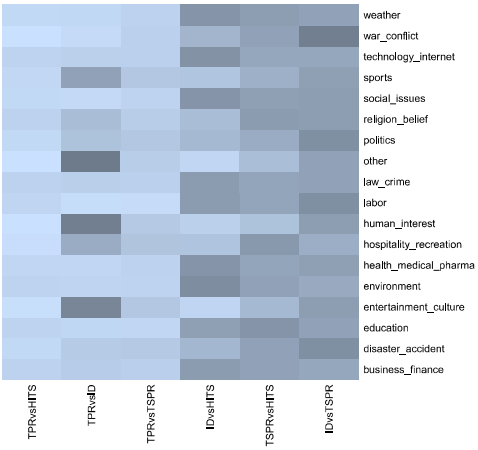


Fig. 6. Kendall's τ correlation of rank lists, the darker the color the higher the value.

Figure 6 presents a colormap exhibiting the τ values obtained from the correlation of TPR with the HITS, ID, TSPR and among them. TPR presented a higher agreement with the ranked list resulting from the ID algorithm. Particularly the topics: Entertainment_Culture, Human_Interest and other; presented the highest agreement between TPR and ID. When comparing the averaged correlation for the five topics among the rest of the evaluated algorithms we observed that the highest agreement was presented between the ID and TSPR algorithms. These correlations show that the level of agreement across rank list is topic dependant. It also exhibits that the rank lists obtained with TSPR differs from those exposed by HITS, ID and TSPR. In order to calculate the performance of these ranking algorithms, the following subsections presents and recommendation-based evaluation.

6.3. Performance in Recommendation Task

In order to evaluate the validity of the results provided by our approach we computed the performance recommendation task suggested by Weng et al [32]. We adapted this task to the case of the retweet graph as follows:

Let R be the set of existing re-tweeting relationships.

1: Randomly choose $|R|$ existing "retweeting" relationship formed among twitterers;
2: **for all** $r \in R$ **do**

3: let u_a and u_b be the user posting a tweet, and the original post's author, in the "retweet" relationship r;
4: randomly choose 10 users that u_a is not retweeting, denote this set as S;
5: remove r to generate a new network in which user u_a does not retweet u_b;
6: apply different algorithms to measure the influence of u_b and all the users in S in the new network, based on which u_a is recommended whether to "retweet" u_b;
7: compare the quality of the recommendation by different algorithms;
8: **end for**

Definition 3. *Definition: Let k be a ranked list recommended by any of the algorithms, and u_i a user. The ranks are ordered in ascending order; a lower number means is higher in rank; the higher in the rank the higher the recommendation. Let $k(u_i)$ be the rank of u_i in k (a higher rank corresponds to a higher-numbered rank in k). The quality of the recommendation is based on the value of Q_k which is measured as $Q(k) = |u_i$ such that $ui \in S$, and $k(u_i) > k(u_b)|$ (i.e. the number of users in the new network S which show a higher rank that the retweeted user b); where u_b is the retweeted user which was removed in step 5 of the recommendation task. In this case, the higher the value of $Q(k)$, the lower the quality of the algorithm.*

According to Figure 7, which presents the Q values for all the 18 topics; the lowest Q values where achieved by TPR, followed by TSPR, ID and HITS. This figure also exhibits that although the highest performance is consistently achieved by TPR and TSPR, where the first outperforms the latter, the performance of these algorithms is topical dependent. Particularly TPR exhibited a low performance for the War_Conflict topic; this could be due to users topical-dependent behaviour for retweeting. Table 3 presents the average Q values over all the 18 topics. The average of the quality metric for the performance task favours the TPR over both Indegree and the TSPR algorithm.

Table 3
Averaged Q values

TPR	HITS	ID	TSPR
4.3	6.13	5.48	5.2

As opposed to TSPR, which weights the graph in equal proportions (based on the number of users belonging to a given topic), TPR weights the graph based

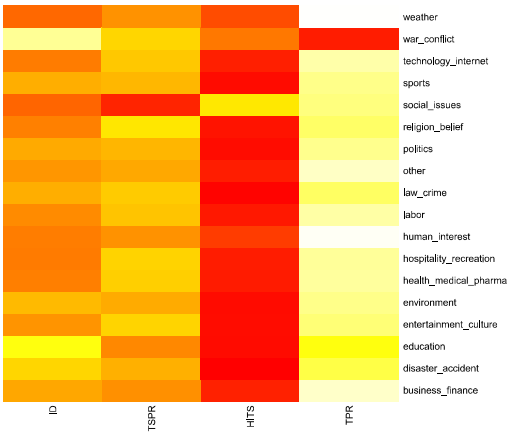


Fig. 7. Q values for all topics, the darker the color the higher the value. The lower the value the better the performance of the algorithm.

on the topical relevance of individual users. Therefore, TPR definition of topical relevance captures the notion that the more a user i retweets from user j about a topic t the higher the relevance of user j to this topic.

These results suggest that users act as proxy of topical influence by means of retweets relations. Moreover the variation in performance for all 18 topics and four different algorithms suggests that users' topical-dependent retweeting behavior can impact information diffusion, in particular in this case, influence in retweeting networks.

The following section presents a use case in which TPR is used for visualising top ranked influential user.

7. VisInfluence - Visualise Influential Users and Content on Twitter

In this section we introduce the VisInfluence platform, which is a web-based interface for visualising topical influential users from the Twitter graph. The main component of the tool, the influence chart is inspired from seismographs, which are devices that measure ground movements and seismic waves from sources such as earthquakes, volcanic eruption and tectonic movements. The influence chart indicates the influence users have on a specified topic as line graphs, aligned parallel to each other.

The interface consists of: i) a set of user controls (top and bottom); ii) an influence chart (left); iii) hashtag and entity tag clouds (right); and iv) links . The in-

fluence chart (left) is built using Processing.js[14], while the tag clouds are built as HTML strings, rendered on DIV elements. The communication with the back end is achieved via web services, which expose methods that continuously re-evaluated influence scores based on recently retrieved tweets. The communication between the front-end and the back-end is via AJAX queries and JSON responses.

Upon initialization, VisInfluence prompts the user to select a topic of interest from a drop-down list of the 18 available topics on a dialogue. These topics correspond to the topical categorisation provided by the OpenCalais service (see Subsection 3.3).

When a user selects the topic for which influential Twitter users need to be extracted, VisInfluence queries for the top ranked influential users and thereby presents them via the dynamic influence chart (Figure 8 (left)). Twitter users are presented as a blue circle on the left, whereas their influence is presented as a line chart next to them. The line chart dynamically changes (the chart shifts toward the right hand side of the screen) as the influence changes.

Users can select the period for which the dataset should be analysed as well as the time rates in which the ranking algorithm should be recalculated. For example, a user can choose to aggregate a month's Twitter data and feed back to the visualizations presenting per day influential users. These selections are done by two sliders: 'Aggregation time' and 'Update rate'.

As can be seen from the Figure 8, VisInfluence also presents dynamic tag clouds, and interactive lists extracted from the influential users' semantic profiles. The tag clouds (for presenting entities and hashtags) and the related links update at the same rate as that of the line chart. This presents a coherent view of the underlying analysis and the dataset as it progresses over time. The users, however, can choose to 'pause' the visualization by clicking on the pause/play controls at the bottom of the screen. As a pre-configuration step, users can select the number of users they would like to visualize on the interface as well as the number of historical 'readings' they would like to keep.

Users can choose to select a different topic from the top of the interface from a drop down menu and all the visualisations are then updated to the current topical context. Users appearing to gain a higher influence rank at a particular point in time are represented as a red point on the line chart.

[14]http://processingjs.org/, a JavaScript port of the Processing Visualization Language

Although VisInfluence has been developed for monitoring real-time changes in the influence scores, the updates appearing on the influence charts depend upon: i) the rate of changes in the Retweet Graph; and ii) the updates of the influence score models obtained with TPR . These two dependencies can generate a time-lag for any updates in the visualisation. Although VisInfluence is a system being currently developed, we have plans to make a beta version available for users to interact with it online.

8. Related Work

Following the semantic-social network model [20], Wagner and Strohmaier introduce the Tweetonomy model [29], which is a formalisation of social awareness streams. This model adopts a theoretical approach similar to the one presented by Mika. However, the Tweetonomy model presents a more complex and dynamic structure than traditional folksonomies.

The analysis of user-generated content extracted from social media sites is an active research area. In particular, studies related to Twitter dataset have investigated questions related to network and community structure. For example, Krishnamurthy et al [16] present a characterisation of the Twitter social network, which includes patterns in geographic growth and user's social activity. In their work, they suggest that frequent updates might be correlated with high overlap between friends and followers. Java et al [13], present an analysis of Twitter and suggest that the differences in users' network connection structures can be explained by the following types of user activities: information seeking, information sharing and social activity. They use the HITS algorithm [15] for detecting users intent.

Recent work has investigated the derivation of user profiles based on the semantic relations extracted from Twitter personal awareness streams. Rowe and Stankovic [28], generate semantic user profiles based on DBPedia resources derived from the users' tweets. Using these profiles as a topical interest representation they use a machine learning approach for proving alignments between tweets and an events. They use this alignment for recommending events to users. Cano et al [7], presented a model for deriving a context-based semantic user profiling based on a user's Twitter personal awareness stream. They enriched Twitter content using Zemantha and OpenCalais services, using entities, hashtags, and keywords. Abel et al [1], en-

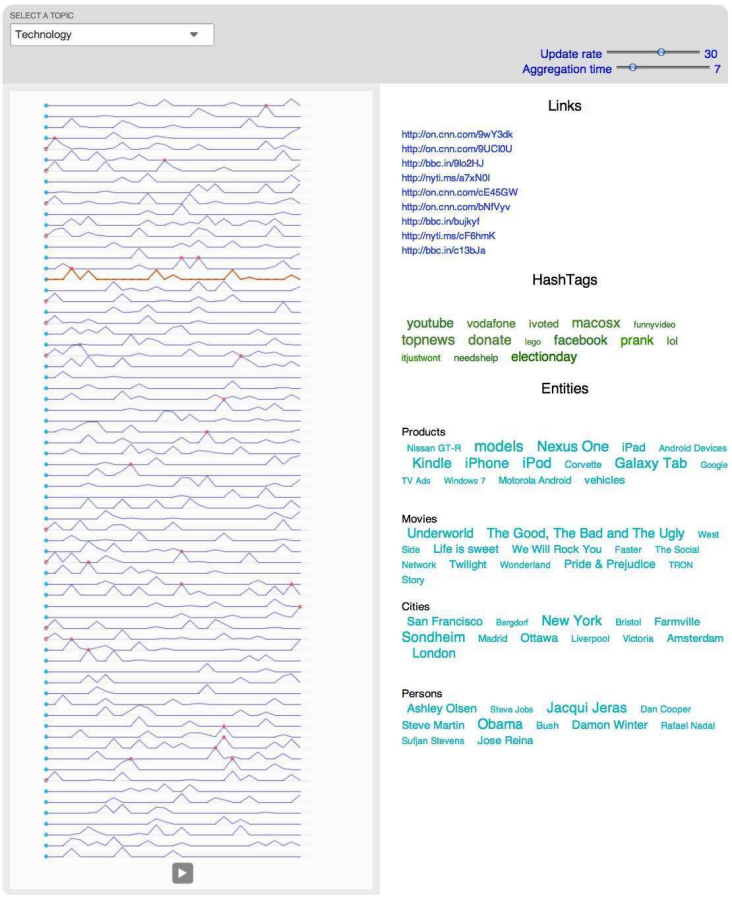


Fig. 8. The VisInfluence interface, presents influential users for the topic Technology, as well as the related entities of type

rich tweets' and news articles' content using concepts, and entities extracted using the OpenCalais service. They build semantic user profiles and provide similarity metrics for recommending news articles based on these profiles. Rather than recommending content to users we make use of users' semantic profiles for deriving influential users within a Twitter graph.

Other studies have explored the conversational subgraph generated on the Twitter graph. Ritter et al [25], model dialogues in Twittter using an unsupervised approach. Their model aims to identify strong topic clusters within noisy conversations. Rowe et al [26] [2], present an approach for predicting discussions on the Social Semantic Web. After identifying discussion seed posts, they characterise the content and user-features of a post by studying their effects in predicting the level of discussion that this post can generate. In our work rather than be interested on attention over the conversational graph, we study attention as a source for deriving influence within the Twitter's retweet graph.

The analysis of users' influence on the Social Web has been studied in both the Information Retrieval and Semantic Web communities. Studies in blogs have investigated the reading and posting behaviour for deriving a blogger's influence of the public [19] [9] [10].

The problem of finding topic-based influential users is closed to the problem of finding experts. In the later problem, Stankovic et al [27] propose the subject, and type homogeneity metrics, for finding experts based on a user's topical traces left on the Linked Data cloud. However their metric does not consider social network relationships.

Some of the closest work to our own includes the work of Cha et al [8]. They study users influence based on the retweet graph of the Twitter graph, by measuring the degree of followers, retweets and mentions of

a user. In the Web ecology project [15] Twitter users are ranked using different features including for example the number of followers, and the average content per tweet. TunkRank [16] describes user's influence as the expected number of users who will read a tweet from them, either through a following or a retweet relation, however they do not focus on topic or entity sensitive ranking and only propagate influence over follow links. They rank user influence based on the indegree algorithm. Our work extends this work by providing a metric for studying topical and entity-based influence of users following a pageRank algorithm.

Boyd et al[3] have highlighted that retweets are prone to present variations in style, which can lead to ambiguity in and around authorship since a message morphs as they are passed along. However other work have proved that retweets are a good source for analysing topical influence propagation in social networks. For example Welch et al [31] investigates the semantics of the retweet and following relationships from deriving influential users. Applying the PageRank [24] and Topic-Sensitive PageRank [11] algorithms they show that transitivity of topical relevance is better preserved over retweet links, and that retweeting a user is a stronger indicator of topical interest than following him. They argue that the propagation of topical influence through *following* links is problematic since traversing a single following link dramatically reduces the probability of topical relevance.

Weng et al, present the TwitterRank algorithm [32] for deriving influential users on Twitter based on their following relationships. They extend the PageRank algorithm, by introducing a lda-topic-based transition probability which takes into account the topical similarity between users in a following relationship. Our work extends this work by studying the retweet graph and providing metrics based on semantic profiles for deriving topical and entity based influence.

There has been an increasingly growing interest in providing visual access to microblogs over the past few years[17]. The massive influence of microblogging platforms like Twitter and Facebook has sparked a need for ways to identify highly influential individuals as well as content. Browser plugins, widgets, add-ons and third party applications like TweetDeck[18], Seesmic[19], Twitbin[20] and so on allow users to follow trending topics, generate new content, follow live posts and so on. Several interfaces (such as TweetDeck) allow users to follow multiple topics at the same time by arranging posts in pre-defined columns.

Though most of these applications allow direct access to individual tweet instances there have been a few tools that encode such tweets into visual items, thereby creating visual abstractions of twitter posts. Most commonly used techniques for visualizing tweets involve presenting tweets as visual clusters, on geographical maps, timelines, networks or other visualisations: TweetStats[21] enables users to visualize their tweet clusters as tag clouds, timelines, Tweet Density and so on; Trendsmap[22] and Twittearth[23] provides a geographical visualization of real time tweets that are being posted; [12] provides an interesting 3 dimensional visualization of user activities, relationships, communications, message transitions and message flows in order to identify trends and relationships trends in 3D space.

Closer to our intended goal of presenting information generated from processing twitter user influence, TunkRank[24] provides a list of most influential users; Klout[25] presents an interface for users to visualize their influence over time, influential topics, their followers categorized according to influence and which users they have an influence on. In our extends this work by providing an overall visualisation of top ranked users based on topical and entity-based influence.

9. Conclusions and Future Work

In a world where information can bias public opinion it is essential to analyse the propagation and influence of information in large-scale networks. In this work, we have presented an analysis of social influence of the Twitter Graph. We have focused on the discovery of top ranked topical influential users based on the retweet-relationships subgraph of the Twitter graph. By generating semantic profiles we have extracted the

[15] The Web Ecology Project, http://webecologyproject.org

[16] TunkRank, http://tunkrank.com

[17] 17 ways to visualize the twitter universe has a few interesting examples, http://flowingdata.com/2008/03/12/17-ways-to-visualize-the-twitter-universe/

[18] http://www.tweetdeck.com/

[19] https://seesmic.com/

[20] http://www.twitbin.com/

[21] http://tweetstats.com/status/pulse2dotcom

[22] http://trendsmap.com/

[23] http://www.twittearth.com/

[24] http://tunkrank.com/score/top

[25] http://klout.com/home

topical and entity relevance of a retweet triple consisting of a retweeter user, the retweeted message and the original author.

We introduced a variation of the PageRank algorithm for analysing users' topical and entity influence based on the topical/entity relevance of a retweet relations expressed by these semantic profiles. Our experimental results shows that except for one topic (War_Conflict), TPR consistently outperforms all TSPR, ID and HITS.

While these results suggest that users act as proxy of topical influence by means of retweets relations; it also highlights the relevance of the users' topical-dependent retweeting behaviour, which can impact the performance of these algorithms. This opens new questions regarding the impact of user's topic based behaviour in information diffusion, and in this case, in the study of influence in retweeting networks.

Future work includes the consideration of topical and entity relevance for modifying the transition matrix. In this case the probability of a user retweeting certain post will depend on his retweet relations' topical relevance. We also plan to apply sentiment analysis on the tweets content for modelling influence based on topical-sentiment and entity-sentiment features in the Twitter Graph.

We have also presented a dynamic visualisation interface which enriches the context in which a user have been ranked as influential. By using the users' semantic profiles, it shows the related entities, hashtags and keywords of a particular user, and the aggregation of these information when a particular influential user hasn't been selected. In this way it also acts as a top ranked trending concepts visualisation.

10. Acknowledgements

A.E.Cano is supported by CONACyT, grant 175203. S. Mazumdar is funded by Samulet (Strategic Affordable Manufacturing in the UK through Leading Environmental Technologies), a project partially supported by TSB and from the Engineering and Physical Sciences Research Council.

References

[1] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. In D. C. e. In P. Brusilovski, editor, *User Modeling and User-Adapted Interaction (UMUAI) Special Issue on Personalization in Social Web Systems*, pages 1–42, 2011.

[2] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *Proceedings of the 10th International Semantic Web Conference*, volume I of *ISWC'11*, pages 35–50, Berlin, Heidelberg, October 2011. Springer-Verlag.

[3] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), Proceedings of the 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[4] S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a web in your pocket. *In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 21:37–47, 1998.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, 1998.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks : the international journal of computer and telecommunications netowrking*, 33(1-6):309–320, June 2000.

[7] A. E. Cano, S. Tucker, and F. Ciravegna. Follow me: Capturing entity-based semantics emerging from personal awareness streams. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages*, volume 7 of *ESWC '11*, pages 33–44, May 2011.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Weblogs and Social*, 2010.

[9] K. Gill. How can we measure the influence of the blogosphere? In *Proceedings of the international conference on World Wide Web, workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, WWW '04, 2004.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[11] T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference*, WWW '02, pages 517–526, New York, NY, USA, 2002. ACM.

[12] M. Itoh. 3d techniques for visualizing user activities on microblogs. In *Frontier Computing. Theory, Technologies and Applications, 2010 IET International Conference on*, pages 384 –389. IEEE, 2010.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[14] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):668–677, 1999.

[16] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, WOSP '08, pages 19–24, New York, NY, USA, 2008. ACM.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the 8th international conference on World Wide Web*, WWW '99, pages 1481–1493, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[19] Y.-M. Li, C.-Y. Lai, and C.-W. Chen. Identifying bloggers with marketing influence in the blogosphere. In *Proceedings of the 11th International Conference on Electronic Commerce*, ICEC '09, pages 335–340, New York, NY, USA, 2009. ACM.

[20] P. Mika. Ontologies are us: A united model of social networks and semantics. *Journal of Web Semantics*, 5(1):5–15, March 2007.

[21] S. Milstein, B. Lorica, R. Magoulas, G. Hochmuth, A. Chowdhury, and T. O'Reilly. Twitter and the Micro-Messaging Revolution. Technical report, O'Reilly Radar Report, October 2008.

[22] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, USA, i edition, 1995.

[23] M. Naaman, J. Boase, and C. H. Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[25] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[26] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ESWC'11, pages 405–420, Berlin, Heidelberg, 2011. Springer-Verlag.

[27] M. Stankovic, W. Breitfuss, and P. Laublet. Linked-data based suggestion of relevant topics. In *Proceedings of I-SEMANTICS concerence*, I-SEMANTICS '11, pages 49–55, New York, NY, USA, September 2011. ACM.

[28] M. Stankovic, M. Rowe, and P. Laublet. Mapping tweets to conference talks: A goldmine for semantics. In *Proceedings of the 3rd International Workshop on Social Data on the Web (SDoW2010) Workshop at the 9th International Semantic Web Conference (ISWC2010) - ISWC 2010 Workshops*, volume I of *ISWC '10*, 2010.

[29] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the 3rd International Semantic Search Workshop*, SemSearch '10, pages 6:1–6:10, New York, NY, USA, april 2010. ACM.

[30] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[31] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 327–336, New York, NY, USA, 2011. ACM.

[32] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.