Engineering Applications of Artificial Intelligence 51 (2016) 37–49

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Discovering similar Twitter accounts using semantics

Gerasimos Razis, Ioannis Anagnostopoulos *

Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia 35100, Greece

ARTICLE INFO

Available online 12 February 2016

Keywords:
Similarity network
Social semantics
Twitter entities

ABSTRACT

On daily basis, millions of Twitter accounts post a vast number of tweets including numerous Twitter entities (mentions, replies, hashtags, photos, URLs). Many of these entities are used in common by many accounts. The more common entities are found in the messages of two different accounts, the more similar, in terms of content or interest, they tend to be. Towards this direction, we introduce a methodology for discovering and suggesting similar Twitter accounts, based entirely on their disseminated content in terms of Twitter entities used. The methodology is based exclusively on semantic representation protocols and related technologies. An ontological schema is also described towards the semantification of the Twitter accounts and their entities.

1. Introduction

Microblogging is a form of Online Social Network (OSN) which attracts millions of users on daily basis. Twitter is one of these microblog services, where its users vary from citizens to political persons and from news agencies to large organizations. Obviously, some users are more influential than others, while many tend to have similar interests. We have created InfluenceTracker[1], a publicly available website where anyone can rate and compare the recent activity and influence of any Twitter account.

The aim of this paper is threefold. Firstly, we propose an improvement over a previous work of us, which is used for calculating the importance and influence of a Twitter account. This improvement incorporates a quality measurement that reflects other users' feel or preference over the examined tweets. Secondly, we propose an ontology and its related semantic mechanisms/technologies which allow us to semantify similarity features (mentions, replies, hashtags, photos, and URLs) as Linked Data. Finally, we propose a methodology for rating the similarity of different Twitter accounts. This methodology is entirely based on the contents of the tweets generated by the accounts, and more specifically based on (i) the three basic Twitter entities, (mentions, hashtags and URLs), and (ii) the web domains that host the URLs. All the necessary information for the implementation of this methodology was retrieved using exclusively SPARQL queries from the graph generated from our proposed ontology.

The remainder of this paper is organized as follows. In the next section, we provide an overview over the related work on semantifying and generating RDF graphs from Twitter data, on semantic modeling and recommendation in Twitter, as well as on measuring influence in Twittersphere. In Section 3, we describe our approach in terms of data semantification and modeling, as well as how we rate the influence of a Twitter account. In Section 4, we analytically present the implemented on-line service and the ontology behind it that transforms raw data from the Twitter API into an RDF graph. Section 5 describes our approach towards similarity recommendation for Twitter accounts. In order to gain insight into this methodology we describe and a case study. In Section 6 we evaluate and discuss the results of the case study, and we further evaluate our methodology against subjective ratings from 22 evaluators. Finally, Section 7 provides the conclusions of our work by summarizing the derived outcomes, while providing considerations on our future directions.

2. Related work

This section provides an overview over the related literature on discovering influential users, and on related modeling and recommendation techniques in terms of content-based and data-driven approaches.

2.1. Measuring influence in Twitter

The calculation of the impact a user has on social networks, as well as the discovery of influencers in them is not a new topic. It

* Corresponding author. Tel.: +30 22310 66937.
E-mail address: janag@dib.uth.gr (I. Anagnostopoulos).
[1] http://www.influencetracker.com.

http://dx.doi.org/10.1016/j.engappai.2016.01.015
0952-1976/

covers a wide range of sciences, ranging from sociology to viral marketing and from oral interactions to Online Social Networks (OSNs). In the related literature, the term "influence" has several meanings and it is differently considered most of the times.

Romero et al. (2011) utilized a large number of tweets containing at least one URL, their authors and their followers. Their aim is to calculate how influential or passive the Twitter users are. The produced influence metric depends on the "Follower–Followee" relations of the users, as well as their retweeting behavior. The authors state that the number of followers a user has, is a relatively weak predictor of the maximum number of views a URL can achieve. As our work has shown (Razis and Anagnostopoulos, 2014a), the number of followers an account has does not guarantee the maximum diffusion of information in Twitter. This is because, in order to achieve high-levels of diffusion, your followers should not only be active, but they should also have a high-probability of retweeting, thus transmitting the messages they receive to their followers.

The work described in (Cha et al., 2010) proposes a methodology where for each Twitter user, three different types of influence are introduced. These types are "Indegree" (number of followers), "Retweet" (number of user generated tweets that have been reweeted) and "Mention" (number of times the user is mentioned in other users' tweets). A necessary condition for the computation of these influence types is the existence of at least ten tweets per user. The authors claim that "Retweet" and "Mention" influence correlate well with each other, while the "Indegree" does not. Therefore, they come up with the conclusion that users with high "Indegree" influence are not necessarily influential.

A topic-oriented study on the calculation of influence in OSNs is presented by Weng et al. (2010). The authors propose an algorithm which takes into consideration both the topical similarity between users and their link structure. It is claimed that due to homophily, which is the tendency of individuals to associate and bond with others having similar interests, most of the "Follower–Followee" relations appear. This work also suggests that the active users are not necessarily influential.

Another approach which defines influence in terms of copying what the directly related account does is presented in (Goyal et al., 2010). In this work, the authors propose an "influenceability" score, which represents how easily a user is influenced by others or by external events. It is built on the hypothesis that a very active user performs actions without getting influenced by anyone. The users of such a type are considered as responsible for the overall information dissemination in the network.

Boyd et al. (2010) stated that retweeting can be also characterized as a conversational infrastructure. According to the authors, a conversation "exists" either during a retweet where some new information is added to the initial message, or when a single tweet is retweeted multiple times. The latter is interpreted by the authors as an action to invite new users into the conversation.

Barbieri et al. (2013) developed a framework for modeling the spread of influence on OSNs by discovering the most influential users and by analyzing their social activity and interconnections inside the communities they belong to.

The authors in King et al. (2013) have proposed the t-index metric which aims in measuring the influence of a user on a specific domain. It is based on the H-index and denotes the number of times a user's tweet on a specific domain has been retweeted. The authors state that because someone's tweets are influential in one domain does not necessarily mean that they are also in others.

A graph-based approach for the identification of influential users in OSNs is presented in Sun et al. (2012). A created graph represents the relationships among the tweets and the users. The more implicit or explicit relationships among tweets exist for a user, the more influential the user is.

Zhao et al. (2014) propose a framework for measuring influence which is based on the sentiment of the messages exchanged among users. The framework has been applied in health communities and it measures the way the variations in the sentiment of users who have received health support can influence others.

All the related studies have shown that the most active users or those with the most followers are not necessarily the most influential. This fact has also been spotted by our work (Razis and Anagnostopoulos, 2014a). As described in Section 3, our Influence Metric depends on several factors, where the account activity is only one of them. Simply put, as the authors in (Srinivasan et al., 2014) state, enormous influence may spring from lesser known persons, while the "celebrities" may not be influencers.

Contrary to the aforementioned studies, for the calculation of our Influence Metric we neither set a lower threshold on the number of the user-generated tweets, nor we only utilize a specific subset of tweets that fulfill certain criteria (e.g. those containing URL etc.). All the Twitter accounts can be used as seed for the calculation of our Influence Metric, thus differentiating our work in respect to the related literature.

2.2. Semantic modeling and recommendation in Twitter

As semantics and linked data continuously rise, more works relevant to semantic modeling in OSNs appear. The authors in Celik et al. (2011) and Abel et al. (2011) propose frameworks for enriching Twitter messages with semantics. The first work involves the identification of semantic relationships between entities by analyzing Twitter posts. These semantic links are between persons, products, events and other entities and are utilized in order to provide suggestion to the users. The latter aims in modeling the users' profiles based on their microblogging activities in order to link Twitter posts with news articles from the web.

The work presented in Shinavier (2010) introduces a semantic data aggregator, which combines a collection of compact formats for structured microblog content with Semantic Web vocabularies. Its main purpose is to provide user-driven Linked Data. The main focus of this work is on microblog posts and specifically on their creators, their content and their associated metadata.

Another framework which utilizes semantic technologies, common vocabularies and Linked Data in order to extract and mine microblogging data regarding scientific events from Twitter is proposed in (De Vocht et al., 2011). The authors attempt to identify persons and organization related to them based on time, location and topic categorization.

Although ontologies and semantic technologies have been used in other works, none of them capture and model such a wide range of information, spanning from the Twitter related characteristics of the accounts to the entities found in the posted messages. In a previous work of ours (Razis and Anagnostopoulos, 2014b) we proposed an ontological schema towards semantification provision of Twitter analytics.

Finally, content-based and data-driven approaches have been used for estimating a Twitter user's location (Cheng et al., 2010), as well as the interestingness in terms of diffusion and the content of the tweets (Naveed et al., 2011). In a previous work of ours (Anagnostopoulos et al., 2015), we utilize the data retrieved from Twitter in order to investigate the query suggestion provision that can be extracted from large graphs, having no prior knowledge of them. Towards this direction, an algorithmic approach is introduced for creating a dynamic query suggestion set which consists of the most viral and trendy Twitter entities (hashtags, mentions and URLs) with respect to a user's provided query input.