Affected People’s Needs Detection after the East Japan Great Earthquake  
- Time Series Analysis using LDA -

Takako Hashimoto∗, Basabi Chakraborty†, Supavadee Aramvith‡, Tetsuji Kuboyama§ and Yukari Shirota§

∗Chiba University of Commerce, Chiba, Japan  
Email: takako@cuc.ac.jp
†Iwate Prefectural University, Iwate, Japan  
Email: basabi@iwate-pu.ac.jp
‡Chulalongkorn University, Bangkok, Thailand  
Email: supavadee.a@CHULA.AC.TH
§Gakushuin University, Tokyo, Japan  
Email: kuboyama@tk.cc.gakushuin.ac.jp, yukari.shirota@gakushuin.ac.jp

Abstract—After the East Japan Great Earthquake happened on Mar. 11, 2011, many affected people who lost houses, jobs and families fell into difficulties. Governmental agencies and NPOs supported them by offering relief supplies, foods, evacuation centers and temporary houses. When various supports were offered to affected people, if Governmental agencies and NPOs could detect their needs appropriately, it was effective for supporting them. This paper proposes the method to extract affected people’s needs from Social Media after the Earthquake and analyze their needs changes over time. We target the blog that expressed thoughts, requirements and complaints of affected people, and adopt the Latent Dirichlet Allocation (LDA) that is one of popular techniques for topic extraction. We then compare the analysis result with affected people’s actual situation and real events and evaluate the effectiveness of our method. In addition, we evaluate the effectiveness as the method that can help decision making for providing appropriate supports to affected people.

I. INTRODUCTION

Three years have already passed after the East Japan Great Earthquake happened on Mar. 11 in 2011. After the earthquake, many affected people who lost houses, jobs and families fell into difficulties. Governmental agencies and NPOs supported them by offering relief supplies, foods, evacuation centers and temporary houses. However, since it was an unprecedented disaster, supports have been done without confidence. It was not sure if the supports were provided appropriately for answering affected people’s needs.

It is difficult to evaluate if the supports were provided appropriately for affected people’s needs. To analyze affected people’s needs over time, this paper proposes the method to detect affected people’s needs from the blog that expresses affected people needs/thoughts/opinions after the earthquake. The target blog is called "Banya Nippou" provided by the NPO "SAVE IWATE" that is one of the biggest NPO in Iwate prefecture[1]. The blog shows affected people’s thoughts, requirements and so on acquired by NPOs. From the blog, we extract affected people needs and analyzed them over time by adopting LDA (Latent Dirichlet Allocation)[3]. LDA is one of major topic extraction methods. We assume affected people’s needs as topics, so that we can detect affected people’s needs changes over time.

We then compare the analyzed results with actual feeling about affected people’s situation by NPO and evaluate the effectiveness of our method. In addition, we discuss if the method is effective for decision making for providing appropriate supports to affected people. This paper can show the effectiveness of needs detection by LDA, and the possibility of developing the decision making support system based on the topic extraction.

This paper is organized as follows. Section II presents related work on topic extraction in social media. In Section III, the target blog provided by the NPO "SAVE IWATE" is introduced. Section IV explains our proposed method using LDA for detecting affected people’s needs changes. Section V discusses the effectiveness of needs detection by our method, and the possibility of developing the decision making support system. Finally, Section VI concludes this paper.

II. RELATED WORK

Most related works for detecting topics from social media, extract keywords from messages posted by users and compute topics based on occurrence numbers or co-occurrences between keywords. Asur et al.[5] investigated trending topics on Twitter. They proposed a simple model based on the number of tweets and found that the resonance of the content with the users of the social network plays a major role in causing trends. Liu et al.[6] and Cao et al.[7] focus on web video analysis. Especially, Cao et al. cluster video tags into groups to get small events and then link these events into topics based on textual and video similarity. Radinsky et al.[8] proposed Temporal Semantic Analysis (TSA), a semantic relatedness model, that captures the words temporal information. They targeted words in news archives (New York Times, etc.) and used the dynamic time warping technique to compute a semantic relation between pre-defined words. Wang et al.[9] proposed time series analysis which has been used to detect similar topic patterns.
Pennacchiote et al. [10] targeted the twitter and extracted topics by LDA. They found the document that consists of a set of tweets that were posted by one user can achieve better outcomes for topic extraction than the document that consists of just one tweet.

Our proposed method also focuses on extracting topics based on co-occurrences between keywords from social media. We already proposed methods using network community mapping and Latent Semantic Analysis [12], [13]. This paper utilizes LDA method as well as Wang [9] and Pennacchiote [10]. We target topics related to the East Japan Great Earthquake and flexibly visualize topic transition over time. In addition, we evaluate the experimental results (affected people reactions after the earthquake) by the interviews to NPO people. A quantitative evaluation with user study has been done in our research. It is the novelty of our work.

III. TARGET BLOG AND EXPECTATION FROM NPO

This paper targets the blog titled “Bannya Nippou” [2] provided by SAVE IWATE that is one of the biggest NPO in Iwage Prefecture. Save Iwate was formed immediately after the East Japan Great Earthquake. They have supported affected people by providing information about the well-being of affected people, investigating about disaster area, offering necessity goods and supporting affected people’s life. To provide appropriate support to affected people at an appropriate timing, they wanted to know affected people’s needs changes adequately. For example, they soon started the support center for affected people after the quake. This support was really necessary and effective. However, the problem is to take the decision of closing the center at the right time. As time advances after the quake, they have felt that the necessity on different supports (e.g. providing clothes) was gradually reducing. They have to decide to end the support. However, such a decision-making was very difficult for NPO members. And they had to know how closing of the center would influence affected people. Therefore, we believe that it is useful to know affected people’s thought by an information technology tool. If the tool can analyze affected people’s needs changes and visualize them, it can be a kind of the decision making support system.

"Bannya Nippou" is provided by SAVE IWATE and expresses thoughts, complaints and emotions of affected people. Figure 1 is a screen example of the "Bannya Nippou" blog. The blog was started immediately after the earthquake and still continue to express affected people’s thoughts. It is appropriate to adopt our method to "Bannya Nippou" messages as the 1st step for "Decision Making Support System for Affected People Support".

IV. NEEDS DETECTION USING LATENT DIRICHLET ALLOCATION

A. Proposed Method

Our proposed method consists of the following 5 steps (Figure 2).

- Step A: Data Crawling
- Step B: Language Processing
- Step C: Topic Extraction
- Step D: Time Series Topic Detection
- Step E: Visualization

Fig. 1. Screen Example of Bannya Nippou

Note that the Framework was already shown in our previous work [12], [13]. In this paper, in Step C and Step D, we utilize the Latent Dirichlet Allocation (LDA) model for topic detection.

1) Step A: Data Crawling:

STEP A acquires messages $D = \{d_i\}$ from the target social media. In this paper, this step crawled blog messages from the Bannya Nippou from Jun. 3, 2011 to Jul. 27, 2012.

One message is defined as one document $d_i$ and the step retrieves it as the following tuples:

$$d_i = (MID_i, Posted_i, Title_i, Content_i).$$

Here, $MID_i$ is an ID of each document, $Posted_i$ is a date-time that each document was posted, $Title_i$ is a title of each document and $Content_i$ is a text of each document.
2) Step B: Language Processing:

The step B extracts keywords $KW = \{kw_i\}$ that are nouns/verbs/adjectives/adverbs from Content of each $d_i$ by morphological analysis. We utilized Mecab [16] which is a Japanese morphological analyzer.

$$kw_i = (\text{MID}_i, \text{Posted}_i, \{w_{ij}\}).$$

Here, $\{w_{ij}\}$ is a list of extracted keywords from document $d_i$.

3) Step C: Topic Extraction:

Step C extracts topics focusing on word correlations. In our method, the posted date is delimited by an appropriate distribution for document $i$ may be $(0,0.2,0.4,0.1)$. Suppose that the word generating distribution for the topic economics is supposed to be $(0.3,0.2,0.1,0.4)$. Suppose that the word generating distribution is a multinominal distribution. If the input sentence is “stock finance yield stock exchange stock”, the probability of the sentence becomes as follows:

$$0.1 \times 0.3 \times 0.2 \times 0.4^3 \times 0.1.$$  

The multinominal and Dirichlet distributions are defined in machine learning textbooks [14].

One document is supposed to have several topics. The topic distribution for document $i$ may be $(0.7,0.3)$ (in the case of an economic related document) or $(0.2,0.8)$ (in the case of a disaster related document). To express the possible various distributions, we use Dirichlet distribution by using a hyper parameter $\alpha$. On the same way, we define per-topic word distribution by using another hyper parameter $\beta$. The used symbols are as follows:

- $\alpha$ is the parameter of the Dirichlet prior on the per-document topic distributions,
- $\beta$ is the parameter of the Dirichlet prior on the per-topic word distribution,
- $\theta_{ij}$ is the topic distribution for document $i$,
- $\phi_k$ is the word distribution for topic $k$,
- $z_{ij}$ is the topic for the $j$th word in document $i$, and
- $w_{ij}$ is the specific word.

The $w_{ij}$ are the only observable variables, and the other variables are latent variables. $\phi$ is a Markov matrix of which size is $KV$ ($V$ is the dimension of the vocabulary) each row of which denotes the word distribution of a topic. The LDA generative process for a corpus $D$ consisting of $M$ documents each of length $N_i$ is as follows where $K$ denotes the number of topics:

1) Choose $\theta_i \text{Dir}(\alpha)$, where $i \in 1, \ldots, M$ and $\text{Dir}(\alpha)$ is the Dirichlet distribution for parameter $\alpha$.
2) Choose $\phi_k \text{Dir}(\beta)$, where $k \in 1, \ldots, K$.
3) For each of the word positions $i,j$, where $j \in 1, \ldots, N_i$, and $i \in 1, \ldots, M$:
   a) Choose a topic $z_{ij}. \text{Multinominal}(\theta_{ij})$.
   b) Choose a word $w_{ij}. \text{Multinominal}(\phi_{z_{ij}})$.

We want to obtain an estimate of $Z$ that gives high probability to the words that appear in the corpus. $z_{ij}$ represents the topic for the $j$th word in document $i$. This problems becomes a maximum a posteriori estimation of $P(W,Z,\Theta,\Phi|\alpha,\beta)$. By an integration concerning $\theta$ and $\phi$, the expression becomes a simple one, $P(W,Z|\alpha,\beta)$. Therefore, we want to obtain $Z$ so that $P(Z|W,\alpha,\beta)$ is maximum. The $W$ is given data. The cost of the calculation is too high because the estimation space size is the number of topics ($K$) to the power of the dimension of the vocabulary $(V)$, $K^V$. Namely each word has $K$ options independently. So instead of that, a random walk search method by Gibbs sampling is widely used[17]. We used the R packaged offered by The Comprehensive R Archive Network abbreviated as CRAN titled lda: Collapsed Gibbs sampling methods for topic models developed by Jonathan Chang[18]. In LDA method, the number of topics should be decided in advance. Generally, the number of topics is regarded as from 5 to 10. To decide the appropriate number of topics is quite important. The number of topics that indicates topic changes clearly will be defined by trial and error.

4) Step D: Time Series Topic Transition Detection:

This step evaluates extracted topics over time and analyzes similarities between them to recognize topic transitions. Regarding similarities between topics, in our previous work, we utilized the graph edit distance[12] and LSA[13](Analysis). But this paper employs LDA and computes estimated topics’ probabilities over time to analyze time series topic transition.

5) Step E: Visualization:

This step visualizes extracted topics and topic transitions by network graphs and so on. It becomes possible to recognize topics occurred and their transitions.

B. Experimental Result

Step A crawled "Bannya Nippou" messages from Jun. 3rd of 2011 to Jul. 27th of 2012, and got $D = 502$ messages and 2741 unique keywords. Table I shows some message example and Table II shows some keyword example.

Then, for derived keywords by Step B, Step C extracted topics using LDA. We tried to extract several numbers (from 5 to 10) of topics by LDA, and in this experiment, we decided the number of topics as 5 and did 25 iterations. Table III shows
main keywords in the 5 topics. We evaluated the meanings of 5 topics as follows:

1) Affected people wanted to go to temporary house, and felt fear about information and government
2) Affected people felt complaints about temporary houses and livelihood, and required jobs and information
3) Affected people wanted to receive continued support of supplies and foods and had hope about new life
4) Affected people wanted lives back, and felt fear about jobs and health
5) Affected people wanted to stay with children and grandchildren, enjoy events, and take mental care

To evaluate the quality of estimated topics, we computed the perplexity that is monotonically decreasing in the likelihood of the data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance. Generally, the perplexity $PP$ is computed as follows:

$$PP = 2^H(p) = \sum_{x} \log p(x)$$

Here, $X = \{w_1, w_N\}$ is words in the document. $N$ is the number of words in the document. In the experiment, the perplexity was 3.58, so that the document set of the target blog was recognized as the better model.

Then, STEP D computed occurrence ratios of above 5 topics over time and STEP E visualized them. Table IV shows topic occurrence ratios and Figure 3 visualizes time series topic transitions.

Table IV. Topic Occurrence Ratio by Month

<table>
<thead>
<tr>
<th>period</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun-11</td>
<td>0.67</td>
<td>0.18</td>
<td>0.06</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Jul-11</td>
<td>0.51</td>
<td>0.35</td>
<td>0.04</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Aug-11</td>
<td>0.27</td>
<td>0.56</td>
<td>0.06</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>Sep-11</td>
<td>0.34</td>
<td>0.57</td>
<td>0.00</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>Oct-11</td>
<td>0.16</td>
<td>0.54</td>
<td>0.02</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>Nov-11</td>
<td>0.20</td>
<td>0.36</td>
<td>0.02</td>
<td>0.04</td>
<td>0.38</td>
</tr>
<tr>
<td>Dec-11</td>
<td>0.11</td>
<td>0.27</td>
<td>0.05</td>
<td>0.04</td>
<td>0.53</td>
</tr>
<tr>
<td>Jan-12</td>
<td>0.31</td>
<td>0.20</td>
<td>0.04</td>
<td>0.08</td>
<td>0.37</td>
</tr>
<tr>
<td>Feb-12</td>
<td>0.16</td>
<td>0.21</td>
<td>0.15</td>
<td>0.19</td>
<td>0.28</td>
</tr>
<tr>
<td>Mar-12</td>
<td>0.12</td>
<td>0.17</td>
<td>0.15</td>
<td>0.32</td>
<td>0.25</td>
</tr>
<tr>
<td>Apr-12</td>
<td>0.04</td>
<td>0.08</td>
<td>0.24</td>
<td>0.40</td>
<td>0.24</td>
</tr>
<tr>
<td>May-12</td>
<td>0.04</td>
<td>0.05</td>
<td>0.21</td>
<td>0.46</td>
<td>0.25</td>
</tr>
<tr>
<td>Jun-12</td>
<td>0.06</td>
<td>0.02</td>
<td>0.24</td>
<td>0.50</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Topic 1 has been shown in the early stage of the earthquake (from June to August, 2011). Topic 2 has increased from August to November, 2011. Topic 3 has occurred in 2012. Topic 4 has also been shown to emerge greatly in 2012. Especially, it has increased after March of 2012. Topic 5 has
increased during November, 2011 to January 2012. It seems to be related to year end and new year specialties.

From the above graph, we can conclude that topic 1 indicated that affected people who lived in the evacuation centers wanted to move to the temporary houses and had complaints about the lack of information after the earthquake. Topic 2 indicated the needs soon after moving to the temporary houses. We suppose that even after people moved to the temporary houses, the livelihood was inconvenient and information was not enough yet. Topic 3 indicated people’s needs after one year of the event, which seems to be related to continued support. One year after the earthquake, various supports were gradually close, so that affected people wanted to continue support for supplies and foods. Topic 4 also increased after one year of the event and is supposed to be related to people’s fear about jobs, stability of life and health. Topic 5 was related to the year-end and new year specialties, so that people concerned a time with children and grandchildren. In addition, they seemed to feel the necessity of a mental health care.

V. Effectiveness of Extracted Needs and Possibility for Decision Support System for NPO

In this section, we evaluate whether the analyzed results indicate actual affected people’s situation and needs changes. We asked the SAVE IWATE people to check the results and give comments on them.

Table V shows the actual affected people situation and real events after the earthquake provided by the SAVE IWATE people.

<table>
<thead>
<tr>
<th>Month</th>
<th>Affected People Situation</th>
<th>Real Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar-11</td>
<td>Move to Evacuation Center</td>
<td>Grate Earthquake Occurred</td>
</tr>
<tr>
<td>Apr-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May-11</td>
<td>Secondary EvacuationMove to other Evacuation Centers</td>
<td></td>
</tr>
<tr>
<td>Jun-11</td>
<td>Secondary EvacuationMove to other Evacuation Centers</td>
<td></td>
</tr>
<tr>
<td>Jul-11</td>
<td>Started moving to Temporary Houses</td>
<td>Bon Vacation,</td>
</tr>
<tr>
<td>Aug-11</td>
<td>Most Affected People moved to Temporary Houses</td>
<td></td>
</tr>
<tr>
<td>Sep-11</td>
<td>Unemployment Insurance Lapsed, Fear about Jobs</td>
<td></td>
</tr>
<tr>
<td>Oct-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov-11</td>
<td>Complaint about Cold</td>
<td>Winder Season</td>
</tr>
<tr>
<td>Dec-11</td>
<td>Complaint about Cold in Inland’s Temporary Houses</td>
<td>End of Year</td>
</tr>
<tr>
<td>Jan-12</td>
<td>Stay with FamilyNeeds related to Winter</td>
<td>New Year</td>
</tr>
<tr>
<td>Feb-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar-12</td>
<td>Expectation to New Life</td>
<td>Graduation</td>
</tr>
<tr>
<td>Apr-12</td>
<td>Expectation to New LifeFear about Rebuilding</td>
<td>New Semester, Entrance into School</td>
</tr>
<tr>
<td>May-12</td>
<td>Notification about the support center closing</td>
<td></td>
</tr>
<tr>
<td>Jun-12</td>
<td>Requirement about Continued Support</td>
<td>Notification about the support center closing</td>
</tr>
<tr>
<td>Jul-12</td>
<td>Stop Cloth Delivery</td>
<td></td>
</tr>
</tbody>
</table>

After the East Japan Great Earthquake happened on Mar. 11, 2011, a lot of people moved to the evacuation center. Some of them moved to other evaluation centers again as the secondary evacuation. Finally, they started moving to the temporary houses in July. During July and August, most of them could move to the temporary houses. In September, since the unemployment insurance lapsed, affected people felt fear about jobs and their future.

In Winter, affected people were chilled to the bone, and they started feeling frustrated by the temporary houses. Actually, the quality of temporary houses was not stable, so that people who lived in low-quality temporary houses particularly complained a lot. In addition, they needed goods for winter.

In March and April of 2012, it was a time of entrance into school or a new semester, so that people were motivated positively, and also felt fear about rebuilding. In June, since the SAVE IWATE decided to close the support center, affected people were supposed to feel fear about continued support.

Figure 3 also added actual events and affected people’s situation. We could find that the timings of topic changes strongly were related to affected people situation changes. We can say that needs extracted by our method can express affected people’s situation changes.

Next, we discuss whether needs extracted by our proposed method can be utilized for decision making for NPO activities. Extracted needs could express affected people’s situation well, so that we believe that by our method, NPO people can confirm the thought/needs of affected people. For example, NPO people can objectively recognize affected people’s thought/needs by knowing complaints about the temporary houses and expectations for the year-end and New Year and needs for mental care by our method. Information provided by our method can support NPO activities for reducing affected people’s fear, so that we can say that our proposed method can be the decision support tool for NPO members.

VI. Conclusion

This paper proposed the method using LDA to detect affected people’s needs from the blog that expresses affected people needs/thoughts/opinions after the earthquake. In our method, affected people’s needs were recognized topics and extracted over time by LDA, so that the method could visualized time series affected people’s needs changes.

In addition, we evaluated the results of our method with NPO members by comparing the with real events and actual affected people’s situation. We could confirm extracted needs indicated affected people’s situation well and our proposed method can be useful for the decision making support tool for NPO members.

We are going to conduct experiments more for other data sources and evaluate results with NPO members to show our method’s effectiveness. We also plan to archive our results as the disaster record. To explore possibilities for decision support system, we improve our method to analyze topics in real time, such as Dynamic Topic Models [19], Topic Tracking Model [20], and so on.
ACKNOWLEDGMENT

The authors would like to thank SAVE IWATE that provided blog data, affected people’s situation information and helped the evaluation. This work was supported by JSPS KAKENHI Grant Number 24500180.

REFERENCES

[18] CRAN web page: http://cran.r-project.org/