Visualizing Collective Attention Using Association Networks

Kazutoshi SASAHARA
Department of Complex Systems Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, JAPAN
sasahara@nagoya-u.jp

Received 30 July 2015
Revised manuscript received 27 November 2015

Abstract The socialization of the Web changes the ways we behave both online and offline, leading to a novel emergent phenomenon called “collective attention” in which people’s attention is suddenly concentrated on a particular real-life event. Visualizing collective attention is fundamental to understand human behavior in the digital age. Here we propose “association networks” to visualize usage-based, term-association patterns in a large dataset of tweets (short text messages) during collective attention events. First, we train the word2vec model to obtain vector representations of terms (words) based on semantic similarities, and then construct association networks: given some terms as seeds, the associated terms are linked with each other using the trained word2vec model, and considering the resulting terms as new seeds, the same procedure is repeated. Using two sets of Twitter data—the 2011 Japan earthquake and the 2011 FIFA Women’s World Cup—we demonstrate how association networks visualize collective attention on these events. Provided the Japan earthquake dataset, the association networks that emerged from the most frequently used terms exhibit distinct network structure related to people’s attention during the earthquake, whereas one that emerged from emotion-related terms, such as great and terrible, shows a large connected cluster of negative terms and small clusters of positive terms. Furthermore, we compare association networks in different datasets, using the same seed terms. These results indicate the proposed method to be a useful tool for visualizing the implicit nature of collective attention that is otherwise invisible.

Keywords: Association Network, Collective Attention, Skip-gram Model, Social Media, Twitter, Word2vec.

§1 Introduction

The socialization of the Internet is rapidly progressing and is allowing people to spontaneously exchange an unprecedented amount of information online. The details of such online activity are recorded moment by moment and accumulated digitally. Now, social media is beyond a mere useful communication tool; it can connect people and information instantly and encourage their interactions, from which an online-offline cycle that changes peoples' behavior in real life is emerging. Example includes the well-known "Arab Spring," a series of democratic political movements and "part-time job terrorism," where part-time employees posted videos about their pranks to a social media site, which caused online flaming. Social media such as Twitter and Facebook mediated a vast amount of information and emotion, and significantly affected the details of the incidents, for better or worse.

Using social big data, human behavior and social phenomena in an overly connected world are actively being explored in the field of computational social science.^{1,2)} Using the publicly available Twitter data, this paper focuses on "collective attention," an emergent phenomenon characteristic in an overly connected world. Collective attention is a phenomenon that occurs when a particular event happens in the real world or on the Internet, and people's attention suddenly concentrates on that event through social media. For example, previous research has revealed that collective attention can be captured by bursty increases and unstable oscillations of the number of posts on Twitter.^{3,4)} This phenomenon is thought of as a macroscopic reflection of human nature and a key to studying human behavior in the Internet age. For deep understanding of collective attention, it is important to explore not only apparent properties like bursts of tweets, but also latent ones, for which we have to look into semantic structures underlying collective attention. Our understanding of collective attention remains at the current level due to a lack of effective methods for quantifying semantic structure using social data. In recent years, however, effective methods have been proposed in the field of statistical semantics, in which semantics is treated as local contextual information of words in texts,⁵⁾ and some of these methods are applicable to social data.

Using recent progress from statistical semantics, we propose a novel visualization method for semantic structures of collective attention. The latent properties of collective attention are observed using this method. We also compare the proposed method with several existing methods to demonstrate its advantages and disadvantages.

§2 Related Work

This section introduces related work and the key idea of this study.

2.1 Twitter and Social Sensor

Twitter is a social networking service (SNS) by which people transmit information and communicate by posting short text messages called tweets, which are limited to 140 characters. The real-time nature and the high information

diffusivity are the most unique properties of Twitter as a social medium. Thanks to these properties, Twitter users can behave like talkative sensors and are often called “social sensors.” Several previous studies have addressed Twitter as a social sensor network, and we describe three examples.

Sakaki et al. monitored tweets involving earthquake-related keywords and used the data to train a statistical model for real-time detection of earthquake events. As a result, they were successful with 96% accuracy in detecting earthquake events of the Japan Meteorological Agency of scale-three seismic intensity or greater.⁶⁾ Zhao et al. collected tweets related to US National Football League (NFL) games and conducted a real-time event detection during these games based on the frequencies of tweets and predefined keywords. They achieved a detection accuracy of 90% in the most successful case.⁷⁾ Similar studies also focused on bursty increases of tweet counts (or keyword counts) in the context of event detection.^{8–10)} Takeichi et al. analyzed tweet and retweet time series during Japanese professional baseball games and demonstrated that the degree of concurrency between tweet and retweet bursts correlated with winning or losing a game.¹¹⁾

2.2 Twitter and Collective Attention

As an active social sensor network, Twitter is known to show different collective reactions that depend on the types of real-world events. Several studies analyzed such behavior in the context of collective attention. Lehmann et al. analyzed tweets including hashtags (tweet topics), and showed that the generative patterns of tweet bursts were divided into four classes depending on the tweet topics.⁴⁾ For example, tweets related to a new movie exhibited a symmetric pattern: tweet count gradually increased before the day of the movie release, reached a peak on that day, and then gradually decreased. Tweets related to a presidential speech exhibited a bursty increase only on the day of the speech. These temporal patterns are related to what topics people paid attention to and

Fig. 1 Collective Attention and Its Intensity (2011)

how.

Sasahara et al. focused on the fact that tweet count series obey circadian rhythms in normal situations, but when a significant event happens in the real world, bursts and unstable oscillations of tweet count occur.⁴⁾ This approach detected collective attention events such as major disasters, sports, and events related to politics, science, culture, and customs. Figure 1 shows the intensity of collective attention events in Japan in 2011. In that year, the strongest collective attention corresponded to the Japan earthquake, while the second strongest was for the AFC cup, and the third was for the FIFA women's cup.

The Japan earthquake and the FIFA women's cup are good examples for examining latent properties of collective attention, because details of tweets regarding these events have been revealed.⁴⁾ Thus, we used these as examples of collective attention events in our experiments.

§3 Association Networks

We focus on the semantic relationship of words in tweets in order to visualize latent properties of collective attention. First, we explain the basic part of the proposed method—the word vectorization using word2vec. Then, we describe a network-based visualization of the learning results of the trained word2vec.

3.1 Word Vectorization Using Word2vec

Word2vec was proposed by Mikolov et al. as a method for quickly transforming words in a corpus to low-dimensional vectors that are available for semantic computation.^{12,13)} In the following, we explain how word2vec constructs word vectors in the skip-gram model.

The skip-gram model is a language model that predicts neighboring words of word w_t (c words before and after w_t). Word2vec uses a three-layer feed-forward neural network, as shown in Fig. 2, to learn the skip-gram model. Given word w_t in the input layer, the connection weights between layers are adjusted so that the neural network can better predict word w_{t+j} . The learning of the neural network proceeds by iterating this procedure. Consequently, if the distributions of neighboring words are similar, which means that these words are similar in

Fig. 2 Example of the skip-gram model (Created using a previous study¹³⁾ as a reference).

semantics, the resulting word vectors also become similar. Because the size of the hidden layer in word2vec is quite small compared to that of the input layer, if a corpus has one million different words and the size of the hidden layer is 200, a million-dimensional vector (1-of-K format) can be compressed to a 200-dimensional vector.

More formally, the objective function of the skip-gram model for a sentence or a sequence of words $w_1, w_2, \dots, w_T$ is given by the equation:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) \quad (1)$$

$p(w_{t+j}|w_t)$ in Eq. (1) is computed by the softmax function below:

$$p(w_O|w_I) = \frac{\exp(v_{w_O}'^\top v_{w_I})}{\sum_{w=1}^W \exp(v_w'^\top v_{w_I})} \quad (2)$$

where v_w and v_w' are the input and output vectors of word w , respectively, and W is the number of different words in a corpus. The computational cost of Eq. (2) becomes large as W increases. To reduce the computational cost, word2vec takes advantage of several techniques, including the negative sampling.

As mentioned, word2vec can construct vectors that reflect word meanings based on word usage because it learns the conditional probability regarding word sequence. Consequently, the arithmetic operations of word meanings and the inference of similar words are possible with the resulting word vectors,^{12, 13)} which is difficult to achieve by other language models. For example, LDA (Latent Dirichlet Allocation)¹⁴⁾ based on the Bag-of-Words model does not use contextual information of words, so it cannot construct word vectors that reflect their semantics. Thus, word2vec is suitable for visualizing the latent nature of collective attention in terms of word association in social data.

3.2 Construction Method of Association Networks

We describe the procedures to visualize the learning results of word2vec. As described below, the preprocessing of tweet segmentation yields both words and compound words, so we refer to both as terms without distinguishing them. The similarity between term 1 of vector v_{w_1} and term 2 of vector v_{w_2} is measured by cosine similarity:

$$s = \cos(v_{w_1}, v_{w_2}) = \frac{v_{w_1} \cdot v_{w_2}}{|v_{w_1}| |v_{w_2}|} \quad (3)$$

By computing Eq. (3) given a seed term w , we can list terms whose local context is similar to that of w in order of cosine similarity.

Provided similarity threshold s_{th} , we can construct a network from term vectors with the following steps. First, we list terms whose cosine similarity to w is greater than or equal to s_{th} , and then select the top N most similar terms. If there are multiple seeds, the same procedure is applied to all seeds. Next, by using the selected terms as new seeds, we again list terms with cosine similarity

to w that is greater than or equal to s_{th} , and then select the top N most similar terms. Terms obtained by such “association chains” are used as nodes. The next step concerns how to make links between nodes. If term w_2 is selected when term w_1 is a seed, we deem these similar in semantics and link them together. If multiple terms are selected when w_1 is a seed, we connect w_1 to all these terms. We use a force-directed layout¹⁵⁾ to visualize the resulting network. We call this an “association network” because the procedures correspond to playing a kind of word association game using the knowledge representations about terms of the trained word2vec.

We used $s_{th} = 0.6$ and $N = 20$ (the parameter dependency is discussed in Sec. 5.6) in the following experiments. The number of terms that meet $s_{th} = 0.6$ can be less than 20 in some cases, and no links are made if no terms meet $s_{th} = 0.6$.

§4 Dataset

4.1 Data Collection

We analyzed the user timeline data that were continuously collected from Twitter by snowball sampling. First, we selected 10 users with many followers as seeds from among famous scholars, entrepreneurs, public entertainers, and athletes, and then collected all the available user timelines. Next, from other users who retweeted posts by the seed users, we collected all the available user timelines. By considering all the collected users as new seeds, the same procedure was repeated. The tweet crawling was carried out using the Twitter REST API^{*1} for about a year beginning in April, 2011, which yielded 500M tweets from 400 thousand users (mostly Japanese). From these, we selected tweets from May 11 to 15 (related to the 2011 Japan earthquake) and from July 16 to 18 (related to the 2011 FIFA women’s cup) as datasets for the following experiments. The data size is 6,101,930 tweets for the Japan earthquake and 7,497,877 tweets for the FIFA women’s cup.

4.2 Data Preprocessing

We conducted word segmentation of the Japanese tweets using MeCab,^{*2} a Japanese morphological analysis system. To improve the accuracy of word segmentation, we updated the NAIST Japanese Dictionary^{*3} in advance by introducing entry words from Japanese version of Wikipedia,^{*4} emoticons in the Japanese input system Mozc, and five other emoticons (T-T, ^_^; , ^_Π^ , ^o^ , ^_^) used previously.⁴⁾ As mentioned, the results of the word segmentation include compound words as well as words, so we refer to both of them as terms. By applying word2vec to these datasets, we obtained the vector representations of terms in tweets. During the preprocessing, corrupted and machine-dependent

^{*1} <https://dev.twitter.com/docs/api/>

^{*2} <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

^{*3} <http://naist-jdic.sourceforge.jp>

^{*4} <http://naist-jdic.sourceforge.jp>

emojicons were excluded from the input data. To compute the skip-gram model, we used the word2vec program developed by Mikolov et al.^{*5} The size of the hidden layer was 200, the context size c was 5, and the default values were used for other parameters.

§5 Experiments

We performed experiments on the proposed method using tweets about the 2011 Japan earthquake and the 2011 FIFA women's cup. The purpose of experiments is to observe how the structure of association networks reflects different collective attention events. In the experiments, we used three types of seeds because the use of different seeds on the same dataset may result in different association networks, which may represent different aspects of collective attention. Furthermore, we examined the parameter dependency of association networks and compared this method with similar methods.

5.1 Association Networks Emerged From Popular Terms

We begin by visualizing the 2011 Japan earthquake dataset using association networks. According to the previous research,⁴⁾ the most frequently used terms in the aftermath of the Japan earthquake (14:46, March 11, 2011) are as follows: “地震” [earthquake] (1), “大丈夫” [OK] (2), “無事” [safe] (4), “津波” [tsunami] (5), “電話” [phone] (8), and “避難” [evacuation] (9). Text in brackets denotes English translations, while the numbers in the parentheses denote the frequency rank. Figure 3 shows association networks that emerged from these six seed terms related to the Japan earthquake. The most frequently used term, “地震” [earthquake], had connections to the past major earthquakes such as “茨城県沖地震” [Ibaraki earthquake] and “チリ沖地震” [Chile earthquake] and finally reached to “津波” [tsunami], thereby forming a large cluster of terms related to earthquakes and tsunami. This result makes us imagine that people could have evoked past major earthquake and tsunami events, which they saw in the situation they were facing.

The other seed terms did not connect with each other and formed characteristic but local semantic networks. For example, “電話” [phone] was connected to “メール” [e-mail] and “Skype,” and “大丈夫” [OK] was connected to “どうしてますか” [How are you doing?] and “元気ですか” [Are you doing fine?]. In addition, there were several interesting connections, e.g., ones by spelling inconsistencies such as “電話” and “でんわ,” (both meaning phone) and “大丈夫” and “だいじょうぶ” (both meaning OK), as well as connections by frequent typos such as “避難” and “非難” (which have the same pronunciation but different meanings). Interestingly, there were association leaps in the network of the seed “無事” [safe], in which “両家” [two families] and “これで家族” [now we are family] connected to the seed. This network reflects a semantic context other than the Japan earthquake, which implies that not only negative tweets dominated the Twitter timeline on the day of the earthquake.

^{*5} <https://code.google.com/p/word2vec/>

Fig. 3 Association networks emerged from popular terms (2011 Japan earthquake). Seed terms are “地震” [earthquake], “大丈夫” [OK], “無事” [safe], “津波” [tsunami], “電話” [telephone], and “避難” [evacuation]. The term size is proportional to in-degree.

5.2 Association Networks of Emotional Adjectives

With the same dataset, we constructed and observed association networks using six emotional adjective seeds: “すごい” [great], “やばい” [cool/too bad], “こわい” [scary], “おもしろい” [interesting], “たのしい” [fun], and “ひどい” [awful]. Figure 4 illustrates that clusters that emerged from these adjectives were all connected by association chains. It is noteworthy that the association expanded very widely from the negative terms, “こわい” [scary] and “ひどい” [awful], and the neutral ones, “すごい” [great] and “やばい” [cool/too bad], thereby forming a gigantic cluster. Mutual links were formed in this cluster between terms like “恐ろしい,” a synonym of “こわい,” and “凄まじい,” a synonym of “すごい,” as well as earthquake-related terms like “今日も眠れない” [cannot sleep today again] and “よしん” [aftershock]. In colloquial use, “やばい” can be positive or negative depending on the context, but it was used negatively in most cases on this occasion. Massively generated connections between negative terms are thought to be a visual representation of people’s macro-level depression caused by the Japan earthquake.

5.3 Association Networks of Emoticons

In Japanese written communication, emoticons are frequently used to transmit emotional information that is slightly different from what adjectives can represent.⁴⁾ We constructed association networks by using five types of emoticon seeds: “T_T”, “^_^;”, “(´Д`”, “^o^”, and “^__^”. The right-hand side of Fig. 5 shows that the negative seed emoticons “T_T”, “^_^;”, and “^__^” were connected through the intermediary of other negative emoticons and a Kanji character, such as “*_*” and “泣,” thereby forming a large cluster. There is another large cluster on the left-hand side of Fig. 5 that had a slightly negative emoticon “(´Д`)” as the center, with many of its parts created by the miss-segmentation of tweets in preprocessing. The miss-segmentation occurred because the Japan earthquake dataset had a variety of novel emoticons derived from the existing ones; e.g., “((((°Д°))) ガクガクブルブル” and “ヤシマ作戦 (ノД`”, which were not covered by the Japanese dictionary. However, it is evident that these slightly negative emoticons were used in a similar context to “(´Д`”. In contrast, term association from the positive emoticon “^o^” did not expand, which connected to only a small number of terms.^{*6}

Fig. 5 Association networks of emoticons (2011 Japan earthquake).
Seed terms are “T_T”, “^_^;”, “(´Д`”, “^o^”, and “^__^”.
The term size is proportional to in-degree.

As we have seen so far, association networks that emerged from emotion-related terms are thought to provide qualitatively different information from those emerging from popular terms. The association network with popular seed terms seems to be suitable for overviewing the nature of target collective attention, while one with emotion-related seed terms seems to provide an exaggerated representation of prevailing collective emotion. The cause of this difference is discussed in Sec. 5.6.

^{*6} “フジサン” is part of the emoticon “／^o^＼フジサン,” which is an expression of interjection.

5.4 Semantic Relationship of the Seed Terms

We examined semantic relationships of the seed terms that we used. In Fig. 6, these terms are two-dimensionally arranged by multidimensional scaling. This figure shows that the upper left of this space is related to a term's negativity whereas the lower right is related to a term's positivity. We also find that “地震” [earthquake], “津波” [tsunami], and “避難” [evacuation] were arranged nearby each other but far from the arrangement of nearby terms “電話” [phone], “無事” [safe], and “大丈夫” [OK]. The emotional adjectives and emoticons were grouped separately, but in each of these groups, similar terms were arranged nearby; e.g., “すごい” and “やばい,” or “T_T” and “^_^;”. In the Japan earthquake dataset, we confirmed the emergence of a semantic relationship between terms that reflects the collective attention paid to this disaster.

Fig. 6 Semantic relationship of terms related to the Japan earthquake in a vector space

5.5 Association Networks in Different Contexts

In light of the experimental results, the use of the same seeds for different datasets may result in different association networks. We examined this by using the dataset of the 2011 FIFA women's cup. This event was held in Germany, and in the final match Japan beat the United States in a penalty shootout. Thus, we constructed association networks with the seeds “日本” [Japan] and “アメリカ” [the United States]. The result is shown in Fig. 7(A). It is not necessary for the football-related terms to connect massively, because “日本” [Japan] and “アメリカ” [the United States] are country names, but in reality, terms related to the FIFA women's cup emerged from these seeds by association chains, forming a single cluster, as shown in Fig. 7(A). Although there are several unrelated terms, such as “我が国” [my country] and “自動車産業” [motor industry], most are associated with football, such as “ファイティングスピリット” [fighting spirit] and “三位決定戦” [third-place match]. This result shows that collective attention on the FIFA women's cup emerged.

Fig. 7 Association networks emerged from the seeds “日本” [Japan] and “アメリカ” [America] in different contexts. (A) 2011 Japan earthquake, (B) 2011 FIFA women’s cup.

With the same seeds, we next constructed association networks from the Japan earthquake dataset used so far, as shown in Fig. 7(B). In this figure, there were two almost independent clusters; one that emerged from “日本” [Japan] and another from “アメリカ” [the United States], although “世界中” [all the world] bridged these clusters. In the context of the Japan earthquake disaster, “日本” [Japan] connected to “うらやむ” [envious] and “最貧” [the poorest], whereas “アメリカ” [the United States] connected to foreign country names other than Japan. As expected, we confirmed that the use of the same seed terms may result in different association networks depending on the types of collective attention events.

5.6 Parameter Dependencies of Association Networks

Here, we describe the parameter dependency of association networks and the adequacy of the parameter settings that we used. Figure 8 shows the number of links in association networks that emerged from the seed “地震” [earthquake] as a function of s_{th} . When we limited link destinations to the top 20 terms in similarity rank ($N = 20$), the number of generated links that meet $s_{th} \leq 0.5$ was 6,681, which suggests that the resulting networks were too large for visual exploration purposes. If $N = 200$, the resulting networks are also too large and not adequate for visualization purposes. In contrast, if $s_{th} \geq 0.7$, no association networks are constructed since there are no terms that meet this similarity condition. If $s_{th} = 0.6$, the number of generated links is 681, and such a network size is adequate for the purpose of visual exploration. These

Fig. 8 Number of links in association networks as a function of similarity threshold s_{th} (semilogarithmic scale).

Fig. 9 Number of links in association networks as a function of the repeat count of association chains (semilogarithmic scale).

properties were similarly observed in the cases of other seed terms. Because of these empirical facts, we used $s_{th} = 0.6$ and $N = 20$ for experiments.

We limited the repeat count of association chains to two times for the empirical fact shown in Fig. 9. This figure shows the number of generated links in association networks that emerged from the seeds “たのしい” [fun], “こわい” [scary], “^o^”, and “T_T” as a function of the repeat count of association chains (the parameter setting is the same as above). In all the seeds, the number of links increased roughly exponentially, so if the repeat count of association chains was more than three, the network size was too large to visually examine the results. Furthermore, there is a potential problem: as the repeat count of association chains increases, the semantic association would deviate from the

meaning of the original seed terms. To maintain the visibility of association networks, the repeat count of association chains should be two. However, if we statistically assess the global structure of association networks, three or more would be valuable for association chains.

This figure also shows that the number of generated links (i.e., the degree to which term association expands) was different at several times, depending on the seeds. For example, if the repeat count of association chains is two in the Japan earthquake dataset, the number of links generated from “こわい” [scary] is about six times larger than that of “たのしい” [fun], and the links generated from “T.T” is twice as many as that of “^o^”. The formation of distinct term clusters may result from different tendencies of association chains in positive and negative terms, which reflects the nature of collective attention.

5.7 Comparison with Other Visualization Methods

The word cloud may be the simplest way for visualizing text information in social data. For example, given a dataset of Japanese tweets, we perform the segmentation of tweets into terms in order to examine their occurrence frequencies, and show terms whose sizes are in proportion to the frequencies. The summary of a dataset can be made visually appealing by changing fonts, colors, and directions of terms. The word cloud, however, can only show the frequency information of terms, so it cannot tell us about the context in which terms are used and the relationship of terms.

The co-occurrence network is another popular method for visualizing text information. There are various ways of constructing the co-occurrence network, but one of the most standard ways is as follows. We list adjacent term pairs (term bigrams) in tweets and consider each as two nodes connected with a link. After identifying all nodes and links with this procedure, we construct a co-occurrence network. Figure 10 illustrates the largest connected component of the co-occurrence network constructed from the Japan earthquake dataset, in which only links with occurrence frequency more than 200 are depicted (the actual size is therefore much larger).

The co-occurrence network is useful for summarizing an entire dataset. For example, the figure reveals special connections that reflect a particular situation of the Japan earthquake and the succeeding disaster, such as between “みんな” [everyone] and “大丈夫” [OK] and between “義援金” [monetary donation] and “サイト” [(web)site]. What the co-occurrence network illustrates, however, is the frequency information of term bigrams, and as seen in this example, relatively obvious associations are often emphasized, from which we cannot find useful information about collective emotion. With the association network, one can explore term association patterns through trial and error by choosing seed terms, whereas there is no such degree of freedom in the co-occurrence network. Although any visualization method has advantages and disadvantages, the association network has many preferable features for visualizing latent properties of collection attention from social data.

Fig. 10 Example of the co-occurrence network (2011 Japan earthquake). The largest connected component is shown. The term size is proportional to degree.

§6 Conclusion

We have proposed a visualization method for latent properties of collective attention that emerges on social media by incorporating methods from statistical semantics and network science. In principle, the association network makes visible meaningful patterns and unexpected associations of terms because it is created by association chains using term vectors based on semantics. The experimental results confirm that this method was successful in visualizing latent properties of collective attention, such as term association patterns reflect-

Fig. 11 User distribution with respect to posted tweet counts (logarithmic scale).

ing semantic structures of collection attention events, term clusters related to subtopics, and an exaggerated expression of dominant emotion. Neither the language model based on the Bag-of-Words nor the co-occurrence network can hardly yield these results. Visualizing semantic association chains using social data means characterizing real-world events by spontaneous, subjective reports from a large unspecified number of people.

A point to be aware of when using this method is a danger of bias in social data. Figure 11 shows the distribution of users with respect to the number of generated tweets in our datasets. As shown, the users have a fat-tailed distribution. This clearly shows bias in tweet generation: for a few days, the majority of users posted only one or two tweets, while a few users heavily tweeted. In our experiments, no major problem was recognized in understanding the resulting association networks, but depending on target events and data collection methods, inappropriate association networks could be constructed for data bias, which may eventually lead to a wrong conclusion. This could be a weak point of the association network. Whatever the visualization method is, dealing with data bias is a common problem, and special attention is needed when using the association network as a tool for discovery.

The method of choosing adequate seed terms is another important issue, but because it depends on properties of collective attention events, a general solution might not exist. As demonstrated in the experiments, we need to select a better seed term after constructing and examining association networks with various possible seed terms. In other words, such a process of explanatory data analysis is a process of understanding semantic structures of collective attention. Applied properly, the association network is useful to gain insights from social data, although the mentioned problems exist. This method is effective for visualizing the invisible and can thus be a new tool for exploring human behavior in the massive data flow era.

Acknowledgements

This work was supported by The Hori Science and Arts Foundation.

References

- 1) Lazer, D., Pentland A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Alstyne, M. V., "Computational Social Science," *Science*, 323, 5915, pp. 721–723, 2009.
- 2) Miller, G., "Social Scientists Wade Into the Tweet Stream," *Science*, 333, 6051, pp. 1814–1815, 2011.
- 3) Lehmann, J., Gonçalves, B., Ramasco, J. and Cattuto, C., "Dynamical Classes of Collective Attention in Twitter," in *Proc. of the 21st International Conference on World Wide Web*, pp. 251–260, 2012.
- 4) Sasahara, K., Hirata, Y., Toyoda, M., Kitsuregawa, M. and Aihara, K., "Quantifying Collective Attention from Tweet Stream," *PLoS ONE*, 8, 4, e61823, 2013.
- 5) Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T., "Statistical Semantics: Analysis of the Potential Performance of Keyword Information Systems," *Bell System Technical Journal*, 62, 6, pp. 1753–1806, 1983.
- 6) Sakaki, T., Okazaki, M. and Matsuo, Y., "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," in *Proc. of the 19th International Conference on World Wide Web*, pp. 851–860, 2010.
- 7) Zhao, S., Zhong, L., Wickramasuriya, J. and Vasudevan, V., "Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games," in *Proc of CoRR*, 2011.
- 8) Weng, J., Yao, Y. and Leonardi, E., "Event Detection in Twitter," in *Proc. of the 5th International AAAI Conference on Weblogs and Social Media*, pp. 401–408, 2011.
- 9) Lamos, V. and Cristianini, N., "Nowcasting Events from the Social Web with Statistical Learning," *Transactions on Intelligent Systems and Technology*, 3, 4, pp. 72:1–72:22, 2012.
- 10) Panisson, A., Gauvin, L., Quaggiotto, M. and Cattuto, C., "Mining Concurrent Topical Activity in Microblog Streams," in *The 4th Workshop on Making Sense of Microposts (#Microposts2014)*, pp. 3–10, 2014.
- 11) Takeichi, Y., Sasahara, K., Suzuki, R. and Arita, T., "Twitter as Social Sensor: Dynamics and Structure in Major Sporting Events," in *Proc of the 14th International Conference on the Simulation and Synthesis of Living Systems (ALIFE14)*, pp. 778–784, 2014.
- 12) Mikolov, T., Chen, K., Corrado, G. and Dean, J., "Efficient Estimation of Word Representations in Vector Space," in *Proc. of International Conference on Learning Representations (ICLR)*, 2013.
- 13) Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J., "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems*, 26, pp. 3111–3119, 2013.
- 14) Blei, D. M., Ng, A. Y and Jordan, M. I., "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, 3, pp. 993–1022, 2003.

- 15) Bastian, M., Heymann, S. and Jacomy, M., “Gephi: An Open Source Software for Exploring and Manipulating Networks,” in *Proc of the Third International Conference on Weblogs and Social Media*, pp. 361–362, 2009.

Kazutoshi Sasahara, Ph.D.: He received his Ph.D. in Multidisciplinary Sciences from The University of Tokyo in 2005. He worked as a researcher at several institutions including RIKEN BSI, The University of Tokyo, and UCLA. Since 2012, he is an Assistant Professor in Department of Complex Systems Science, Nagoya University. His research interests are in artificial life, behavioral biology, and computational social science.