Visualizing Impression-Based Preferences of Twitter Users

Tadahiko Kumamoto, Tomoya Suzuki, and Hitomi Wada

Chiba Institute of Technology, Narashino, Chiba 275-0016, Japan
kumamoto@net.it-chiba.ac.jp

Abstract. Twitter is extremely useful for connecting with other users, because, on Twitter, following other users is simple. On the other hand, people are often followed by unknown and anonymous users and are sometimes shown tweets of unknown users through the tweets of the users they follow. In such a situation, they wonder whether they should follow such unknown users. This paper proposes a system for visualizing impression-based preferences of Twitter users to help people select whom to follow. The impression-based preference of a user is derived based on the impressions of the tweets the user has posted and those of the tweets of users followed by the user under consideration. Our proposed system enables people to select whom to follow depending on whether or not another user adheres to the user's own sensibilities, rather than on whether or not another user provides valuable information.

1 Introduction

A number of social networking services (SNS), such as Twitter and Facebook, are actively used. Twitter is superior to other SNSs as a tool for connecting with famous people, on-screen talent, and strangers, as well as with friends and acquaintances, because it makes following any user simple. Many users post daily tweets with topics ranging from political and economic events to events concerning themselves. Obtaining such information routinely requires following users who are posting the information. However, the type of tweets that users normally post can be judged by reading a large number of users' tweets carefully. Some users may always tweet only somber messages, and other users may always express anger in their tweets. Many users prefer not to receive negative tweets.

This paper proposes a system for visualizing impression-based preferences of Twitter users to help people select whom to follow. Usually, Twitter users view tweets posted by someone they follow voluntarily, or post tweets with the awareness that the tweets are viewed by their followers. Therefore, we consider that the impression-based preferences of Twitter users can be derived based on the impressions of the tweets they have posted and those of the tweets the users they follow have posted. People can use the proposed system to determine the type of tweets Twitter users usually view or post by visually checking the impression-based preferences of the users. The system also extracts keyphrases,

G. Meiselwitz (Ed.): SCSM 2014, LNCS 8531, pp. 209–220, 2014.

or characteristic character strings, from each set of tweets and presents the commonalities and differences among the sets. This helps people identify the topics that interest the users.

When people specify the account name of a Twitter user as input to our proposed system, the system uses the Twitter API [1] to collect tweets posted by the specified user and her or his following users. Then, it rates each tweet based on three distinct impressions, using the impression mining method that the authors have proposed [2]. The target impressions are limited to those represented by three bipolar scales of impressions [3], "Happy – Sad," "Glad – Angry," and "Peaceful – Strained." The strength of each impression is computed as an "impression value," *i.e.* a real number between one and seven denoting a position on the corresponding scale. For example, on the scale "Happy – Sad," the score one indicates "Happy," the middle score four denotes "Neither happy nor sad," and the score seven equals "Sad." If the impression value of a tweet is 2.5, then the average person will experience an intermediate impression between "Comparatively happy (2)" and "A little happy (3)" from reading the tweet. In addition, the system uses the Yahoo! Keyphrase Extraction API [4] to extract keyphrases from each set of tweets. Last, the system uses the Google Chart API [5] to visualize the results of the previous processing and presents it to the user. Using our proposed system, people can visually grasp the impression-based preferences of their specified users, that is, the type of tweets the specified users usually view or post and the topics that interests them.

2 Related Work

Many systems have been developed for using information on Twitter effectively including a system that detects trends over the Twitter stream [6], a system that recommends news articles based on Twitter-based user modeling [7], and a system that detects earthquakes by monitoring tweets [8].

Studies of ways to recommend users as candidates to follow are also ongoing, with the goal of supporting the making of connections with other Twitter users. Weng et al. proposed a Pagerank-like algorithm, called TwitterRank, for identifying influential Twitter users and recommending them as users to follow [9]. Sadilek et al. proposed a system for suggesting users to follow by inferring friendship in the physical world [10]. Pennacchiotti et al. proposed a method that suggests users who have a similar latent interest based on information extracted from their tweets [11]. In Japan, a method for recommending users to follow was proposed based on how many of the user's tweets were registered in "Favorites" regarding a topic [12]. On the official Twitter website, several users are suggested as "Who to follow" based on whom other users follow and other criteria. In addition, users can easily follow popular users in the field of a topic by clicking on "Popular accounts" on the official website and selecting a topic that interests them. Our proposed system enables people to select whom to follow depending on whether or not they adhere to their own sensibilities, rather than on whether or not they provide valuable information, by incorporating information on their sensibilities or impression-based preferences of users.

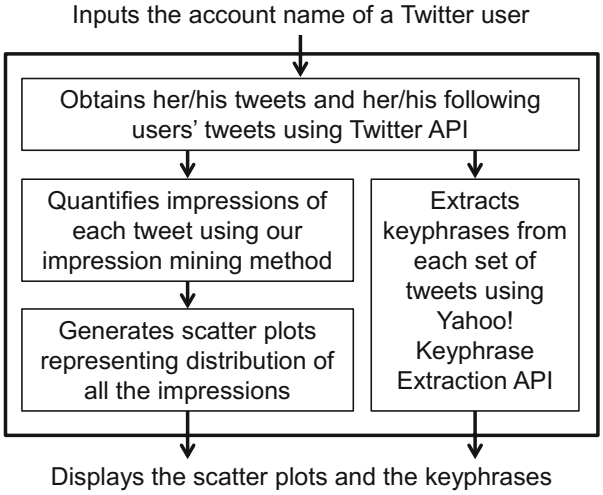


Fig. 1. Architecture of our proposed visualization system

In research on affective computing and sentiment analysis, studies on extracting subjective information, called sentiment, emotion, or impression, from text data, such as reviews, news articles, and web pages, are ongoing, and their results have been applied to various task domains, for example, sentiment analysis [13], information visualization [14], annotation of impression tags [15], and the like. However, these studies only classify text data into emotion classes or attach impression tags to text data, without quantifying impressions of those data.

3 Design of the Visualization System

3.1 System Architecture

We show the architecture of our proposed system in Figure 1. First, the system's users are asked to specify a Twitter user by entering her or his account name. Using the Twitter API, the system then obtains tweets posted by the specified user and tweets posted by her or his following users. These two sets of tweets are analyzed in the following two ways:

In the first method, the system computes three impression values of each tweet using the impression mining method we proposed previously [2]. This method generates word unigrams from an input tweet and obtains values representing the effect of each word unigram, based on the impression lexicons we constructed from a newspaper database. The method then computes and outputs the three impression values of the tweet based on these values. This method is highly accurate even for unlearned data, and our experimental results show that the average root-mean-square errors (RMSE) for unlearned data were 0.69, 0.49, and 0.64 on the respective scales. Last, the proposed system plots the impression

Table 1. Regular expressions to remove specific character strings

Character strings to be removed	Regular expressions
URLs	`/(https?:\/\/[\x21-\x7e]+)/i`
Twitter account names	`/@(\w+)/i`
Face marks	`/(\(.*?\))/`
Letter "w" concatenated twice or more	`/(w|w){2,}/i`

values on three two-dimensional planes, each spanned by two of the three scales. In each 2D plane, tweets posted by the specified user and her or his following users are explicitly differentiated by red and blue plots, respectively.

In the second method, the proposed system extracts keyphrases from each set (tweets of the specified user and tweets of her or his following users) using Yahoo! Keyphrase Extraction API. Then, the system compares keyphrases extracted from one set with those extracted from the other set, and presents the common and different keyphrases of the sets. Consequently, the system's users become aware of the strong and weak interests of the specified user.

The following subsections will describe in detail the main parts of the system.

3.2 Collecting Tweets on User and Home Timelines

Users are asked to input a Twitter user's account name to the system. First, the system obtains user and home timelines (User TL and Home TL, respectively) of the specified user from Twitter using the Twitter API. In this paper, the former timeline consists of the tweets the specified user posted, and the latter timeline consists of the tweets the users, or followees, whom the specified user has followed posted. A user's User TL can be obtained using the API for getting User TLs. On the other hand, only the Home TL of a user authenticated in advance can be obtained using the API for getting Home TLs; the Home TLs of arbitrary users are not available. The system must reconstruct the Home TL of an arbitrary user artificially. First, the system gets an ID list of the specified user's followees using the Twitter API, and then it checks the newest date and hour posted in each ID. Any ID in which the newest date and hour is older than the base date set in advance is removed from the ID list. The base date is set to one, representing "yesterday" as a default value, after which it can be changed freely by the system's users. Next, the system gets the User TL of each followee who has an ID in the remaining ID list, and then reconstructs the Home TL of the specified user. Note that tweets for which the date and hour are older than the base date are not collected.

3.3 Removing Noise

The system removes character strings, such as face marks and Uniform Resource Locators (URLs), from every tweet, because the current version of the impression mining method cannot analyze these character strings adequately. Character

strings to be removed and regular expressions for removing them are enumerated in Table 1.

3.4 Quantifying Impressions of Tweets

Determining Target Impressions. We designed six bipolar scales suitable for representing impressions of news articles, "Happy – Sad," "Glad – Angry," "Interesting – Uninteresting," "Optimistic – Pessimistic," "Peaceful – Strained," and "Surprising – Common." First, we conducted nine experiments, in each of which 100 subjects read ten news articles and estimated their impressions on a scale from one to five for each of 42 impression words. These 42 impression words were manually selected from a Japanese thesaurus [16] as words that can express impressions of news articles. Next, factor analysis was applied to the data obtained in the experiments, and the 42 words were divided into four groups, negative words, positive words, two words that were "uninteresting" and "common," and two words that were "surprising" and "unexpected." In the meantime, after cluster analysis of the data, the 42 words were divided into ten groups. The results of the two analyses were used to create the six bipolar scales mentioned above. We showed that impressions on the "Surprising – Common" scale differed greatly among individuals in terms of their perspective. We also showed that processing according to the background knowledge, interests, and characters of individuals was required to deal with the impressions represented by the two scales "Interesting – Uninteresting" and "Optimistic – Pessimistic." Therefore, we decided not to use these three scales at the present stage, and adopted the remaining three scales, "Happy – Sad," "Glad – Angry," and "Peaceful – Strained."

Constructing Impression Lexicons. An impression lexicon plays an important role in computing impressions of text data. In this paper, we describe the implementation of a method for automatically constructing an impression lexicon.

First, two contrasting sets, each consisting of multiple reference words, are used to construct an impression lexicon for each scale. Next, we let the set of reference words that expresses an impression at the left of a scale be S_L, and we let the set of reference words that expresses an impression at the right of the scale be S_R. Articles including one or more reference words in S_L or S_R are extracted from a newspaper database, and the number of reference words belonging to each set is counted in each article. For this we use the 2002 to 2006 editions of the Yomiuri Newspaper Text Database as the newspaper database. Then, we let A_L be the articles that each contains a number of reference words belonging to S_L larger than the number of reference words belonging to S_R, and we let the number of articles in A_L be N_L. We let A_R be the articles that each contains a number of reference words belonging to S_L smaller than the number of reference words belonging to S_R, and we let the number of articles in A_R be N_R. Next, all words are extracted from each of A_L and A_R except for

Table 2. Specifications of our impression lexicons

Scales	# of entries	W_L	W_R
Happy – Sad	387,428	4.90	3.80
Glad – Angry	350,388	4.76	3.82
Peaceful – Strained	324,590	3.91	4.67

Table 3. Reference words prepared for each scale

Scales	Reference words
Happy	tanoshii (happy), tanoshimu (enjoy), tanosimida (look forward to), tanoshigeda (joyous)
– Sad	kanashii (sad), kanashimu (suffer sadness), kanashimida (feel sad), kanashigeda (look sad)
Glad	ureshii (glad), yorokobashii (blessed), yorokobu (feel delight)
– Angry	ikaru/okoru (get angry), ikidooru (become irate), gekidosuru (get enraged)
Peaceful	nodokada (peaceful), nagoyakada (friendly), sobokuda (simple), anshinda (feel easy)
– Strained	kinpakusuru (strained), bukimida (scared), fuanda (be anxious), osoreru (fear)

particles, adnominal words[1], and demonstratives, and the document frequency of each word is measured. Then, we let the document frequency in A_L of a word w be $N_L(w)$, and we let the document frequency in A_R of a word w be $N_R(w)$. The revised conditional probabilities of a word w are defined as follows.

$$P_L(w) = \frac{N_L(w)}{N_L}, \qquad P_R(w) = \frac{N_R(w)}{N_R}$$

In these equations, only articles that satisfy the assumptions described above are used to calculate $P_L(w)$ and $P_R(w)$.

Finally, the impression value $v(w)$ of a word w is calculated using these $P_L(w)$ and $P_R(w)$ as follows.

$$v(w) = \frac{P_L(w) * W_L}{P_L(w) * W_L + P_R(w) * W_R}$$

$$W_L = \log_{10} N_L, \qquad W_R = \log_{10} N_R$$

That is, a weighted interior division ratio $v(w)$ of $P_L(w)$ and $P_R(w)$ is calculated using these formulas and stored as an impression value of w in the scale "S_L – S_R" in an impression lexicon. Note that W_L and W_R denote weights, and that the larger N_L and N_R are, the heavier W_L and W_R are.

[1] This part of speech exists only in Japanese, not in English. For example, "that," "so called," and "of no particular distinction" are expressed using adnominal words in Japanese.

The numbers of entries in the impression lexicons constructed as above are shown in Table 2, together with the values of W_L and W_R obtained. Further, the two contrasting sets of reference words[2] used in creating the impression lexicons are enumerated in Table 3 for each scale. These words were determined after some trial and error and are based on two criteria: (i) a word is a verb or adjective that expresses either of two contrasting impressions represented by a scale, and (ii) as far as possible, the word does not suggest other types of impressions.

Computing Impression Values of News Articles. For each scale, the impression value of a news article is calculated as follows. First, the article is segmented into words using "Juman" [17][3], one of the most powerful Japanese morphological analysis systems, and an impression value for each word is obtained by consulting the impression lexicon constructed for the scale. Seventeen rules that we designed are then applied to the Juman output. For example, there is a rule that a phrase of a negative form such as "sakujo-shi-nai (do not erase)" should not be divided into a verb "shi (do)," a suffix "nai (not)," and an action noun "sakujo (erasure)," but should be treated as a single verb "sakujo-shi-nai (do-not-erase)." There is also a rule that an assertive phrase such as "hoomuran-da (is a home run)" should not be divided into a copula "da (is)" and a noun "hoomuran (a home run)," but should form a single copula "hoomuran-da (is-a-home-run)." Further, there is a rule that a phrase with a prefix, such as "sai-charenji (re-challenge)" should not be divided into a prefix "sai (re)" and an action noun "charenji (challenge)," but should form a single action noun "sai-charenji (re-challenge)." All the rules are applied to the Juman output in creating impression lexicons and computing the impression values of articles. Finally, an average of the impression values obtained for all of the words except for particles, adnominal words, and demonstratives is calculated and presented as the impression value of the article.

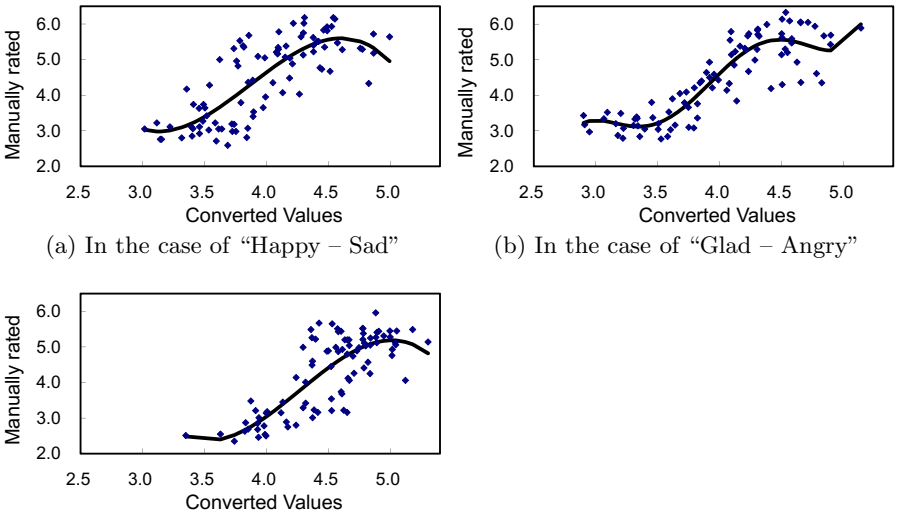
Correcting Computed Impression Values. We considered that some gaps would occur between impression values computed by an unsupervised method such as the one we used and those of the readers. Therefore, we conducted experiments with a total of 900 people participating as subjects and identified the gaps that actually occurred.

First, we conducted experiments with 900 subjects and obtained data that described correspondence relationships between news articles and impressions to be extracted from the articles. That is, the 900 subjects were randomly divided into nine equal groups, each group consisting of 50 males and 50 females, and 90 articles selected from the 2002 edition of the Mainichi Newspaper Text Database[4]

[2] These words were translated into English by the authors.

[3] Since there are no boundary markers between words in Japanese, word segmentation is needed to identify individual words.

[4] This database is different from the Yomiuri newspaper database we used in creating impression lexicons.



(a) In the case of "Happy – Sad"

(b) In the case of "Glad – Angry"

(c) In the case of "Peaceful – Strained"

Fig. 2. Scatter diagrams and regression equations

were randomly divided into nine equal parts. Then, each subject was asked to read the ten articles presented in a random order and rate each of them using three seven-point bipolar scales presented in a random order. The scales we used were "Happy – Sad," "Glad – Angry," and "Peaceful – Strained," and the subjects were asked to assess, on a scale from one to seven, the intensity of each impression, represented by each scale, from reading a target article. After the experiments, for each scale, we calculated an average of the 100 values rated for every article. We regarded this average as the impression value to be extracted from the article. Note that in these experiments, we presented only the first paragraphs of the original news articles to the subjects. This procedure was based on the fact that people can understand the outline of a news article by just reading the first paragraph of the article.

Next, impression values for the first paragraphs of the 90 articles were computed using the method we implemented in 3.4, where the first paragraphs were identical to those presented to the subjects in the experiments. Note that according to the definition of our equations, these impression values are close to one, when impressions on the left of a scale are felt strongly, and are close to zero, when impressions on the right of a scale are felt strongly. We therefore used the following formula to convert the computed value into a value between 1.0 and 7.0.

$$Converted = (1 - Computed) * 6 + 1$$

Next, for each scale, we drew a scatter diagram to identify the potential correspondence relationship between these converted values and the averages

Table 4. Regression equations designed for impression data of ninety articles

Scales	Regression equations (x: converted values)
Happy – Sad	$-1.6355586x^3 + 18.971570x^2 - 70.68575x + 88.5147$
Glad – Angry	$2.384741939x^5 - 46.87159982x^4 + 363.6602058x^3 - 1391.589442x^2 + 2627.06261x - 1955.3058$
Peaceful – Strained	$-1.7138394x^3 + 21.942197x^2 - 90.79203x + 124.8218$

obtained in the experiments, as illustrated in Figure 2. We can see from any of the scatter diagrams that the impression values manually rated by the subjects are positively correlated with those automatically computed by the method we implemented. In fact, from the case at the top of the figure, their coefficients of correlation are 0.76, 0.84, and 0.78, which are all high. This not only means that as an overall trend, the underlying assumption of this paper is confirmed, but also indicates that the correspondence relationships can be represented by regression equations.

Next, we applied regression analysis to the converted values and the averages, where the converted values were used as the explanatory variable, and the averages were used as the objective variable. Various regression models, such as linear function, logarithmic function, logistic curve, quadratic function, cubic function, quartic function, and quintic function, were used in this regression analysis on a trial basis. As a result, the regression equation with the highest coefficient of determination was determined to be an optimal function denoting the correspondence relationship between the converted values and the averages in each scale. This means that for each scale, the impression value of an article was obtained more accurately by correcting the value computed by the method we implemented using the corresponding regression equation.

The regression equations obtained here are shown in Table 4 and are already illustrated on the corresponding scatter diagrams in Figure 2. Their coefficients of determination were 0.63, 0.81, 0.64, respectively, which were higher than 0.5 in all scales. This means that the results of regression analysis were good. In addition, we can see from Figure 2 that each regression equation fits the shape of the corresponding scatter diagram.

The impression mining method described above is applied to tweets, and three impression values of each tweet are computed.

3.5 Extracting Keyphrases

Keyphrases are extracted from the tweets from which noise was removed. The system extracts keyphrases from two sets of tweets, those of the specified user and those of the specified user's followees using the Yahoo! Keyphrase Extraction API. Last, twenty keyphrases extracted from each set of tweets are presented to the system's user in a tag cloud form.

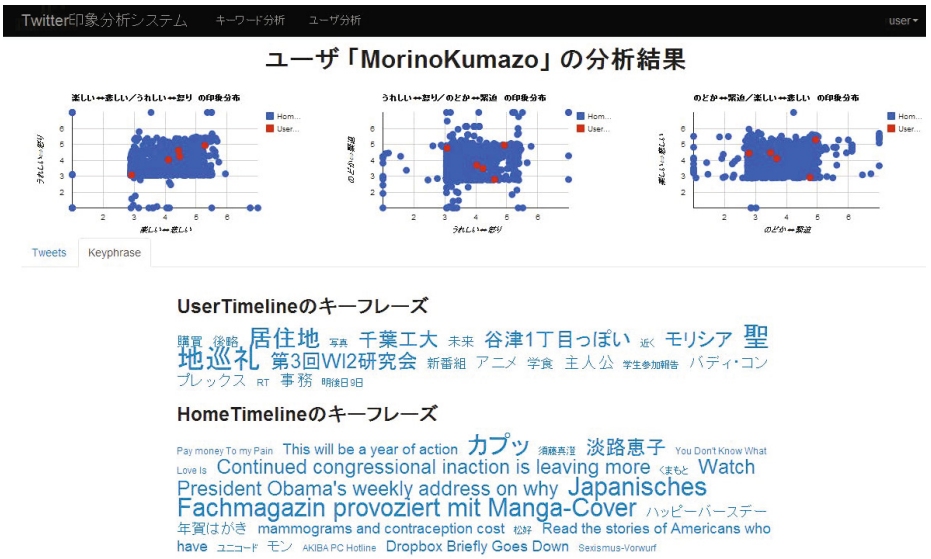


Fig. 3. Results of user analysis

3.6 Generating Scatter Plots

The system generates three scatter plots from the impression values of the tweets using the Google Chart API to visualize the impression-based preference of a Twitter user. That is, a plot for impression values in "Happy – Sad" and "Glad – Angry," one for "Glad – Angry" and "Peaceful – Strained," and one for "Peaceful – Strained" and "Happy – Sad" are generated. In each scatter plot, tweets on a User TL are indicated by red plots, and those on a Home TL by blue plots.

4 Implementation as Web Application

Our proposed system has been implemented as a web application system. This section provides a snapshot to show how the proposed system works.

A snapshot of the screen displayed when we specified *@MorinoKumazo*, one of the authors, as the target user is shown in Figure 3. Three scatter plots are displayed at the top of the screen. Scatter plots for impression values in "Happy – Sad" and "Glad – Angry", for impression values in "Glad – Angry" and "Peaceful – Strained," and for impression values in "Peaceful – Strained" and "Happy – Sad" are on the left, in the middle, and on the right, respectively. Keyphrases extracted from User and Home TLs are displayed separately in tag cloud form at the lower part of the screen. The system's user can view the tweets that were classified into the impression scale she or he selected, instead of using keyphrases.

5 Conclusion

In this paper, we presented a web application system for visualizing impression-based preferences of Twitter users. When a person specifies the account name of a Twitter user as input to the system, the system checks and visualizes the type of tweets the specified user usually views or posts. The target impressions are limited to those represented by three bipolar scales of impressions: "Happy – Sad," "Glad – Angry," and "Peaceful – Strained." With the system, a person can easily grasp the impression-based preferences of the specified user. The system also extracts twenty keyphrases from each of the user and home timelines of the specified user and presents the commonalities and differences of the two timelines. This helps people to identify the topics that interest the specified user.

Our future work is as follows. Since the impression mining method we used in the proposed system was designed for quantifying impressions of news articles [2], the effectiveness of the method for tweets has not been verified. Many ungrammatical sentences, short sentences consisting of one or two words, and Twitter-dependent expressions, such as face marks and Internet slang words, are observed in tweets. We therefore consider that the current lexicon-based approach to impression mining is not suitable for such tweets. Now we are planning to design and develop an impression mining method suitable for tweets. Impression scales should also be redesigned according to impressions to be extracted from tweets. In addition, we will design a followee recommendation system by expanding the proposed system.

Acknowledgment. This work was supported by JSPS KAKENHI Grant Number 24500134 and donations from Masaharu Fukuda.

References

1. Documentation — Twitter Developers, `https://dev.twitter.com/docs`
2. Kumamoto, T., Kawai, Y., Tanaka, K.: Improving a Method for Quantifying Readers' Impressions of News Articles with a Regression Equation. In: Proc. of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon, USA, pp. 87–95 (2011)
3. Kumamoto, T.: Design of Impression Scales for Assessing Impressions of News Articles. In: Yoshikawa, M., Meng, X., Yumoto, T., Ma, Q., Sun, L., Watanabe, C. (eds.) DASFAA 2010. LNCS, vol. 6193, pp. 285–295. Springer, Heidelberg (2010)
4. Yahoo! Keyphrase Extraction API – Yahoo! Developers, `http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html`
5. Google Charts – Google Developers, `https://developers.google.com/chart/`
6. Mathioudakis, M., Koudas, N.: TwitterMonitor: Trend Detection over the Twitter Stream. In: Proc. of the ACM SIGMOD International Conference on Management of Data, Indianapolis, USA, pp. 1155–1158 (2010)
7. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Twitter-Based User Modeling for News Recommendations. In: Proc. of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, pp. 2962–2966 (2013)

8. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proc. of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, pp. 851–860 (2010)
9. Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank, Finding Topic-Sensitive Influential Twitterers. In: Proc. of the Third ACM International Conference on Web Search and Data Mining, New York, USA, pp. 261–270 (2010)
10. Sadilek, A., Kautz, H., Bigham, J.P.: Finding Your Friends and Following Them to Where You Are. In: Proc. of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, USA, pp. 723–732 (2012)
11. Pennacchiotti, M., Gurumurthy, S.: Investigating Topic Models for Social Media User Recommendation. In: Proc. of the 20th International Conference Companion on World Wide Web, Hyderabad, India, pp. 101–102 (2011)
12. Watabe, S., Miyamori, H.: Twitter User Recommender: The User Recommendation System Using the Favorite Function of Twitter. In: Proc. of the Forum on Data Engineering and Information Management, No. B3-4, Kobe, Japan (2012)
13. Pang, B., Lee, L.: Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In: Proc. of the Annual Meeting on Association for Computational Linguistics, Morristown, USA, pp. 115–124 (2005)
14. Lin, K.H.-Y., Yang, C., Chen, H.-H.: Emotion Classification of Online News Articles from the Reader's Perspective. In: Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 220–226 (2008)
15. Kiyoki, Y., Kitagawa, T., Hayama, T.: A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning. ACM SIGMOD Record 23(4), 34–41 (1994)
16. Ohno, S., Hamanishi, M. (eds.): Ruigo-Kokugo-Jiten. Kadokawa Shoten Publishing Co.,Ltd., Tokyo (1986)
17. Kurohashi, S., Nakamura, T., Matsumoto, Y., Nagao, M.: Improvements of Japanese Morphological Analyzer JUMAN. In: Proc. of the International Workshop on Sharable Natural Language Resources, Nara, Japan, pp. 22–28 (1994)