

Identification and Recommendation of Authorities on Different Topics Based on Twitter

¹Roque D. Contreras-Chacón, ²Jack F. Bravo-Torres, ³Jennifer A. Yépez-Alulema, ⁴Diego A. Cuji-Dután and ⁵Paul E. Vintimilla-Tapia

¹Centro de Investigación e Innovación en Ingeniería, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
rcontrerasc @ups.edu.ec

²Centro de Investigación e Innovación en Ingeniería, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
Jbravo @ups.edu.ec

³Centro de Investigación e Innovación en Ingeniería, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
Jyepetz @ups.edu.ec

⁴Centro de Investigación e Innovación en Ingeniería, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
Dcuji @ups.edu.ec

⁵Centro de Investigación e Innovación en Ingeniería, Universidad Politécnica Salesiana, Calle Vieja 12-30 y Elia Liut, Cuenca, Ecuador.
Pvintimilla @ups.edu.ec

Abstract

The social platform Twitter is one of the most famous world microblogs. Monthly, it hosts hundreds of millions of visitors who publish a large amount of data that is consumed by others. Some of these publications help to others to generate knowledge. Users who generate content in a particular area, helping people to increase their knowledge, are classified as authorities in that field. This content could be exploited by students to improve their academic performance. In this paper, we present a system to identify and recommend authorities in a particular area based on machine learning, supervised and unsupervised techniques. To verify accuracy and our model, we developed several tests within a particular topic. The results show that the system can identify authorities with an average accuracy of $[78.82 \pm 2.51]$ %, representing a high degree of confidence.

1. Introduction

The emerging network society [1,2], based on communications skills and social interaction provided by the omnipresence of the Internet, is creating new development opportunities in different fields of citizens life, such as, economics, politics, education, and research. Global communications networks have been and are being deployed around the world. Consequently, this deployment leads us to think about the Internet as the linchpin of the new social structure.

Therefore, developments in the field of information and communication technology (ICT), electronic technology, and wireless communications systems are generating great environments of digital information and they are giving to their users free movement and access to information resources. Similarly, growing interest in social networks—a group of technologies that uses Internet applications to allow people sharing and creating information through social networks— provides new forms of socialization and networking among people located hundreds of kilometers away from each other [3]. For instance, billions of people around the world use their social networks to share and post feelings, personal experiences, and activities that they do every day [4]. Thus, social media is used by people to communicate and interact with each other despite their location or time zone [5].

In this context, educational processes cannot be limited to time and space constraints of traditional classes. This new social and digital environment is changing the way that students learn. Learning networks—based on interaction in social networks and information sharing— might stimulate teaching efficiency and learning processes.

E-learning environments require the support of technological tools, such as wikis, blogs, and social online learning. Considering that knowledge changes steadily, the interchange of this knowledge using social networks has become a stronghold in learning process [6]. Furthermore, the huge amount of data, generated by social networks users, hides different knowledge to be discovered. Mainly, decision makers (in industry, education, politics, and more) have to dig into this data to understand how their professional environment works by disclosing this knowledge [7].

Due to social media users produce and consume information, they are also known as “prosumers”. These users’ activity causes that messages related with a specific topic can reach dozens of thousands every day. In order to validate these information (knowledge), it is relevant to identify true authorities in a desired topic. For example, regarding to Twitter, there are different models, algorithms and metrics that are used to recognize people considered as authorities in a subject [8]. In addition, it is required to know how to process unstructured messages to determine if they have information regarding to a

particular field. So, it is necessary to search for text in accordance to a given word or taking in account the context. Also, the number of words used in a message has to be processed in a right manner to select relevant posts [9]. The aforementioned points should be considered to process messages gathered from social networks related to any topic or any purpose for which they will be used.

Thus, learning networks can be formed between students and authorities encouraging efficiency of teaching and learning processes. This paper aims to analyze and classify authorities and contents for students in different topics. Specifically, it will analyze authorities profiles and content quality of tweets posted by themselves. Hence, this collected information will be used to generate recommendations to students about these topics.

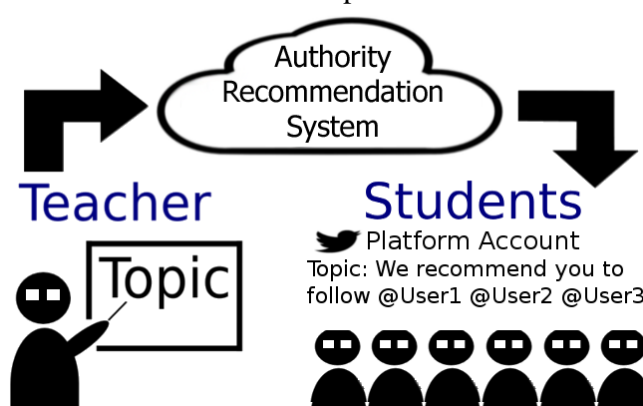


Fig. 1: Conceptual structure of the recommendation system

This paper is organized as follows. We begin, in Section II, with an overview of previous approaches to identify authorities on topics within social networks. Then, we provide an overview of the architecture of our proposal. Section III contains results of our experiments. Conclusions and future work are given in Section IV.

2. Related Work

People are surrounded by a huge amount of information that we need to share and exchange with others. Arrival of social networks has opened a new communication channel for this propose. However, shared information is generated by users who may be authorities or not in each of the topics. This situation strongly restricts its use without priori analysis.

Particularly, social platforms, called microblogs, have captured attention for its remarkable feature of posting short messages in real time, which is suitable for multiple devices, especially mobile devices. A giant of these kind of social platforms is Twitter¹, which is one of the most popular data sources in the world. This platform allows publication of short text messages up to 140

¹Twitter: number of monthly active users 2015. <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>. Accessed September 20, 2016

characters in real time. It has a labeling system that allows us to categorize a publication and speed up our search.

The exploitation of Twitter contents can be beneficial for virtual learning platforms. In particular, there are numerous authority users who are constantly publishing relevant and reliable information regarding to a particular topic. These publications may help students to deepen various subjects, increasing their knowledge and ensuring greater academic performance.

In a literature review, we can find several approaches that address the problem of identifying authorities on topics in microblogs. In 2010, Twitter launched a service for recommendation of whom to follow, based on the analysis of a keyword, users profile, their activity, and other aspects. This service, called Who To Follow, uses graph theory to achieve its goal. However, this approach has several limitations. For this reason, different algorithms, based on machine learning techniques, have been proposed [10]. The Framework TwitterRank [11] employs machine learning algorithms and graph theory to identify the most influential users in a fixed set of topics, previously identified. There is a limitation in this approach since the identified issues are fixed, so you cannot use it to identify authorities in varying themes. In [12], the authors propose an algorithm based on a machine leaning unsupervised technique, specifically, a probabilistic clustering based on a set of characteristics. Initially, tweets are extracted based on a keyword; then using probabilistic clustering algorithm, a set of authority users is created in the area, to finally be filtered to select the best k . This algorithm reported good results, allowing to add new users based on their contribution to the topic rather than their online activity. Other researchers have focused on knowledge of the crowd to determine users who are authorities in certain topics. Twitter allows its users to create lists of other users who want to receive notifications [13].

3. System Architecture

Our system is designed to provide greater support to formal education. Therefore, its architecture consists of two main parts: a learning management system, where information about the course and various resources provided by teachers and students are stored and shared; and a recommender system, which analyzes topics covered in the course, chooses authorities that are presents in social networks and, additionally, provides recommendations to students. Fig. 1 shows the overall process followed by our proposal. In this paper, we only discuss the recommender system. In the remaining part of this Section, we will analyze the structure of the recommendation system.

A. Authority Recommendation System

The main goal of the system is to recommend the top Twitter users who have expertise in a particular topic. To do so, we rely on the homophily feature, which many social networks have, including Twitter.

The homophily is a strong organizing principle of social systems and refers to the tendency of individuals in a social system to link with others who are similar to them rather than those who are less similar. For example, users may be similar to each other if they live in the same city, they share same topics, they have related jobs, etc.

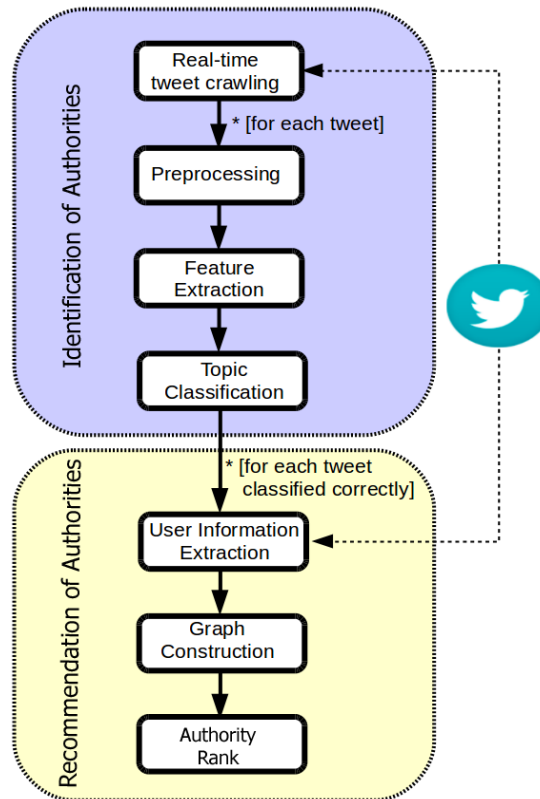


Fig. 2: Processes of identification and recommendation of authorities

We assume that many authorities on a particular topic may be friends of each other in Twitter due their similarities. Thus, the system finds a particular authorities community in the social platform and then recommends the best ones. Consequently, the system is composed of two main procedures: i) Identification of Authorities and ii) Recommendation of Authorities. Fig. 2 illustrates the system processes.

Identification of Authorities

This procedure is responsible for finding the possible authorities in a particular topic among Twitter users. We assume that a user is a topic authority if he/she publishes something related with the topic. Therefore, this procedure finds every person who tweets something about a particular topic and tags him/her as a possible authority (i.e. this procedure returns a set of possible topic authorities).

A supervised machine learning classification method is used to determine whether a tweet is related with a topic. Specifically, we are using a text classification method where tweets are the documents and topics are the labels.

Many text classification methods are based on a Bag of Words (BOW) for each class. A BOW is a set of keywords or key-phrases related with a topic; those keywords or key-phrases are usually composed only by adjectives, nouns and proper nouns. Thus, we need to build a BOWs for each topic. To do so, we are using an unsupervised approach for key-phrases extraction such as Tf-Idf and TextRank.

Finally, every user, who owns a tweet classified on a topic, will be added in the set which contains the authorities in the particular topic. The procedure of Identification of Authorities is composed of the following activities:

1. Real-time Tweet Crawling: This activity is responsible for catching tweets in real-time through the Twitter Streaming API. To capture tweets that might be related with a specific topic the process can filter tweets with basic keywords. The activity keeps metadata related with the user who is owner of the publication and with the tweet itself. Metadata related with the user are the following: id, creation date, screen name, description, and number of publications, followers and friends. Metadata related with the tweet are the following: creation date, number of retweets and favorites, hashtags, user mentions, URLs and the text content of the publication.
2. Preprocessing: This activity processes each captured tweet. It processes the description of the user and the text content of the tweet by eliminating user mentions, URLs, stop words and punctuations. Also, words that only make sense in the context of Twitter are eliminated. For instance, the word RT is used to let the users know that the tweet is a duplicate, so it should be removed.
3. Feature extraction: This activity builds the feature vector for each tweet. Construction of a feature vector is based on two BOWs. The first one corresponds to the keywords usually used in the description of the users that publish something related with a topic; the second one corresponds to the key-phrases usually used in the text content of tweets related with a topic. Each entry of the feature vector represents whether the user's description or the tweet's text-context that contains a word of the BOWs. Also, the feature vector includes other normalized data such as number of retweets, favorites, hash tags, user mentions and URLs.
4. Topic Classification: given the feature vector s and the topic classification model M , this activity determines the topic of the tweet. If the tweet is related to topic t , then the tweet owner is added to a set of authorities in topic t . We use a trained support vector machine to classify each topic.

Recommendation of Authorities

This procedure is responsible for ranking the authorities, obtained by the procedure of Identification of Authorities, to recommend k top authorities.

Ranking of authorities is based on graph theory. There are several methods that rank vertices according to the structure of the graph. PageRank, SimRank or Twitter Rank are some of the many methods that we can use to rank authorities. In our case, we use Page Rank.

The procedure of Recommendation of Authorities is composed by the following activities:

1. **User Information Extraction:** This activity extracts the following and follower relationships between each pair of users contained in the set of authorities returned by the procedure of Identification of Authorities. The relationships identified are added to a set of relationships.
2. **Graph Construction:** This activity builds a subgraph of the Twitter social graph, denoted with $G = (V, E)$, which is composed of two sets: V and E . The elements of V are the vertices of G (i.e. the authorities found). The elements of E are the directed edges of G (i.e. the relationship between the authorities).
3. **Ranking of Authorities:** This activity executes a rank method (PageRank) over the built social graph. Once the vertices are ranked, the activity returns the k top users.

Once the process returns the top users of the generated subgraph, the system publishes those users on the Twitter account of the platform to recommend them to the students of the courses.

4. Evaluation and Results

In this section, we present the results of the tests executed for the Identification of Authorities procedure.

Firstly, we defined a specific topic. Then, initial keywords related with the topic were selected. Next, a process of live download of tweets was executed for three days in a row. Fifteen thousand of tweets related with the topic were captured by the end of the process.

Two thousand randomly selected tweets from the previously captured set were employed as the initial subset to train a classification model to identify possible authorities in the topic. Random selection allows us to be neutral of the location where the tweets were published.

Authorities in the topic were in charge of the labelling task of the selected tweets. Only two hundred and seventy tweets (13.5 %) were labeled as tweets related with the topic. So, we assume that many Twitter users employed the

initial keywords in their tweets even if they were not related with the topic. The final subset was reduced to six hundred tweets, were two hundred and seventy tweets are related with the topic and rest are not.

Finally, one hundred simulations of the training and test of the classification model were executed using the final subset. As a result, the classification model has an average accuracy value of $[78.82 \pm 2.51]$ %.

5. Conclusions and Future Work

In this paper, we have presented the architecture of a recommender system that selects authorities on a particular topic based on their publications on Twitter. Experiments show that our system can identify authorities with an average accuracy of 78.8% and with a standard deviation of ± 2.51 %.

Currently, we are developing several tests to compare our proposal with other platforms and other recommended algorithms presented in the literature review.

Acknowledgment

This work has been sponsored by the Consorcio Ecuatoriano para el Desarrollo del Internet Avanzado (CEDIA) and the Universidad Politécnica Salesiana.

References

- [1] Castells G.M., La sociedad red, Madrid Alianza Editorial, (1996).
- [2] Castells M., Internet y la Sociedad Red, Lección inaugural del curso de Doctorado sobre la sociedad de la información y el conocimiento 2001-2002 de la Universitat Oberta Catalunya.
- [3] Krishnan K., Rogers S., Social data analytics. Waltham, MA: Morgan Kaufmann (2015), 76-91.
- [4] Zafarani R., Abbasi M., Liu H., Social media mining. New York, N.Y, Cambridge University Press (2014).
- [5] Adedoyin-Olowe M., Gaber M., Stahl F., A Survey of Data Mining Techniques for Social Network Analysis, Journal of Data Mining & Digital Humanities (2013).
- [6] Zaina L., Ameida T., Torres G, Can the Online Social Networks Be Used as a Learning Tool? A Case Study in Twitter, Communications in Computer and Information Science (2014), 114-123.
- [7] He W., Zha S., Li L., Social media competitive analysis and text mining: A case study in the pizza industry, International Journal of Information Management 33(3) (2013), 464-472

- [8] Pal A., Counts S., Identifying topical authorities in microblogs, Proceedings of the fourth ACM international conference on Web search and data mining (2011).
- [9] Bird S., Klein E., Loper E., Natural language processing with Python. Beijing: O'Reilly (2009).
- [10] Gupta P., Goel A., Lin J., Sharma A., Wang D., Zadeh R., Wtf: The who to follow service at twitter, Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee (2013).
- [11] Weng J., Lim E., Jiang J., He Q., Twitterank: finding topic-sensitive influential twitterers, Proceedings of the third ACM International conference on Web search and data mining (2010).
- [12] Pal A., Counts S., Identifying topical authorities in microblogs, international conference on Proceedings of the fourth ACM on Web search and data mining, (2011).
- [13] Ghosh S., Sharma N., Benevenuto F., Ganguly N., Gummadi, K., Cognos: crowd sourcing search for topic experts in microblogs. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (2012).

