Characterizing Topic-Specific Hashtag Cascade in Twitter Based on Distributions of User Influence

Geerajit Rattanaritnont, Masashi Toyoda, and Masaru Kitsuregawa

The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{aomi,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp
http://www.u-tokyo.ac.jp

Abstract. As online social networks become extremely popular in these days, people communicate and exchange information for various purposes. In this paper, we investigate patterns of information diffusion and behaviors of participating users in Twitter, which would be useful to verify the effectiveness of marketing and publicity campaigns. We characterize Twitter hashtag cascades corresponding to different topics by exploiting distributions of user influence; cascade ratio and tweet ratio. The cascade ratio indicates an ability of users to spread information to their neighborhoods, and the tweet ratio measures how much each user participates in each topic. We examined these two measures on a real Twitter dataset and found three major diffusion patterns over four topics.

Keywords: Data mining, information diffusion, social network

1 Introduction

In addition to real world communication, people can now keep in touch with each other on social networking sites such as Facebook, Twitter, and MySpace. People connecting to online social networks can share interests and activities with their friends, and even make new friends all over the world. The resulting networks grow rapidly and gained significant popularity on the Internet.

Many researchers have studied various aspects of online social networks such as network structure, user relationships, and information flow between users. In this paper, we perform a research on Twitter's user networks to understand patterns of information diffusion and behaviors of participating users. This would be useful to verify whether publicity campaigns are successful according to marketing strategies, such as aiming to spread mentions on new products to a large number of people or to a specific group of fans.

We analyze the diffusion of information according to topics of most frequently used hashtags and find the characteristics across them. To study a large amount of data, we consider two probability distributions of user influence: cascade ratio

and tweet ratio. The cascade ratio indicates an ability of a user to cascade information to his neighborhoods and the tweet ratio measures how much each user involves in each topic. We examine these two measures on a real Twitter dataset.

The Twitter dataset used in this paper is crawled from March 11, 2011 to July 11, 2011. It consists of 260 thousand users and 783 million tweets. We select top 100 frequently used hashtags from the dataset and categorize them according to topics. We found that the majority fall into four topics which are earthquake, politics, media, and entertainment. We then further investigate patterns of information diffusion by clustering hashtags based on distributions of those measures. Our results show that there are three patterns of hashtag cascade among four different topics.

The rest of this paper is organized as follows. Section 2 introduces related work on information diffusion. Section 3 explains the dataset. In Section 4, we describe two proposed distributions and investigate the characteristics of information diffusion over four major topics. Then we conduct further analysis by using clustering algorithm in Section 5. Finally, we conclude this paper and future work in Section 6.

2 Related Work

Information diffusion in online communities has been studied for a decade. Gruhl *et al.* [4] studied the dynamics of information propagation in weblogs. They investigated characteristics of long-running topics due to outside world events or within the community. Leskovec *et al.* [6] also studied information propagation in weblogs. They proposed a simple model that mimics the spread of information in blogspace and is similar to propagation found in real life.

Instead of blogsphere, Liben-Nowell *et al.* [7] traced the spread of information at individual level and found that information reach people in a narrow deep pattern, continuing for several hundred steps. Similarly, Sun *et al.* [10] conducted an analysis on information diffusion in Facebook and discovered that large cascade begins with a substantial number of users who initiate short chains.

In most recent years, as Twitter becomes one of the most popular microblogging services and allows us to obtain its data via Twitter API, it gains much interest from many researchers [2, 3, 5, 8, 9, 11–13]. Romero *et al.* [9] studied information spread in Twitter and showed that controversial political topics are particularly persistent with repeated exposures comparing to other topics. Bakshy et al. [1] exploited information cascade to identify influencers in Twitter. Unlike others, we study characteristics of information diffusion over different topics in Twitter in term of cascade ratio and tweet ratio. Because we will understand how people interact with each other and how information is cascaded. Rather than identifying influencers as in [1, 3, 5, 12], we can utilize this work to verify success of viral marketing strategy.

Table 1. Examples of hashtags in each topic

Topic	Examples	Total
Earthquake	jishin, genpatsu, prayforjapan, save_fukushima, save_miyagi	21
Politics	bahrain, iranelection, wiunion, teaparty, gaddafi	32
Media	nicovideo, nhk, news, fujitv, ntv	13
Entertainment	madoka_magica, akb48, atakowa, tigerbunny, anohana	11

3 Twitter Dataset

We crawled the Twitter dataset from Twitter API from March 11, 2011 when the Great East Japan Earthquake took place to July 11, 2011. Our data collection consists of user profiles, timestamp and tweet contents including retweets. We started crawling from famous Japanese users. We firstly got timelines of these users, then repeatedly expanded the set of users by tracing retweets and mentions in their timelines. We then obtained 260 million users as active users and 783 million tweets. Instead of friend-follower graph, we regard directed links among users when user A has at least one retweet from or mention to user B and call this relationship as outgoing neighborhood. We extracted 31 million links by considering only active users.

To study information cascade according to different topics, we treat a hashtag as a representative of the topic users talk about. We select top 100 frequently used hashtags from the dataset and manually categorize them into topics. We found that the majority belong to four major topics which are earthquake, politics, media, and entertainment. Table1 shows examples of hashtags in each topic. Earthquake topic is about the Great East Japan Earthquake, while politics topic is related to political issues and events all over the world especially the uprising events in the Middle East. Media topic is represented by communication channels, such as, television networks. Lastly, entertainment topic refers to television programs, movies and artists.

4 User Influence Distributions

4.1 Cascade Ratio

Cascade ratio determines the proportion of how much a user can influence his neighborhoods to spread a hashtag comparing to all users who used the same hashtag. We captured the cascade by tracing the time each user firstly used a given hashtag. Thus, cascade score of a user is the number of his immediate incoming neighborhoods that reposted the hashtag after him. The cascade ratio cr of a user u posting hashtag h is then defined as below:

$$cr(u,h) = \frac{C(u,h)}{U(h)} \ . \tag{1}$$

where $C(u,h)$ is the cascade score of the user u posting the hashtag h and $U(h)$ is a set of all users using h.

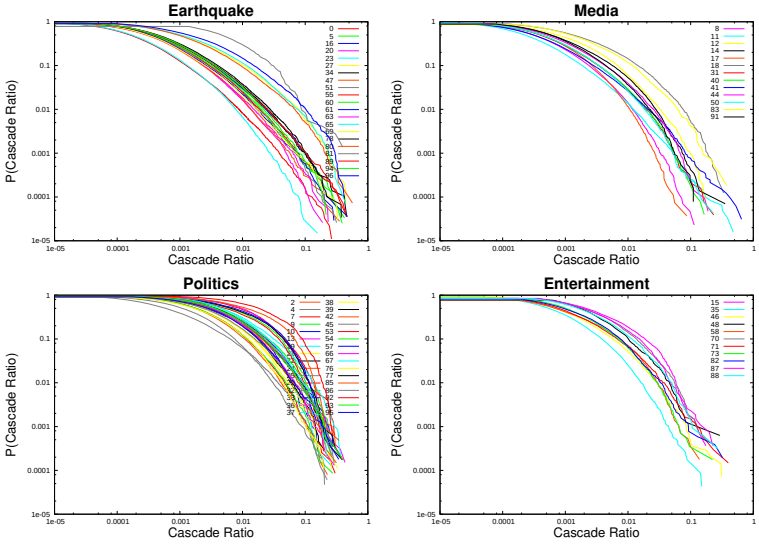


Fig. 1. Cascade ratio distributions of all hashtags in each topic

Fig.1 illustrates the probability distributions of cascade ratio of all hashtags according to four topics. The x-axis is cascade ratio and the y-axis is the number of occurrences of cascade ratios normalized by total number of users using a given hashtag. The plot is in log-log coordinate and calculated as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x.

According to Fig.1, the earthquake and the media topics start to fall down at small cascade ratio. That means a number of people in these two topics have relatively low cascade ratio. It implies that people used those hashtags independently not because of seeing the hashtags in their friends' tweets. In other words, the hashtags themselves are hot topics or general words so that people know them already. For example, "jishin" in the earthquake category means earthquake in Japanese language and "nhk" in the media category is Japan's national public broadcasting organization. In contrast, the politics and the entertainment topics fall down at higher cascade ratio. We can say that a number of people have relatively high cascade ratio. When users post a hashtag in these topics, many of their friends will also post it after them. It means that there exist groups of users responding frequently each other, such as discussion communities on some political topics or fans of popular artists. For instance, "sgp" in the politics category is a non-profit organization for conservative women activists and "akb48" in the entertainment category is a popular Japanese female idol group.

For easy to see the difference among four topics, Fig.3a shows the point-wise average distributions of the cascade ratio for each topic. We see that 90% of all users who use hashtags in the earthquake and the media topics have cascade ratio

less than 0.005 which means they directly influence less than 0.5% of all users, while 2.7% and 0.8% for the politics and the entertainment topics respectively.

4.2 Tweet Ratio

The second measure is tweet ratio which determines how much each user engages in each topic. The tweet ratio tr of a user u posting hashtag h is then simply defined as below:

$$tr(u,h) = \frac{T(u,h)}{\sum_u T(u,h)} \ . \tag{2}$$

where $T(u,h)$ is the number of tweets containing the hashtag h posted by the user u.

Fig.2 shows the probability distribution of tweet ratio of all hashtags according to four topics. The x-axis is tweet ratio and the y-axis is the number of occurrences of tweet ratios normalized by total number of users using a given hashtag. Each line is plotted in log-log coordinate and calculated as a cumulative distribution function.

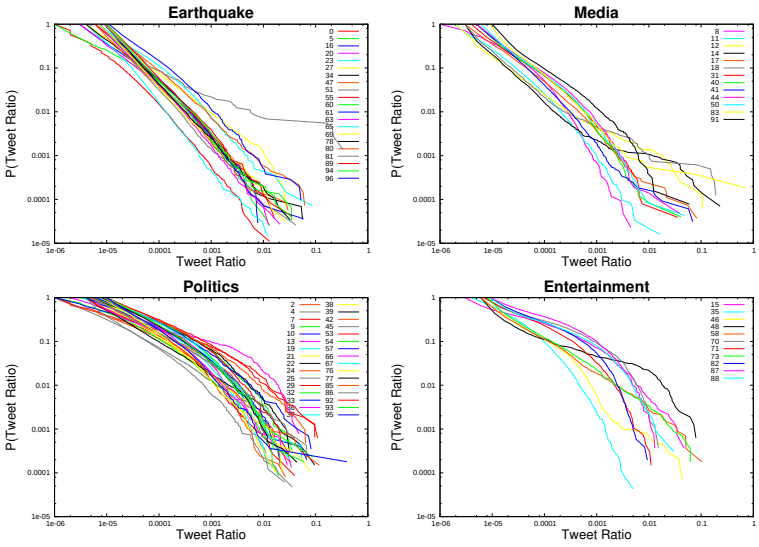
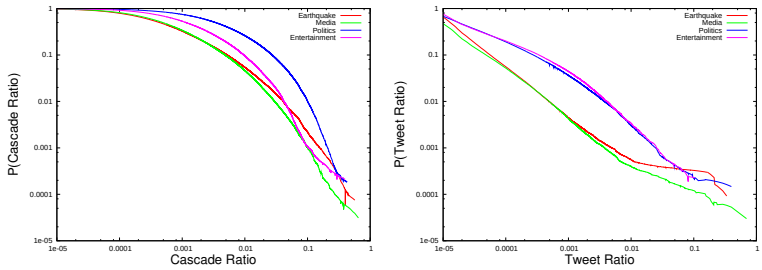


Fig. 2. Tweet ratio distributions of all hashtags in each topic

We see that the earthquake and the media topics follow the power-law, which means a number of people in the these topics have comparably low tweet ratio. They posted tweets containing hashtags but repeated to use those hashtags very few times. Alternatively, the politics and the entertainment topics are more curved. More people repetitively used same hashtags many times than people in the first two topics.



(a) Cascade ratio (b) Tweet ratio

Fig. 3. Point-wise average distributions of four topics

Table 2. Proportion of each topic in each cluster when $k = 4$

No. of hashtags	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Total
Earthquake	1	13	4	3	21
Media	0	11	1	1	13
Politics	24	1	7	0	32
Entertainment	3	0	5	3	11

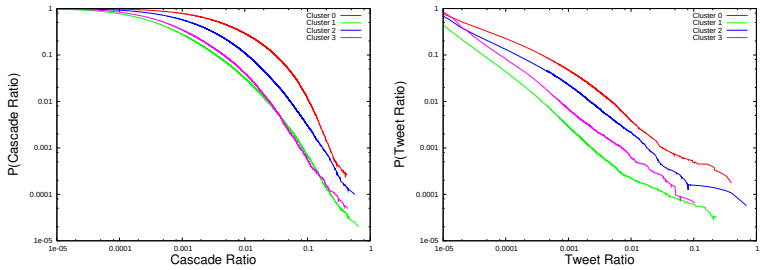
Additionally, Fig.3b demonstrates the point-wise average distributions of tweet ratio for each topic. We see that they are separate clearly into two groups. In more details, 90% of all users who use hashtags in the earthquake and the media topics have tweet ratio less than 0.00005, while 0.0003 in case of the politics and the entertainment topics.

5 Information Cascade Clustering

In this section, we further investigate the characteristics of information diffusion by using clustering algorithm. We performed k-means clustering based on the distributions of both cascade ratio and tweet ratio. Each hashtag is represented as a vector of values at n points in each distributions. For each hashtag, we selected 92 points proportional to the log scale from each distribution. We use Euclidean distance as a distance measure. Table2 illustrates the proportion of four topics assigned to each cluster when we choose the number of clusters as k = 4 as equal to the number of topics.

Fig.4 shows the point-wise average distributions of cascade ratio and tweet ratio according to four clusters. There are three main patterns among four clusters. Cluster 0 has high cascade ratio and high tweet ratio, cluster 1 has low cascade ratio and low tweet ratio, and cluster 2-3 are in the middle between them.

The majority of hashtags in cluster 0 are from the politics category, the majority of hashtags in cluster 1 are from the earthquake and the media categories, and the majority of hashtags in cluster 2 and 3 are from the entertainment category. However, we can see that cluster 2 includes various topics not only the



(a) Cascade ratio (b) Tweet ratio

Fig. 4. Point-wise average distributions of four clusters

entertainment but also the earthquake, the media, and the political topics. For example, there are seven hashtags from the earthquake assigned into these clusters such as "iwakamiyasumi" and "hinan", which are related to nuclear power plants and safety information. They are thus considered to be used in longer period comparing to other earthquake hashtags in cluster 1. Moreover, there are seven hashtags from the politics category put into cluster 2. Almost of them are some of hashtags related to the Middle East such as "bahrain" and "egypt". Since they are country names, they are probably not limited only to political topic and thus more general than other political hashtags in cluster 0.

6 Conclusion

In this paper, we studied the characteristics of information diffusion according to topics of most frequently used hashtags in Twitter. Here we focused on four major topics which are earthquake, politics, media, and entertainment. According to the distributions of cascade ratio and tweet ratio, we have found that people in earthquake and media topics have less influence and use the same hashtags fewer times than people in politics and entertainment topics. Then, we also performed k-mean clustering based on both cascade ratio and tweet ratio. The results showed that we have three main patterns among four topics. The earthquake and the media topics have low cascade ratio and low tweet ratio, the politics topic has high cascade ratio and high tweet ratio, and the entertainment topic is in the middle among them. Finally, as future work, we need to explore other possible features, such as a number of followers and information cascade in term of geographical and temporal aspects, to characterize entertainment hashtags.

References

1. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an Influencer: Quantifying Influence on Twitter. In: 4th International Conference on Web Search and Data Mining, pp. 65–74. ACM (2011)

2. Castillo, C., Mendoza, M., Poblete, B.: Information Credibility on Twitter. In: 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K. P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: 4th International Conference on Weblogs and Social Media, pp. 10–17. AAAI (2010)
4. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion Through Blogspace. In: 13th International Conference on World Wide Web, pp. 491–501. ACM (2004)
5. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media?. In: 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)
6. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Patterns of Cascading Behavior in Large Blog Graphs. In: 7th SIAM International Conference on Data Mining, pp. 551–556. SIAM (2007)
7. Liben-Nowell, D., Kleinberg, J.: Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data. In: The National Academy of Sciences, pp. 4633–4638. PNAS (2008)
8. Meeder, B., Karrer, B., Sayedi, A., Ravi, R., Borgs, C., Chayes, J.: We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter. In: 20th International Conference on World Wide Web, pp. 517–526. ACM (2011)
9. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In: 20th International Conference on World Wide Web, pp. 695–704. ACM (2011)
10. Sun, E., Rosenn, I., Marlow, C., Lento, T.: Gesundheit! Modeling Contagion through Facebook News Feed. In: 3rd International Conference on Weblogs and Social Media, pp. 146–153. AAAI (2009)
11. Scellato, S., Mascolo, C., Musolesi, M., Crowcroft, J.: Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. In: 20th International Conference on World Wide Web, pp. 457–466. ACM (2011)
12. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: Finding Topic-Sensitive Influential Twitterers. In: 3rd International Conference on Web Search and Data Mining, pp. 261–170. ACM (2010)
13. Wu, S., Hofman, J. M., Mason, W. A., Watts, D. J.: Who Says What to Whom on Twitter. In: 20th International Conference on World Wide Web, pp. 705–714. ACM (2011)