# Mass and Social Media Corpus Analysis after the 2011 Great East Japan Earthquake

Shosuke Sato
Tohoku University
6-6-11 Aoba, Aramaki
Aoba-ku, Sendai, Japan
ssato@dcrc.tohoku.ac.jp

Michiaki Tatsubori
IBM Research - Tokyo
1623-14 Shimo-tsuruma
Yamato, Kanagawa, Japan
mich@acm.org

Fumihiko Imamura
Tohoku University
6-6-11 Aoba, Aramaki
Aoba-ku, Sendai, Japan
imamura@tsunami2.civil.tohoku.ac.jp

## ABSTRACT

In this paper, we outline our analysis of mass media and social media as used for disaster management. We looked at the differences among multiple sub-corpuses to find relatively unique keywords based on chronologies, geographic locations, or media types. We are currently analyzing a massive corpus collected from Internet news sources and Twitter after the Great East Japan Earthquake.

## Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: [Text analysis]; H.3.5 [Online Information Services]: Web-based services; H.3.6 [Library Automation]: Large text archives

## General Terms

Algorithms, Experimentation, Languages

## Keywords

social media, web news, corpus analysis, disaster management

## 1. INTRODUCTION

Having complete, accurate and up-to-the-minute situational awareness is essential in disaster response management, and the Internet has become a crucial source of such information. In addition to understanding the natural disasters that cause physical damage and injuries, the human decision makers need to consider the social phenomena such as the rehabilitation and rebuilding process for the affected areas and how to guide the organizational and individual efforts. Thanks to pervasive Internet technologies and mobile devices, the information from the Internet can help track many social phenomena.

To support decision makers, we started a project for disaster management based on Internet media text analysis. The main focus of the project is text analysis methods for a large Web data archive from the 2011 Great East Japan Earthquake [7]. The analysis focuses on mass media (Internet news articles) and social media (Twitter). We collected massive datasets and have been applying chronological and territorial analysis of their corpora for each type of media. We presented a preliminary report of our results for the Internet news analysis at a domestic workshop (in Japanese) [5].

One of the more revaling approaches in our analysis was to look at the differences among multiple sub-corpora to find relatively unusual keywords in terms of chronology, locations, or types of media. To detect these variations, we devised metrics for statistically assessing the importance of each keyword [4] based on its TF-IDF weight [3, 1], and then compared the values among multiple sub-corpuses. For example, we can detect keyword trends characteristic of the Twitter corpus, and contrast those trends with the Internet news corpus.

In this paper, we outline goals and approach of our project. Section is an outline of our project and Section 3 shows some experimental results from the Internet news corpus analysis for the 2011 Great East Japan Earthquake. Related work is in Section 4, and Section 5 concludes this work-in-progress paper.

## 2. PROJECT OUTLINE

Many researchers reported that social media such as Twitter and Facebook were used effectively after the 2011 Great East Japan Earthquake [6, 8]. While this was partially due to the unavailability of alternative websites and to the confused reports regarding the nuclear power station disaster, we believe that the primary reasons were simply the popularity of such websites combined with functional characteristics that made them suitable for use in the emergency situation.

The initial goal of our project was to investigate the characteristics of social media responses during and after a disaster. We hope to leverage these characteristics to extract valuable information from the social media to support disaster responses. Unfortunately, the basic nature of social media is too much similar to mass media [2]. This is because mass media have a huge impact on the way people think, and thus strongly affect the content of social media, obfuscating social media-specific information.

To see the true characteristics of social media, it is crucial to understand the effects coming from traditional mass media, while also considering the chronology and geography. We developed a method for statistically highlighting relatively unusual (and thus important) keywords in a sub-corpus that is compared to another sub-corpus [4]. While this was only used to recognize the chronological phases of a corpus, we can extend it for more general purposes when comparing keywords in multiple sub-corpuses.


*WWW 2012 Companion,* April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

| Ranking | 10 hours | 100 hours | 1,000 hours |
|---|---|---|---|
| 1 | Earthquake | Earthquake | Evacuation |
| 2 | Afternoon | Afternoon | Fukushima |
| 3 | Miyagi | Fukushima | Nuclear power plant |
| 4 | Tsunami | Tsunami | Affect |
| 5 | Tokyo | Nuclear power plant | The Great East Japan Earthquak |
| 6 | Happen | Morning | Support |
| 7 | Fukushima | Prefecture | Prefecture |
| 8 | Offshore | Blackout | Tsunami |
| 9 | Tohoku | Evacuation | Disaster Area |
| 10 | Damage | Damage | Earthquake |

Figure 1: Ranking of frequency of keywords after the 2011 Great East Japan Earthquake for each time slot.

| Ranking | 10 hours | 100 hours | 1,000 hours |
|---|---|---|---|
| 1 | Elementary school | Whip-round | Election |
| 2 | Facility | Group | Compensation |
| 3 | Atomic power | Explosion | Becquerel |
| 4 | The said company | Train service suspension | Candidate |
| 5 | Power generation | Hydrogen | Charge |
| 6 | Cooperation | Volunteer | Incumbent |
| 7 | Cooling | Kyoto | Vote |
| 8 | Airport | Everyone | Majesty |
| 9 | Nuclear power plant | Monetary donation | Legislative seat |
| 10 | Committee | Exposure | Fresh candidate |

Figure 2: Ranking of weighted importance of keywords after the 2011 Great East Japan Earthquake for each time slot.

## 3. WEB MEDIA CORPUS ANALYSIS

To extract the keywords, we first use the well-known TF-IDF (Term Frequency / Inversed Document Frequency) index, which is often used to weight the importance of each term (keyword). This weight is a statistical metric used to evaluate how important each term is to a document within a collection or corpus. The TF-IDF index for a term is calculated as

$$TF - IDF(t_i, d_j) = TF(t_i, d_j) \cdot IDF(t_i)$$

$$IDF(t_i) = log_{10} \frac{N}{DF(t_i)},$$

where

$t_i$: a term,

$d_j$: document containing $t_i$,

$N$: the number of documents,

$TF(t_i, d_j)$: how many times a given $t_i$ appears in $d_i$, and

$DF(t_i)$: how many documents contain $t_i$.

By using two or more sub-corpora, we have different TF-IDF indexes for the same keyword as it appears in each sub-corpus.

Then we inversely weight the TF-IDF index of a keyword in a target corpus using the TF-IDF index of the same keyword in a base corpus. For a chronological series, the indexes of the keywords in a corpus within 10 hours to 100 hours after the earthquake can be weighted using the indexes of the keywords in another corpus within the first 10 hours. Figure 1 shows the frequency of keywords and Figure 2 shows the actual keyword importance ranking after the Great East Japan Earthquake, extracted from our preliminary results [5].

## 4. WORK IN PROGRESS

We are currently analyzing a Twitter corpus and comparing it with a corpus of Internet news and to another corpus of official press releases from Japanese government sources. We will present our latest results at the workshop and seek feedback from the experts in attendance.

## 5. REFERENCES

[1] A. Aizawa. An information-theoretic perspective of TFâĂŞIDF measures. Information Processing & Management, 39(1):45–65, Jan. 2003.

[2] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In WWW 2010 , Proceedings, pages 591–600, 2010.

[3] G. Salton. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5):513–523, 1988.

[4] S. Sato, H. Hayashi, K. Inoue, and T. Nishino. Visualizing chronological behavior of disaster social aspect based on web news articles on disasters and crises. Journal of Visualization, 29(7):17–26, 2009. (in Japanese).

[5] S. Sato, F. Imamura, and H. Hayashi. Basic analysis of the web news corpus broadcasted the 2011 great east japan earthquake disaster. Journal of Social Safety Science, (15):303–311, 2011. (in Japanese).

[6] F. N. Shigyo. The Great East Japan Earthquake: How Net Users Utilized Social Media? The NHK Monthly Report on Broadcast Research, 61(8):2–13, Aug. 2011.

[7] M. Simons, S. E. Minson, A. Sladen, F. Ortega, J. Jiang, S. E. Owen, L. Meng, J.-P. Ampuero, S. Wei, R. Chu, D. V. Helmberger, H. Kanamori, E. Hetland, A. W. Moore, and F. H. Webb. The 2011 magnitude 9.0 Tohoku-Oki earthquake: mosaicking the megathrust from seconds to centuries. Science (New York, N.Y.), 332(6036):1421–5, June 2011.

[8] Y. N. Yoshitsugu. Roles of social media at the time of major disasters observed in The Great East Japan Earthquake: twitter as an example. The NHK Monthly Report on Broadcast Research, 61(7):16–23, 2011.