Real-time Tweet Classification in Disaster Situation

Fujio Toriumi
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku,
Tokyo, Japan 113-8654
tori@sys.t.u-tokyo.ac.jp,

Seigo Baba
The University of Tokyo
baba@crimson.q.t.u-tokyo.ac.jp,

ABSTRACT

During a disaster, appropriate information must be collected quickly. For example, residents along the coast require information about tsunamis and those who have lost their houses need information about shelters. Twitter can attract more attention than other forms of mass media under these circumstances because it can quickly provide such information. Since Twitter has an enormous amount of tweets, they must be classified to provide users with the information they need. Previous works on extracting information from Twitter focused on the text data of tweets. However, in some cases, text mining has difficulty extracting information. For example, it might be difficult for text mining to group tweets with URLs. On the other hand, by assuming that users who retweet the same tweet are interested in the same topic, we can classify tweets that are required by users with similar interests based on retweets. Thus, we employ the tweet classification method that focuses on retweets. In this paper, we demonstrated that our method works quickly in disaster situations and that it can quickly classify the required information based on the needs in disaster situations and is helpful for collecting information under them.

1. INTRODUCTION

During such catastrophic natural disasters as earthquakes, tsunamis, and typhoons, victims and survivors must correctly and quickly collect information about shelters, dangerous areas, and safety advice. Relief workers also need information about volunteers, relief goods, and providing food for evacuees. In other words, the required information changes based on the situations and times of those involved. However, such mass media sources as TV, newspapers, and radio offer general information instead of specifically focusing on more urgently needed information with the time lag. On the other hand, social media are attracting a great deal of attention since they can provide such real-time localized information. The purpose of this study is to realize real-time information sharing systems via twitter for a disaster situation.

In particular, many reports argue that Twitter, one of the most influential social media, is useful for sharing information during disasters. Mendoza et al. analyzed events related to the 2010 earthquake in Chile and characterized Twitter in the hours and days following it [4]. Miyabe et al. surveyed how people used Twitter after the 2011 Great East Japan Earthquake [5]. Sakaki et al. developed a novel earthquake reporting system that promptly notifies people of seismic activity by considering each Twitter user as a sensor [6]. In this paper, we also address Twitter as a source of local information. Previous works about extracting information from Twitter focused on the text data of tweets. In other words, they were based on text mining. García et al. used a vector space model and Latent Dirichlet Allocation to obtain similar keywords [3].

In some cases, text mining has difficulty extracting information. For example, it may be difficult for text mining to deal with tweets that have URLs or very short ones. Therefore, Baba et al. proposed a tweet classification method that focuses on retweets without text mining [1]. We employed the retweet-based clustering methods for real-time tweet classification. In this paper, we applied the retweet-based clustering methods to each time period of after disaster, to evaluate whether the method can be used in the real-time systems. We also analyze the obtained information to clarify what kind of information is required in each time period.

2. TWEET CLUSTERING METHOD

In this paper, we use the log data of tweets written in Japanese that were posted and officially retweeted for 20 days from March 5 to 24, 2011. This period includes the Great Eastern Japan Earthquake that occurred on March 11, 2011. The log data contain 30,607,231 tweets. We selected the 34,860 tweets that were retweeted more than 100 times to focus on how the information was spread and shared.

In this study, we employed the retweet-based clustering method[1] for the tweet classification. When many users retweet both tweets A and B, they probably share a common interest in them and the topics are similar. In other words, two tweets whose similarity of retweeting users is high might share a topic. Therefore, linking such tweets creates a retweet network that connects topic-similar tweets.

Then, the network clustering method is applied to extract clusters that contain similar tweets. We simply employed

WWW'16 Companion, April 11–15, 2016, Montreal, Quebec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889365.

Table 1: Applied in real time

Period	RT ¿100	Time	clusters
0-1 h	293 tweets	2 min	15 clusters
2-3 h	600 tweets	6 min	46 clusters
7-8 h	423 tweets	4 min	35 clusters
10-19 h	1255 tweets	6 min	86 clusters
48-60 h	2807 tweets	4 min	154 clusters

Table 2: Information of obtained clusters in each period

P