修　士　論　文

A Study on Patterns of Information Cascades in Microblogs based on Distributions of Users' Influence and Posting Behaviors

マイクロブログにおけるユーザの影響力および投稿行動の分布に基づく情報伝播パターンに関する研究

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-106439　Geerajit Rattanaritnont

指導教員　喜連川 優　教授

平成24年 2月 8日

Abstract

As online social networks become extremely popular in these days, people communicate and exchange information for various purposes. We realize that different activities tend to have different ways information spread on the network. Knowing patterns of information cascade would help organizations to examine behaviors of public relation campaigns.

In this thesis, we perform a research on Twitter's user network to understand patterns of information cascade and behaviors of participating users in various topics. We verify whether different topics really have different cascade patterns or not by exploring four measures, which are cascade ratio, tweet ratio, time interval, and exposure curve. We conduct experiments on a real Twitter dataset. We consider Twitter hashtags as representatives of topics and obtain six major topics, which are earthquake, media, politics, entertainment, sports, and idiom.

We firstly study the pattern of hashtag cascades in each topic by using statistical approach, then further investigate the relationship between cascade patterns and topics by using clustering algorithm, and lastly verify the effectiveness of each measure due to the clustering results.

Our experiments show that hashtags in different topics have different cascade patterns in term of cascade ratio, tweet ratio, time interval, and exposure curve. For example, the earthquake topic has low cascade ratio, low tweet ratio, short lifespan, and high persistence, while the political topic has high cascade ratio and high persistence. However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We discover that such kind of hidden relationship between topics can be surprisingly revealed by using only four measures rather than considering tweet contents. Finally, among four measures we explored, our results also showed that cascade ratio and time interval are the most effective measures to distinguish cascade patterns in different topics, while tweet ratio and exposure curve from the related work are not effective as we expected.

Acknowledgment

First and foremost, I would like to thanks Kitsuregawa-sensei to give me a big chance to be a member of this laboratory. I am very proud to be a student under his supervision. I also would like to thanks Toyoda-sensei for always being a great teacher. He has spent a lot of his precious time giving very useful suggestions. This thesis and the rest of my publications could not be achieved without him. I am really appreciate his contribution to a student like me. Moreover, I would like to thanks Nakano-sensei to support me in both research and social life.

Next, my sincere thanks to Yokoyama-san, Yoshinaga-san, Kaji-san, Ito-san, and Yang-san for their kind supports. All of their comments are very meaningful to my research. I am also thankful for Takaku, Nakamura, Fujikawa, Tei, and Shi for being good friends and sharing both good and hard times together.

And, last but not least, big thanks to my family for warm encouragement that helps me overcoming many difficult times and accomplishing the goals in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays people can 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. Information is then said to be cascaded over the Internet. For example, people in Japan spread "Operation Yashima" on Twitter to conserve electricity due to the Great East Japan Earthquake. This kind of situation is an emergency and needs to be reached a large number of people within short time. Unlike other activities, for instance, Fukushima Daiichi Nuclear Power Plant faced failures according to the Great East Japan Earthquake. Because this is a serious problem and cannot be solved immediately, much of discussion and concerns are continually talked by people including experts.

Since different activities tend to have different ways information spread on the network, studying patterns of information cascade would help organizations to examine behaviors of public relation campaigns. Therefore, in this thesis, we perform a research on Twitter's user network to understand patterns of information cascade and behaviors of participating users in various topics such as earthquake and political topics. We verify whether different topics really have different cascade patterns or not by exploring four measures, which are cascade ratio, tweet ratio, time interval, and exposure curve. The cascade ratio determines how much people can influence their friends, the tweet ratio determines how much people talk in each topic, the time interval determines how long a topic is still popular in the network, and lastly the exposure curve determines how easy people are influenced by their friends. We consider Twitter hashtags as representatives of topics and conduct experiments 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 500 frequently used hashtags from the dataset and categorize them according to topics. We found that the majority fall into six topics which are earthquake, media, politics, entertainment, sports, and idiom. We firstly study the pattern of hashtag cascades in each topic by using statistical approach. We then further investigate the relationship between cascade patterns and topics by using clustering algorithm. Our results show that hashtags in different topics have different cascade patterns in term of cascade ratio, tweet ratio, time interval, and exposure curve. For example, the earthquake topic has low cascade ratio, low tweet ratio, short lifespan, and high persistence, while the political topic has high cascade ratio and high persistence. However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We discover that such kind of hidden relationship between topics can be surprisingly revealed by using only four measures rather than considering tweet contents. Finally, among four measures we explored, our results also show that cascade ratio and time interval are the most effective measures to distinguish cascade patterns in different topics, while tweet ratio and exposure curve from the related work are not effective as expect.

The rest of this thesis is organized as follows. Chapter 2 introduces related work on information diffusion in online blogging and social networking services. Chapter 3 explains the dataset. In Chapter 4, we describe four measures of users' influence and posting behaviors, and investigate the characteristics of information diffusion over six major topics. Then we conduct further analysis by using clustering algorithm in Chapter 5. Finally, we conclude this paper and future work in Chapter 6.

Chapter 2

Related Work

Information diffusion in online community has been studied for a decade. Gruhl *et al.* [6] studied the dynamics of information propagation in weblogs. They developed a model to observe characteristics of discussion topics generated by outside world events and resonances within the community. For individual level, they proposed another model based on the theory of infectious diseases to identify particular users who potentially effect the spread of information. Adar *et al.* [1] developed a tool to visualize the flow of individual URLs over a blog network. Leskovec *et al.* [11] also studied information propagation in weblogs in term of temporal and topological aspects. In temporal aspect, they found that blog posts have weekly periodic behavior and the popularity of them decays by a power law instead of exponential function. In topological aspect, most of cascades are star shape, that is, a single post contains several links, but is not itself linked from others.

Instead of blogsphere, Watts [18] established a simple model of information cascades on random networks. He found that the cascades follows a power-law distribution when the connectivity of the network is sparse, but corresponds to a bimodal distribution when the connectivity is dense. Leskovec *et al.* [9] observed the propagation of person-to-person recommendations on an e-commerce site for viral marketing purpose. Kempe *et al.* [7] proposed a framework for selecting a subset of individuals who are able to drive a large cascade of adoptions in social networks. Newman *et al.* [14]

studied the mechanism of computer viruses spread on email networks. Liben-Nowell *et al.* [12] traced the spread of Internet chain letters at a person-to-person level. Leskovec *et al.* [10] also described temporal patterns of news cycle by tracking the dynamics of information diffusion between media sites and blogs. Baksy *et al.* [3] studied the propagation of contents in Second Life, an online virtual world. They further constructed a model of social influence based on the adoption rate. Sun *et al.* [16] conducted an analysis on information diffusion in Facebook. They discovered that large cascade begins with a substantial number of users who initiate short chains opposite to theoretical literature, assuming that it starts from a small number of users who generate large chains. By using zero-inflated negative binomial regressions, they also found that users' demographics and their profiles cannot be exploited to predict maximum diffusion size of each initial user.

In most recent years, as Twitter becomes one of the most popular micro-blogging services and allows us to obtain its data via Twitter API, it gains much interest in various aspects. Kwak *et al.* [8] conducted a quantitative study on topological characteristics of Twitter, and its role as a new information sharing medium, such as temporal behavior of trending topics and user participation. Cha *et al.* [5] focused on a concept of user influence. They analyzed three measures to identify influential users, which are indegree, retweets, and mentions. Weng *et al.* [19] proposed another measure called TwitterRank to find influential users. They adapted PageRank algorithm by considering topical similarity between users and their friends.

In addition to determining influential users, Castillo *et al.* [4] proposed a method to automatically judge messages posted on Twitter whether they are credible facts or false rumors. Meeder *et al.* [13] developed another method to infer link creation times by using only a single snapshot of network and user account creation times.

Furthermore, there are several researches on information diffusion in Twitter, on which we mainly focus in this thesis. Wu *et al.* [20] investigated the production, flow, and consumption of information on Twitter among celebrities, bloggers, media, organizations, and ordinary users. Romero *et al.* [15] investigated the ways information diffuses on different topics. They defined exposure curve as a characteristic of information diffusion and found that controversial political topics are particularly persistent comparing to other topics. We then utilize the exposure curve together with other proposed measures to find patterns of information diffusion across topics. We will explain about this in detail in Section 4.4. Rather than understanding how information itself spreads, it can be exploited for various purposes. Bakshy et al. [2] defined information cascade as a measure of user influence. They tried to predict individual influence by using both cascade size and user profiles. Scellato et al. [17] studied whether information cascades in geographically global or local area. They took advantage of this finding to improve cache replacement policies of mulitimedia files in a Content Delivery Network.

Although various measures are studied to explain the patterns of information cascade, there are possibly more standard measures to distinguish them in different topics, for instance, earthquake and political topics. Besides, it is still unclear which measure are the most effective. We thus explore four simple measures, which are cascade ratio, tweet ratio, time interval, and exposure curve, to express the cascade patterns and finally verify the effective of each measure in our experiments.

Chapter 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 and mentions. 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. As a result, we obtained 260 million users and 783 million tweets. Our interested users, network, and hashtags are defined respectively in following sections.

3.1 Users

In this thesis, we consider users who have at least one tweet during the period of dataset. Therefore, we have 260 thousand users as active users.

3.2 Network

Because retweet-mention relationship provides stronger sense of user interaction than just friend-follower relationship, 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. Hence, we can imply that user A subscribes to user B's updates. We extracted 31 million links by considering only active users.

3.3 Hashtags

In order to study information cascade according to different topics, we treat a hashtag as a representative of the topic users talk about. Although URL is another choice, we choose hashtag over URL because it provides the sense of topic more comprehensive than URL. In other words, URL is too specific. One topic can be indicated by a large number of URLs.

Table 3.1: Examples of hashtags in each topic

Topic	Examples	Total
Earthquake	jishin, genpatsu, prayforjapan, save_fukushima, save_miyagi	48
Media	nicovideo, nhk, news, fujitv, cnn	46
Politics	bahrain, iranelection, wiunion, teaparty, gaddafi	94
Entertainment	madoka_magica, akb48, atakowa, tigerbunny, anohana	65
Sports	hanshin, f1jp, dragons, sbhawks, cwc2011	20
Idiom	nowplaying, shoutout, followme, justsaying, pickone	35

We select top 500 frequently used hashtags from the dataset and manually categorize them according to topics. Moreover, to provide meaningful distributions in the rest of this thesis, we focus only on hashtags that have at least 1,000 participating users. Consequently, we found that the majority belong to six major topics each of which has at least 20 hashtags. They are earthquake, media, politics, entertainment, sports, and idiom topics. Table 3.1 shows examples of hashtags in each topic. All of top 500 frequently used hashtags and their corresponding topics are listed in Appendix A.

First, earthquake topic is mainly about the Great East Japan Earthquake. Second, media topic is represented by communication channels, such as, television networks, news channels, and video sharing websites. Most of them are Japanese channels, e.g., "nhk" and "fujitv" hashtags. Third, politics topic is related to political issues and events all over the world. Approximately half of them refers to the uprising events in the Middle East, e.g., "bahrain" and "gaddafi" hashtags. Forth, entertainment topic refers to television programs, musics, and artists. The majority are again Japanese animations, e.g., "madoka_magica" and "tigerbunny" hashtags. Fifth, sports topic corresponds to sports teams and tournaments. Most of them are Japanese baseball teams, e.g., "hanshin" and "dragons" hashtags. Finally, idiom topic is a popular phrase used as Twitter culture, e.g., "shoutout" and "justsaying" hashtags. Although it is still unclear that the idiom topic should be really treated as the topic or not, we include this in our work because it was studied by Romero *et al.* [15].

Chapter 4

Distributions of Users' Influence and Posting Behaviors

In this section, we define three distributions of users' influence and posting behaviors, which are cascade ratio, tweet ratio, and time interval. Besides, we exploit existing exposure curve [15] as an additional distribution. Then we observe patterns of hashtag cascade in different topics by using four distributions above. The definition and the analysis result of each distribution will be explained in following section respectively.

4.1 Cascade Ratio

Cascade ratio determines the proportion of how much a user can influence his/her neighborhoods to spread a hashtag comparing to all users who used the same hashtag. Before seeing how to calculate the cascade ratio, it is necessary to understand the definition of more basic element, which is cascade score. The cascade score of a user is a number of his/her immediate incoming neighborhoods that reposted the hashtag after him/her. For example, our user network is shown in Fig.4.1. A node and a directed edge in the graph represents a user and a link of our network respectively. When user A has

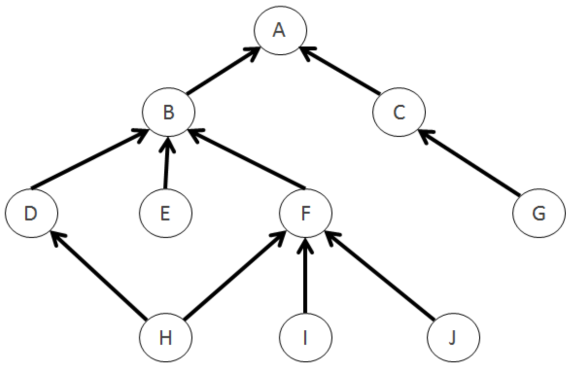


Figure 4.1: An example of user network

link from user B, it means user B has ever retweeted from or mentioned to user A at least one time. We can imply that user B has subscribed for user A's updates.

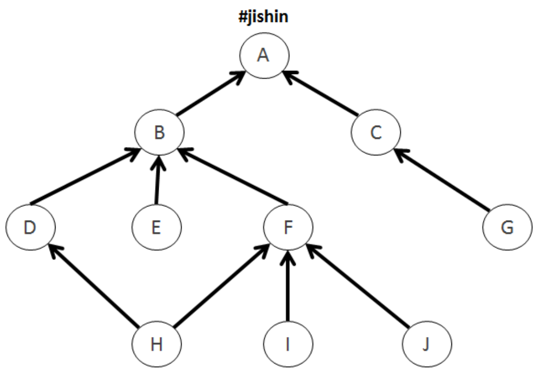
We then captured the cascade by tracing the time each user firstly used a given hashtag. The cascade score of a user is defined as a number of his/her immediate incoming neighborhoods that reposted the hashtag after him/her. Given the "jishin" hashtag, we assume that the cascades take place over the user network as in Fig.4.2. User A starts to post "jishin" at $t = 1$, then user B, C, and F post "jishin" after user A at $t = 2$. Because user A has incoming links from only user B and C, the cascade score of user A is two which refers to user B and C. Next, at $t = 3$, user E starts to use "jishin". In this case, although user B has incoming links from both user E and F, only user E posts the given hashtag after user B. The cascade score of user B is thus one which refers to user E.

The cascade ratio cr of a user u posting a hashtag h is now defined as below:

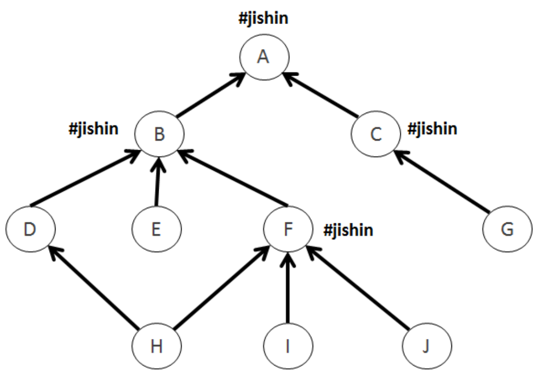
$$cr(u, h) = \frac{C(u, h)}{U(h)} \tag{4.1}$$

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.

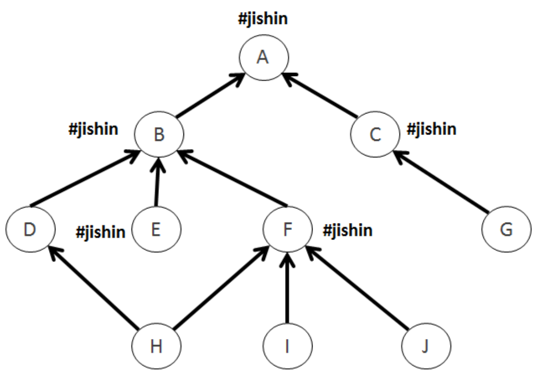
Fig.4.3 illustrates the probability distributions of cascade ratio of all hashtags according to six topics which are earthquake, media, politics, entertainment, sports, and idiom respectively. x is cascade ratio and y is the number



(a) $t = 1$



(b) $t = 2$



(c) $t = 3$

Figure 4.2: An example of hashtag cascade

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.

Each line remains horizontally at the beginning and then starts to fall down at each cascade ratio assigned to a user. Between any two points, the higher slope is, the more users have those corresponding cascade ratio values. However, it is difficult to conclude the characteristics of each topic because the distributions in each topic have wide range of values. Fig.4.4 shows point-wise average cascade ratio distributions. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is 90% confidence interval. In addition to the point-wise average distributions, we calculate the 90% bootstrap confidence intervals to test a null hypothesis. Our null hypothesis is that the particular topic has no difference in cascade ratio from a set of all hashtags. We sample n hashtags at random and calculate the point-wise average distribution, where n is the number of hashtags in the particular topic. After resampling the sample 1000 times with replacement, we have the histogram of 1000 sample means at each point. We then pick off the 5^{th} and 95^{th} percentiles as the 90% confidence intervals.

According to Fig.4.4, only 90% confidence interval of the entertainment topic includes its average distribution. In this case, we cannot reject the null hypothesis. That means we cannot conclude by 90% confidence level that the entertainment topic has no difference in cascade ratio from the set of all hashtags. In contrast, 90% confidence intervals of the earthquake, media, politics, sports, and idiom topics do not contain their corresponding average distributions. Therefore, we can reject the null hypothesis and conclude by 90% confidence level that the earthquake, media, politics, sports, and idiom topics have statistically significant difference in cascade ratio from the population. The earthquake, media, sports, and idiom topics have relatively low cascade ratio. People participating in these topics used hashtags independently not because of seeing from their friends' tweets. On the contrary, the political topic has relatively high cascade ratio. When people posted political hashtags, many of their friends started to post the same hashtags after them.

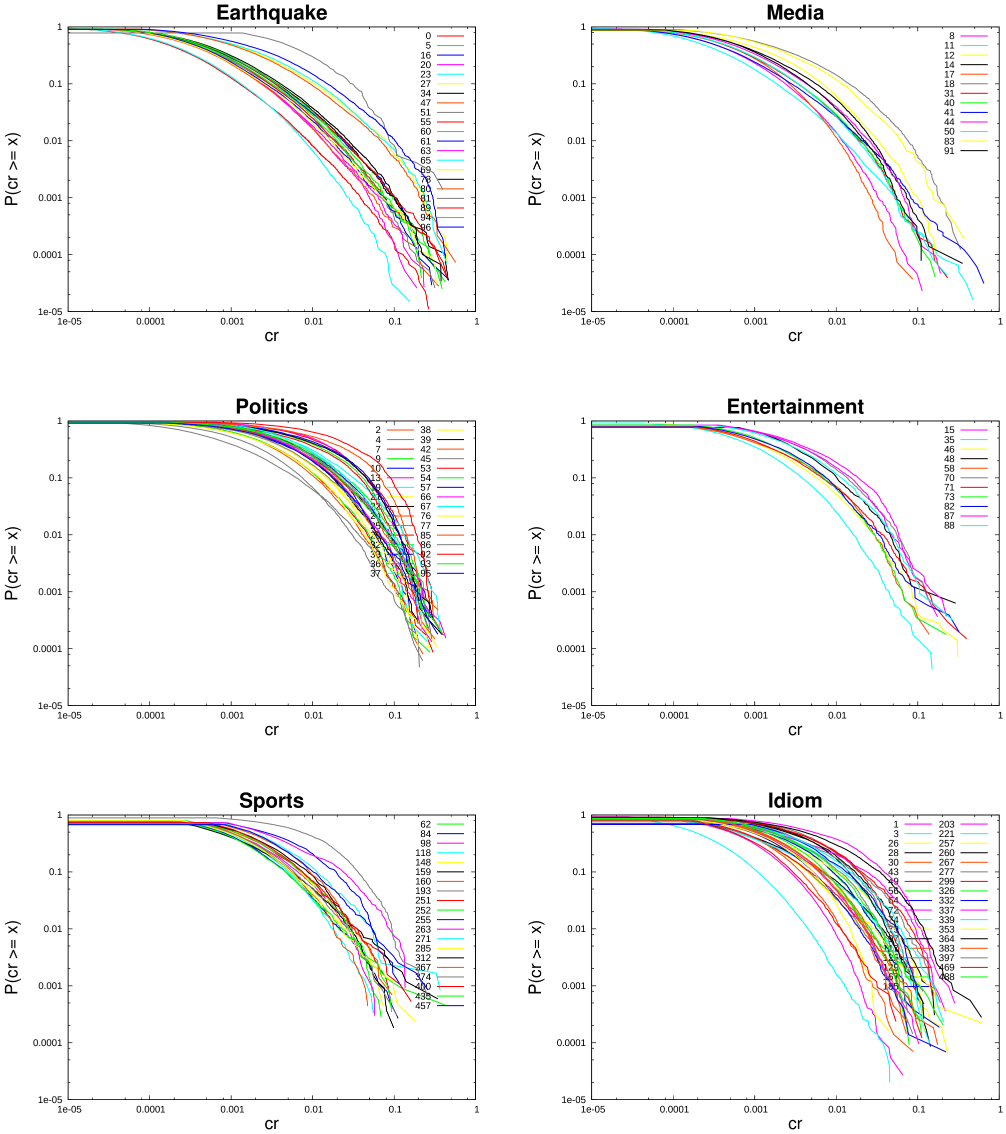


Figure 4.3: Cascade ratio distributions of all hashtags in each topic

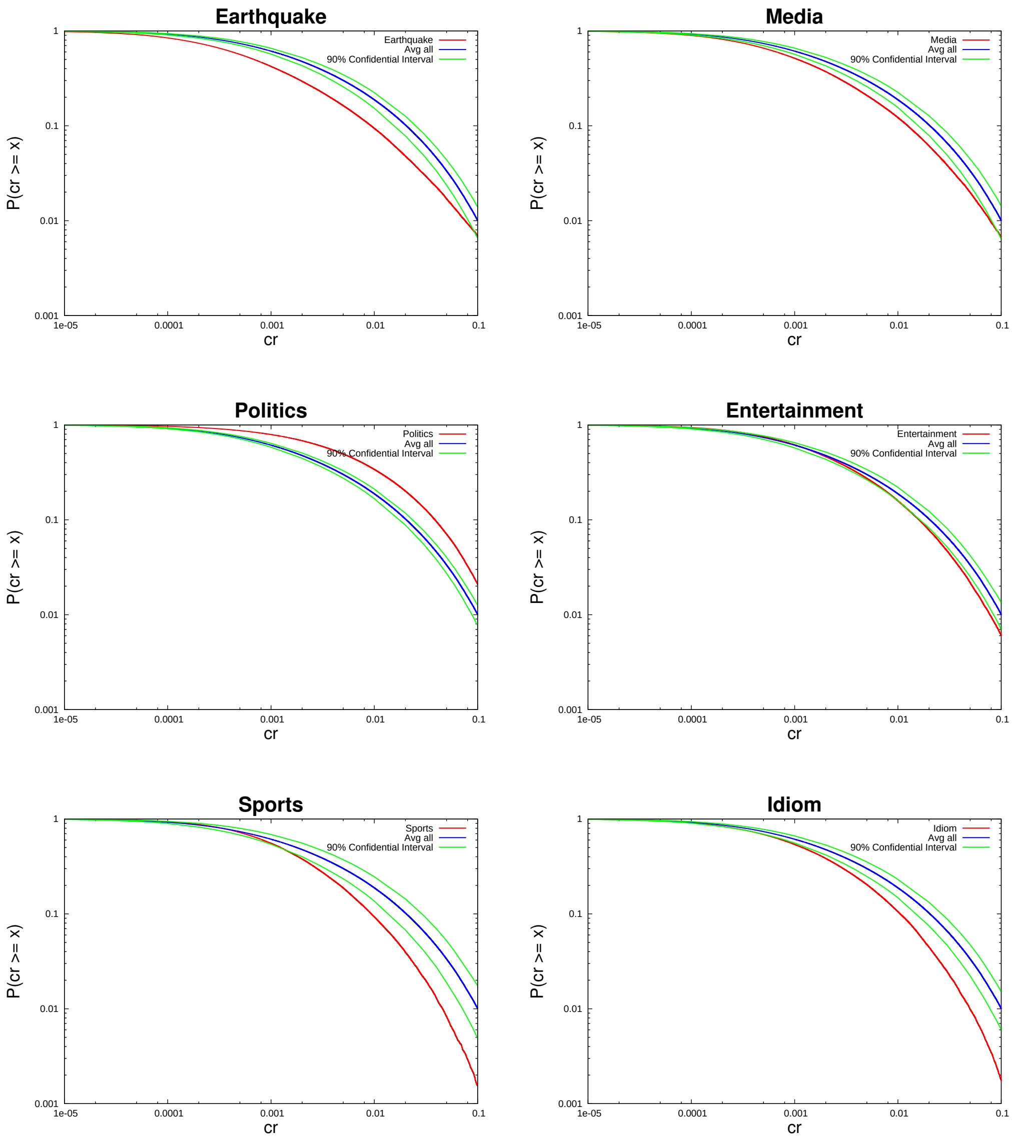


Figure 4.4: Point-wise average cascade ratio distributions of each topic

4.2 Tweet Ratio

The second measure is tweet ratio, the proportion of how many times a user uses a hashtag comparing to all tweets of the same hashtag. The tweet ratio tr of a user u posting a hashtag h is then simply defined as below:

$$tr(u,h) = \frac{T(u,h)}{\sum_u T(u,h)} \tag{4.2}$$

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

Fig.4.5 shows the probability distributions of tweet ratio of all hashtags in each topic. x is tweet ratio and y 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, where y or $P(x)$ is the probability at a value greater than or equal to x.

Fig.4.6 illustrates point-wise average tweet ratio distributions. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is the 90% confidence interval. We see that only 90% confidence interval of the political topic includes its average distribution. That means we cannot conclude by 90% confidence level that the political topic has no difference in tweet ratio from the population. Alternatively, 90% confidence intervals of the earthquake, media, entertainment, sports, and idiom topics do not contain their corresponding average distributions. As a result, we can conclude by 90% confidence level that the earthquake, media, entertainment, sports, and idiom topics have statistically significant difference in tweet ratio from the population. The earthquake, media, and idiom topics have relatively low tweet ratio. People in these topics repeated to use same hashtags very few times. On the other hand, the political topic has relatively high tweet ratio. People repetitively posted same hashtags about the political topic many times.

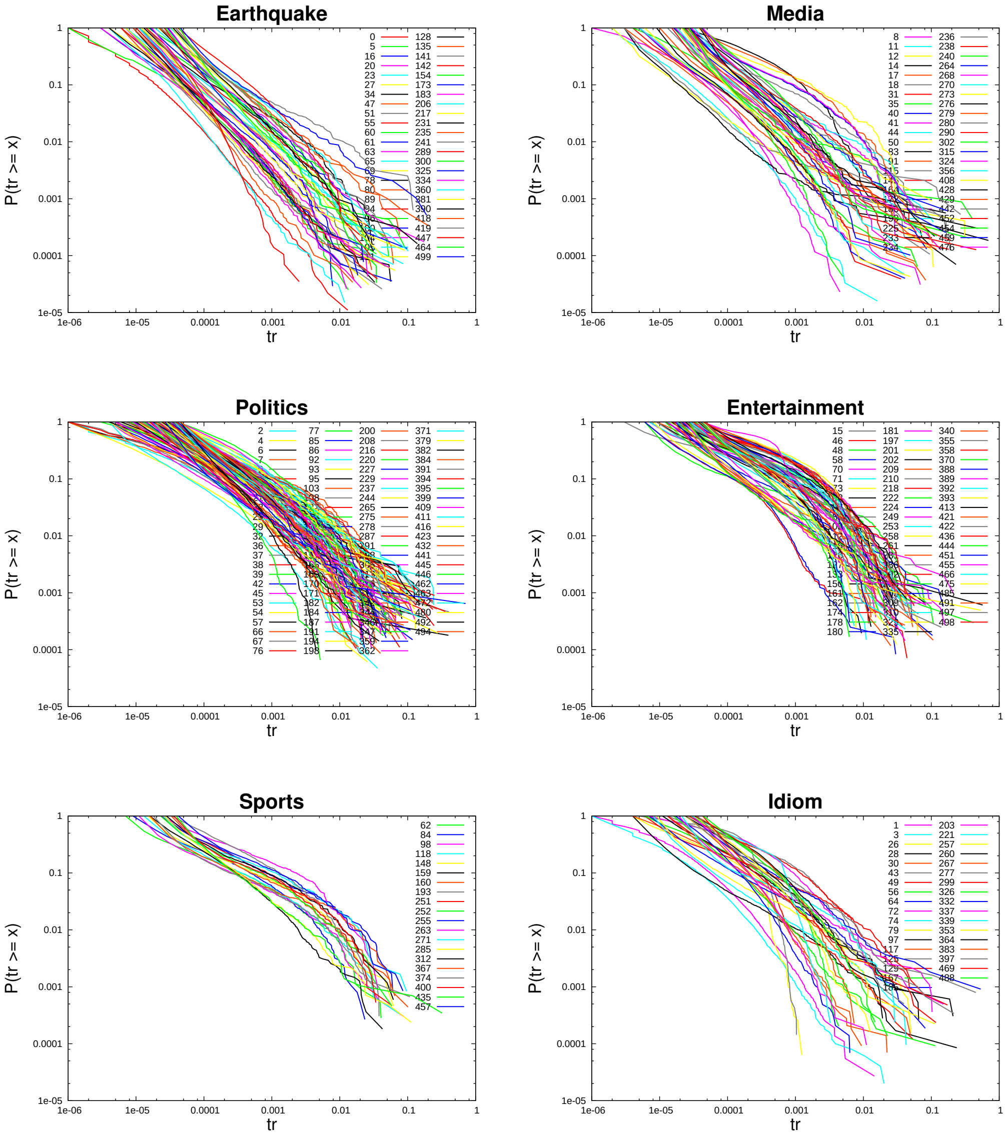


Figure 4.5: Tweet ratio distributions of all hashtags in each topic

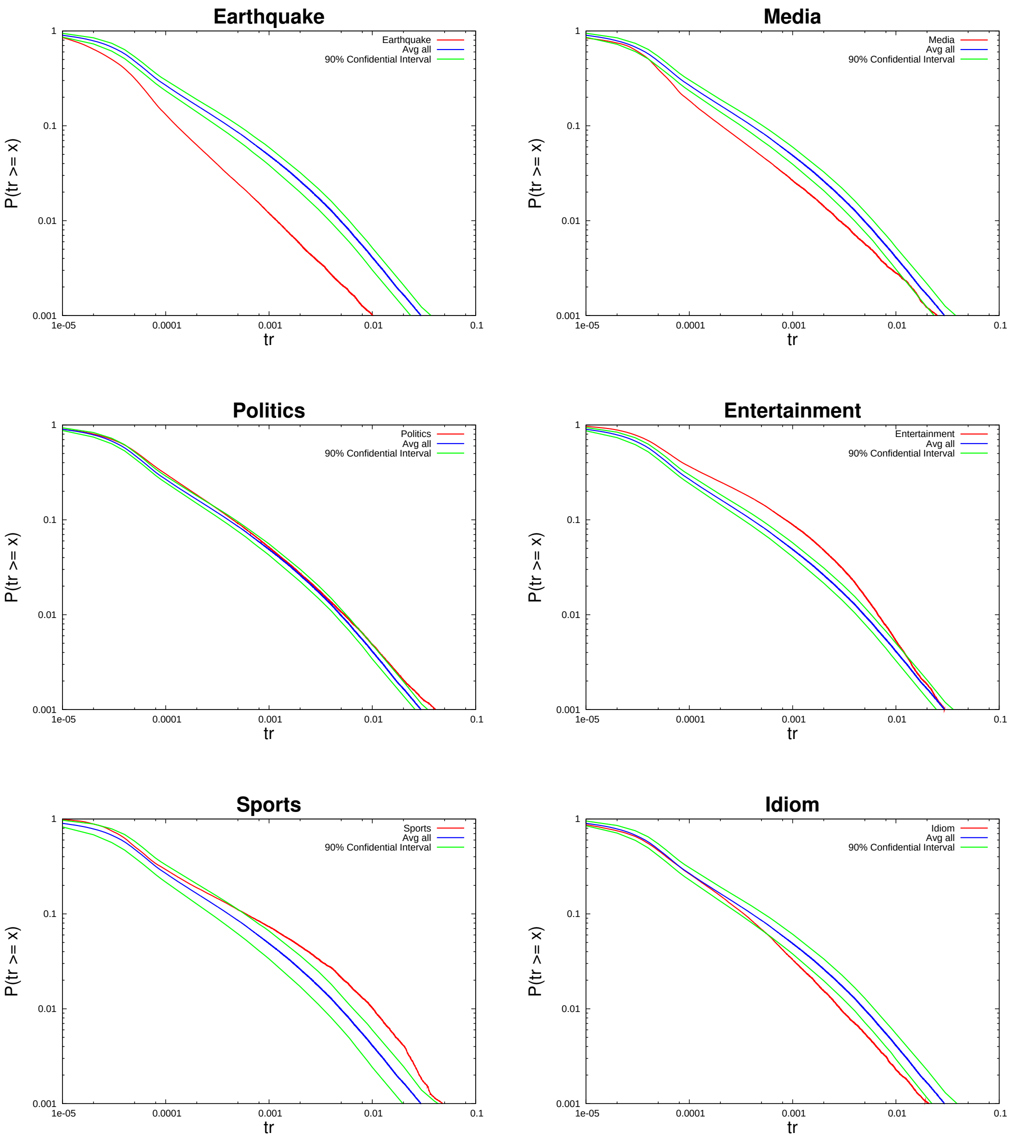


Figure 4.6: Point-wise average tweet ratio distributions of each topic

4.3 Time Interval

The third measure is time interval which is time of each usage of a hashtag from its first appearance. The time interval ti of a tweet tw containing a hashtag h is then straightforwardly defined as the difference in time between tw and the first tweet of h.

Fig.4.8 demonstrates the probability distributions of time interval of all hashtags in each topic. x is time interval in hour(s) and y is the number of occurrences of time intervals normalized by total number of tweets comprising a given hashtag. Each line is plotted as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x.

Fig.4.9 shows point-wise average time interval distributions. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is the 90% confidence interval. We see that 90% confidence intervals of the media, politics, and idiom topic include their corresponding average distributions. That means we cannot conclude by 90% confidence level that the media, politics, and idiom topics have no difference in time interval from the population. Contrarily, 90% confidence intervals of the earthquake, entertainment, and sports topics do not contain their average distributions. Hence, we can conclude by 90% confidence level that the earthquake, entertainment, and sports topics have statistically significant difference in time interval from the population. The earthquake topic falls down at first period. A large number of tweets were posted soon after the topics were raised to Twitter and gradually decreased when time passed. We can imply that people talked very much about the Great East Japan Earthquake during that time and in turn rarely said about it when the situation was back to normal. Conversely, the entertainment and sports topics lay in a diagonal. The number of tweets did not change according to time. People continually talked about these topics during the period of time. Although the average distribution of the entertainment topic in Fig.4.9 lies in a diagonal, some individual distributions in this topic are sawtooth according to Fig.4.8. We can say that they have periodic behavior. For example, Fig.4.7 represents the time interval distribution

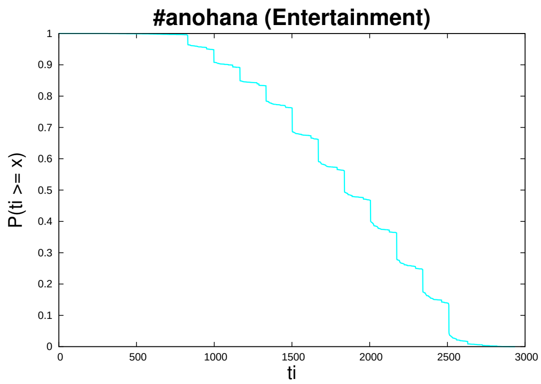


Figure 4.7: Time interval distribution of "anohana" hashtag

of "anohana" hashtag which are Japanese animation that on-air once a week on a television channel. According to Fig.4.7, there are approximately three peaks in each 500 hours or one peak a week. It is likely that fans of this animation also talked much about it on the on-air day.

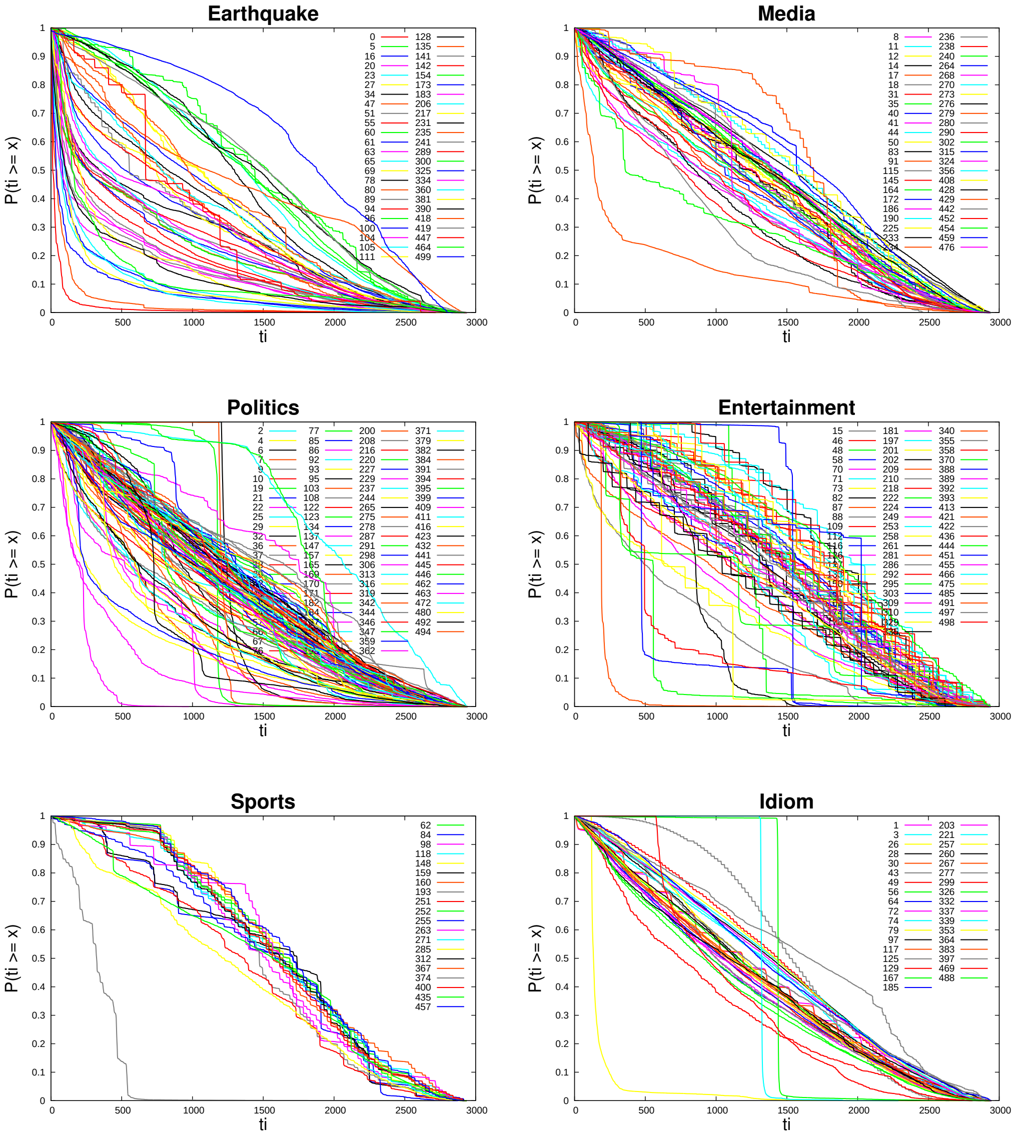


Figure 4.8: Time interval distributions of all hashtags in each topic

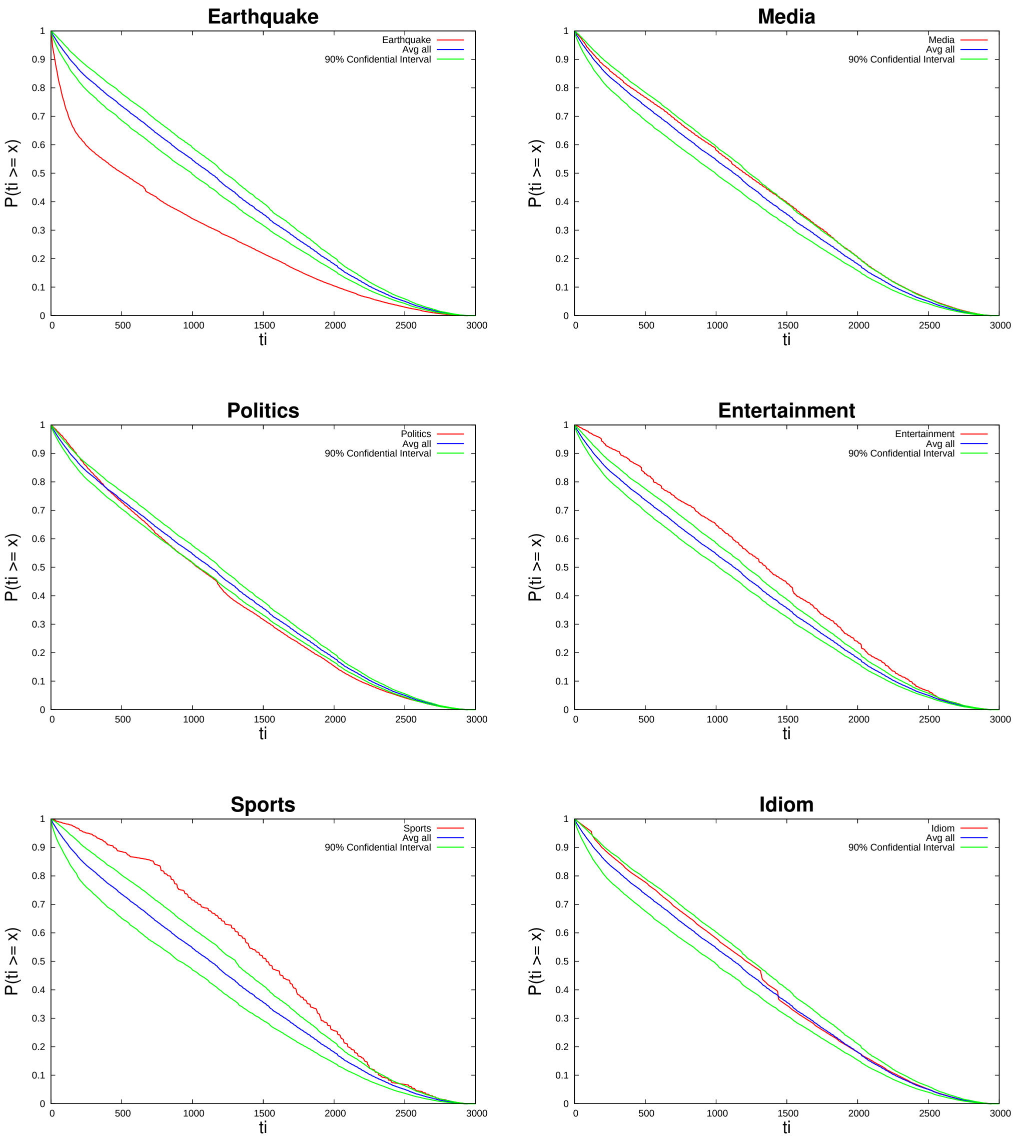


Figure 4.9: Point-wise average time interval distributions of each topic

4.4 Exposure Curve

The last measure is exposure curve proposed by Romero *et al.* [15]. It is another way to represent the relationship between users' influence and hashtag cascade. We begin with basic definition of k-exposed before the exposure curve itself. A user is k-exposed to hashtag h if he/she has k outgoing neighboorhoods who posted h at the time he/she has not used h. According to this definition, one user can be more than one k-exposed during our observation. For example, this time, our user network is shown in Fig.4.10. A node and a directed edge in the graph represents a user and a link of our network respectively.

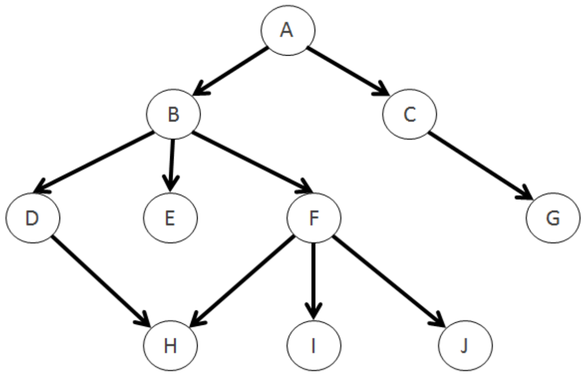
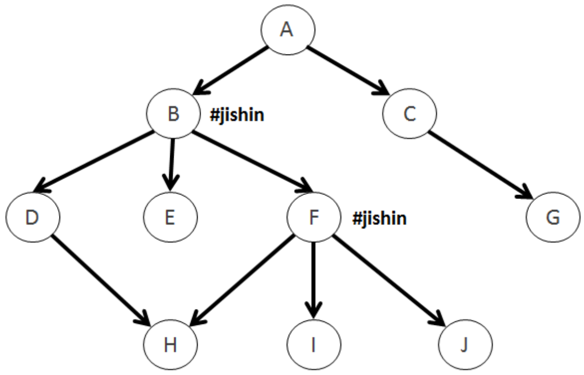
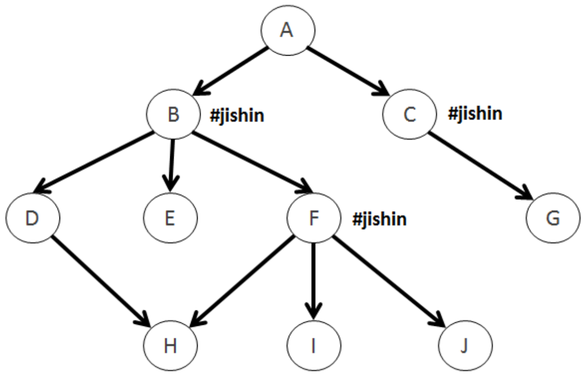


Figure 4.10: An example of user network

We then captured the cascade by tracing the time each user firstly used a given hashtag as same as in case of the cascade ratio. Again, given the "jishin" hashtag, we assume that the cascades take place over the user network as in Fig.4.11. At $t = 0$ when all of users do not start to use "jishin" and so do their outgoing neighborhoods. We can say that all of them are 0-exposed. Then, user B and F begin to post "jishin" at $t = 1$. Because user A has outgoing link to user B and has not used "jishin" yet, user A is 1-exposed. Contrarily, user B has outgoing link to user F but has already used "jishin". As a result, use B is not 1-exposed. Next, at $t = 2$, user C starts to use "jishin". Because user A also has outgoing link to user C and has not posted "jishin" yet, user A at this moment is 2-exposed. We



(a) $t = 1$



(b) $t = 2$

Figure 4.11: An example of hashtag cascade

can see that user A became 0-exposed, 1-exposed, and 2-exposed during the observation.

The exposure curve $P(k)$ is now defined as below:

$$P(k) = \frac{I(k)}{E(k)} \tag{4.3}$$

where $I(k)$ is the number of users who started to post the hashtag h right after becoming k-exposed and $E(k)$ is the number of users who were k-exposed at some time.

Fig.4.12 demonstrates the exposure curves of all hashtags in each topic. x is k-exposed and y the probability $P(k)$ that a user u will use a given hashtag h right after becoming k-exposed.

Fig.4.13 depicts point-wise average exposure curves. The red line is the point-wise average exposure curve of a particular topic, the blue line is the point-wise average exposure curve of all hashtags, and the green line is the 90% confidence interval. We see that 90% confidence intervals of the media and idiom topic include their corresponding average distributions. That means we cannot conclude by 90% confidence level that the media and idiom topics have no difference in exposure curve from the population. On the contrary, 90% confidence intervals of the earthquake, politics, entertainment, and sports topics do not contain their average distributions. Hence, we can conclude by 90% confidence level that the earthquake, politics, entertainment, and sports topics have statistically significant difference in exposure curve from the population. The peaks of the curves, are at $k = 4$ for the earthquake topic and $k = 2$ for the entertainment and sports topics. That means the maximum probability that people will start to post a hashtag about the earthquake topic is when four neighborhoods used that hashtag before them as well as two neighborhoods in case of the entertainment and sports topics. Besides, since the political topic has no peak, we can say that the number of neighborhoods who used a given hashtag do not affect people participating in this topic to start to use the same hashtag. Nevertheless, we here focus on shape of the curve rather than identifying whether the curve is higher or lower than the average. The curve $P(k)$ of the earthquake and political topics do not change as k increases. These two topics are thus high persistent. In turn, the curve $P(k)$ of the entertainment and sports topics fall down rapidly after the peaks. The probability that a user will start to use a hashtag decreases as k increases. We can say that these two topics are low persistent.

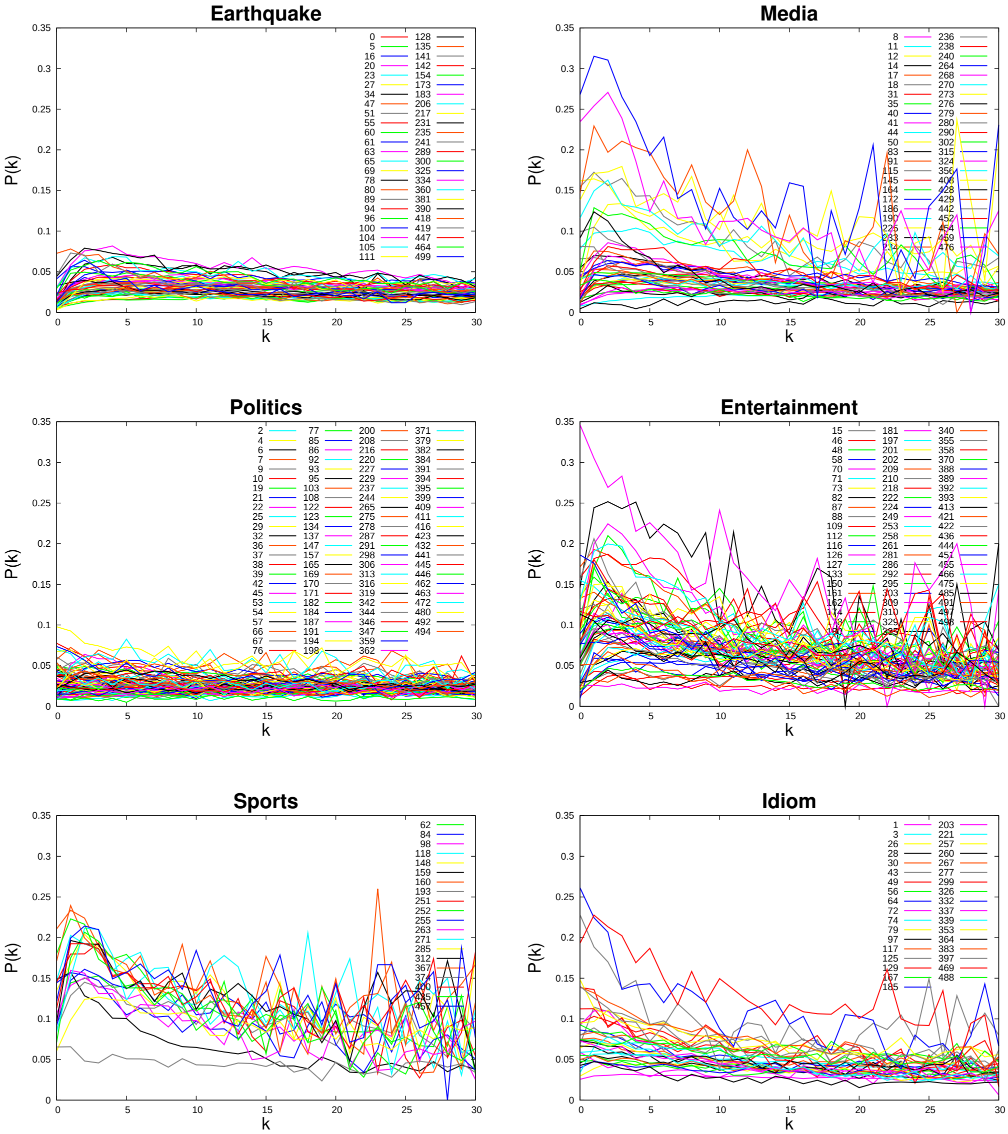


Figure 4.12: Exposure curves of all hashtags in each topic

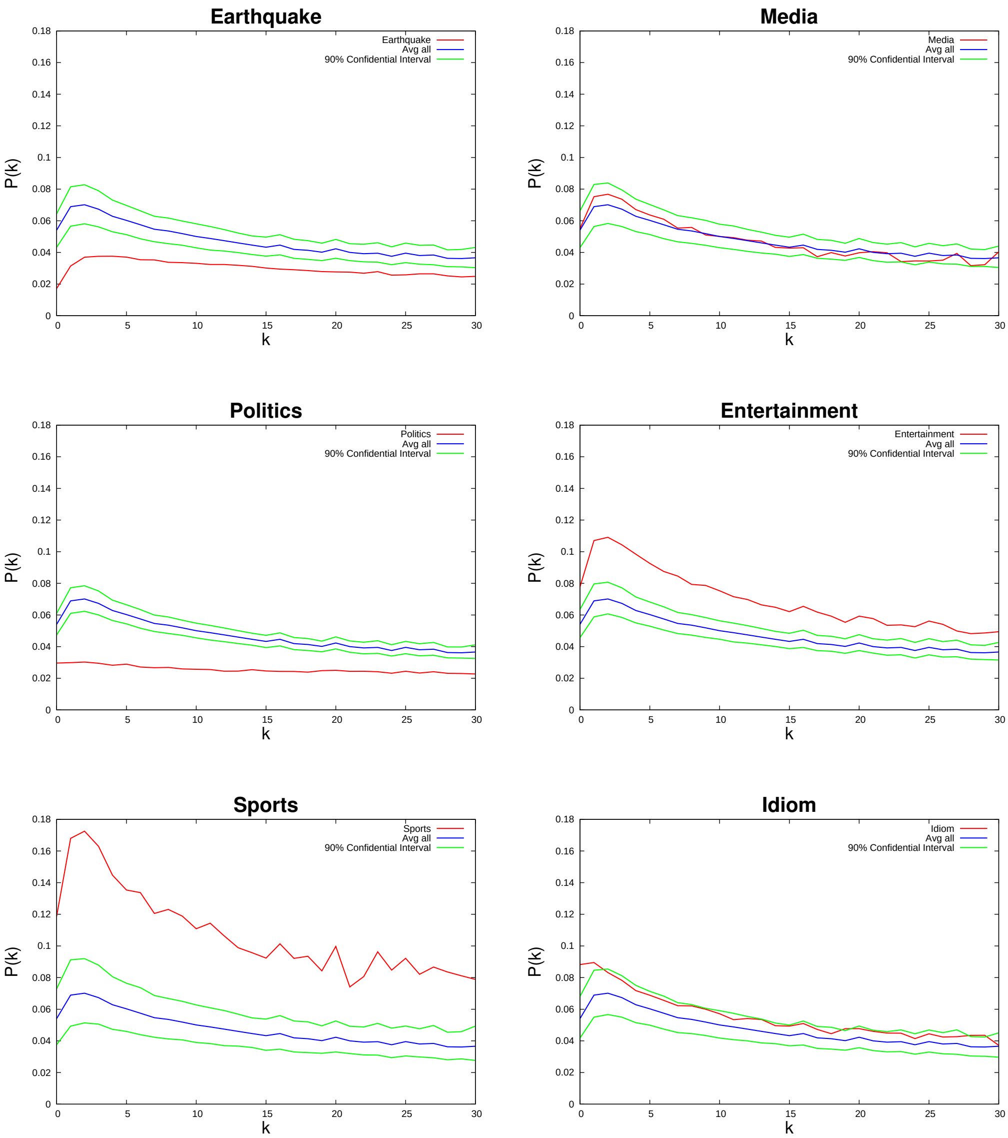


Figure 4.13: Point-wise average exposure curves of each topic

4.5 Patterns of Topic-Sensitive Hashtag Cascades

By using cascade ratio, tweet ratio, time interval, and exposure curve, we summarize patterns of hashtag cascades according to six major topics as in Table 4.1. "H" means high, "L" means low, and - means No statistically significant difference from the population.

We have five patterns of hashtag cascades over six major topics. That is, the media and idiom have the same patterns. Please note that we extracted the following patterns from the average distributions of each topic. In next chapter, we will further study hashtag cascades by blinding out the topics they are assigned and using an automatic algorithm to find their patterns.

Table 4.1: Patterns of hashtag cascades in each topic

Topic	Cascade ratio	Tweet ratio	Time interval	Exposure curve
Earthquake	L	L	L	L
Media	L	L	-	-
Politics	H	-	-	L
Entertainment	-	H	H	H
Sports	L	H	H	H
Idiom	L	L	-	-

Chapter 5

Patterns of Information Cascade across Topics

In this chapter, we further investigate the relationship between cascade patterns and popular topics in Twitter and examine the effectiveness of each measure we described in Chapter 4. We perform k-means clustering based on the distributions of cascade ratio, tweet ratio, time interval, and exposure curve. Each hashtag is represented as a vector of values captured from n points in each distribution as shown in Fig.5.1. For each hashtag, we select 93 points proportional to the log scale.

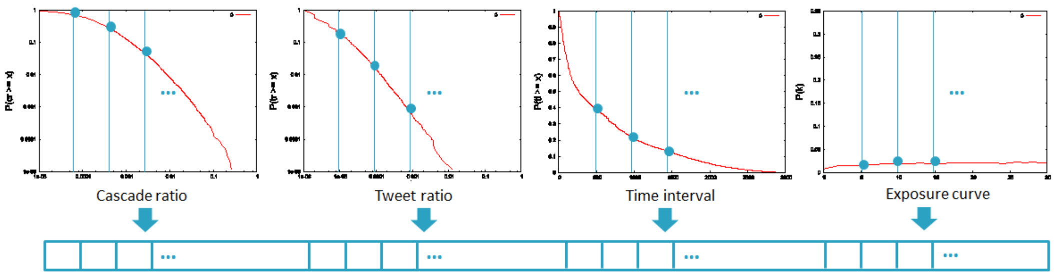


Figure 5.1: A feature vector of "jishin" hashtag for k-means clustering

We use Euclidean distance as a distance measure and randomly assign each hashtag to a cluster at initialization. Considering six major topics in our study, we vary the number of clusters as $k = 6, 7, 8$. Since k-means algorithm provides different results depending on the initialization, we perform five trials for each k and evaluate clustering results by using normalized mutual information (NMI). Instead of other evaluation measures such as purity and F measure, it can be used to compare clustering quality with different numbers of clusters. The normalized mutual information is then defined as below:

$$NMI(\Omega, c) = \frac{I(\Omega; c)}{[H(\Omega) + H(c)]/2} \tag{5.1}$$

where $I(\Omega; C)$ is the mutual information of clusters and topics, $H(\Omega)$ is the entropy of clusters, and $H(C)$ is the entropy of topics.

$$\begin{aligned} I(\Omega; C) &= \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} && (5.2) \\ &= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} && (5.3) \end{aligned}$$

where ω_k is the number of hashtags assigned to cluster k, c_j is the number of hashtags in topic j, and N is the total number of hashtags.

$$\begin{aligned} H(\Omega) &= -\sum_k P(\omega_k) \log P(\omega_k) && (5.4) \\ &= -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} && (5.5) \end{aligned}$$

$$\begin{aligned} H(C) &= -\sum_j P(c_j) \log P(c_j) && (5.6) \\ &= -\sum_j \frac{|c_j|}{N} \log \frac{|c_j|}{N} && (5.7) \end{aligned}$$

For each trial, we compute NMI to evaluate clustering results as shown in Table 5.1. We then pick up the trial that provides the highest NMI at each k.

Additionally, we are able to investigate the effectiveness of each measure on the clustering results by using NMI. We perform clustering by relying on all of four measures, and leaving one measure out at each experiment.

Table 5.1: NMI of each trial when $k = 6, 7, 8$

Trial	$k = 6$	$k = 7$	$k = 8$
1	0.300813	0.301415	0.28647
2	0.287168	0.311266	0.2962
3	0.29966	0.293756	0.270053
4	0.296523	0.300847	0.266965
5	0.283182	0.277615	0.310082

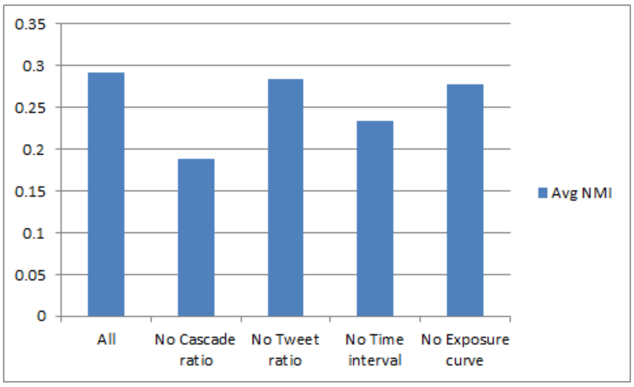


Figure 5.2: Average NMI of each approach when $k = 6$

Fig.5.2 demonstrates the average NMI of five trials in each approach when $k = 6$. We can see that NMI decreases when cascade ratio or time interval are not used. Therefore, cascade ratio and time interval are said to be the most effective measures to characterize hashtag cascade, while tweet ratio and exposure curve even proposed in the existing work are not effective as we expect. According to Table 4.1, we can obtain the same result by using only cascade ratio and time interval.

Table 5.2-5.4 illustrate clustering results of those trials when $k = 6, 7, 8$ respectively. We always have six major clusters according to the results. That is, we ignore cluster 6 when $k = 7$ and cluster 6,7 when $k = 8$ because they contain the small number of hashtags. Besides, the proportion of hashtags in each topic assigned to each cluster are similar for $k = 6, 7, 8$. We then choose the result of $k = 6$ to consider throughout this chapter.

Table 5.2: Clustering result when $k = 6$

No. of hashtags	c0	c1	c2	c3	c4	c5
Earthquake	25	9	1	5	8	0
Media	1	20	1	12	10	2
Politics	0	4	47	2	26	15
Entertainment	0	10	5	39	5	6
Sports	0	2	0	17	0	1
Idiom	1	16	1	7	10	0

Table 5.3: Clustering result when $k = 7$

No. of hashtags	c0	c1	c2	c3	c4	c5	c6
Earthquake	25	9	1	5	8	0	0
Media	1	20	1	12	10	2	0
Politics	0	2	47	1	26	15	3
Entertainment	0	10	5	37	5	6	2
Sports	0	2	0	17	0	1	0
Idiom	1	16	1	5	10	0	2

Table 5.4: Clustering result when $k = 8$

No. of hashtags	c0	c1	c2	c3	c4	c5	c6	c7
Earthquake	25	9	1	5	8	0	0	0
Media	1	19	1	12	11	2	0	0
Politics	0	2	47	1	26	15	3	0
Entertainment	0	10	5	37	5	6	2	0
Sports	0	2	0	17	0	1	0	0
Idiom	1	16	1	5	10	0	2	0

Fig.5.3-5.6 show Point-wise average distributions of each cluster when $k = 6$ based on cascade ratio, tweet ratio, time interval, and exposure curve subsequently. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is the 90% confidence interval. We then summarize patterns of hashtag cascade in each cluster in Table 5.5.

We can see that hashtags from the same topic or the topics having similar patterns of cascade are assigned into the same cluster. According to Table 5.2, the majority of the earthquake topic are assigned into cluster 0.

Moreover, the cascade pattern of this cluster in Table 5.5 is the same as the pattern of the earthquake topic in Table 4.1. This is similar to the media and idiom topics in cluster 1 and the sports topic in cluster 3.

However, some of them even from the same topic have different behaviors and thus put into other clusters. For example, the hashtags in the earthquake topic are mainly divided into cluster 0, 1, and 4. The hashtags in cluster 0 are directly related to the Great East Japan Earthquake such as "jishin", "save_miyagi", and "84ma" (Operation Yashima). On the other hand, the earthquake hashtags in cluster 1, which the majority of the media topic are assigned to, are hashtags such as "iwakamiyasumi" (a journalist who spread information about nuclear power plant after the accident at Fukushima Daiichi Nuclear Power Plant) and "nicojishin". We can see that they are somehow related to the media topic. Likewise, the earthquake hashtags in cluster 4, which its major members are the political topic, are hashtags such as "save_fukusima" and "cnic" (Citizen's Nuclear Information Center). Because they are about the nuclear power plant which needs the Japanese government to concern and take actions on, they are said to be political-related.

In the same way as the media hashtags, they are primarily split into cluster 1, 3, and 4. The hashtags in cluster 1 are Japanese television media such as "fujitv", "nhk", and "tvasahi", while the media hashtags in cluster 3 are Japanese Internet media such as "r_blog" (Rakuten blog), "ameblo" (Ameba blog), and "2chmatome". Furthermore, the media hashtags in cluster 4, which its major members are again the political topic, are hashtags such as "aljazeera", "wikileaks", and "alarabiya". Since these kind of media mainly serve political news, they are thus said to be political-related too.

Lastly, the entertainment and sports hashtags are largely assigned into the same cluster, cluster 3. The entertainment hashtags here are Japanese animations and artists such as "tigerbunny" and "akb48" respectively, while the sports hashtags are Japanese baseball teams such as "hanshin" and "dragons". It is probably that both of them are hobbies, gain much interest from their fans and thus share common behaviors.

Due to the above analysis, it is interesting that we can discover hidden relationship between topics by using only four measures rather than seeing tweet contents.

Table 5.5: Patterns of hashtag cascades in each cluster when $k = 6$

Topic	Cascade ratio	Tweet ratio	Time interval	Exposure curve
Cluster 0	L	L	L	L
Cluster 1	L	L	-	-
Cluster 2	H	H	-	L
Cluster 3	L	H	H	H
Cluster 4	-	L	-	L
Cluster 5	H	H	L	L

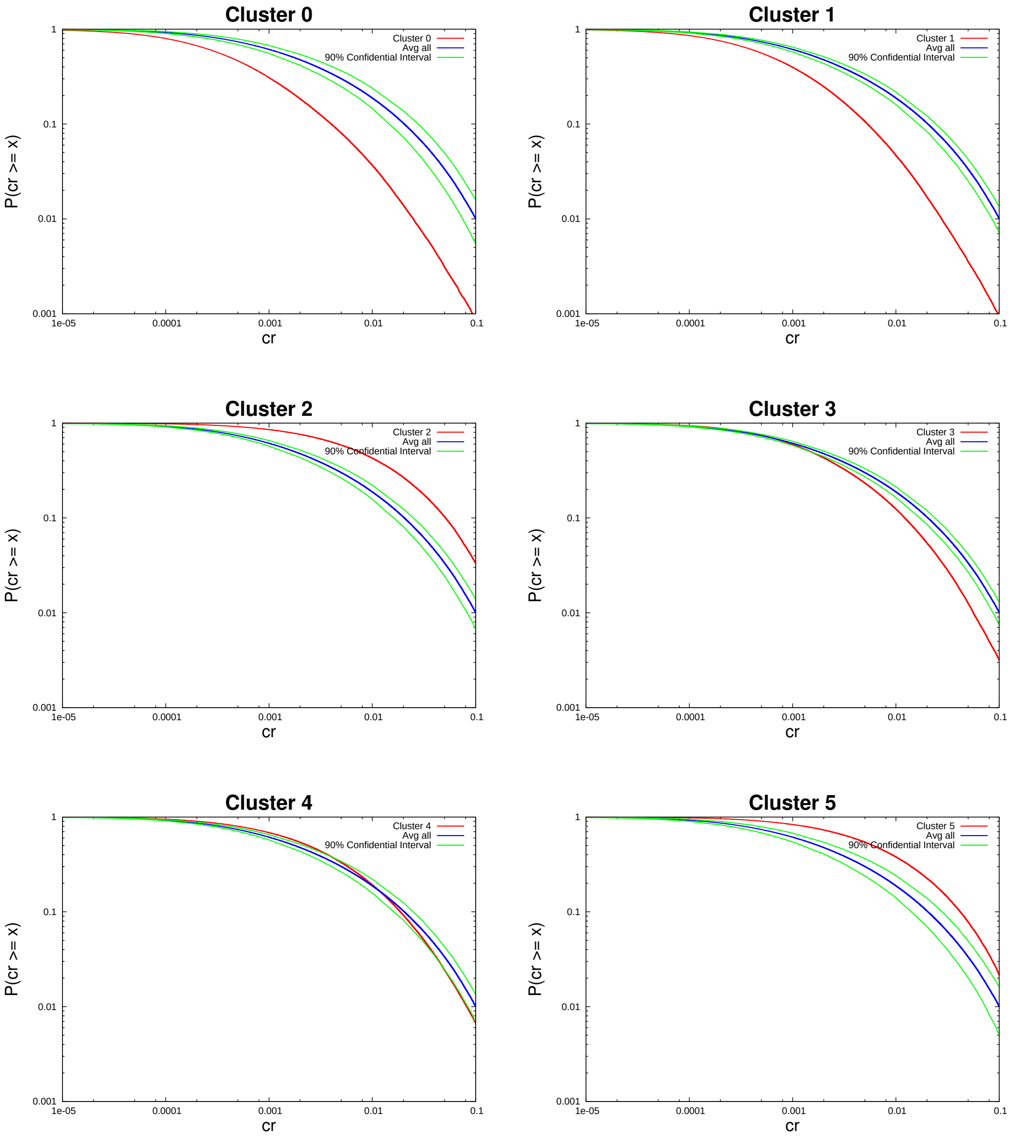


Figure 5.3: Point-wise average cascade ratio distributions of each cluster when $k = 6$

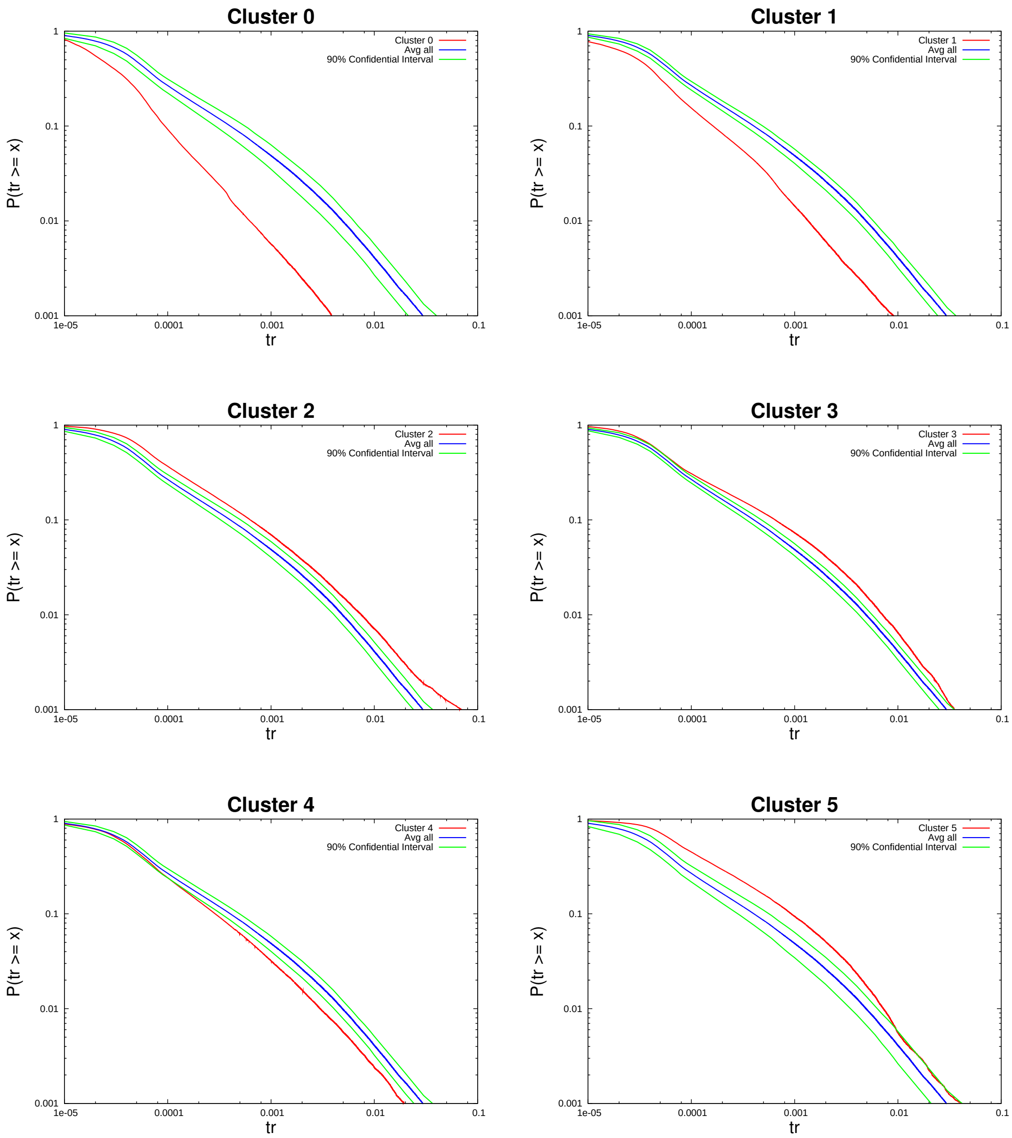


Figure 5.4: Point-wise average tweet ratio distributions of each cluster when $k = 6$

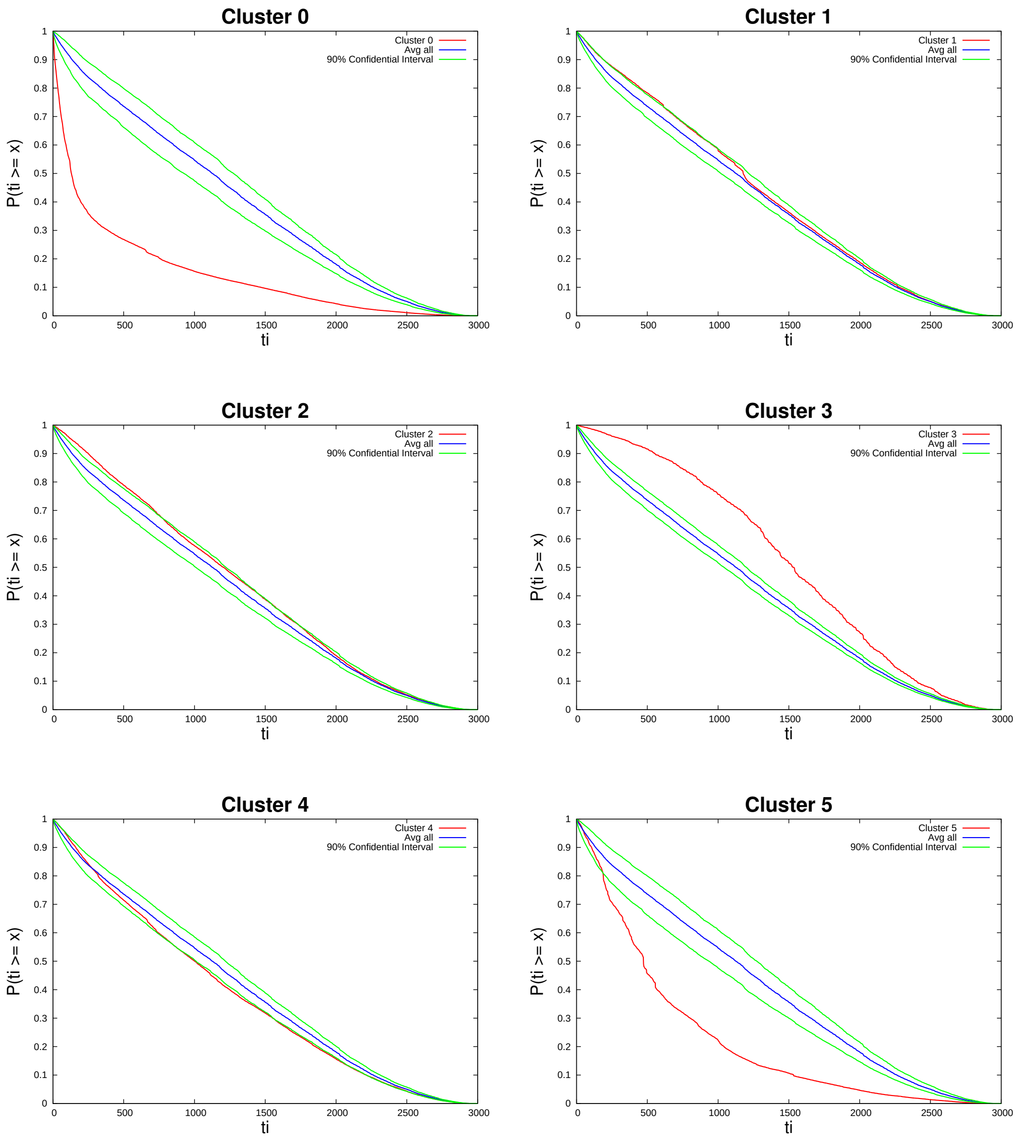


Figure 5.5: Point-wise average time interval distributions of each cluster when $k = 6$

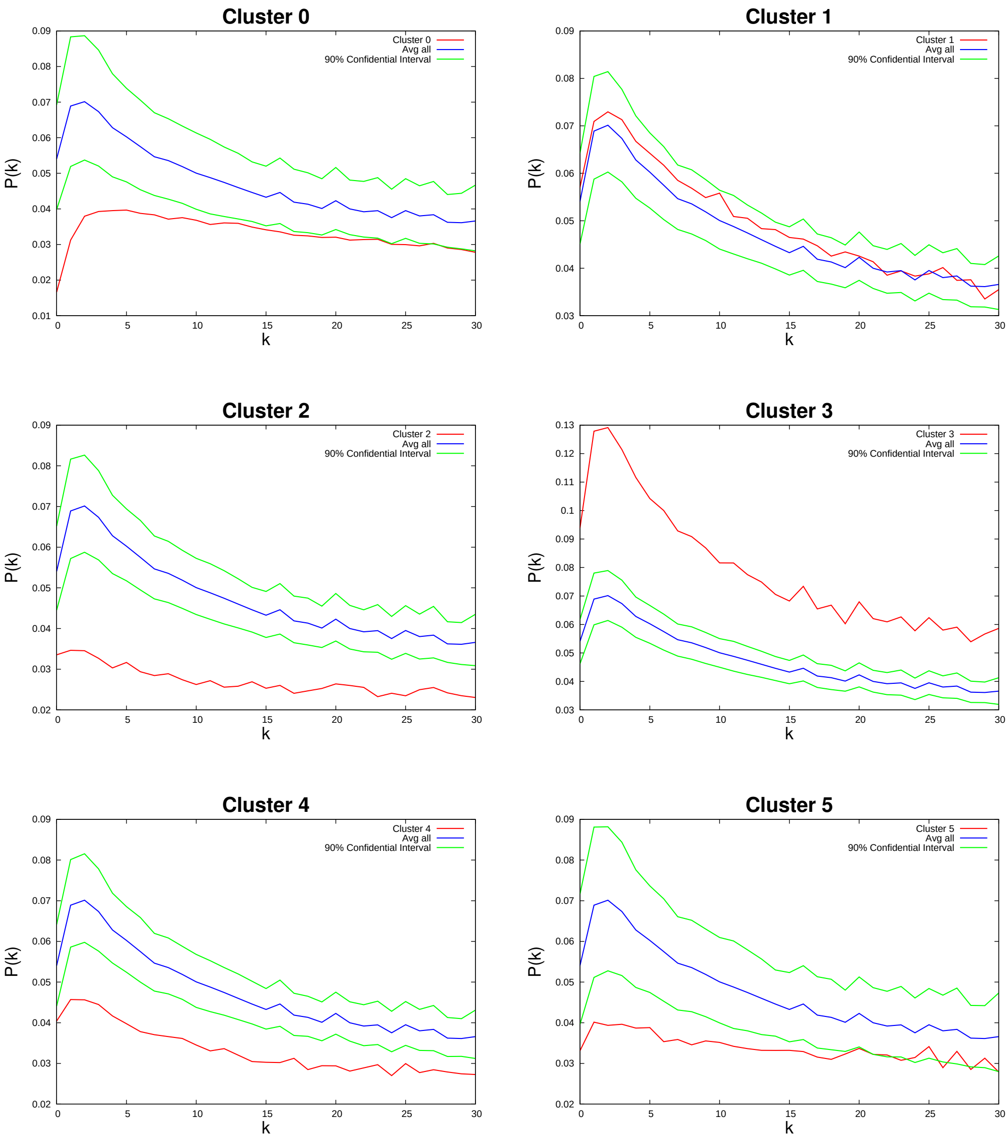


Figure 5.6: Point-wise average exposure curves of each cluster when $k = 6$

Chapter 6

Conclusion

6.1 Conclusion

We studied the patterns of information cascade in six popular topics in Twitter, which are earthquake, media, politics, entertainment, sports, and idiom. We found that different topics mostly have different patterns of hashtag cascades in term of cascade ratio, tweet ratio, time interval, and exposure curve. For example, the earthquake topic has low cascade ratio, low tweet ratio, short lifespan, and high persistence, while the political topic has high cascade ratio and high persistence.

However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We discover that such kind of hidden relationship between topics can be surprisingly revealed by using only four measures rather than considering tweet contents.

Besides, among four measures we explored, we came up with the conclusion that cascade ratio and time interval are the most effective measures to distinguish cascade patterns in different topics, while tweet ratio and exposure curve from the related work are not effective as we expected.

6.2 Future Work

Finally, as future work, we need to explore other useful characteristics such as expert level of individual users. For example, some users have high tweet ratio in one topic but low tweet ratio in others. These kind of users seem to be experts in a particular topic. Moreover, we need to investigate other clustering algorithms and other similarities whether they still provide the same results or not.

Bibliography

[1] E. Adar, L.A. Adamic, Tracking Information Epidemics in Blogspace. In 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI2005), pp. 207–214, 2005.

[2] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an Influencer: Quantifying Influence on Twitter. In 4th ACM International Conference on Web Search and Data Mining (WSDM2011), pp. 65–74, 2011.

[3] E. Bakshy, B. Karrer, L.A. Adamic, Social Influence and the Diffusion of User-Created Content. In 10th ACM Conference on Electronic Commerce (EC2009), pp. 325–334, 2009.

[4] C. Castillo, M. Mendoza, B. Poblete, Information Credibility on Twitter. In 20th International Conference on World Wide Web (WWW2011), pp. 675–684, 2011.

[5] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), pp. 10–17, 2010.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information Diffusion Through Blogspace. In 13th International Conference on World Wide Web (WWW2004), pp. 491–501, 2004.

[7] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence through a Social Network. In 9th ACM SIGKDD Knowledge Discovery and Data Mining (KDD2003), pp. 137–146, 2003.

[8] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a Social Network or a News Media?. In 19th International Conference on World Wide Web (WWW2010), pp. 591–600, 2010.

[9] J. Leskovec, L.A. Adamic, B.A. Huberman, The Dynamics of Viral Marketing, ACM Transaction on the Web, 1(1):5, May 2007.

[10] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the Dynamics of the news Cycle. In 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2009), pp. 497–506, 2009.

[11] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst, Patterns of Cascading Behavior in Large Blog Graphs. In 7th SIAM International Conference on Data Mining (SDM2007), pp. 551–556, 2007.

[12] D. Liben-Nowell, J. Kleinberg, Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data. In the National Academy of Sciences, 105(12):4633–4638, March 25, 2008.

[13] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, J. Chayes, We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter. In 20th International Conference on World Wide Web (WWW2011), pp. 517–526, 2011.

[14] M.E.J. Newman, S. Forrest, J. Balthrop, Email Networks and the Spread of Computer Viruses. Physical Review E, 66(035101), 2002.

[15] D.M. Romero, B. Meeder, J. Kleinberg, 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 (WWW2011), pp. 695–704, 2011.

[16] E. Sun, I. Rosenn, C. Marlow, T. Lento, Gesundheit! Modeling Contagion through Facebook News Feed. In 3rd International AAAI Conference on Weblogs and Social Media (ICWSM2009), pp. 146–153, 2009.

[17] S. Scellato, C. Mascolo, M. Musolesi, J. Crowcroft, Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. In 20th International Conference on World Wide Web (WWW2011), pp. 457–466, 2011.

[18] D.J. Watts, A Simple Model of Global Cascades on Random Networks. In the National Academy of Sciences, 99(9):5766–5771, April 30, 2002.

[19] J. Weng, E.-P. Lim, J. Jiang, Q. He, TwitterRank: Finding Topic-Sensitive Influential Twitterers. In 3rd ACM International Conference on Web Search and Data Mining (WSDM2010), pp. 261–170, 2010.

[20] S. Wu, J.M. Hofman, W.A. Mason, D.J. Watts, Who Says What to Whom on Twitter. In 20th International Conference on World Wide Web (WWW2011), pp. 705–714, 2011.

Publications

1. G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. Characterizing Topic-Specific Hashtag Cascade in Twitter Based on Distributions of User Influence. In 14th Asia-Pacific Web Conference (APWeb2012), 2012. (To be appeared)

2. G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. A Study on Relationships between Information Cascades and Popular Topics in Twitter. In 4th Forum on Data Engineering and Information Management (DEIM2012), 2012.

3. G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. A Study on Characteristics of Topic-Specific Information Cascade in Twitter. In Forum on Data Engineering (DE2011), Vol.111, No.361, DE2011-51, pp.65–70, 2011.

4. G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa. An Analysis on Information Cascade for Detecting Influencers in Twitter. In 2nd Symposium on Social Computing, poster session, (SoC2011), 2011.

Appendix A

List of Top 500 Frequently Used Hashtags

In addition to six major topics as explained in Chapter 3, there are several other topics over top 500 frequently used hashtags. Since these topics have less than 20 hashtags, they are out of our scope. Moreover, many hashtags cannot be matched into any topics or they have less than 1,000 participating users, they are again excluded from our interesting dataset. Table A.1 shows the number of hashtags in each topic. Then, Table A.2 lists the name of top 500 frequently used hashtags and their corresponding topics.

Table A.1: The number of hashtags in each topic

Topic	Total	Topic	Total
Earthquake	48	Technology	7
Politics	94	Games	4
Media	46	Country-City	14
Entertainment	65	Economics	5
Sports	20	Religious	7
Idiom	35	None	155

Table A.2: List of top 500 frequently used hashtags

Rank	Hashtag	Topic	Rank	Hashtag	Topic
1	jishin	Earthquake	37	civ2010	Political
2	ff	Idiom	38	tahrir	Political
3	bahrain	Political	39	saudi	Political
4	nowplaying	Idiom	40	teaparty	Political
5	libya	Political	41	tbs	Media
6	genpatsu	Earthquake	42	nhk_news	Media
7	tcot	Political	43	israel	Political
8	egypt	Political	44	30thou	Idiom
9	nicovideo	Media	45	tvasahi	Media
10	syria	Political	46	gaddafi	Political
11	p2	Political	47	madoka_magica	Entertainment
12	nhk	Media	48	earthquake	Earthquake
13	news	Media	49	atakowa	Entertainment
14	jan25	None	50	ohayo	Idiom
15	2ch	Media	51	ntv	Media
16	agqr	Entertainment	52	jisin	Earthquake
17	fukushima	Earthquake	53	bot	None
18	fb	Media	54	gop	Political
19	newsjp	Media	55	lulu	Political
20	kuwait	Political	56	save_miyagi	Earthquake
21	japan	Earthquake	57	shoutout	Idiom
22	yemen	Political	58	tlot	Political
23	iran	Political	59	akb48	Entertainment
24	prayforjapan	Earthquake	60	fail	None
25	feb17	None	61	eqjp	Earthquake
26	seiji	Political	62	save_ibaraki	Earthquake
27	100factsaboutme	Idiom	63	hanshin	Sports
28	iwakamiyasumi	Earthquake	64	tsunami	Earthquake
29	followmejp	Idiom	65	teamfollowback	Idiom
30	iranelection	Political	66	genpatu	Earthquake
31	np	Idiom	67	ksa	Political
32	fujitv	Media	68	q8	Political
33	wiunion	Political	69	quote	None
34	feb14	None	70	anpi	Earthquake
35	save_fukushima	Earthquake	71	nitiasa	Entertainment
36	pixiv	Media	72	tigerbunny	Entertainment

Rank	Hashtag	Topic	Rank	Hashtag	Topic
73	mf	Idiom	109	cdnpoli	Political
74	jho_ogiri	Entertainment	110	aoex	Entertainment
75	followme	Idiom	111	imagine	None
76	follow	None	112	saigai	Earthquake
77	obama	Political	113	akb	Entertainment
78	palestine	Political	114	gcc	None
79	hinan	Earthquake	115	tree_twinavi	None
80	sougofollow	Idiom	116	cnn	Media
81	nicojishin	Earthquake	117	dommune	Entertainment
82	pf_anpi	None	118	followfriday	Idiom
83	anohana	Entertainment	119	tigers	Sports
84	googlenewsjp	Media	120	soundtracking	None
85	f1jp	Sports	121	twitkiss	None
86	sgp	Political	122	swag	None
87	humanrights	Political	123	prochoice	Political
88	hanairo	Entertainment	124	royalwedding	Political
89	precure	Entertainment	125	meigen	None
90	311care	Earthquake	126	pickone	Idiom
91	twitpict	None	127	cho_ag	Entertainment
92	tvtokyo	Media	128	dogdays	Entertainment
93	ocra	Political	129	311pet	Earthquake
94	gaza	Political	130	tfb	Idiom
95	miyagi	Earthquake	131	jesus	Religious
96	un	Political	132	neko	None
97	nuclearjp	Earthquake	133	winning	None
98	mustfollow	Idiom	134	nichijou	Entertainment
99	carp	Sports	135	pakistan	Political
100	fact	None	136	j_j_helpme	Earthquake
101	sendai	Earthquake	137	traindelay	None
102	prolife	None	138	daraa	Political
103	akiba	None	139	mar15	None
104	elxn41	Political	140	uae	Country-City
105	ibaraki	Earthquake	141	14feb	None
106	iwaki	Earthquake	142	save_iwate	Earthquake
107	us	Country-City	143	edano_nero	Earthquake
108	qanow	None	144	bookmeter	None

Rank	Hashtag	Topic	Rank	Hashtag	Topic
145	nokill	None	181	joqr	Entertainment
146	nikkei	Media	182	bieberfact	Entertainment
147	hsus	None	183	yf	Political
148	tunisie	Political	184	iwate	Earthquake
149	dragons	Sports	185	tunisia	Political
150	truth	None	186	justsaying	Idiom
151	mc1242	Entertainment	187	dig954	Media
152	cat	None	188	kokkai	Political
153	peta	None	189	okaeri	None
154	usa	Country-City	190	may27	None
155	iwakamiyasumi2	Earthquake	191	jugem_blog	Media
156	piston2438	None	192	congress	Political
157	mogsnap	None	193	qatar	Country-City
158	arab	Political	194	baystars	Sports
159	china	Country-City	195	p21	Political
160	f1	Sports	196	aspca	None
161	giants	Sports	197	tweetbatt	None
162	kirakira	Entertainment	198	aaabc	Entertainment
163	a_ch	Entertainment	199	misrata	Political
164	islam	Religious	200	tes3inat	None
165	niconews	Media	201	jnsc	Political
166	obl	Political	202	denpa_girl	Entertainment
167	twitter	None	203	mogra	Entertainment
168	lol	Idiom	204	followdaibosyu	Idiom
169	twisters	None	205	rt	None
170	tripoli	Political	206	nice20	None
171	jo	Political	207	kaminoseki	Earthquake
172	politics	Political	208	tokyo	Country-City
173	etv	Media	209	lebanon	Political
174	nuclear	Earthquake	210	music	Entertainment
175	neversaynever	Entertainment	211	aiww	Entertainment
176	twitbackr	None	212	itunes	Technology
177	love	None	213	redeye	None
178	asia	None	214	nemuritsuzuketeshinu	None
179	moshidora	Entertainment	215	uk	Country-City
180	photography	None	216	ogiri_dan	None

Rank	Hashtag	Topic	Rank	Hashtag	Topic
217	sidibouzid	Political	253	sbhawks	Sports
218	fukunp	Earthquake	254	magnhk	Entertainment
219	kaminomi	Entertainment	255	tellme	None
220	dostor2011	None	256	canucks	Sports
221	mubarak	Political	257	mysky	None
222	miteru	Idiom	258	damnitstrue	Idiom
223	sht	Entertainment	259	tokyofm	Entertainment
224	photo	None	260	hhrs	None
225	ske48	Entertainment	261	nowfollowing	Idiom
226	youtube	Media	262	suidou	Entertainment
227	daihyo	None	263	25jan	None
228	benghazi	Political	264	jspocycle	Sports
229	okaeriradio	None	265	nicoch	Media
230	iraq	Political	266	nato	Political
231	gizjp	Technology	267	yokohama	Country-City
232	tepco	Earthquake	268	justsayin	Idiom
233	dsk	Economics	269	bbc	Media
234	ontveg	Media	270	finance_news	Economics
235	nhkgtv	Media	271	wikileaks	Media
236	iwakamiyasumi3	Earthquake	272	lovefighters	Sports
237	mbs	Media	273	haiku	None
238	egyarmy	Political	274	tokyomx	Media
239	aljazeera	Media	275	reformjo	None
240	lgbt	None	276	hcr	Political
241	hdln	Media	277	alarabiya	Media
242	cnic	Earthquake	278	500aday	Idiom
243	iphone	Technology	279	libye	Political
244	android	Technology	280	mpj	Media
245	india	Political	281	r_blog	Media
246	quotes	None	282	jwave	Entertainment
247	ylog	None	283	maigo	None
248	dog	None	284	keizai	Economics
249	anonymous	None	285	hibaku	None
250	eiga	Entertainment	286	npb	Sports
251	edl	None	287	gintama	Entertainment
252	keiba	Sports	288	topprog	Political

Rank	Hashtag	Topic	Rank	Hashtag	Topic
289	softbank	None	325	fpaj	Media
290	ishinomaki	Earthquake	326	84ma	Earthquake
291	r_socialnews	Media	327	childhoodmemories	Idiom
292	jordan	Political	328	tokyonews	None
293	kaiji	Entertainment	329	epictweets	None
294	bethaderej	None	330	c_anime	Entertainment
295	team_naraku	None	331	green	None
296	anime	Entertainment	332	free	None
297	inu	None	333	ifollowback	Idiom
298	itrotter	Game	334	yokote	Country-City
299	hijitsuzai	Political	335	kesennuma	Earthquake
300	whatif	Idiom	336	seenomore	Entertainment
301	shien	Earthquake	337	facebook	None
302	quran	Religious	338	wtf	Idiom
303	2chmatome	Media	339	steinsgate	Game
304	eurovision	Entertainment	340	20peopleilove	Idiom
305	brk	None	341	jin	Entertainment
306	win	None	342	ske	None
307	women2drive	Political	343	osama	Political
308	anybeats	Game	344	zodiacfacts	None
309	imacoconow	None	345	saleh	Political
310	anipoke	Entertainment	346	twkrs	None
311	sg_anime	Entertainment	347	amman	Political
312	nakayoshiex	None	348	kuw	Political
313	chibalotte	Sports	349	butei	None
314	gaddaficrimes	Political	350	nwo	None
315	inthemood	None	351	tochigi	Country-City
316	mynippon	Media	352	god	Religious
317	haiti	Political	353	doncabot	None
318	offline	None	354	nowwatching	Idiom
319	france	Country-City	355	fx	Economics
320	p2b	Political	356	momoclo	Entertainment
321	atheism	Religious	357	wbs	Media
322	jobs	None	358	trustinjapan	None
323	random	None	359	gosick	Entertainment
324	kafi	None	360	minsyu	Political

Rank	Hashtag	Topic	Rank	Hashtag	Topic
361	jpquake	Earthquake	397	ohayopanda	None
362	ojisanplus	None	398	followback	Idiom
363	weinergate	Political	399	akita	Country-City
364	nakba	None	400	homs	Political
365	muchlove	Idiom	401	swallows	Sports
366	wanko	None	402	cambiochat	None
367	33fan	None	403	venezuela	None
368	kyojin	Sports	404	opsafe	None
369	tworship	None	405	art	None
370	theraj	None	406	intw	None
371	sutadora	Entertainment	407	commando	None
372	1u	Political	408	micropoetry	None
373	jgf	None	409	bs11	Media
374	scan_level0	None	410	cairo	Political
375	cwc2011	Sports	411	dead	None
376	zexal	None	412	ippyo	Political
377	cfneed	None	413	jewelpet	None
378	androidjp	Technology	414	utamaru	Entertainment
379	eu	Economics	415	sengokuotome	None
380	sanaa	Political	416	mccann	None
381	megu_game	Game	417	wi	Political
382	save_touhoku	Earthquake	418	soor5	None
383	tpp	Political	419	radiation	Earthquake
384	nw	Idiom	420	gwatcherver2	Earthquake
385	palin	Political	421	support	None
386	sxsw	None	422	ss3malaysia	Entertainment
387	itweetmytunes	None	423	ann	Entertainment
388	installnow	None	424	oman	Political
389	followmeariana	Entertainment	425	ponponpain	None
390	onepiece	Entertainment	426	s_tr	None
391	quake	Earthquake	427	doya	None
392	sudan	Political	428	auspol	None
393	glee	Entertainment	429	niftynews	Media
394	saitokazuyoshi	Entertainment	430	nhkfm	Media
395	algeria	Political	431	travel	None
396	damascus	Political	432	freedom	None

Rank	Hashtag	Topic	Rank	Hashtag	Topic
433	jimin	Political	467	prfm	Entertainment
434	sphere	None	468	bey2ollak	None
435	8ji_sentai	None	469	kissdum	None
436	soccer	Sports	470	feelon	Idiom
437	hw813	Entertainment	471	retweet	None
438	rakutenichiba	None	472	autotranslated	None
439	prettyrhythm	None	473	morocco	Political
440	twitmusic	None	474	kyoto	Country-City
441	fm99	None	475	fm802noa	None
442	turkey	Political	476	kpop	Entertainment
443	foxnews	Media	477	ameblo	Media
444	rapture	Religious	478	giveluv2jp	None
445	asamadetv	Entertainment	479	soundcloud	Technology
446	assad	Political	480	pisces	None
447	scaf	Political	481	ukuncut	Political
448	iwakamiyasumi6	Earthquake	482	hero_message	None
449	hate_korea	None	483	video	None
450	gh	None	484	cafeadictos	None
451	yamagata	Country-City	485	iphonejp	Technology
452	mj	Entertainment	486	xfactor	Entertainment
453	tbsradio	Media	487	ifollowall	None
454	twipple_vote	None	488	copts	Religious
455	videonews	Media	489	smh	Idiom
456	lastfm	Entertainment	490	women	None
457	epic	None	491	ikuji	None
458	seibulions	Sports	492	takajin	Entertainment
459	edchat	None	493	senkyo	Political
460	ameba	Media	494	758ben	None
461	kaigo	None	495	afghanistan	Political
462	march15	None	496	oogiribu_app	None
463	protest	Political	497	kanto	None
464	unitebh	Political	498	nyc	Entertainment
465	teiden	Earthquake	499	is_anime	Entertainment
466	syy	None	500	jp_geiger	Earthquake