Information Fusion 46 (2019) 91–101

Contents lists available at ScienceDirect

Information Fusion

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

Fusing pattern discovery and visual analytics approaches in tweet propagation

Octavio Loyola-González[a], Armando López-Cuevas[b], Miguel Angel Medina-Pérez[*,a], Benito Camiña[a], José Emmanuel Ramírez-Márquez[c], Raúl Monroy[a]

[a] School of Science and Engineering, Tecnológico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán, Estado de México 52926, México
[b] Tecnólogico de Monterrey, Campus Guadalajara, Av. Gral Ramon Corona 2514, Zapopan, Jal 45201, México
[c] Enterprise Science and Engineering Division, School of Systems & Enterprises, Stevens Institute of Technology, USA

ARTICLE INFO

Keywords:
Social networks
Twitter
Pattern recognition
Influence modeling
Visual analytics

ABSTRACT

Over the past several years, social networks have become a major channel for information delivery. At present, social networks are being used to obtain more followers and exert influence over people during political campaigns. However, the propagation of a social network post is dependent on numerous factors. Some of these are known; for example, the post contents, the time when it was posted, and the person or entity by whom it was posted. However, other factors remain unknown, such as what makes a post more successful than others, and how posts from similar profiles evolve and propagate differently over time. The main subject of this work is addressing these types of questions. Our approach relies on a three-fold methodology for studying the influence and propagation of posts: graph-based, semantic, and contrast pattern recognition analysis. The results obtained are complemented by a dynamic visualization that encompasses all of the variables involved. In order to corroborate our results, we collected all posts from the Twitter accounts of the most prominent Mexican political figures and analyzed the influence and propagation of each post issued.

1. Introduction

Social network interactions give rise to highly complex dynamics in terms of information propagation. The information propagation is driven by not only the network connectivity, but also user actions. In this manner, users act as gates that allow the information flow through the network. Whether a user creates new content, replies to other user comments, or propagates repeated information may or may not be influenced by other user actions. The influence exerted by a person or group of persons on a community may drive decisions across different areas. For example, it is very important to determine influence patterns in marketing because this can aid in understanding the adoption of new products and optimizing resource allocation [1–3]. Companies and governments invest resources into tools for analyzing social media trends and population behavior in order to support business and policy decisions. A growing number of research studies in social networks analysis aim to gather insights from these networks. For example, in [4], the authors developed a community embedding framework to support community-level applications and analysis. In [5], the authors developed a method for visualizing the evolution of different topics related to a particular company, which allowed for the detection of different evolving pattern types hidden by the data volume. In recent years, in politics, emergency response, and business, it has become important to account for influential social network nodes in order to influence and reach as many people as possible [6]. As such, several studies have aimed to identify the *influentials* within a society or a community for commercial, political or economic purposes [7–12]. Bennett [13] investigated the ability to influence and then take common actions, while [14] studied the logic of the connections and relationships with human beings.

The advent of social networks has introduced a new framework in which to study, experiment and, for the first time, optimize influence in a manner that was not previously possible. Real-time data can be collected from social networks regarding specific subjects or products from a wide diversity of users and locations [2,15,16]. However, to the best of our knowledge, very limited existing studies quantify influence in social media (and Twitter in particular), with the majority employing a traditional network theoretical perspective. Questions regarding tweets concerning popularity, survivability, strength, and importance are generally devoid of quantitative answers. As such, there is currently no

* Corresponding author.
E-mail addresses: octavioloyola@itesm.mx (O. Loyola-González), acuevas@itesm.mx (A. López-Cuevas), migue@itesm.mx (M.A. Medina-Pérez), jmarquez@stevens.edu (J.E. Ramírez-Márquez), raulm@itesm.mx (R. Monroy).

https://doi.org/10.1016/j.inffus.2018.05.004
Received 9 December 2017; Received in revised form 4 April 2018; Accepted 13 May 2018

method available to quantitatively compare different tweets in terms of their "importance". Yet, there is also no approach for allowing understanding information diffusion and individual influence as a function of a tweet (or set of tweets).

The main contribution of this work is the proposal of *a comprehensive method to compare, analyze, and differentiate among diverse tweets in terms of influence, visual analytics, and semantic analysis, using a mining contrast patterns approach*. The novelty of this method relies on combining different techniques into a unified framework in order to analyze influence. It is based on the fact that the proposed framework provides a data visualization analytics approach that allows for comparison, identification, and clustering of the tweets and also provides dynamic interaction, in the sense that the user can manipulate the visualization nodes with the cursor to obtain additional information, such as the tweet text, user ID, and other details. Furthermore, this framework explains the behavior of each Twitter account by using a language that is closer to social network experts; specifically, using contrast patterns represented by feature and value combinations, according to the sentiment analysis provided in the data acquisition. These and other characteristics result in our framework being an improvement over previous methods based on sentiment analysis.

The methods implemented in this study are as follows: (i) quantitative characterization of the influence of a tweet, (ii) a data visualization analytics approach that allows for the comparison, identification, and clustering of *influentials* as a function of tweets and the sentiments associated with them, and (iii) contrast pattern-recognition analysis based on semantic features and influence measures, which allows for identification distinct activity patterns and, based on those patterns, differentiation among different Twitter accounts.

The remainder of the paper is organized as follows. Section 2 reviews related works, while Section 3 describes the methods used to construct our model. Section 4 explains our visualization model, and Section 5 describes in detail the methods used for mining contrast patterns as well as the features used for this task. In Section 6, we demonstrate the results of our proposed method utilizing real-life examples. Finally, Section 7 discusses the advantages and drawbacks of the method, and presents conclusions drawn from this work.

2. Previous work

In recent years, significant advances have been made in natural language processing (NLP) research, owing to the use of deep learning techniques for solving specific NLP tasks [17–19]. One area that has presented particular advances on sentiment analysis is the task of classifying sentiments within text written in natural language. For example, in [20], the authors used a deep convolutional neural network to perform aspect extraction, which is a sentiment analysis subtask. In this work, the authors use a 100 billion word corpus from Google News, which employs a bag-of-words-based model known as CBOW. In [21], the authors proposed a new tensor fusion network to perform multimodal analysis of text and spoken language by using long-short-term memory (LSTM) networks. In order to achieve this, they used the Multimodal Opinion Sentiment Intensity (CMUMOSI) dataset, which is an annotated dataset of video opinions from YouTube movie reviews. An excellent recent review on sentiment analysis can be found in [22], in which the authors discuss NLP tasks that need to be solved in order to achieve human-like performance in sentiment analysis. Many of the papers mentioned above use an approach based on graph theory, although their classification results have not been proven using a large social network such as Twitter.

Twitter is a micro-blogging web service whereby users can share information through a network. On Twitter, a user can *tweet* or post as many micro-blogs as desired (again with the restriction that each text must be a maximum of 140 characters long). Twitter users can *follow* other users in order to read their tweets; in this manner, a directed network is constructed. Furthermore, there is a *retweet* function to for re-sending another user's tweet through the network, as well as a *favorite* function to mark a specific tweet as a favorite. Moreover, Twitter offers a *reply* function, which makes possible for a user to respond to a tweet with another tweet (also known as *mention*), thereby forming a conversation. The number of mentions containing the same *username* represents the extent to which that user enrolls other users to discuss a topic.

Information propagation in Twitter is dependent on several factors, such as the number of followers, network topology, user influence, time of day, and events such as elections or natural disasters, among numerous others [1,10,11,23–25]. The influence of a user will determine how rapidly and far a tweet will reach other users, but also how long that tweet will survive on the net. Most efforts aimed at describing influence in Twitter have focused on the user follower network and rate of user tweets-retweets. However, one feature is not accounted for: mentions or replies. In contrast, our approach proposes a metric based on the number of tweet replies (offspring) and the number of generations reached (replies to replies), in combination with the traditional retweet status and favorite count. Such a metric is incorporated into a model in order to quantify the propagation of a specific tweet on Twitter, which is referred to as *engagement*. The *engagement* concept can be useful when comparing the propagation of two tweets.

Several studies have measured twitter user influence by means of analytic models [24,26,27]. In [28], the authors modeled Twitter information propagation with the survival theory in order to develop additive and multiplicative risk models under which network inference problems could be solved. In [29], specific topic-aware social influence propagation was modeled by means of independent cascade and linear threshold models consisting of modeling authoritativeness, influence, and relevance under a topic-aware perspective. The authors devised methods to learn the model parameters from a dataset of past propagations. Stochastic models for predicting propagation have also been proposed [30,31]. The authors in [32] used a complex network theory to characterize user influence. Another study proposed an algorithm that assigns relative influence and passivity scores to every user [33]. In [34], the authors proposed that users who *retweet* often (target users) are hub-like nodes, because they collect and distribute tweets. In order to determine these, they constructed two graphs: one representing the propagation paths of a retweet, and the other a superposition of multiple propagation paths of retweets. Therefore, an overlapped graph represents users to be followed. The authors in [23] noted periodic traffic behavior and demonstrated that post popularity drops off following a power law. They also produced a simple "epidemic model" that captures most topological characteristics of the social network. Epidemic models are mathematical models that aim to describe the manner in which a disease propagates through a population; in the case of social networks, information can be modeled as a disease that propagates through the network. In [35], the authors studied the spread of retweet cascades using machine learning techniques. Although several studies have characterized Twitter influence, few or no works have quantified the influence of a tweet as an energy function of its replies.

3. Methods

The methodology implemented in this work consists of the following steps: (i) data acquisition and preprocessing, (ii) energy computation, (iii) visualization of the overall user-related activity (galaxy of tweets), (iv) visualization of specific tweet-related activity (constellation of tweets), (v) feature extraction, and (vi) mining contrast patterns. Each of these processes is described in the subsections that follow and the methodology pipeline is illustrated in Fig. 1, which represents each of the processes used in this study. The methodologies for (ii) and (iii) are novelties and both are described in a detailed manner below.