International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 077-082
© MIR Labs, www.mirlabs.net/ijcisim/index.html

On Analysis and Visualization of Twitter Data

Fathelalem Ali[1], Yasuki Shima[2]

[1] Faculty of International Studies, Meio University,
1220-1 Bimata, Nago, Okinawa 905-8585, Japan
ali@mail.meio-u.ac.jp

[2] Institute of Visual Informatics (IVI), UKM,
43600 Bangi , Selangor , Malaysia
sssyasuki@gmail.com

***Abstract*: Provision of big data analysis in a customer-friendly applicable form, with ease and affordable cost to a wide range of customers and businesses is still a big challenge for data scientists and engineers. In this study, present a framework for analysis and its visualization. We analyze Twitter messages related to a one-year span in a specific geographical area, Okinawa Island, in Japan. Our approach includes considering three types of data; basic data, further derived or extended data, and a priori information about subjects of interest. For analysis results and visualization we arrange data in a three-dimension framework of time, quality and volume. We map different elements of the data, such as number of tweets per user, time, span of time stayed in the Island, geographical location and content of messages. Based on the elements of the data within the framework, twitters are grouped and analyzed. A visual representation of analysis is presented.**

Keywords: Twitter Messages, big data Analysis, data Visualization,

I. Introduction

Big data is remarkable opportunities for businesses to achieve innovative methods and bring more robust services and strengthen decision making. Many customers claim huge data and information and but still lack knowledge and know-how to make the best use of such data and information. Management and analysis of big data require enormous time and expense. Moreover, it is not easy to quantify and qualify the results of analysis and adapt that to the business decision and process. That it difficult to decide on the investment and cost to unspecified value-added results of the analysis and co-relate that to the decisions and later returns and profitability [1],[2].

Looking at the current trends in data science field, remarkable advances have been and being made in information acquisition, transmission, and analysis techniques[3],[4]. However, still the provision of big data analysis in an easy to utilize form, with ease and affordable cost to a wide range of customers and businesses is a challenge that is still far from being met[5],[6].

In this work we consider analysis and visualization of big data. We apply our analysis approach to Twitter data. Several works been done on analysis of Twitter messages. Earlier work includes Ritterman et al. [7], who showed that looking at Twitter messages can improve the accuracy of market forecasting models by providing early warnings of external events like the H1N1 outbreak.

In this study, we use a 3-attributes frame work to guide and simplify analysis and visualization of results related to big data. As early as in 2001 and forward, *Volume*, *Variety* and *Velocity* are three defining properties or elements of big data [8]. *Volume* refers to the amount of data, *variety* refers to the types of data, and *velocity* refers to the speed of data processing. In regard, the analytics techniques ought to evolve to include those elements and scale up to more complex analytics [9] .

Adapting ideas of earlier works, in this work we use 3-attributes of *Volume*, *Quality,* and *Time* to accommodate analysis and visualization of the data available. Earlier to the step we use a 3-elements approach to sort and incorporate data for analysis.

Using those approaches, we analyze twitter data related to a one-year span in Okinawa Island area, the most southern prefecture in Japan.

For the initial analyzing step, look into the initial basic data, then exploit for extended data or information using analytical or mathematical processing. We also seek some a priori information or data that might be related to the subject of interest. We map and analyze the messages according to three attributes, *Volume*, *Quality* and *Time* of elements of data. We start by arranging the data and information available with regard to volume, quality and time attributes. Further elements of data are exploited.

The following sections describes our research framework, defining different data and exploit values available. the mapping of different elements to the three attributes of volume, quality and time, visualization and analysis. Analysis and results, are presented and discussed. Finally we conclude with future work directions.

II. Frame work for Sample Data and Analysis

A Framework

We adopt a 3-element framework for Analysis in order to extract an applicable knowledge from the data collection. The three elements or components of analysis go as follow.

I) *Basic Information* or data (B): includes basic data available about specific subject.

II) *Extended Information* or data (E): information or data can be extracted from the basic data using mathematical or statistical processing

III) *A priori information* or data (P): information or data available from other sources and can be related to the subject or environment of the basic data.

B Sample Data and Measurement Period

- Data : Twitter messages originated from Okinawa Island, in one year span.
- Date: September 1, 2014 - August 31, 2015
- Total Number of tweets : 117,435 tweets
- Number of Tweeters: 12,079
- Target Tweets: Tweets that originated at the specified period in Okinawa Island.

Okinawa is a tourist spot, where thousands of tourists visit every day. It has a population of 1.3 Million. More than 7 Millions visited the Okinawa in 2014, where around 87% were from the Main land of Japan [10].

Figure 1 shows the location of Okinawa as the most southern Prefecture in Japan

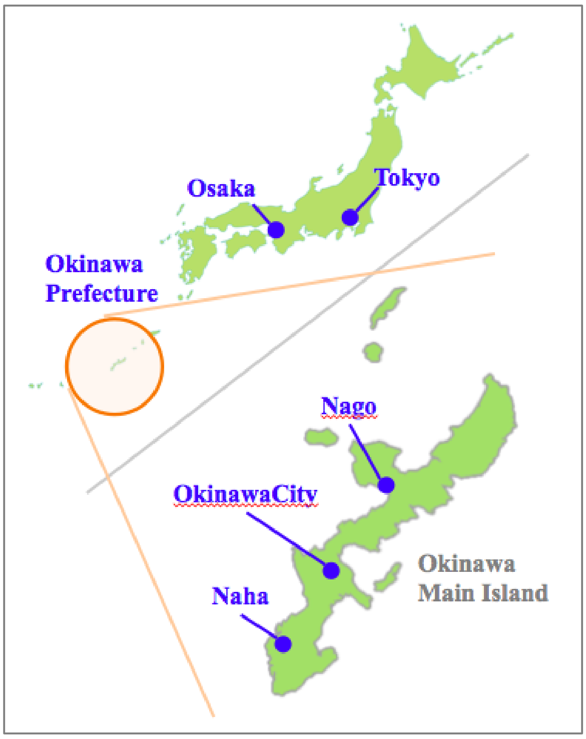


Figure 1. Okinawa Main Island, the origin of Twitter tweets being examined in this Study

C Data gathering and analysis tool and source

We have applied our analysis and visualization approach with the framework mentioned earlier to Twitter messages that gathered from the following tool and source.

Tool to gather information : "CORONA" tool .
CORONA tool extracts tweet data (text), user name, time, coordinates.

CORONA Providing source：TIDA

（http://www.tida-okinawa.com/modules/pico/）

Data obtained are as follows.
·POST_DATE: post time
·LATITUDE: Latitude
·LONGITUDE: longitude
·USER_NAME: user name
·BODY_TEXT: tweet text

Table 1 below shows elements of tweets data used in our analysis.

Table 1. Tweets data sample

POST_DATE	LATITUDE	LONGITUDE	USER_NAME
2015/7/23 12:12	26.65018	127.8171	@XXXX
BODY_TEXT			
おはよう沖縄 Good morning Okinawa #水納島 @ 水納島ビーチ			

III. Data and Analysis

A. Data

Our approach defines three types of data

I) Basic Data

The basic data considered here come as follows

a) Tweets content
b) Location (tweets location GPS data)
c) Time of tweets

II) Extended Data

We exploit initial Basic data elements (Table 1) to add further qualitative or quantitative data as Extended data. In the step we added the following data.

a) Number of Tweets per Tweeter.
b) Tweets average per day (*tAvg*)
c) Tweets Span: Number of days between first and last tweets (*Span*).
d) Total distance between locations of tweets.

III) A priori Data

We considered the following information for our analysis [8].

a) Okinawa area is a tourism place.
b) Most of tourists stay 2~7 days (96.8 %)
c) Tweets Span: Number of days between first and last tweets (*Span*).
d) Tourism Spots and hotels.

B. *Analysis*

I) Initial Analysis (tweeters based)

Considering average stay of a priori data (ii), and average of daily tweets, and time Span of extended data (ii) and (iii) above, the users (tweeters) were classified according to the following criteria.

a) *Short-Active-Span(SAS)*: Tweeting for 2~7 days and the average of tweets per day is 3 or more (*Eq*.(1))

$$(2 <= Span <= 7)\ \text{AND}\ (tAvg >= 2.5) \tag{1}$$

b) *Long-Active-Span(LAS)*: Tweeting for 8~180 days and the average of tweets per day is 2.5 or more (*Eq*.(2))

$$(7 < Span <= 180)\ \text{AND}\ (tAvg >= 2.5) \tag{2}$$

c) *Other-NonActive-Span(ONA)*: Tweeters other that (1) or (2) above (*Eq*.(3)).

$$(Span > 180)\ \text{OR}\ (Span < 2)\ \text{OR}\ (tAvg < 2.5) \tag{3}$$

Accordingly we obtained the following three groups of tweeters, with their population density percentages shown in Table 2 below.

a) *Short-Active-Span Tweeters*: A great deal of them are assumed tourists, through a look at their other features, such as time of the stay in the Island, or/and distance moved around while tweeting.

b) *Long-Active-Span Tweeters*: Those are likely not tourists being for a while in the area. However, they represent less than 1% of all the tweeters.

c) *Other-Non-Active-Span Tweeters*: Those non active tweeters or those stay for longer than 6 months in the area. They are assumed to be mostly resident tweeters or those who were not active, tweeting less frequent, and hence provide less information to trace their behavior or grasp their features easily. Most of those are likely not tourists, and they represent 92% of the tweeters population.

Table 2. Tweeters Grouping, according to their *first* and *last* tweets time (Span)

Tweets Time Span (*TTS*)	**2〜7 days (*SAS*)**	**8〜180 days (*LAS*)**	**Non Active, Longer (*ONA*)**
Tweeters Percentage	7.0%	0.8%	92.0%

C. *Mapping Tweets Information*

As suggested earlier, elements of the initial and extended data are mapped on a 3-attribute space, namely X:*Volume*, Y:Quality and Z:Time. Here are some of the elements their attribute mapping:

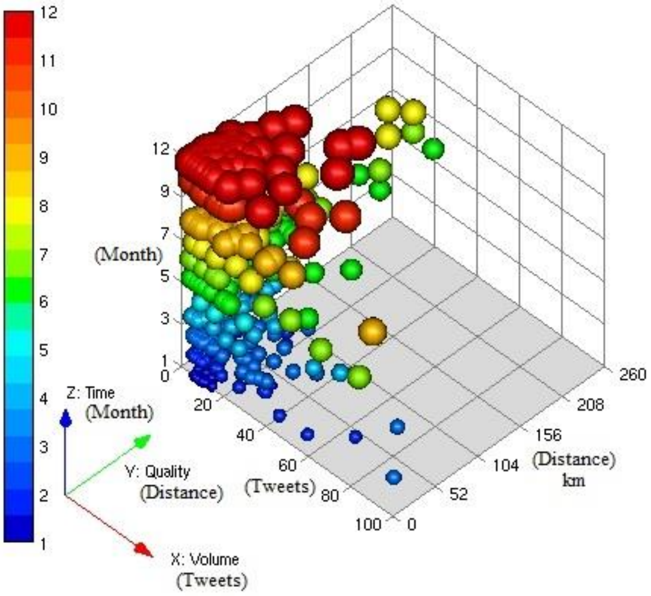
- Time of Tweets (Time)
- Number of tweets (Volume)
- Distance between tweets locations (Quality)

The next section explains our approach to provide a qualitative and quantitative visualization of information co-relating tweets data text, time, location and volume.

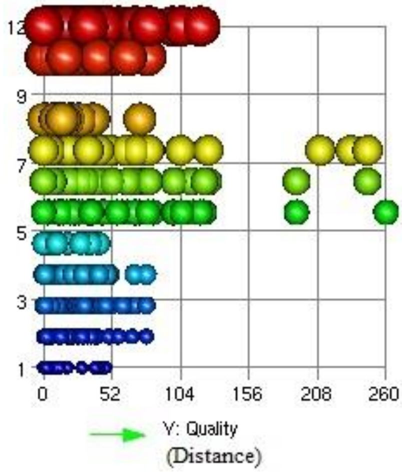
IV. Visualization

Our visualization main concept is to express the elements in regard with our 3-attribues-frame work. The graph (a) in Fig 2 shows *three* elements for *Short-Active-Span Tweeters*. It would be possible to rotate the 3-D graph to look at different angles to exploit co-relation between different elements.

The graph (b) at the bottom of fig.2. shows a look at x and y axes, with a focus on the distance travelled during different months.



(a)

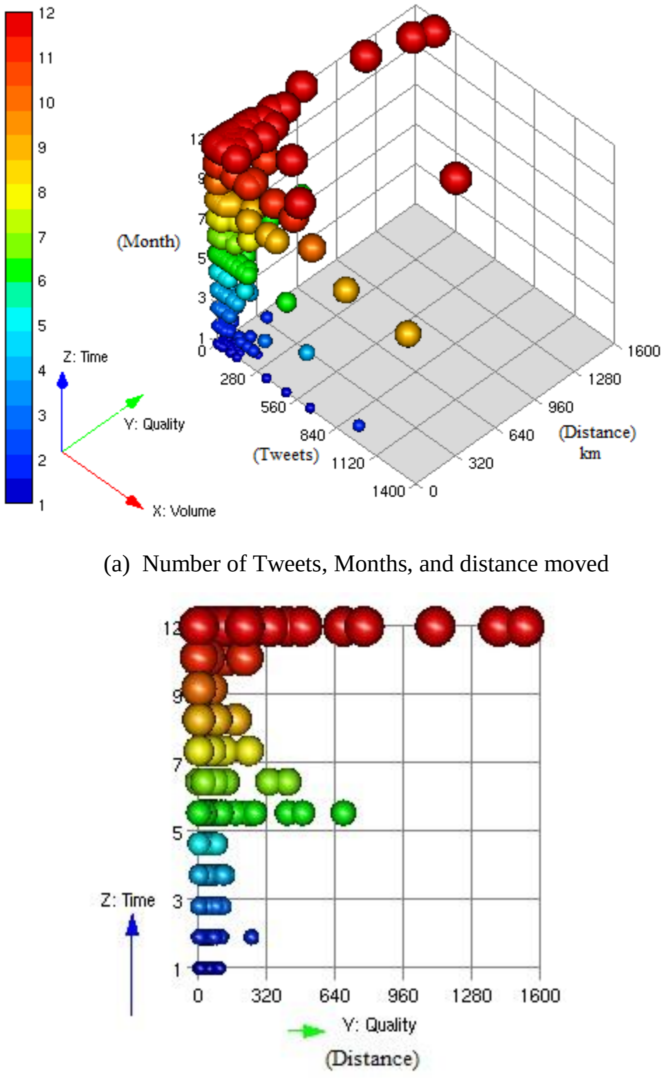


(b)

Fig. 2. Distance between tweets locations, number of tweets and time of posting, for *Short spam active* tweeters.

As we can see from Fig 2, for this group of tweeters, they are concentrated in summer time Jun(6) ~ Aug(8), where they travel for longer distances. For sake of clarity, Fig 2 (bottom) shows a distance-time plane of the graph.

Fig 3 below shows the same elements (distance, number of tweets) for *Other-NoneActive* group. The graph reveals that this group tweeters move wider during December. That suggest they were residents who tends to be active during the year end.



(a) Number of Tweets, Months, and distance moved

(b)Tweets along *distance* and *time*

Fig. 3. Distance between tweets locations, number of tweets and time of posting, for *Other-NoneActive* group .

D. Consolidation Analysis (tweets text based)

Considering tourists spots in Okinawa and its islands [10], we have conducted a text mining for each group of the tweeters groups specified in the previous section. The top frequent keywords as well as the appearance of benchmark keywords of the tourists famous spots are explored. Table 3 below shows the top 17 keywords for spots frequently visited by tourists.

Table 3. Top Frequent Keyboards for the three Groups of tweeters

Frequent Tourists Spots [10]	7 days	8-180 days	Other
	12829	28131	52161
Churaumi Aquarium	6.4%	0.7%	5.0%
Kokusaidori	7.9%	2.2%	4.3%
Shurijyo Castle	3.5%	0.4%	1.3%
American village	1.3%	0.5%	1.7%
Manzamo	1.6%	0.1%	2.0%
Zanpamisaki	0.6%	0.3%	0.6%
Outlet Mall	0.5%	0.3%	0.5%
Kourijima	3.5%	0.5%	2.1%
Kousetsuichiba	1.0%	0.1%	0.5%
heiwadouri St.	0.2%	0.1%	0.1%
DFS	0.6%	0.2%	0.4%
Heiwakinenkouen	0.0%	0.1%	0.0%
Naminouegu	0.1%	0.0%	0.0%
Kaicyudouro 4Is.	1.2%	0.5%	1.4%
Main Island Is.	2.6%	5.8%	5.6%
Ishigaki Island	14.3%	16.4%	9.2%
Miyako Island	8.7%	3.0%	7.2%
Total	53.9%	31.1%	41.9%

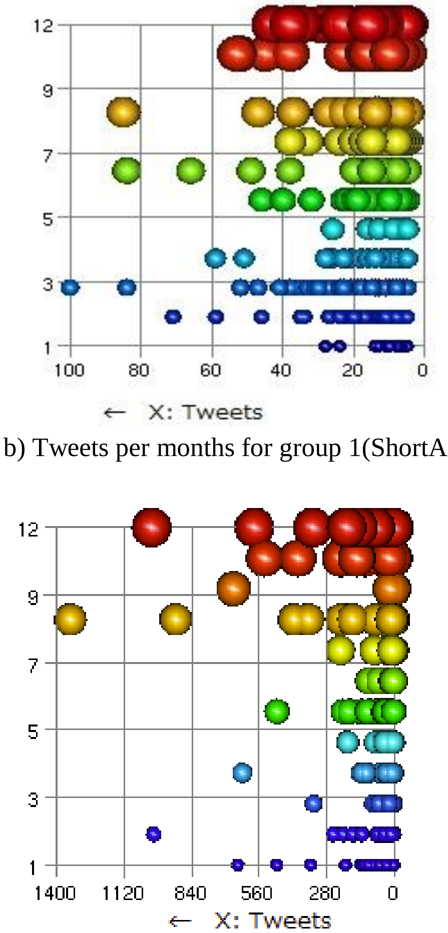
As appears from the texting mining results of Table 3, The short stay tweeters used the tourists related keywords most frequent (53.9%) compared to other groups (31.1 or 41.9%).

V. Results and Discussion

We demonstrated mapping the initial *basic data* elements and further *extended data* to *Volume*, *Quality* and *Time* attributes. Further elements could be calculated and added using existing elements at different attributes. For example, the "Distance" element has been calculated making use of the *location* of tweets and *time* for each posting, aggregating distance traveled between consecutive tweets for each tweeter.

A priori information about the tourists spots as provided by related agencies, was used to consolidate and improve the reliability of findings from the basic and extended data.

The flexible multi-dimensional visualization enables with ease looking at different angles of interests. Fig. 4 shows a look focusing on number of tweets distributed over months. One thing is that we can see late months of the year have relatively less tweets. Note that for the especially for the *Other-None Active* tweeters, the same period witnessed more distance between tweets, as appears in the Fig 3(b).



(a) b) Tweets per months for group 1(ShortActive)

(b) Tweets per months for group 3(Other)

Fig. 4.**N**umber of tweets and time of tweets, for *Short-Active-Spam* tweeters(a), and *OtherNonActive* tweeters(b).

Text mining results of frequent keywords within tweets showed more frequency for the short-visit active group, as shown in graph of Figure 5 (data extracted from Table 3.).

Statistics published Okinawa prefecture government provided lists of places visited most by tourists[10]. That information was used as a guideline to verify and justify attitude of stay and locations as well as words used frequently within tweets of, and particularly for short-stay active tweeters, that can be with a high degree of confidence can be reflected by the first group (up to 7 days of stay).

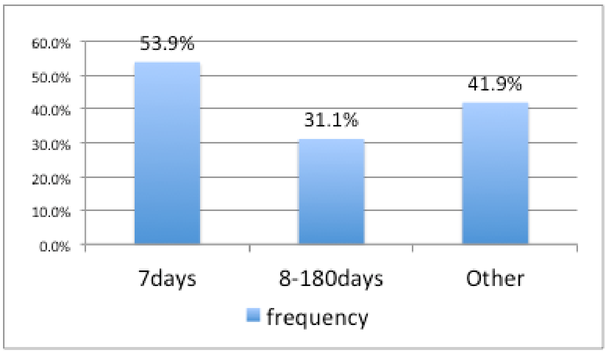


Fig. 5. Accumulated average keywords frequency for three groups as per span of stay.

VI. Conclusion

Our approach in analysis categorized and used three types of data *basic*, *extended* and other a *priori* data or information. Further in and analysis and visualization, we used a 3-attribute framework to guide data analysis and visualization. Our approaches for analysis and visualization are applied to Twitter data collected on one-year span, and related to specific area of origin, Okinawa Island, Japan.

The approach presented as a guideline for arranging different elements of the data, that can be applied for a variety of big data types. The analysis in our approach starts with re-arranging data elements by adding further elements that widens the alternatives of various angle-of-looks to analyze data by different point of interests for customers.

We have used the Twitter data collected from a 0ne-year span in Okinawa Island area, in Japan. We carried out mapping of different data elements on 3-attributes of Volume, Quality and Time. Despite the fact that Twitter has a quite diverse set of users [7], it was possible to feature groups of users looking at some elements in the 3-attributes display.

The approach introduced here is meant to give a general way for data scientists to develop generic standard analysis tools and models that would yield more values and insights to satisfy different customers..

On the other hand, The visualization goes well with the 3-attribute approach, and expected to provide the customer with a flexible way to look at data analysis and make the best of it at different plans and settings.

Different people or businesses can look at different aspects and angles and make their observations in their context of interest.

VII. Future work

Our future work includes development of analysis tools that would in cooperate other external elements, such as weather information, festivals and events data that can add further elements to the initial data set.

Also we are working on refining and automating visualization process with wider options of applications.

More experiments are also considered in various areas of big data to verify the framework and approaches mentioned here.

The authors are also working on refining and application the frames and approached suggested here to a variety of areas of big data.

References

[1] Shirota Makoto: Current Status and Challenges of Big Data in Japan" Publications of Nomura Research Institute, (2012). (In *Japanese*).

[2] Yayama Koji "Trends of in big data utilization in the United States," JETRO / IPA, Japan, (2014). (In *Japanese*).

[3] Huina Mao, Xin Shuai, Apu Kapadia: Loose Tweets: An Analysis of Privacy Leaks on Twitter, Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. PP. 1-12 (2011).

[4] Yong Shi: Big Data History, Current Status, and Challenges Going Forward, The Bridge, Winter2014, pp.6-11 (2014).

[5] Y.Okazaki, T.Tsuruga : For economic and price analysis using big data. pp 9--13(2015) . http://www.boj.or.jp/research/brp/ron_2015/data/ron150625a.pdf, 2015.9

[6] Johan Bollen, Huina Mao, Xiao-Jun Zeng : Twitter mood predicts the stock market. 2011, http://arxiv.org/pdf/1010.3003v1.pdf , 2015.9

[7] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. 1st International Workshop on Mining Social Media, 2009.

[8] [15] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" , Meta Group Publication, File 949(2001).

[9] Zikopoulos, P.C., deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., and J. Giles. Harness the Power of Big Data—The IBM Big Data Platform. New York: McGrawHill, 2013.

[10] Okinawa Prefecutre statistics report on tourism, 2014: http://www.pref.okinawa.jp /site/bunka-sports/kankoseisaku/kikaku/report/youran/documents/toukei.pdf, as of 2015.12A.

Author Biographies

Fathelalem Ali is a professor of information engineering, at the Department of Management and Information Sciences, Meio University, Japan, where he joined as faculty in 2000.
He had obtained a B.Sc. in Mechanical Engineering from the University of Khartoum, in 1988, Master of Engineering in Electrical and Information Engineering, and Ph.D. of Information Engineering in Complex Intelligent Systems, from University of the Ryukyus, Japan, in 1997 and 2000, respectively.

He has been working in a multidisciplinary research and education environment and has been actively involved with collaborative research with several universities and institutes in Japan, Malaysia, US, France and Sudan. His research work has been mainly in computational intelligence, Internet technologies and applications. He had developed ICT related courses and curricula for university students as well as for training and capacity building agencies. His recent research interests are in data science and ICT in education and learning.

Yasuki Shima had graduated form the Department of Management and Information Science, Meio University, Japan in 2002, and obtained a Master degree from the same university in 2006, majoring in Management and Information Science (Information Design). He is working with INDEX Asia Co. Ltd, Japan, as a senior researcher. He has been enrolled at the Institute of Visual Informatics (IVI), UKM, Malaysia, as a Ph.D. student in 2015. Shima’s work and research interests include big data analysis and visualization. He is a core member of a research team of R&D project pursued by INDEX group.