Information Propagation Routes between Countries in Social Media

Didier Henry
University of French West Indies
Pointe A Pitre, France
didier.henry@univ-antilles.fr

Erick Stattner
University of French West Indies
Pointe A Pitre, France
erick.stattner@univ-antilles.fr

Martine Collard
University of French West Indies
Pointe A Pitre, France
martine.collard@univ-antilles.fr

ABSTRACT

Today, social media are one of the fastest ways to have access to information related to several topics. Indeed, a diffused information on these supports can travel thousands of kilometres in only few seconds contrary to an article posted on a news site. Despite the fact that a large variety of studies have been conducted to understand how fast and how scale information spreads in social media, we observe that they have not yet been interested in the geographical aspect. In this paper, we perform a geographical and temporal analysis of Twitter trends spread between May and June 2017. We introduce interesting patterns which deal with the paths taken by information between countries. In addition, we observe relevant results by taking into account the topic. Finally, we conclude and give perspectives of research of this work.

KEYWORDS

social network; geographical propagation; information diffusion; data analysis

ACM Reference Format:

Didier Henry, Erick Stattner, and Martine Collard. 2018. Information Propagation Routes between Countries in Social Media. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191569

1 INTRODUCTION

Nowadays, social media have an important place in the propagation of information. These platforms of exchange are used by several millions of people in the world who post millions of messages every single day. Information contained in messages can reach millions of people at the same time in just a few seconds. In order to quantify the impact of messages, several researchers were interested in parameters which influence the information propagation. Indeed, there is a lot of work in this field which has shown that the structure of the user local network[8], the number of communities[22] to which users are linked, the user position in the network[13], and the user category[4] affect the information flow in terms of speed and volume. Other studies have focused on the diffused message. In the literature, we found a lot of researchers who argue that the message polarity[11], the message topic[18] and the message content[19] have also an influence on its diffusion.

Despite the fact that studies go deeper and deeper into understanding the factors that allow the quantification of the impact of information in terms of volume and speed, the geographical dimension which is susceptible to influence the information propagation is still little studied. It is an important parameter to take into account because a posted information can potentially impact people according to its starting location. We argue that better knowledge on this subject could help to propose models able to predict geographic location of people impacted by information. Such models could be useful in marketing domain to target a specific audience. In another field, by knowing the next paths taken by the information, rumours could be attenuated or stopped in a specific geographical location. For all these reasons, we found interesting and innovative to study information propagation routes between countries.

In this paper, we perform a geographical and temporal analysis of 2923 Twitter trends spread between May and June 2017. We introduce interesting observations deal with the paths taken by information between countries by taking into account the topic. We argue that all these observations could be used for modelling task.

The rest of the article is organised as follows: Section 2 presents related works conducted on information diffusion and geographical data analysis. Section 3 describes the proposed approach. Section 4 details the observations. Finally, section 5 and presents our future directions.

2 RELATED WORK

Our work is at the crossroads of two lines of research: studies of diffusion information in social media, and geographical analysis of social media data.

Information diffusion in social media: A wide variety of studies have been conducted on social media to identify parameters which influence the diffusion in terms of volume[24](i.e., how much), speed[23](i.e., how fast). Some researchers have focused on the user social network. For instance, Cheng et al.[5] note that the local network structure of the original broadcaster has an impact on the reach of information. In addition, De Meo et al.[7] remark that weak ties increase the reach of the broadcast message. From the point of view of propagation speed, Vega-Oliveros et al.[21] indicate that individuals at the centre of the network have the ability to spread information faster than others. Other works have focused on the spread message to understand the diffusion process. For instance, Naveed et al.[16] observe that negative messages spread faster than positive ones. According to another observation point, some researchers[11] note that a positive information spreads more widely than neutral and negative ones. Other works have shown that the message topic may influence the diffusion. Indeed, Son et

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191569

al.[19] notice that social criticisms or problems against companies spread faster than promotions or discounts. Myers et al.[15] remark that messages about entertainment, business, or health have a wider reach than those on art, education or work. Finally, several studies[3, 20] note that characters or symbols occurring in the message text (like URL, hashtag) may also have an impact on the diffusion.

Geographical analysis of social media data: Some researchers [6] have proposed an approach to estimate a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geo-spatial cues. In the same idea, O'Connor et al. [9, 17] use the vocabulary variation of geotagged tweets in their model to classify a tweet into a geographic region. Other researchers[1] have worked on the Facebook network and have introduced an algorithm to predict the physical location of a user, given the known location of his friends. Another work[2] on geo-spatial properties of YouTube videos have indicated the highly local nature of video views. A study[14] close to our work have shown that user close in spatial distance tend to be close in time adoption of hashtags in Twitter. In another work [10], researchers have studied the origins and pathways of Twitter trends in 63 main US locations. However, to the best of our knowledge not any work in this field has focused on the information propagation routes between countries.

3 PROPOSED APPROACH

Our aim is to found interesting patterns which deal with the emergence of a trend in a specific country. We argue that observations on real data could be useful to propose realistic models to simulate the appearance of trends in countries.

In a first step, we have collected every five minutes, the list of 50 Twitter trending topics for 62 countries thanks to the Twitter API between May 29 2017 and June 15 2017. Based on the idea that once is chance, twice is a coincidence, third time is a pattern, we have selected trends which appear at least in the trend list of 4 countries in a three-day interval. Our dataset contains 2923 trends which match to this criterion.

In a second step, we have chronologically ordered the emergence of trends in each country (see Figure 1). Then, we have classified automatically trends according to their origin, i.e. the country or countries where the trend appears first. In addition, we have shorted manually trends according to the 8 following topics: people, games, movies/tv, music, news, politics, sport, and technology. We have added the category others for trends about other topics.

In a third step, we have extracted 3 parameters for each country in a trend pathway: the *adoption time*, the *infection time* and the *impact radius*. The first two concern the temporal dimension while the third is related to the geographical aspect.

The *adoption time* is the elapsed time between the appearance of a trend in the origin country or countries and the appearance of this trend in a country in the trend path.

The *infection time* is the elapsed time between the first appearance of the trend in the country list of trends and the moment she disappears from the list.

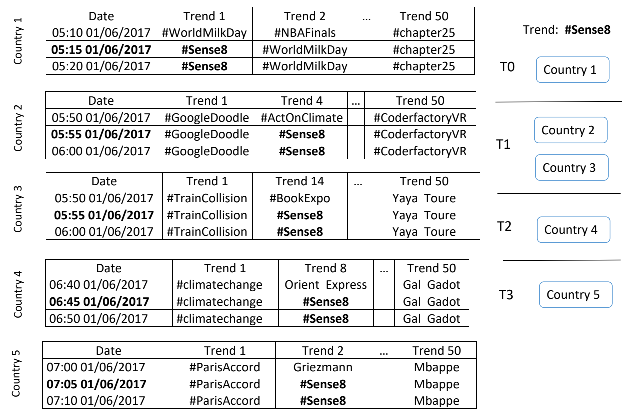

Figure 1: Methodology of trends extraction.

The *impact radius* is the geographic distance between the capital of the country at the origin of the trend and the capital of a country in the trend path.

4 OBSERVATIONS

Firstly, we observed the data distribution according the trend topic and the trend origin country. We remarked that Vietnam, the United States, Belarus and the United Kingdom are at the origin of half of the trends. In addition, we noticed sport seem an important Twitter trending topic.

Secondly, we observed the length of trends path, in other words, the number of countries impacted in function of the origin of the trend and according to the topic. For instance, we remark that trends related to technology seem impacted more countries than other topics (see Figure 3). Moreover, we notice that the number of countries affected by a trend seem influenced by the trend origin country (see Figure 2). Thus, the length of trends path varies both according to the topic and the origin of the trend. Therefore, these two parameters should be taken into account in models. In addition, we noticed that some countries never adopt first trends on a specific subject. For instance, we remarked that only 40% of countries are at the origin of trends about people (see Figure 4).

Then, we were interested in the temporal aspect of the data. We observed the adoption time and the infection time for according to the topic. We remark that the adoption time is different according to the topic. For instance, we note that the adoption time is less important for trends linked with games than other topics (see Figure 5). By observing the infection time, we notice that there are variations according to the topic. In addition, we see that the infection time of countries at the origin of the trend is greater than other countries in the trend path for some topic like people, movies, politics and sport (see Figure 6). Thus, the adoption time and the infection time are also two parameters which should be integrated in models.

Next, we observed the evolution of the radius impact according the topic. We remark that this parameter also varies according to the topic (see Figure 7). Although we do not present the observations of the adoption time, the infection time and the radius

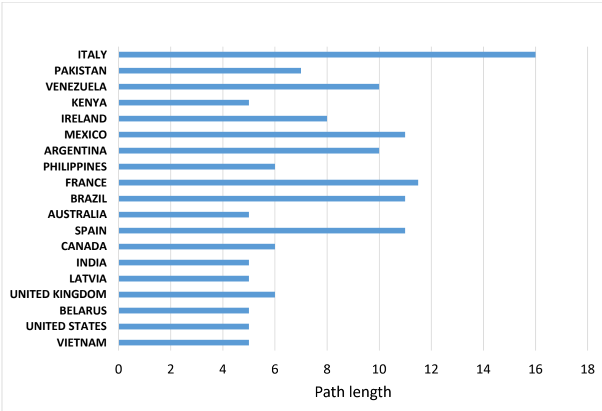

Figure 2: Median path length of trends according to the origin country (number of trends > 1%).

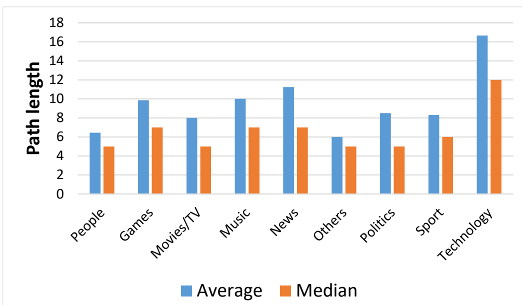

Figure 3: Median and average path length of trends according to the topic.

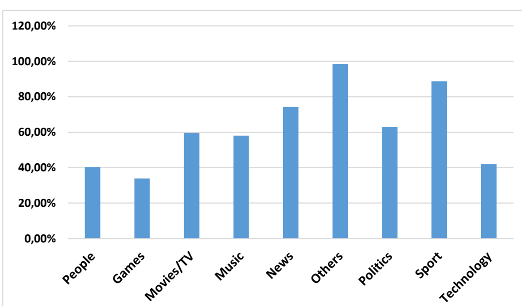

Figure 4: Percent of countries which adopt first a trend according to the topic.

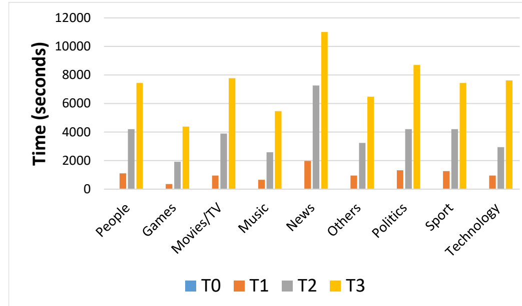

Figure 5: Median adoption time according to the topic.

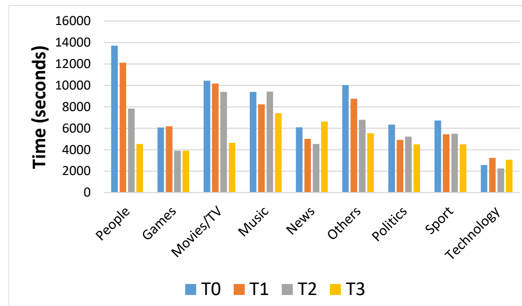

Figure 6: Median infection time according to the topic.

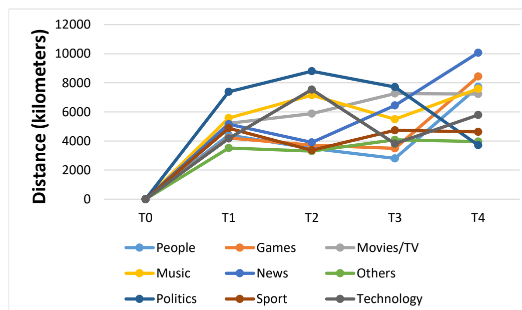

Figure 7: Median radius impact according to the topic.

impact according to the origin of the trend, we remark the value of these parameters is dependent on the origin country. Therefore, the adoption time, the infection time and the radius impact should be considered for modelling task.

Finally, we were interested in the path taken by the information. We have used Apriori and Fournier algorithms found in the SPMF tool[12] in order to discover frequent itemsets and sequential pattern. We did several experiments on the data. We present here, observations for trends appear in Vietnam, the United State, Belarus and the United Kingdom with the minimum support set to 40%, without take into consideration the topic. We remark that often some countries: Belarus, Latvia, Puerto Rico, US and Vietnam are often impacted together no matter the subject and the place of origin of the trend (see Table 1).

Then, we extracted sequential patterns on these countries. We introduce here the results obtained by set the minimum support to 20%. We observe that there are temporal sequences in adoption of trends (see Table 2), in other words there are frequent paths of trends apparition between countries.

5 CONCLUSION AND PERSPECTIVES

To the best of our knowledge, this work is the first study in the field of the information propagation in social media between countries.

Table 1: Itemsets with support min ≥ 40%.

Origin of the trend	Itemsets
Belarus	**{Belarus,Latvia,Vietnam}** 55% **{Belarus,Puerto Rico,Vietnam}** 65% **{Belarus,US,Vietnam}** 48%
United Kingdom (UK)	{Belarus,Puerto Rico,UK} 47% {Belarus,Puerto Rico,Vietnam} 47% {Belarus,Vietnam,UK} 84% {Puerto Rico,UK,Vietnam} 56% {Belarus,Puerto Rico,UK,Vietnam} 47%
United States (US)	{Belarus,Latvia,US} 44% **{Belarus,Latvia,Vietnam}** 44% {Belarus,Puerto Rico,US} 49% **{Belarus,Puerto Rico,Vietnam}** 49% **{Belarus,US,Vietnam}** 85% {Latvia,US,Vietnam} 46% {Puerto Rico,US,Vietnam} 58% {Belarus,Latvia,US,Vietnam} 43% {Belarus,Puerto Rico,US,Vietnam} 49%
Vietnam	**{Belarus,Latvia,Vietnam}** 46% **{Belarus,Puerto Rico,Vietnam}** 53% **{Belarus,US,Vietnam}** 46%

Table 2: Sequences with support min ≥ 20%.

Origin of the trend	Sequences
Belarus	T0: Belarus Latvia 23% T0: Belarus T1: Vietnam 29% T0: Belarus Puerto Rico Vietnam 31% T0: Belarus Vietnam 67%
United Kingdom (UK)	T0: UK T1: Belarus 29% T0: UK T2: Belarus 25% T0: UK T1: Vietnam T2: Belarus 20% T0: UK T1: Vietnam 53%
United States (US)	T0: US T1: Belarus 30% T0: US T2: Belarus 22% T0: US T1: Vietnam 48% T0: US Vietnam 38%
Vietnam	T0: Vietnam Puerto Rico 29% T0: Vietnam US 26% T0: Vietnam T1: Belarus 31%

In this paper, we have addressed the diffusion of information between countries. Our contributions can be summarised as follows.

(1) We have proposed an approach to extract a path adoption between countries related to a trending topic on Twitter.
(2) A first study has allowed to observe that the path adoption length of a trend varies according to the topic and the origin of the trend. Moreover, we have seen that some country never adopt first trends on a specific subject.
(3) In a second step, we have noticed that the infection time, the adoption time, and the radius impact of trend varies according to the topic and the origin of the trend. In addition, we have remarked that countries which adopt a trend first seem to have a greater infection time than others on some topics.
(4) Thirdly, we have noted that some countries are often impacted together no matter the subject and the place of origin of the trend. Furthermore, we have observed that there are frequent sequences temporal in the adoption of a trend by countries.

In a forthcoming study, we plan to propose a model which takes into account the topic of the trend, the place of origin of the trend, the adoption time, the infection time and the radius impact in order to predict the countries impacted by the trend and the order of adoption.

REFERENCES

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*. ACM, 61–70.

[2] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. 2012. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 241–250.

[3] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 675–684.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM* 10, 10-17 (2010), 30.

[5] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 925–936.

[6] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 759–768.

[7] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2014. On Facebook, most ties are weak. *Commun. ACM* 57, 11 (2014), 78–84.

[8] P Alex Dow, Lada A Adamic, and Adrien Friggeri. 2013. The Anatomy of Large Facebook Cascades.. In *ICWSM*.

[9] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1277–1287.

[10] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. 2013. Traveling trends: social butterflies or frequent fliers?. In *Proceedings of the first ACM conference on Online social networks*. ACM, 213–222.

[11] Emilio Ferrara and Zeyao Yang. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1 (2015), e26.

[12] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. 2016. The SPMF Open-Source Data Mining Library Version 2. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 36–40.

[13] Caroline Haythornthwaite. 1996. Social network analysis: An approach and technique for the study of information exchange. *Library & information science research* 18, 4 (1996), 323–342.

[14] Krishna Y Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 667–678.

[15] Seth A Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 33–41.

[16] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, 8.

[17] Brendan O'Connor, Jacob Eisenstein, Eric P Xing, and Noah A Smith. 2010. Discovering demographic language variation. (2010).

[18] Daniel M. Romero, Brendan Meeder, and Jon M. Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*. ACM, 695–704.

[19] Insoo Son, Dongwon Lee, and Youngkyu Kim. 2013. Understanding the Effect of Message Content and User Identity on Information Diffusion in Online Social Networks.. In *PACIS*. 8.

[20] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 177–184.

[21] Didier A Vega-Oliveros, Luciano da F Costa, and Francisco A Rodrigues. 2016. Rumor propagation with heterogeneous transmission in social networks. *arXiv preprint arXiv:1610.01012* (2016).

[22] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting successful memes using network and community structure. *arXiv preprint arXiv:1403.6199* (2014).

[23] Shaomei Wu, Chenhao Tan, Jon M Kleinberg, and Michael W Macy. 2011. Does Bad News Go Away Faster?. In *ICWSM*. Citeseer.

[24] Jiang Yang and Scott Counts. 2010. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *ICWSM* 10 (2010), 355–358.