Topic Extraction from Millions of Tweets using Singular Value Decomposition and Feature Selection

Takako Hashimoto*, Tetsuji Kuboyama† and Basabi Chakraborty‡
* Chiba University of Commerce, Chiba, Japan
E-mail: takako@cuc.ac.jp
† Gakushuin University, Tokyo, Japan
E-mail: kuboyama@tk.cc.gakushuin.ac.jp
‡ Iwate Prefectural University, Iwate, Japan
E-mail: basabi@iwate-pu.ac.jp

***Abstract*—Social media offers a wealth of insight into how significant topics—such as the Great East Japan Earthquake, the Arab Spring, and the Boston Bombing—affect individuals. The scale of available data, however, can be intimidating: during the Great East Japan Earthquake, over 8 million tweets were sent each day from Japan alone. Conventional word vector-based social media analysis method using Latent Semantic Analysis, Latent Dirichlet Allocation, or graph community detection often cannot scale to such a large volume of data due to their space and time complexity. To overcome the scalability problem, in this paper, high performance Singular Vector Decomposition (SVD) library $redsvd$ has been used to identify topics over time from the huge data set of over two hundred million tweets sent in the 21 days following the Great East Japan Earthquake. While we begin with word count vectors of authors and words for each time slot (in our case, every hour), authors' clusters from each slot are extracted by SVD and k-means. And then, the original fast feature selection algorithm named CWC has been used to extract discriminative words from each cluster. As a result, authors' clusters recognized as topics as well as issues of conventional social media analysis method for big data can be visualized overcoming the scalability problem.**

I. Introduction

Social media offers a wealth of insight into how significant topics–such as the Great East Japan Earthquake, the Arab Spring, and the Boston Bombing–affect individuals. The scale of available data, however, can be intimidating: during the Great East Japan Earthquake, over 8 million tweets per day were sent from Japan alone. Discovering such an event, and classifying tweets relevant to the event, remains an ongoing area of research. Many techniques such as graph based methods [1], Latent Semantic Analysis (LSA) [2] and Latent Dirichlet Allocation (LDA) [3] have been proposed so far, but none of them scales adequately to millions of tweets.

In this paper, to overcome the scalability problem, high performance Singular Vector Decomposition (SVD) library $redsvd$ [4] has been utilized to identify topic clusters over time from the huge data set of over two hundred million tweets sent in the 21 days following the Great East Japan Earthquake, and to confirm the feasibility of topic extraction from big data. Then, CWC [5], a fast feature selection technique is then used to extract discriminative words from the clusters.

The main contributions in this work are as follows:

- to improve the conventional social media analysis method for big data using high performance SVD library $redsvd$ and the original fast feature selection technique CWC.
- to identify topics after the Great East Japan Earthquake from large twitter data.
- to discuss issues of conventional social media analysis method for big data.

We already developed the time series social media analysis technique for blog data related to the Great East Japan Earthquake [6]. But our previous technique targeted just around one thousand blog data. This work targets over 200 million Tweets, so that we have to develop new method for big data.

The paper is organized as follows. Section II introduces related work on social media analysis. Section 1 describes our method using high performance SVD library and the original feature selection technique CWC. Section IV demonstrates experimental results of our method. Section V discusses issues on the conventional social media analysis method. Finally, Section VI concludes this paper and offers directions for future research.

II. Related Work

Most social media analysis methods comprise of the following basic template:

1) Form matrices (or bipartite graphs) of connections between authors (or documents) and words over time.
2) For each matrix, form clusters and adopt a topic modeling technique such as LDA, or k-means [7] algorithm with dimensionality reduction such as LSA or adopt a network community extraction method in case of bipartite graphs.
3) For each cluster, define important keywords to represent the contents (LDA also produces keyword importance scores)

Generally, this conventional method lacks scalability. Existing data mining technique target thousands of items, not millions. For example, Fujino et al. [8] analyzed tweets over time based on LDA, but the number of their targeted tweets was only around 200K. Paul et al. [9] proposed a topic model based on LDA and targeted over 100 million tweets. However,

they had to filter them first to reduce data until it reached to appropriate data size (around 5000 tweets). Zhao et al. [10] analyzed twitter and news article using LDA. At first, the number of targeted tweets was 1 million, but they also filtered the data to reduce its size. Kazama et al. [11] targeted 200 million tweets related to the Great East Japan Earthquake, and tried to analyze them by LDA based technique. However, they employed parallel processing to tackle big data. parallel processing is one of the solutions for handling big data, but to make big data analysis easier, high performance data mining technique is quite necessary.

To alleviate the scalability problem, high performance SVD library $redsvd$ is used here for clustering and the original technique CWC for feature selection.

III. Proposed Method

The proposed method in this paper follows conventional method as well, but to scale to big data, high performance SVD library $redsvd$ is employed for clustering and CWC is used for feature selection (1).

A. Step 1: Creation of Author-Word Count Matrices

In the first step, following conventional methods, the tweets are grouped by a certain period (e.g. hour) during which they were sent. Then the sequence of author-word count matrices , $\langle A_0, A_1 \ldots, A_t, \ldots, A_T \rangle$ that summarizes the words used in tweets by each author during each time slot are created.

$$A_t = \begin{pmatrix} a_{11} & a_{12} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{32} & \cdots & a_{3n} \\ & \vdots & & \ddots & \\ a_{m1} & a_{m2} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_{ij})_t$$

where $1 \leq i \leq m$ and $1 \leq j \leq n$. The index m is the number of authors and n is the number of words during a time period. The element a_{ij} shows the number of times the i-th author used a particular word w_j during a time period. These time series matrices, $A_0, \ldots, A_T$, are obviously sparse. We assume that any significant event does not happen in the first time period $t = 0$, and let A_0 be the initial matrix representing an ordinary state.

B. Step 2: Clustering

We calculate TF-IDF [12] for $(a_{ij})_t$ and apply $redsvd$ for reducing dimensions of each author-word matrix. $redsvd$ is $C++$ library for solving several matrix decompositions. It can handle very large matrix efficiently, and is optimized for a truncated SVD of sparse matrices. For example, $redsvd$ can compute a truncated SVD with top 20 singular values for a 100K x 100K matrix with 1M nonzero entries in less than one second.

Truncated SVD's formula is as follows:

$$A \approx U_r \Sigma_r V_r^T$$

where U_r is an $m \times r$ matrix of authors, Σ_r is an $m \times r$ rectangular diagonal matrix, and V_r^T is an $r \times n$ matrix of words. By setting a specific rank r, A is approximated as $U_r \Sigma_r V_r^T$. Only the r column vectors of U and r row vectors of V^T corresponding to the r largest singular values Σ_r are calculated.

Then a matrix of the first main component to the n-th main component from U_r is obtained and clusters are formed by k-means, each cluster shows a group of authors.

TABLE I
An example of dataset

F_1	F_2	F_3	F_4	F_5	C
0	1	0	0	0	0
1	1	0	1	0	0
1	0	0	1	1	0
1	0	1	1	1	0
0	0	1	0	0	1
1	1	0	0	1	1
1	0	1	1	0	1
0	1	1	0	1	1

C. Step 3: Feature Selection

For clusters of each time slot, the fast feature selection algorithm CWC is applied.

CWC is an accurate and fast feature selection algorithm for categorical data. Feature selection addresses the problem of finding a small set of features relevant to class labels. Table I shows an example of a dataset (note that CWC can deal with multi-category in general, but we use two category problem here for simplicity). The features are denoted by $F_1, \ldots, F_5$, respectively, and the variable of the class labels for instances is denoted by C.

The single feature F_2 is useless to determine the class label since mutual information $I(F_2, C) = 0$. In the same way, the single feature F_5 is also useless due to $I(F_5, C) = 0$. In contrast, the single feature F_4 is more informative than F_2 and F_5 to determine the class label since $I(F_4, C) = 0.13$. Let us consider the combination of features F_2 and F_5. Then, these features completely determine the class label since $I(\{F_2, F_5\}, C) = 1$, and the negation of exclusive-OR of F_2 and F_5 is equivalent to C.

This example suggests that it is essential to search for combination of features relevant to class labels. The most prospective method to address the problem is called *consistency-based feature selection* [13]. If a subset of features is *consistent*, it implies that the subset completely determines all the class labels.

CWC is one of the fastest consistency-based feature selection algorithms. CWC employs the simplest consistency measure for the criteria of feature selection called *binary consistency measure*. This measure just discriminates whether the subset of features can completely determine all the class labels or not. Recently, we have further improved CWC by incorporating a drastically faster search strategy and adapting it to sparse datasets for handling a massive amount of data.

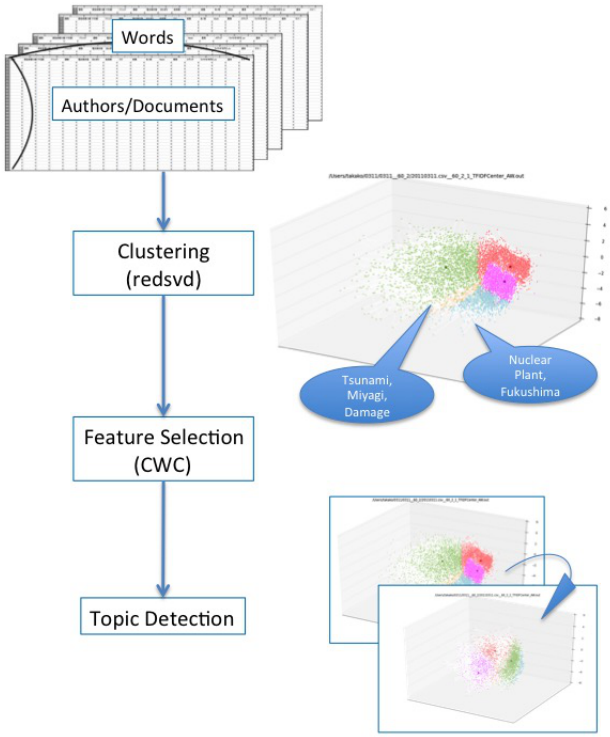

Fig. 1. Conventional Method (a) vs. Proposed Method (b)

IV. Experimental Result

In this section, our experimental results are reported. The experiment is conducted on the MabBook Air 1.7 GHz Core i7 with 8GB memory.

A. target data

Our target data is over 200 million tweets in Japanese that were sent around the time of the Great East Japan Earthquake, starting from March 9, 2011. The social media monitoring company Hottolink [14] tracked users who used one of 43 hashtags (for example, #jishin, #nhk, and #prayforjapan) or one of 21 keywords related to the disaster. Later, they captured all tweets sent by all of these users between March 9th and March 29th. This resulted in an archive of around 200 million tweets, sent by around 1 million users. An average of about 8 million tweets were posted by around 200 thousand authors per day. The average data size per day was around 8GB, and the total data size was over 150GB. (Figure 2). This dataset offers a significant document of users' responses to a crisis, but its size presents a challenge for analysis.

In the following subsections, our experimental result for tweets from 9:00 on March 11 to 24:00 on March 12, a total of 39 hours are shown.

B. Step 1: Creation of Author-Word Count Matrices

In the first step, author-word count matrices are created from the dataset. The fast and customizable Japanese morphological analyzer, MeCab [15] is employed to segment tweets not having spaces to delineate word boundaries,. Author-word count matrices are created for a duration of one hour,e.g, each matrix for an hour on March 11 after 15:00 (the time of the earthquake), contains 600,000-980,000 tweets by 140,000-165,000 authors with over 200,000 words. The total size of each matrix is over 30MB and they were all quite sparse.

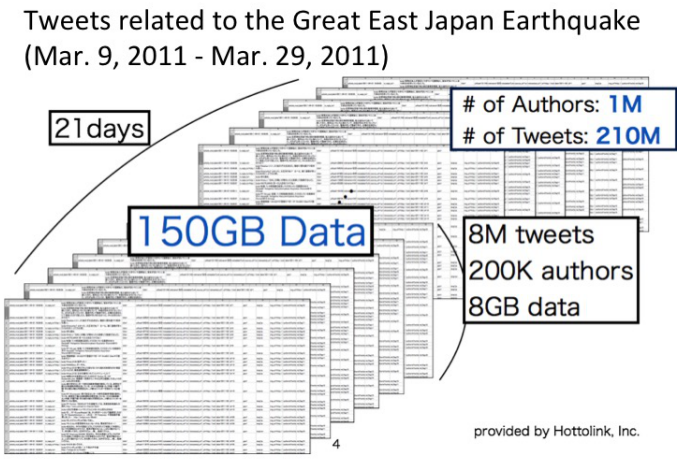

Fig. 2. Target Data: 200 million tweets related to the Great East Japan Earthquake

Table II shows the exact number of authors, words, and total size of each hour's matrix derived from tweets on March 11,

TABLE II
AUTHOR-WORD MATRICES ON MAR. 11

hour (24h)	# of tweets	# of authors	# of words	size of file (MB)
09 - 10	136167	48711	147271	4.6
10 - 11	138491	49101	146940	9.1
11 - 12	148240	52243	149395	9.6
12 - 13	206444	67394	179200	9.5
13 - 14	185175	61513	164897	8.4
14 - 15	351491	103789	163520	12.5
15 - 16	978155	165299	234832	32.5
16 - 17	835257	158711	231822	33.6
17 - 18	745095	154450	228337	32.8
18 - 19	722444	153898	228000	37.2
19 - 20	644618	146167	221226	32.2
20 - 21	621817	142464	225409	30.0
21 - 22	634095	143889	230248	31.1
22 - 23	642385	142940	233102	30.2
23 - 24	629936	138903	229783	29.5

2011.

C. Step 2: Clustering

Then TF-IDF for $(a_{ij})_t$ are calculated and $redsvd$ with rank = 10 has been applied. The performance of redsvd was reasonable. For example, the run-time of redsvd for the matrix during 15:00-16:00 on March 11 (165299 authors $\times$ 234832 words) was less than 10 seconds. We formed clusters by k-means by setting k = 5. From Figure 3, we realize that authors could be divided into five clusters.

D. Step 3: Feature Selection

For five clusters of each time slot, CWC has been adopted for feature selection. *Matthew's Correlation Coefficient* (MCC) [16] is used to order extracted feature words whose score ranges from -1 to 1. The words with high MCC value (> 0) positively express the feature of the cluster while the words with low MCC value (< 0) negatively express the feature of the cluster they belong to. To extract feature words for representation of each cluster, positive words are selected. (All words were originally in Japanese, but translated to English.)

Table III shows the feature selection result during 15:00-18:00 on Mar. 11. According to the feature words in Table III, the topic of each cluster is observed as follows:

- March 11 15:00-16:00
  - $cluster0$: Damage after the quake
  - $cluster1$: Emergency call on the quake
  - $cluster2$: No specific topic
  - $cluster3$: Tsunami warning and evacuation
  - $cluster4$: Message dial for the quake for confirming safety
- March 11 16:00-17:00
  - $cluster0$: No specific topic
  - $cluster1$: Escape from Tsunami with hope
  - $cluster2$: Hoax on Twitter/net
  - $cluster3$: Injury due to the quake
  - $cluster4$: Diffusion of hope, power failure

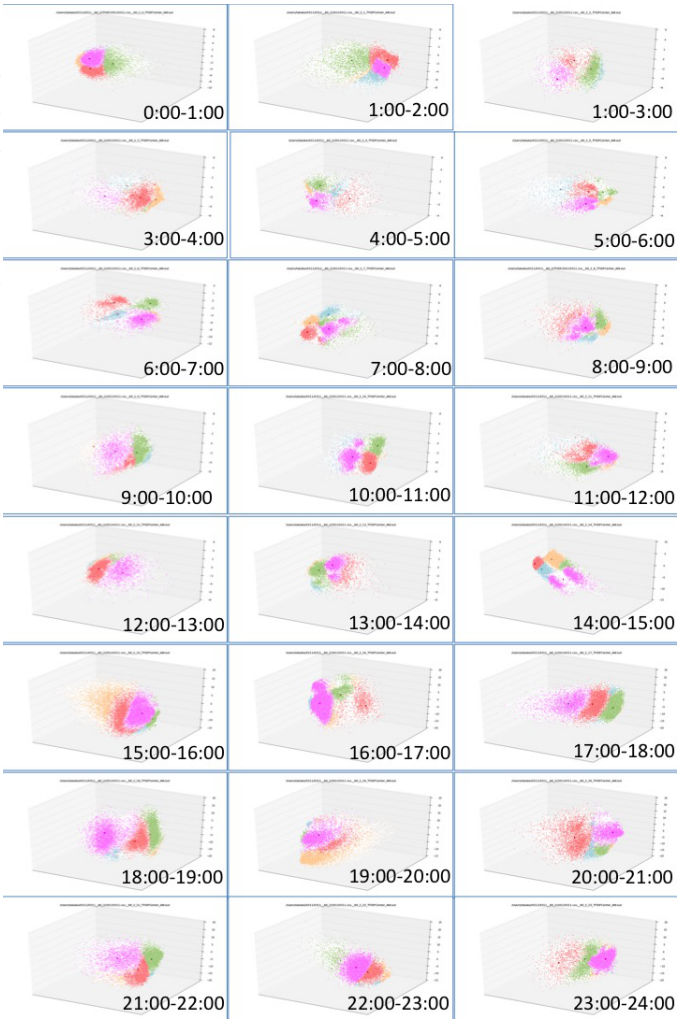

Fig. 3. Clustering Results by SVD and k-means during 0:00-24:00 on Mar. 11 (by hour)

- March 11 17:00-18:00
  - $cluster0$: Diffusion of hope
  - $cluster1$: No specific topic
  - $cluster2$: Diffusion of damaged situation
  - $cluster3$: Diffusion of evacuation situation
  - $cluster4$: Risk of women after the quake

Extracted feature words with positive MCC in the $cluster0$ during15:00-16:00 on March 11 were "Earthquake", "all right", "aftershock", "so", "worry" and son. These words can be interpreted as "after the earthquake, people were worried about the damage of the quake". For the $cluster1$ during 15:00-16:00 on March 11, extracted feature words with positive MCC were "Emergency", "net", "use", "ask", "tsunami warning", "location", "telephone" and so on. This may show that people used the emergency call after the quake. On the other hand, for the $cluster3$ during 15:00-16:00 on March 11, extracted feature words with positive MCC were "Telephone", "tsunami warning", "experience", "confirmation", "evacuation", "contact" and so on. The $cluster4$

TABLE III
FEATURE SELECTION RESULT DURING 15:00-18:00 ON MAR. 11

Time Slot	Cluster #	# of Authors	# of Words	CWC Runtime ($msec$)	# of Feature words	Excerpts from Feature words ( ) shows MCC value
15:00-16:00	0	40354	38822	72298	254	Earthquake(0.2940) all right(0.2277) message(−0.1610) use(−0.1438) use(−0.1312) disaster(−0.1415) net(−0.1250) aftershock(0.1438) Twitter(−0.1082) hope(−0.1094) so(0.1113) worry(0.1046) tsunami(0.0981) kana(0.0876) need(−0.0794) please(−0.0850) confirmation(−0.0896) diffusion(−0.0888) Mr.(0.0868) seismic intensity(0.0892) successfully(0.0845) Tokyo(0.0761) shaking(0.0753) Tohoku(0.0653)
	1	7956	38822	55080	42	Emergency(0.4564) net(0.4518) use(0.4396) ask(0.4356) Bath(0.4239) tsunami warning(0.4138) location(0.3688) telephone(0.3561) RT(0.3555) evacuation(0.3645) absolute(0.3354) everyone(0.3465) possible(0.3178) information(0.3382) so(0.3425) preparation(0.3114) Miyagi(0.3324) possibility(0.2983) it(0.3193) Great Hanshin Earthquake(0.2889) contact(0.3067)
	2	89182	38822	63135	227	Telephone(−0.5044) use(−0.4263) diffusion(−0.4626) disaster(−0.4458) confirmation(−0.4625) safety(−0.4434) hope(−0.4325) earthquake(−0.4915) message(−0.4128) Bathing(−0.3819) net(−0.3850) experience(−0.3799) please(−0.3878) Tsuitta(−0.3916) rice(−0.3596) Great Hanshin-Awaji Earthquake(−0.3533) electricity(−0.3608)
	3	14466	38822	62668	85	Telephone(0.2888) tsunami warning(0.2729) experience(0.2626) confirmation(0.2710) evacuation(0.2688) contact(0.2484) diffusion(0.2472) information(0.2332) earthquake(0.2367) Fire(0.2073) electricity(0.2197) disaster(0.2234) Miyagi(0.2255) tsunami(0.2231) location(0.2033) safety(0.2022)
	4	9614	38822	39734	66	Dial(0.5629) use(0.5281) message(0.5149) emergency(0.5147) Twitter(0.5040) safety(0.4831) net(0.4787) use(0.4643) hope(0.4402) diffusion(0.4162) ask(0.3341) Fukushima Prefecture(0.1608) magnitude 5(0.1559) earthquake(−0.1769) all right(−0.1375) earthquake information(0.1073) aftershocks(−0.1151)
16:00-17:00	0	103114	37659	76145	263	Diffusion(−0.6629) hope(−0.6393) Asakusa(−0.4423) Tokyo(−0.4577) so(−0.4663) power failure(−0.4476) tsunami warning(−0.4131) earthquake(−0.4680) confirmation(−0.4393) net(−0.3967) evacuation(−0.4307) Miyagi(−0.4035) it(−0.4291) information(−0.4093)
	1	8823	37659	47497	40	Hope(0.3876) refuge(0.3885) big tsunami alert(0.3621) outage(0.3606) confirmation(0.3587) hill(0.3449) possibility(0.3438) case(0.3410) BLEMMER(0.3385) Miyagi(0.3391) telephone(0.3293) Intelligence(0.3266) yuan bolt(0.2976) drink water(0.2949) Yun Yan(0.3142) Jin wave(0.3158) may(0.2890) Note(0.2989) earthquake(0.2988) coast(0.2825)
	2	9629	37659	55416	48	Asakusa(0.8184) Gikuhausu(0.8145) Tokyo(0.5900) Who(0.4552) real(0.4023) Search(0.3931) abdomen(0.3840) mackerel(0.3783) hoax(0.3445) important(0.2679) Twitter(0.2565) diffusion(0.2732) location(0.2227) emergency(0.1813) net(0.2001) information(0.1783)
	3	1214	37659	60294	34	Bleeding(0.2760) hemostasis(0.2352) drinking water(0.2456) the main cock(0.2430) roar(0.2089) possible(0.2395) rescue(0.2361) woman(0.2170) Konkurito(0.2226) leakage Bureka(0.2220) advice(0.2260) moment(0.2141) mobile phone(0.2281) . ※ (0.1912) if(0.2413) police(0.2203) Hanshin(0.2108) Supido(0.2128)
	4	32613	37659	56202	111	Diffusion(0.4344) hope(0.3969) earthquake(0.2947) power failure(0.2791) confirmation(0.2699) it(0.2727) so(0.2657) telephone(0.2575) Large tsunami warning(0.2362) evacuation(0.2459) Miyagi(0.2354) tsunami(0.2423) safety(0.2160) disaster(0.2054) like(0.2089) message(0.2019)
17:00-18:00	0	17613	37601	45228	82	Diffusion(0.3575) hope(0.2898) earthquake(0.2680) so(0.2584) maximum(0.2312) ask(0.2281) evacuation(0.2255) shaking(0.2058) time(0.2017) disaster(0.1941) Great Hanshin-Awaji Earthquake(0.1881) it(0.1903) information(0.1859) Free(0.1755) so(0.1786) telephone(0.1744) for(0.1708)
	1	83658	37601	54149	218	Diffusion(−0.5298) earthquake(−0.5160) hope(−0.4457) so(−0.4230) evacuation(−0.3773) please(−0.3654) maximum(−0.3591) disaster(−0.3351) it(−0.3494) because(−0.3209) information(−0.3284) Note(−0.3040) contact(−0.3261) shaking(−0.3180) tsunami(−0.3209) so(−0.3250)
	2	38796	37601	67161	234	Earthquake(0.02471) diffusion(0.01092) contact(0.00663) so(0.00534) hope(0.00485) family(0.00456) it(0.00447) maximum(0.00398) so(0.00389) all right(0.003710) today(0.003611) successfully(0.003512) aftershock(0.003313) worry(0.003314) because(0.002816) after(0.002617) tsunami(0.002519) provides(0.002520) ask(0.002521) shaking(0.002322) like(0.002123) time(0.002024) confirmation(0.002025) information(0.001926)
	3	8035	37601	56479	28	Diffusion(0.3407) evacuation(0.3465) hope(0.3371) disaster(0.3206) so(0.3208) Note(0.2976) shelter(0.2794) ask(0.2896) absolute(0.2669) blankets(0.2592) information(0.2818) the vicinity(0.2640) current(0.2611) risk(0.2511) telephone(0.2720) earthquake(0.2707) location(0.2558) prepared(0.2425) tsunami(0.2656) it(0.2616) confirmation(0.2537) Great Hanshin Earthquake(0.2312)
	4	2581	37601	28272	34	Woman(0.3668) risk(0.3490) absolute(0.3500) shelter(0.3505) crime(0.3340) Note(0.3491) disaster(0.3483) open(0.3408) current(0.3373) If(0.3301) everyone(0.3348) possibility(0.3300) location(0.3287) rescue(0.3065) evacuation(0.3177) use (0.3048) Hanshin Earthquake (0.3151) emergency (0.3138) possible(0.2956)

also had ”Dial”, ”use”, ”message”, ”emergency”, ”Twitter,” ”safety”, ”net2”,”use”, ”hope”, ”diffusion”, ”ask” and so on as extracted feature words with positive MCC. This is also estimated that people used a message dial for confirming safety. However, feature words of $cluster3$ and $cluster4$ are similar with $cluster2$. They can be considered as the same cluster.

Of course, the $cluster4$ in 17:00-18:00 on March 11 showed the topic about the risk of women after the quake, some clusters showed their topics relatively clearly. As the number of clusters are set in advance, the clustering results did not seem to work well in most of the cases.

V. Discussion: Issues on Conventional Social Media Analysis Method

As we described in Section II, generally, the conventional social media analysis method has a scalability problem. Existing data mining technique target thousands of items, not millions. In addition to lack of scalability, we believe there are several problems.

First, the accuracy of clustering (decomposition) techniques is not high, nor can these techniques deliver reasonable performance. Most of the clustering techniques like k-means require the number of clusters to be estimated in advance which lowers cluster quality.

Next, to extract important keywords from clusters, word scoring methods such as TF-IDF [12] or term-score [17] are generally used. However, such scoring methods are based on word occurrence, and high-frequency words tend to be extracted. Therefore, word scoring methods cannot always represent each cluster with high precision.

Third, in this paper, the original technique CWC for feature selection has been utilized, yet even using CWC, it is not easy to extract appropriate words from low quality clusters.

Finally, sometimes these methods identify false similarities between clusters over time.

To overcome these issues, development of new method for social media analysis is required.

VI. Conclusion

This paper proposed an improvement of the conventional word vector-based topic detection method for social media by using high performance Singular Vector Decomposition library $redsvd$ and k-means to identify topic clusters over time from the huge data set of over two hundred million tweets related to the Great East Japan Earthquake. The fast feature selection technique CWC has also been utilized to extract features from each cluster. The proposed technique confirmed the feasibility of topic extraction from big data. From the experiment, though the emergent topics can be observed from the authors' clusters, the issues of conventional topic detection techniques from big data can also be identified as well. To overcome the issues on social media analysis, we plan to develop new social media analysis method that can achieve better performance and accuracy.

Acknowledgment

This paper was supported by the Grant-in-Aid for Scientific Research (KAKENHI Grant Numbers 26280090, and 15K00314) from the Japan Society for the Promotion of Science.

References

[1] S. T. Dumais , *A Graph Analytical Approach for Topic Detection*, Annual Review of Information Science and Technology, 38: 188, doi:10.1002/aris.1440380105, 2005.

[2] H. Sayyadi and L. Raschid, *Latent Semantic Analysis*, ACM Transactions on Internet Technology ,13(2), Article No. 4, November 2013.

[3] D. M. Blei, A. Y. Ng and M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 3 (4-5), pp. 993-1022, doi:10.1162/jmlr.2003.3.4-5.993, 2003.

[4] redsvd, https://code.google.com/p/redsvd/.

[5] K. Shin , D. Fern and S.Miyazaki, *Consistency Measures for Feature Selection: A Formal Definition, Relative Sensitivity Comparison and a Fast Algorithm*, Proc. the Twenty-Second International Joint Conference on Artificial Intelligence, ,pp. 1491-1497, 2011.

[6] T. Hashimoto, T. Kuboyama and Y. Shirota *Topic Detection about the East Japan Great Earthquake based on Emerging Modularity*, Volume 251: Information Modelling and Knowledge Bases XXIV, pp. 110-126, 2013.

[7] J. B. MacQueen, *Some Methods for classification and Analysis of Multivariate Observations*, Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281-297, 1967.

[8] I. Fujino and Y. Hoshino, *A Method for Identifying Topics in Twitter and its Application for Analyzing the Transition of Topics*, Proc. DEIM Forum 2014, C4-2, 2014.

[9] M. J. Paul and M. Dredze, *Discovering Health Topics in Social Media Using Topic Models*, PLoS ONE 9(8): e103408. doi:10.1371/journal.pone.0103408, 2014.

[10] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan and X. Li, *Comparing Twitter and Traditional Media Using Topic Models*, Proc. the 33rd European Conference on Information Retrieval(ECIR 2011), LNCS 6611, pp. 338-349, 2011.

[11] T. Kitada, K. Kazama, T. Sakaki F. Toriumi, A. Kurihara, K. Shinoda, I. Noda and K. Saito, *Analysis and Visualization of Topic Series Using Tweets in Great East Japan Earthquake*, The 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2B3-NFC-02a-1, 2015.

[12] H. C. Wu, R. W. P. Luk, K. F. Wong and K. L., Kwok, *Interpreting TF-IDF term weights as making relevance decisions*, ACM Transactions on Information Systems, 26 (3), doi:10.1145/1361684.1361686, 2008.

[13] Z. Zhao and H. Liu. Searching for interacting features. *In Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1156-1161, 2007.

[14] Hottolink, Inc., http://www.hottolink.co.jp/english.

[15] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/

[16] B. W. Matthews, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochimica et Biophysica Acta (BBA) - Protein Structure, 405 (2), pp.442-451, 1975.

[17] D. M.Blei and J. D. Lafferty, *Text Mining: Theory and Applications*, chapter TOPIC MODELS, Taylor and Francis, 2009.