A Study of Visualization for Hidden Relation between Published Documents and Message from Twitter by means of Sentimental Analysis Approach

Masakazu Kyokane, Kazuaki Ando, Yoshiro Imai
Graduate School of Engineering, Kagawa University,
2217-20 Hayashi-cho, Takamatsu 761-0396 Japan
s14g463@stmail.eng.kagawa-u.ac.jp, {ando, imai}@eng.kagawa-u.ac.jp

ABSTRACT

Recent years have found that information processing for social media can bring several kinds of effective results to us in the domains of industries, education, social science, economics, and so on. Such approaches, sometimes called sentiment analysis and/or opinion mining, are potentially applicable from trend analysis for industrial products and useful business services to embossment of human thinking, behaviour and/or emotion. This time our study focuses on Twitter used by students of some Japanese universities, acquires their tweeting data (messages on Twitter), analyzes sentimental values from the according data, and performs visualizing demonstration.

It also investigates existence of some relations between calculated sentimental values and location attributes of students' universities. At the same time, it tries to discuss whether the above procedure and analysis can visualize conventionally hidden relationship between contents of tweeting messages and characteristics of universities categorization. One of the aims of this study is to help young persons to choose their suitable universities based on relationship visualized in the suitable manners and to provide some useful examples for extraction of sentimental values from categorized groups.

KEYWORDS

Sentiment Analysis, Twitter, Visualization of Hidden Relation, Data Mining.

1 INTRODUCTION

Data mining is one of the useful information processing techniques to extract characteristics or attributes from a huge amount of data from several phenomena. Some special-purpose filtering can extract useful pattern and information from the given data in a relatively short period. With suitable tuning of the filtering parameters, we will be able to obtain necessary relations and characteristics through data mining processes.

As you know, nowadays, we can watch very much various kinds of messages from Social Media, such as Facebook, Twitter, Blogs and so on. We can investigate several meanings embedded in the messages from Social Media through data mining approach. Information processing technique can show us interesting meanings of messages in Social Media by means of a suitable translation scheme from words in the above messages into numerical expression such as probability of occurrence.

This study has been challenging to realize some kind of data mining as an example of visualization of sentimental words extracted from messages of Twitter and relation between sentimental value and the relevant human group. We have defined translation scheme from words in message into sentimental values. And we apply this scheme into data mining, calculation of sentimental values for tweeting messages acquired from Twitter, and demonstration of relation between sentimental values and categorized human group.

The paper will describe an approach to extract sentimental messages from Twitter and its application to investigate characteristics of categorized human groups. And it will mention our trial visualization of potentially hidden relation between sentimental tweeting message and the relevant human groups. It introduces related works about analysis of message from Social media, calculation of sentimental values and visualization of relation between message and categorization in the next section. It illustrates our system configuration and structure of information processing for data mining in the third section. It explains our results of analytical data with our defined translation scheme from message to sentimental values in the fourth section. It reports how to visualize relation between sentimental Tweeting and categorized human groups, and discusses whether proposed approach can provide suitable results to visualize useful relations in the fifth section. And finally it summarizes our conclusion in the last section.

2 RELATED WORK

This section introduces some suitable related works to design and implement our data mining approach for message from Social Media.

Yang Yu and Xiao Wang [1] of Rochester Institute of Technology USA, collected real-time tweets from U.S. soccer fans during five 2014 FIFA World Cup games using Twitter search API and used sentiment analysis to examine U.S. soccer fans' emotional responses in their tweets, particularly, the emotional changes after goals. They found that during the matches that the U.S. team played, fear and anger were the most common negative emotions and in general, increased when the opponent team scored and decreased when the U.S. team scored. Anticipation and joy were also generally consistent with the goal results and the associated circumstances during the games. Their project revealed that sports fans use Twitter for emotional purposes and that the big data approach to analyze sports fans' sentiment showed results generally consistent with the predictions of the disposition theory when the fanship was clear and showed good predictive validity.

Apoorv Agarwal and his team [2] of Columbia University USA, presented results for sentiment analysis on Twitter. They used previously proposed state-of-the-art unigram model as their baseline and reported an overall gain of over 4% for two classification tasks: a binary, positive versus negative and a 3-way positive versus negative versus neutral. They presented a comprehensive set of experiments for both these tasks on manually annotated data that is a random sample of stream of tweets. They tentatively concluded that sentiment analysis for Twitter data was not that different from sentiment analysis for other genres.

Mike Thelwall et. al. [3] from University of Wolverhampton UK, described “An analysis of Twitter may give insights into why particular events resonate with the population.” Their article reported a study of a month of English Twitter posts, assessing whether popular events are typically associated with increases in sentiment strength. Using the top 30 events, determined by a measure of relative increase in (general) term usage, the results gave strong evidence that popular events were normally associated with increases in negative sentiment strength and some evidence that peaks of interest in events had stronger positive sentiment than the time before the peak. It seemed that many positive events were capable of generating increased negative sentiment in reaction to them.

Kumamoto and Tanaka [4][5] in Japan, focused on the impressions people got from news articles, and propose a method for determining impressions of these news articles. Their proposed method consisted of two main parts; one part involved building an `impression dictionary' that described the relationships among words and impressions. Another of the method involved determining impressions of input news articles using such an impression dictionary. The impressions of a news article were represented as scale values in user-specified impression scales, like `sad - glad' and `angry - pleased'. Each scale value was a real number between 0 and 1, and was calculated from the words (common nouns, action nouns, verbs, adjectives, and katakana characters) extracted from an input news article using the above impression dictionary.

With reference of these previous works, we have started to build our system for acquisition of tweeting data, sentiment analysis and visualization of usually hidden relation.

3 SYSTEM CONFIGURATION

This session illustrates a scheme of our system for data processing flow and an important idea how to select sample accounts as well as acquire tweeting messages from Twitter.

3.1 Data Processing Flow of our System

Our system is configured with three major parts, namely acquisition of data from Web (Social Media), information processing scheme (data mining for acquired message from Web), and generation of documents by our system. Figure 1 shows schematic block diagram for our system.

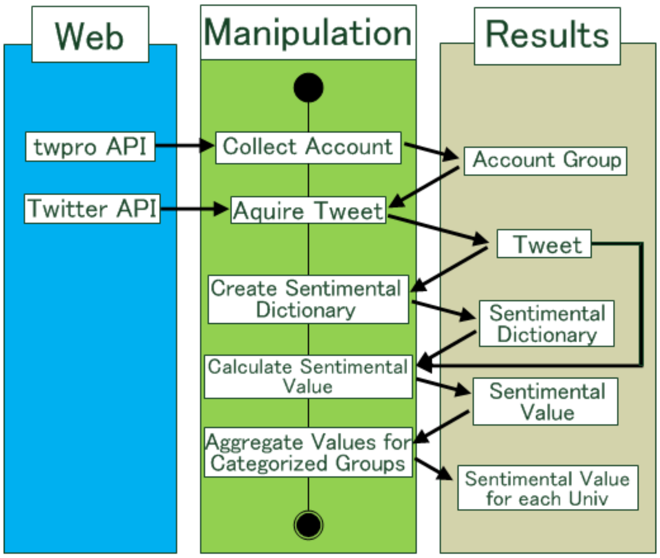

Web
Manipulation
Results
twpro API
Twitter API
Collect Account
Aquire Tweet
Create Sentimental Dictionary
Calculate Sentimental Value
Aggregate Values for Categorized Groups
Account Group
Tweet
Sentimental Dictionary
Sentimental Value
Sentimental Value for each Univ

Figure 1. Schematic Diagram for System Configuration

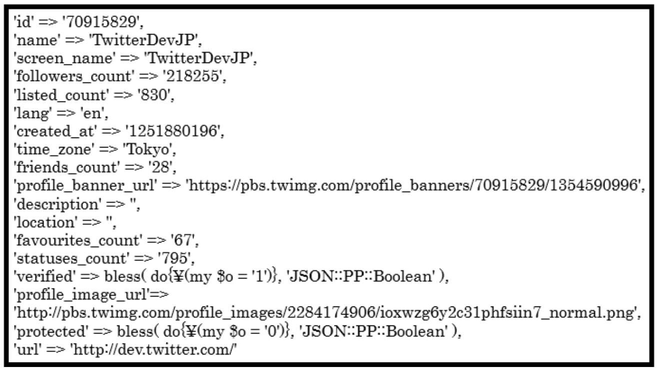
The leftmost part of our system is to connect Web-based Social Media (Twitter as our case) and to acquire amount of data to be analyzed. “twpro API” symbolizes selection facility to decide target group among the whole world. “Twitter API” plays a role of periodic data acquisition for the target group selected with “twpro API”. In our case, data to be acquired is tweeting message from Twitter. The number of account selected is 748. And acquired data from Twitter are approximately 2,280,000 numbers of tweeting messages.

The middle part is to present a series of manipulations for information processing from data into generated results which include some temporary files. These manipulations will be carried out from top to bottom sequentially. They are sometimes performed automatically. This part is the main body of our system and can accomplish some kind of sentiment analysis and generate our interesting results.

The rightmost part is to explain temporarily and/or permanently generated document files from the middle part. Some of them are output destination for one manipulation as well as input source for another one. All of them are to be analyzed in our study and these results possibly create useful evidences for trial visualization of relationship.

3.2 Acquisition of Tweeting Message

At first, we have decided the necessary accounts obtained from Twitter, which is the target of Social media for our sentiment analysis. We utilized an API in order to obtain accounting information and selected target accounts through which we obtain Tweeting message from Twitter. Figure 2 shows an example of retrieval results by using “twpro API” in Figure 1.



```
'id' => '70915829',
'name' => 'TwitterDevJP',
'screen_name' => 'TwitterDevJP',
'followers_count' => '218255',
'listed_count' => '830',
'lang' => 'en',
'created_at' => '1251880196',
'time_zone' => 'Tokyo',
'friends_count' => '28',
'profile_banner_url' => 'https://pbs.twimg.com/profile_banners/70915829/1354590996',
'description' => '',
'location' => '',
'favourites_count' => '67',
'statuses_count' => '795',
'verified' => bless( do{¥(my $o = '1')}, 'JSON::PP::Boolean' ),
'profile_image_url'=>
'http://pbs.twimg.com/profile_images/2284174906/ioxwzg6y2c31phfsiin7_normal.png',
'protected' => bless( do{¥(my $o = '0')}, 'JSON::PP::Boolean' ),
'url' => 'http://dev.twitter.com/'
```

Figure 2. Retrieval Results by means of “twpro API”

Our study is partly to focus on Natural Language Processing, so in this paper, we cannot avoid Japanese expression sometimes.

Social media, especially Twitter, will be playing a typical role in young generation cultures. We have selected the students of Japanese national universities as our target to acquire tweeting messages from Twitter. In such a case, we can consider that users of the relevant accounts must be approximately limited from 18 years old to 22 years old. The merit of our limitation is to assume that the according users used to tweet almost similar meaning for the same word and/or short sentence.

We can efficiently define a common dictionary to support translation from word for target message acquired from Twitter into sentimental values based on the previously selected account users. This is why we have introduced selection of account for users of Twitter by “twpro API”. We can, therefore, accomplish effective sentiment analysis for tweeting messages of selected account of Twitter.

4 SENTIMENT ANALYSIS

The first half of the section describes designing dedicated translation mechanism for our approach from extracted message to suitable words for sentiment analysis. The second one provides calculating sentimental values for selected tweeting message by means of utilization of translation mechanism.

4.1 Creation of Sentimental Dictionary

After acquisition of the target data to be analyzed from Twitter, suitable manipulation is necessary to do filtering words from sentences in acquired message in order to perform the first step of sentiment analysis. Such filtering manipulation can be carried out in the following ways;

(1) getting one line(= taking a sentence from a message)
(2) retrieving predefined character-string(= search parameter) in that line with pattern matching mechanism
(3) accumulating times of detection for searched patterns from line to line
(4) making a correlation between the total results during accumulation and the target message(= one tweeting)
(5) aggregating each message into categorized group according to amount of total results

Table 1. Category and Search Patterns

category	sentiment	search patterns
W_N	anger	murderous, scold, fume
	sadness	lonely, sorrow
	shame	bashfully, shamefully
	disgust	depressed, despair, heavy
	fear	helpless, uneasiness, weird
W_P	joy	smiling, pleasant feeling
	good	love, fascinate, admire
	comfort	cheerful, clean, easy
	excitement	desperate, impatient
	surprise	gulp, stupefy, swoon

It is very important to predefine search patterns based on their sentimental meaning and to categorize tweeting messages according to occurrences of search patterns and their sentimental meanings. Table 1 shows category of sentimental meanings and their examples of search patterns. W_N specifies set of words which do clearly Negative Sentiment, conventionally, and W_P also specifies set of words which express clearly Positive Sentiment, conversely. We must define S_N and S_P according to frequency of appearance in a tweet message about W_N and W_P. We assume that $w_N \in W_N$ and $w_P \in W_P$. S_N is a set of tweeting messages which include w_N and/or w_P if and only if the number of w_N is greater than the number of w_P, and S_P is another set of tweeting messages which include w_N and/or w_P if and only if the number of w_N is less than the number of w_P. $n(S_P)$ is number of S_P, while $n(S_N)$ is number of S_N. Figure 3 shows Venn diagram for Relation between S_N and S_P.

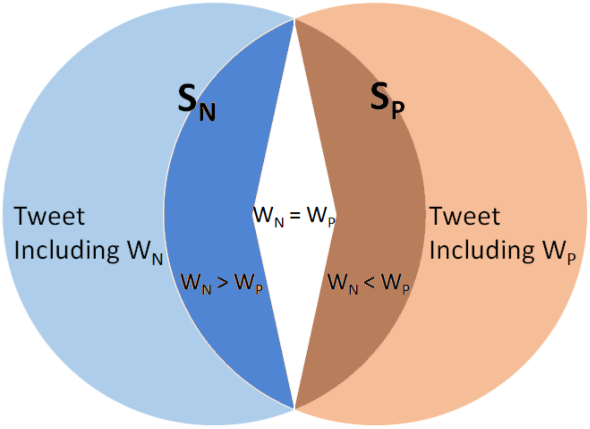


Figure 3. Venn diagram for Relation between S_N and S_P

Now, we can calculate conditional probabilities: $Pr_N(w)$ and $Pr_P(w)$ for a give word, respectively, as follows;

$$\Pr_N(w) = \frac{n(S_N, w)}{n(S_N)} \tag{1}$$

$$\Pr_P(w) = \frac{n(S_P, w)}{n(S_P)} \tag{2}$$

$n(S_N,w)$ is the number for “frequency of appearance” about the word: w in the above set: S_N, and $n(S_P,w)$ is also the number for "frequency of appearance" about the word: w in the above set: S_P. With conditional probabilities (1)(2), we specify the relevant sentimental value: $sv(w)$ as the expression (3) using special additional weights: $weight_N$ and $weight_P$, where $weight_N = log_{10}n(S_N)$ and $weight_P = log_{10}n(S_P)$. These weights are simply defined to grow as $n(S_N)$ and $n(S_P)$ increase.

$$sv(w) = \frac{2 * \Pr_P(w) * weight_P}{\Pr_N(w) * weight_N + \Pr_P(w) * weight_P} - 1 \quad (3)$$

So we can define the sentimental values for given words. In the next subsection, we will show an example of sentimental value dictionary in Table 2.

4.2 Calculation of Sentimental Value

Calculation of sentimental values for extracted tweeting message using dictionary is one of the most important manipulations in our system to perform some kind of sentiment analysis approach. Based on arguments in the previous section, we have defined dictionary for sentimental values to make a correlation between selected words and their sentimental values. Table 2 shows an example part of our sentimental value dictionary. It presents Japanese words at the leftmost, their sentimental values at the middle, and their meanings in English at the rightmost.

Table 2. Dictionary of Sentimental Value

word	value	mean
mondai	-0.2	problem
syakkin	-0.29	debt
heiwa	-0.06	peace
hazimari	0.214	begins
hohoemi	1	smile

Our dictionary of sentimental value now has approximately 9,500 words of items in it. With references of such a dictionary, calculation of sentimental value for one sentence will be carried out, for example, in the following ways, which is shown in Figure 4. This Figure tries to present schematic illustration of mechanism to calculate sentimental value in the sentence-by-sentence manner using English expression instead of real Japanese message.

- At the upper case, Meaning of the first sentence is "I have a problem with my debt," which really includes Japanese word {"mondai", "syakkin"}. Each word can be converted into sentimental value {-0.2, -0.29}. And a total result of sentimental value for the relevant tweeting sentence is -0.49. The result is negative
- At the lower case, the second is "Peace begin with a smile," which really includes Japanese word {"heiwa","hazimari", "hohoemi"}. Each word can be converted into sentimental value {-0.06, 0.214, 1}. And a total result of sentimental value for the relevant tweeting message is 1.154. The result is positive.

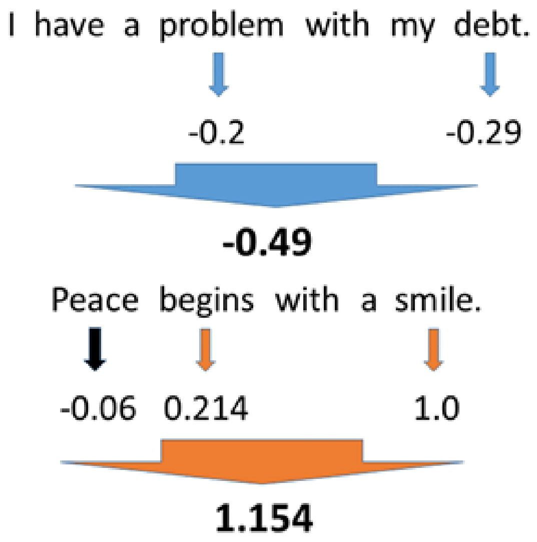


Figure 4. Calculation Example of Sentimental Value based on Dictionary shown in Table 2

Just like the same ways shown in Figure 4, our system can calculate sentimental values for each tweeting message with our sentimental value dictionary. And moreover we can categorize users of accounts based on the criterion into some groups. After that, we calculate the average sentimental value for each categorized group for the sake of visualization of the relation between sentimental values and characteristics of patterns of thinking and behaving in the same categorized group.

5 DISCUSSION

This section presents discussion about results for sentiment analysis and demonstration for trial visualization, namely discuss whether to visualize hidden relation between sentimental value and characteristics of categorized group or not.

5.1 Demonstration to show Hidden Relation

We have focused on tweeting message from young persons, who are the students of 26numbers of Japanese national universities with medical faculty. Table 3 shows categorized sentimental values for the 26 universities using normalized summation of sentimental values (calculated with expression (4)) for selected 20 students.

$$SV(u) = \sum_{i=s}^{i=e} S(u)_i \Big/ n \qquad (4)$$

Table 3. Categorized sentimental values for each university

University ID u	Categorized Sentimental Values SV(u)
01	0.109
02	0.218
03	0.229
04	0.268
05	0.284
06	0.306
07	0.311
08	0.313
09	0.321
10	0.325
11	0.328
12	0.338
13	0.352
14	0.357
15	0.362
16	0.364
17	0.374
18	0.398
19	0.418
20	0.428
21	0.431
22	0.439
23	0.447
24	0.462
25	0.495
26	0.672

SV(u) is categorized sentimental value, defined for each university: *u* and calculated by expression (4). *S(u)ᵢ* is an average of sentimental values per one tweet on one day: *i* by the students of their belonging university: *u*. In (4), an interval of its summation can be specified from “Sampling Start Date: *s*” to “Sampling End Date: *e*”. And *n* is total number of tweets per one day.

So we can provide an example of base for “Rank Correlation” of the relevant 26 numbers of our universities by means of sentimental values from Twitter. Table 3 also shows the above example using our expression (4) in the mode of ascending order. With such an order of each university based on categorized sentimental values, we can probably some capability to visualize normally unrecognized, i.e. so-called ‘hidden’, relationship between trends/phenomena/behaviour of tweeting messages and the relevant categorized universities by means of the intermediation of sentimental values.

We can obtain a published data book for Japanese Universities, named “Daigaku no Jitsuryoku” [6] (in Japanese, its meaning of the book title is capability and current status of university). We have challenged to perform visualization of a hidden relationship between Japanese universities and their possessing students’ tweeting messages with reference of the above published data book. People of university, students and teachers, are very interesting in one of special information about ranking of university, for example, an order for “dropout ratio of students” at each university. So we have decided to demonstrate visualization of relation between categorized sentimental values for each university and dropout ratio of students at the corresponding university

Figure 6 shows a scatter graph of the relation between Categorized Sentimental Value for each University and the corresponding Dropout Ratio of Students at University as a results of the above demonstration. From Figure 6, we can probably recognize some uncertain (we cannot describe suitably) relation between dropout ratio and sentimental value, because Spearman's rank correlation coefficient (ρ) is statistically calculated as $\rho = 0.33$ in this case. It seems that young people, students of university, who present their tweeting message with higher sentiment values, belong to university which has higher scores of dropout ratio. So this result can probably demonstrate some relation between categorized sentimental value and characteristics of patterns of thinking and behaving in the correspondingly categorized group.

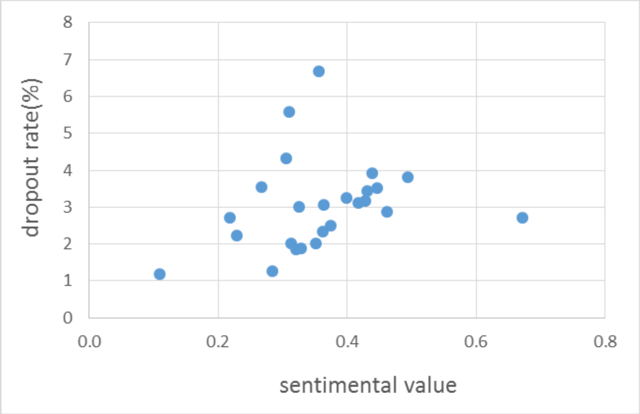


Figure 5. Scatter Graph of Relation between Categorized sentimental values for university and its Dropout ratio

Of course, we have been investigating some feasible reasons for such a relation, but we cannot find globally useful and reasonable interpretation until now. So we do not definitely report that this is a meaningful example for "visualization of hidden relation" by means of sentiment analysis using message from Twitter.

5.2 Trial Visualization of Other Relations

In order to confirm that our approach is available for visualization of generally hidden and normally non-recognized relations, we have applied the same approach to investigate possible relations or dependencies among our categorized sentimental values, period of time to generate tweeting, numbers of students / teachers of universities, and location factors of universities. We can recognize accurate period of time to generate the relevant tweeting messages from Twitter. We can easily obtain correct information about the numbers of students/teachers from published data of statistics. Location factors of universities can be obtained from geometric data (also published).

Coefficients of correlation between sentimental values and period of time to generate tweeting are calculated and summarized as follows: $r = 0.06$ for anytime, and $r = -0.08$ for daytime (from 9:00 o'clock to 18:00 o'clock). We must determine that sentimental values can be independent from period of time to generate the relevant tweeting messages from Twitter.

Secondarily, we have applied the same approach to determine whether or not there exist some possible relations between sentimental values and the number of students or the number of teachers.

Figure 6 shows a scatter graph of relation between our categorized sentimental values for university and the numbers of students in the corresponding university. And Figure 7 shows a graph of the relevant relation between the categorized sentimental values for university and the numbers of teachers in the corresponding university.

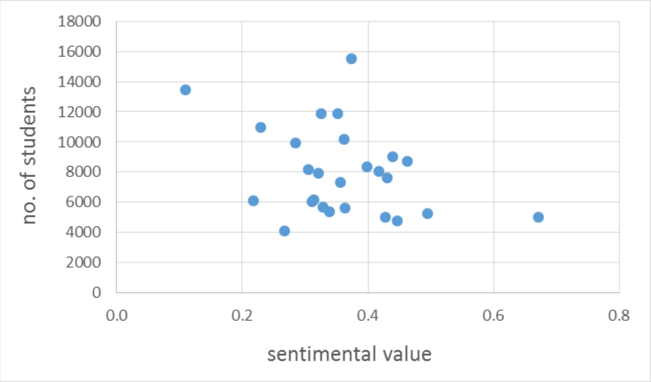


Figure 6. Scatter Graph of Relation between Categorized sentimental value for university and its number of students

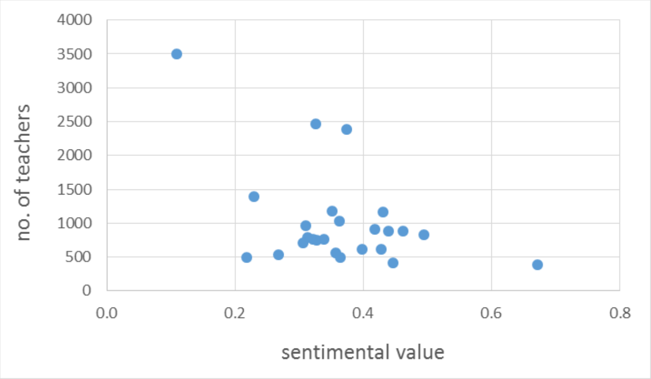


Figure 7. Scatter Graph of Relation between Categorized sentimental value for university and its number of teachers

Coefficients of correlation between sentimental values and the numbers of students / teachers are calculated and summarized as follows: $r = -0.34$ for student number, and $r = -0.46$ for teacher number. We can assume that our categorized sentimental value is negatively-correlated with the numbers of students / teachers of the corresponding universities. The numbers of students and teachers can indicate or express scaling factor of the relevant university. So the above results might visualize some hidden relation between sentimental value and scaling factor of university, namely there are probably existing some dependency of sentimental value and scaling factor of university.

And finally we have applied our approach to determine whether or not there exist some possible relations between categorized sentimental values and population of the corresponding prefecture. Figure 8 shows a scatter graph of relation between our categorized sentimental values for university and the corresponding prefectural population.

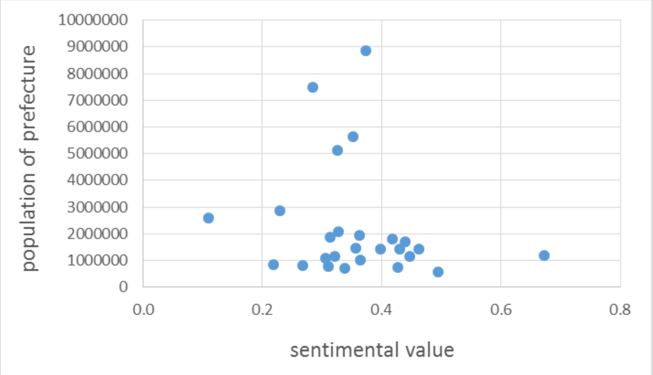


Figure 8. Scatter Graph of Relation between Categorized sentimental value for university and prefectural population

Coefficient of correlation between sentimental values and population of the corresponding prefecture is calculated as $r = -0.18$, and it is considered to be slightly negative relation between them. But we cannot definitely assume that there exists some relation between categorized sentimental value and population of the corresponding prefecture.

6 CONCLUSION

This paper described an approach to collect accounts of Twitter, acquire tweeting messages, create useful sentimental dictionary, calculate sentimental values, and aggregate sentimental values for categorized groups. And it also demonstrates some applications to visualization of some relations, which are normally not recognized and potentially hidden in our conventional environments. So it can show our trial application of sentiment analysis to visualization of hidden relation between categorized sentimental value and characteristics of patterns of thinking and behaving in the same categorized group.

Our aim of this research is to visualize some normally unrecognized relation, namely hidden relation in other words, using approach by big data analysis. This study can bring us amount of information, useful experience and suitable techniques for sentiment analysis for visualization. Simultaneously, it has provided a very important problems for us to resolve in the future.

REFERENCES

1. Yang Yu, Xiao Wang,"World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets," Computers in Human Behavior, Vol.48, pp. 392 - 400, 2015.
2. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data," Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30 - 38, Portland, Oregon, 23 June 2011.
3. Mike Thelwall, Kevan Buckley, and Georgios Paltoglou, "Sentiment in Twitter events," Journal of the American Society for Information Science and Technology Volume 62, Issue 2, pp. 406 - 418, February 2011.
4. Tadahiko Kumamoto, Yukiko Kawai, Katsumi Tanaka, "Design and Evaluation of a Method for Mining Impressions from Text (in Japanese)," The IEICE Transaction D on Information and Systems(Institute of Electronics, Information and Communication Engineers), Vol.J94-D(3), pp. 540-548, March 2011.
5. Tadahiko Kumamoto, Katsumi Tanaka, "Proposal of Impression Mining from News Articles", Knowledge-Based Intelligent Information and Engineering Systems Lecture Notes in Computer Science Volume 3681, pp 901-910, 2005
6. Yomiuri-Shinbun-sya eds. (published by Chuokouron-sya). "Daigaku no Jitsuryoku"(136 pages 2015 ISBN978-4-12-004656-8)
http://www.chuko.co.jp/tanko/2014/ 09/004656.html
http://kyoiku.yomiuri.co.jp/torikumi/jitsuryoku/yomitoku/contents/1-4.php