Tweeting about the Tsunami? - Mining Twitter for Information on the Tohoku Earthquake and Tsunami

Akiko Murakami
IBM Research - Tokyo
1623-14 Shimo-tsuruma, Yamato, Kanagawa
242-8502, Japan
akikom@jp.ibm.com

Tetsuya Nasukawa
IBM Research - Tokyo
1623-14 Shimo-tsuruma, Yamato, Kanagawa
242-8502, Japan
nasukawa@jp.ibm.com

ABSTRACT

On 11th March 2011, a 9.0-magnitude megathrust earthquake occurred in the ocean near Japan. This was the first large–scale natural disaster in Japan since the broad adoption of social media tools (such as Facebook and Twitter). In particular, Twitter is suitable for broadcasting information, naturally making it the most frequently used social medias when disasters strike. This paper presents a topical analysis using text mining tools and shows the tools' effectiveness for the analysis of social media data analysis after a disaster. Though an ad hoc system without prepared resources was useful, an improved system with some syntactic pattern dictionaries showed better results.

1. INTRODUCTION

The 2011 Tohoku earthquake and tsunami were the first large-scale natural disasters in Japan became widespread. Though the wired and mobile phone networks went down just after the accidents, the Internet was widely accessible. Although people in the worst-affected region could not use the Internet because of the power outages, social media still played an important role in exchanging and gathering information and in confirming the safety of family members and friends without phones.

Most of the social media data is open to the public, so we can sense how people are feeling, both for the affected and non-affected people, and learn about their thoughts, demands, and opinions. We used text mining technologies with this social media data to extract and analyze data related to the disaster. These kinds of social media websites revealed peoples' real needs through shortage-related comments. We also identified many problems related to the disaster, such as fraud and rumors. This research is still at an early stage, but our results show the effectiveness of text mining techniques using social media data after a disaster.

2. PREPARATION FOR TEXT MINING

As a part of our urgent response to the disaster, we used our text mining technology, named TAKMI, to analyze social media data.

2.1 Data and Resources

During and after the disaster, Twitter (where each message called a tweet) was the leading social media service that people could use to exchange and gather information. Though we can not access all of this Twitter data, we gathered Japanese tweets with the words "地震"(Earthquake) or "被災"(Disaster-affected) or with hash tags such as "#jisin" and "#jishin" (Romaji forms of earthquake). We also collected tweets with "原発"(Nuclear Plants) after the 14th, since news on the nuclear accident first appeared in the mass media on that date.

We prepared two sets of Twitter data. One set is the data collected from the 13th to the 16th of March 2011, which contains 280,429 tweets. The other set is data collected from the 13th to the 28th, which contains 1,135,495 tweets. Most of the tweets were written in Japanese. Although this was not all of the data from Twitter over those times, it is enough data for trends in the social media after the emergency.

Our text mining system can extract knowledge using domain-knowledge dictionaries, such as technical dictionaries and pattern dictionaries. First we used the system without any domain knowledge, since we wanted to use the tools effectively as soon as the disaster struck. It is hard to prepare such supporting information quickly, so it is natural that a system will initially be used without such specialized knowledge. After the first preliminary analysis, we created some simple pattern dictionaries to extract more knowledge. We will give details about the dictionaries in Section 3.1.2.

2.2 Text Mining Tool

One of the most important parts of the Twitter data is textual data, where we used our text mining technologies. Our TAKMI technology was productized as ICA (IBM Content Analytics)[1]. It provides powerful natural language preprocessing including pattern-matching with a customized syntactic pattern dictionary and interactive mining functions with a visual interface. The interface supports keyword and full-text searches and drilling into the entire documents with various search conditions, thus revealing various types patterns such as document frequencies over time and the relevance of the selected documents.

3. MINING RESULTS USING TEXT MINING TECHNOLOGIES

This section describes some of the main results we obtained from social media analysis using our text mining techniques just after the disasters.

3.1 Identifying Shortage of Supplies

Many types of supplies were in short supply at the time of disaster. The shortages were due to various causes such as damaged transportation infrastructure, shortages of power and water, and so on. The specific shortages changed over time, making it hard to identify the "needs in the disaster area" as the situation evolved. In this section, we show some typical examples of analysis that identified shortages and their trends. First we describe the results with no special resources using simple keyword matches. and then describe the results with syntactic pattern dictionaries.

[1]http://www-01.ibm.com/software/ecm/content-analytics/bundle.html

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

Table 1: Relevant nouns in tweets contains "Shortage"

Noun	Frequency	Correlation [2]
personnel resources	35	48.6
supplies	21	12.2
consideration	25	7.8
heating oil	30	7.3
provision	32	6.0
medicine	25	5.7
electric power	63	5.4
food	31	4.6

3.1.1 *Finding Shortages of Supplies without Analytic Resources*

In this section, we describe the effectiveness of the tools itself in the mining results with no special resources. First we narrowed the tweets with the word "足りない"(shortage) to identify an insufficient supply of something. In the small sample from the 13th to the 16th, there were 1,251 such tweets. Next we searched for correlations among the nouns in this set of the tweets, and extracted the relevant nouns as possible shortages. Table 1 shows this noun list with document (tweet) frequencies and correlation values. The results showed that people ran short of supplies and food, but also of less physical resources (such as humans and electricity). The system found trends of growing needs for supplies. The results show that "personnel resources" and "gasoline" shortages continued throughout this period. In contrast, "electricity", "information" and "food" become less frequent in the second half of the period.

3.1.2 *Finding Shortages of Supplies with a Pattern Dictionary*

The results in Section 3.1.1 showed that the system without resources for specific domains or purposes was able to quickly identify the affected people's needs. However, the results included some noise where irrelevant nouns accidentally appeared with the word "shortage", so the list of nouns included some unneeded words. Since we wanted a more advanced analysis of the data, we prepared a syntactic pattern dictionary to identify things that were in short supply. Here are some example patterns.

- ¡Noun¿ が 買えない
  (cannot buy ¡Noun¿)
- ¡Noun¿ が 売り切れ
  (¡Noun¿ is sold out)

With this pattern dictionary the system extracted nouns representing shortages. From the data spanning the 13th to the 28th, 1,015 nouns were extracted as shortages. The five most frequent nouns were "water", "battery", "rice", "gasoline" and "toilet paper". We also found "natto"(Fermented soybeans) and "yogurt" in the top 20, which were not intuitively obvious as shortages in the disaster areas. Since these are fermented foods requiring electricity for production and storage, these foods were quickly exhausted.

Figure 1 shows some trends of noun frequency over time from the 13th to the 27th of March 2011. We can see that references to fermented foods (yogurt and natto) were increasing. We also found some increases in such goods as cigarettes and ice cream. These are not our essential foods but preferences, so people in the affected areas began to seek them after the situation began to improve. In contract, the essential "water" and "mineral water" were increasingly needed because of concerns about radioactive materials escaping from the damaged nuclear plant.

[2] "Correlation" value is a degree of correlation between document set and an item. [1]

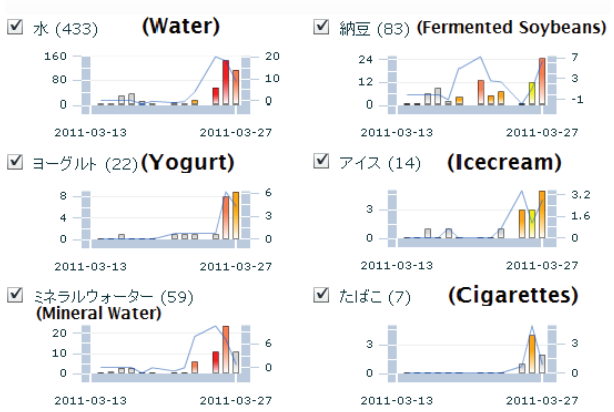

Figure 1: Trend of shortages extracted syntactic pattern dictionaries.

3.2 Identifying Types of Fraud related to Disasters

Relief fraud is one of the major concerns not only for people in an affected area, but also for people elsewhere. At the time of the disaster, various types of fraud were reported throughout Japan. We investigated tweets including the words "詐欺"(fraud) or 詐欺師 "(fraudster)" and "騙す"(cheat) in the data from the 13th to the 28th. There were 33,651 tweets containing these words, and we found "義援金"(donation) was a relevant words so we also searched for the tweets that contained that word adding an additional 3,051 tweets.

Based on the selected tweets we found "貴金属"(jewels) as a frequent reference in these tweets, which helped us identify a widely form of disaster fraud. These criminals went to houses where an old person lived, asking them to donate unused jewelry to help pay for relief efforts.

To identify the areas where this type of fraud was occurring, we extracted the locations of the tweets using named entity extraction. In 37 of these tweets that referenced fraud, donation and jewels there were several locations mentioned, such as Ina, Fukuoka, and Saga. Most of these locations are in the Kyushu area, though some were elsewhere. Even the number of tweets showed a trend in which this type of fraud started in Fukuoka (in Kyushu), Hokkaido and Nagano simultaneously, and then it diffuses across Kyusyu area.

4. CONCLUSIONS

We assessed the effectiveness of our text mining tool for Twitter data from a time of disaster. Though we only used the textual data in the tweets, the Twitter includes other rich information such as the authors, replies and retweet information. There is potential value from network analysis showing how information or rumors spread. In addition, a combination of deep text analytics and network analytics may more clearly reveal human behaviors and needs when disasters strike.

5. REFERENCES

[1] W.-D. Zhu, A. Iwai, T. Leyba, J. Magdalen, K. MacNeil, T. Nasukawa, P. N. (Nita), and K. Sugano. IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content. IBM Redbooks, 2011. ISBN: 0738435287