2012 International Conference on Industrial and Intelligent Information (ICIII 2012)
IPCSIT vol.31 (2012) © (2012) IACSIT Press, Singapore

Extracting Similar and Opposite News Websites Based on Sentiment Analysis

Jianwei Zhang [1], Yukiko Kawai [1] and Tadahiko Kumamoto [2]

[1] Kyoto Sangyo University
[2] Chiba Institute of Technology

Abstract. With the widespread of online news websites, people can browse and retrieve news articles more easily. However, for a contentious news topic, different news websites may have different sentiment tendencies and the tendencies may vary over time. To catch this feature, we construct a sentiment dictionary and develop a system that can extract news articles' sentiments, visually present the sentiment variation over time inside a news website, and compare sentiment correlation between news websites. In particular, the system adopts three-dimension sentiments, that are more suitable for the analysis of news articles than the conventional positive-negative sentiments. The experimental evaluations show the accuracy of sentiment extraction is good, and the observation results show sentiment comparison is effective.

Keywords: sentiment analysis, news analysis, correlation analysis

1. Introduction

Recently, an increasing number of portal news websites, such as Google News and Yahoo! News, have been designed to collect and integrate similar news articles from various news websites. These portal websites provide news browsing, keyword search, and various personalized services. People can thus browse and retrieve news articles more easily.

For some domains such as politics and economy, contentious issues continuously arise in news articles. For a contentious news topic, different news websites may have similar or opposite sentiment tendencies. Moreover, a news website may always persist in consistent sentiment, whereas another news websites may show various sentiments over time. Extracting and presenting this kind of background knowledge is significant for news readers to obtain impartial information.

We construct a sentiment dictionary that stores words and their sentiment values. Based on our previous research [1], three dimensions ("Happy ⇔ Sad", "Glad ⇔ Angry", and "Peaceful ⇔ Strained"), that are proved suitable for news articles, are adopted. Using the constructed sentiment dictionary, we develop a system that can detect and visualize the sentiments of news articles and news websites. The system achieves the following functions:

1. Given a news article (target article), the system can extract its topic and its sentiment.
2. The system can identify the news website (target website) of the target article, and present sentiment variation over time inside the target website related to the topic.
3. The system can calculate the sentiment correlation between the target website and other websites, and consequently extract the websites whose sentiment tendencies are similar or dissimilar to the target website.

Table 1. A sample of sentiment dictionary

Word w	s(w) on Happy⇔Sad	s(w) on Glad⇔Angry	s(w) on Peaceful⇔Strained
prize	0.862	1.000	0.808
cooking	1.000	0.653	0.881
deception	0.245	0.075	0.297
death	0.013	0.028	0.000

Table 2. Original sentiment words

Dimension	Original sentiment words
Happy ⇔Sad	Happy, Enjoy, … (OW_L) Sad, Grieve, … (OW_R)
Glad ⇔Angry	Glad, Delightful, … (OW_L) Angry, Infuriate, … (OW_R)
Peaceful ⇔Strained	Peaceful, Mild, … (OW_L) Tense, Eerie, … (OW_R)

The rest of this paper is structured as follows. Section 2 describes the construction of the sentiment dictionary. Section 3 and Section 4 describe the offline processing and online processing of the system respectively. Section 5 evaluates the accuracy of sentiment extraction and shows the prototype of the system. Section 6 reviews related work. Finally, Section 7 concludes the paper and discusses future work.

2. Sentiment dictionary construction

We then construct the sentiment dictionary, in which each entry indicates the correspondence of a target word and its sentiment values on the three dimensions. A sample of the sentiment dictionary is shown in Table 1. A sentiment value $s(w)$ of a word w on each dimension is a value between 0 and 1. The values close to 1 mean the sentiments of the words are close to "Happy", "Glad", or "Peaceful", while the values close to 0 mean the words' sentiments are close to "Sad", "Angry", or "Strained". For example, the sentiment value of the word "prize" on "Happy ⇔ Sad" is 0.862, which means the word "prize" conveys a "Happy" sentiment. The sentiment value of the word "deception" on "Glad ⇔ Angry" is 0.075, which means "deception" conveys an "Angry" sentiment.

For each of the three dimensions, we set two opposite sets (OW_L and OW_R) of original sentiment words (Table 2). The basic idea of sentiment dictionary construction is that a word expressing a left sentiment on a dimension often occurs with the dimension's OW_L, but rarely occurs with its OW_R. For example, the word "prize" expressing the sentiment "Happy" often occurs with the words "Happy", "Enjoy", etc, but rarely occurs with the words "Sad", "Grieve", etc. We compare the co-occurrence of each target word with the two sets of original sentiment words for each dimension by analyzing the news articles published by a Japanese newspaper YOMIURI ONLINE during 2002 - 2006.

First, for each dimension, we extract the set S of news articles including one or more original sentiment words in OW_L or OW_R. Then, for each news article, we count the numbers of the words that are included in OW_L and in OW_R. The news articles, in which there are more words included in OW_L than in OW_R, constitute the set S_L. Inversely, the news articles, in which there are more words included in OW_R than in OW_L, constitute the set S_R. N_L and N_R represent the numbers of the news articles in S_L and S_R respectively. For each word w occurring in the set S, we count the number of news articles including w in S_L and mark it as $N_L(w)$. Similarly, we count and mark the number of news articles including w in S_R as $N_R(w)$. The conditional probabilities are

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

A sentiment value $s(w)$ of a word w is calculated as follows:

$$s(w) = \frac{P_L(w) * weight_L}{P_L(w) * weight_L + P_R(w) * weight_R} : \qquad weight_L = \log_{10} N_L, \quad weight_R = \log_{10} N_R$$

3. System's offline processing

We implement a news crawler for collecting news articles on my own. News articles are crawled from 25 specified news websites (15 newspapers published in Japan and 10 newspapers' Japanese versions in other countries) every day. Then, the articles are morphologically analyzed to extract proper nouns, general nouns, adjectives, and verbs. The *tf · idf* values of each extracted word in a news article are calculated.

The sentiment value of a news article is also calculated by looking up the sentiment values of the words extracted from it from the sentiment dictionary and averaging them. A news article can obtain a sentiment value ranging from 0 to 1. Considering the comprehensibility and the symmetry, the calculation value is further converted to a value ranging from -3 to 3 by the formula: *conversion value* = $6 * calculation\ value - 3$. When the calculation values are 1, 0.5, and 0, the corresponding conversion values become 3, 0 and -3. The conversion values 3, 2, 1, 0, -1, -2, -3 on a dimension, e.g., "Happy ⇔ Sad", correspond to "Happy", "Relatively happy", "A little happy", "Neutral", "A little sad", "Relatively sad" and "Sad", respectively.

The above processing is done offline. As a result, the collected news articles, the *tf ·idf* values of the words extracted from the news articles, and the sentiment values of the news articles are stored in a database.

4. System's online processing

4.1. Extracting the topic and the sentiment of the target article

Given a news article, the system first extracts keywords and sub-keywords representing the article's topic. The keywords are the top five words with the highest *tf ·idf* values extracted from the target article. The sub-keywords are the top five words with the highest sums of *tf · idf* values in the related articles that include any of the five keywords. Both the five keywords and the five sub-keywords are presented to the user. The user selects the words representing the topic that he or she has concern about. The selected words are later used to retrieve past articles for analyzing sentiment tendencies of news websites related to the concerned topic. The sentiment value of the target article is also calculated by using the sentiment dictionary and converted to a conversion value. The conversion values of the target article on the three dimensions are also presented to the user.

4.2. Presenting sentiment variation inside the target website

The news website of the target article is identified by analyzing the URL of the article. Figure 1 is an example of the sentiment variation over time inside the target website. The news articles including the user-selected words in the target website date back at a regular interval t_i (e.g., one day or two days). At each interval t_i, the articles, on which the *tf · idf* values of the user-selected words are larger than a threshold τ_0, are extracted. Their sentiment values are calculated, converted and averaged as the sentiment values $s(t_i)$ of the target website at the interval t_i. The real horizontal line represents the mean of sentiment values and the dotted horizontal lines represent the standard deviation of sentiment values. By browsing the sentiment variation inside the target website, users can perceive whether the sentiment of the current article (the red point) is consistent with the past sentiments of the target website.

4.3. Presenting sentiment correlation between websites

Another function of the system is to show the correlation of sentiment tendencies between the target website and its counterpart websites (Figure 2). Let $s_X(t_i)$ and $s_Y(t_i)$ be the sentiment values of two websites X and Y at the interval t_i respectively, and we calculate their correlation coefficient $\rho(X, Y)$ as follows:

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(s_X(t_i) - \overline{s_X}) * (s_Y(t_i) - \overline{s_Y})}{\sqrt{\sum_{i=1}^{n}(s_X(t_i) - \overline{s_X})^2} * \sqrt{\sum_{i=1}^{n}(s_Y(t_i) - \overline{s_Y})^2}} : \qquad \overline{s_X} = \frac{\sum_{i=1}^{n} s_X(t_i)}{n}, \quad \overline{s_Y} = \frac{\sum_{i=1}^{n} s_Y(t_i)}{n}$$

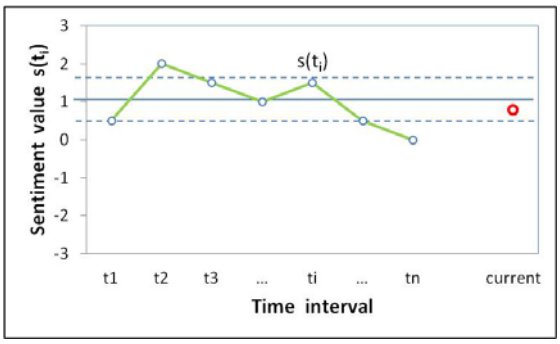

Figure 1. Comparison inside a website

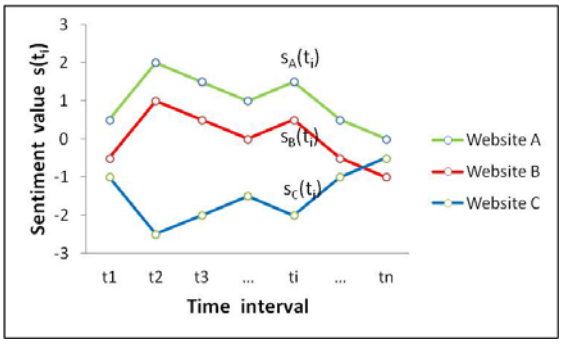

Figure 2. Comparison between websites

Figure 2 shows an extreme example, in which $\rho(A,B)$ is 1 (direct correlation) and $\rho(A,C)$ is -1 (inverse correlation). Based on the calculation results of correlation coefficient, the system can extract sentiment-similar websites by selecting the ones, ρ between which and the target website is larger than a threshold τ_1 (e.g., 0.5) and sentiment-dissimilar websites by selecting the ones, ρ between which and the target website is smaller than another threshold τ_2 (e.g., -0.5).

5. Evaluation and observation

5.1. Evaluation on sentiment extraction accuracy

We specify 10 news domains (Society, Sports, Economy, Synthesis, Politics, Overseas, Life, Entertainment, Culture, and Science) and pick up 10 news articles from each domain. As a result, 100 news articles are selected for evaluating sentiment extraction accuracy. 5 testees are asked to read each of 100 news articles and evaluate how intensely they feel the sentiments on the three dimensions. Each testee can evaluate the sentiment intensity by giving an integer from -3 to 3. For example, for the dimension "Happy ⇔ Sad", 3, 2, 1, 0, -1, -2, -3 represent "Happy", "Relatively happy", "A little happy", "Neutral", "A little sad", "Relatively sad", and "Sad", respectively, The evaluation values from 5 testees for each article and each dimension are averaged as the mean value of the article's sentiment on that dimension. For each of 100 news articles, the conversion value of the sentiment on each dimension is also calculated by using the sentiment dictionary and the conversion formula.

We compare the conversion values (computer's output) of 100 articles with the mean values (testees' evaluation). The average errors between the conversion values and the mean values on "Happy ⇔ Sad", "Glad ⇔ Angry", and "Peaceful ⇔ Strained" are 0.748, 0.746, and 1.128, respectively. Considering the sentiment values have seven levels ranging from -3 to 3, the error of about one level on each dimension indicates that our sentiment extraction accuracy is good.

5.2. Observation on sentiment tendencies

We implement a prototype that extracts the sentiment of a news article, the sentiment variation inside a website, and the sentiment tendencies' correlation between different websites. An example is that a user is browsing a news article reporting the draft of tax increase for the revival of Japanese earthquake. The system first extracts the keywords and the sub-keywords representing the topic and the sentiment of the news article (Figure 3).

After the user selects the concerned words (e.g., "tax increase" and "revival"), the system identifies the website of the article is "NIKKEI" and analyzes the sentiment variation related to "tax increase for the revival" inside the website "NIKKEI" (Figure 4). The overall sentiment tendency about the topic in this website is relatively "Sad", and the sentiment of the current target article (the red point) keeps within the sentiment variation range of the website.

Figure 5 shows the system detects the sentiment-similar website "Asahi" and the sentiment-dissimilar website "Mainichi" on "Happy ⇔ Sad" related to the topic "tax increase for the revival" for the target

website “NIKKEI”. The graph shows the comparison results of sentiment correlation between those websites. The green line represents the sentiment tendency of the target website “NIKKEI”, the blue line represents the sentiment tendency of its sentiment-similar website “Asahi”, and the red line represents the sentiment tendency of its sentiment-dissimilar website “Mainichi”. From the graph, the user can exactly observe that “Asahi” has similar tendencies to “NIKKEI” while “Mainichi” has opposite tendencies to “NIKKEI”.

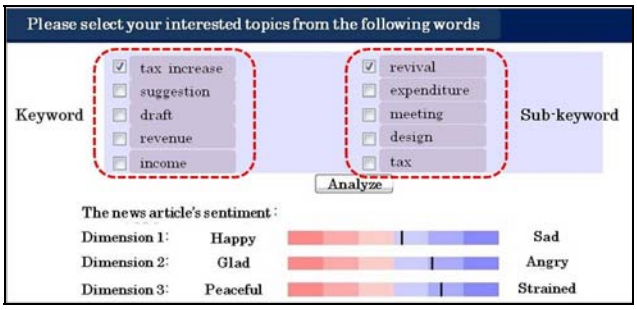

Figure 3. Snapshot of topic and sentiment extraction

Figure 4. Snapshot of sentiment variation inside a website

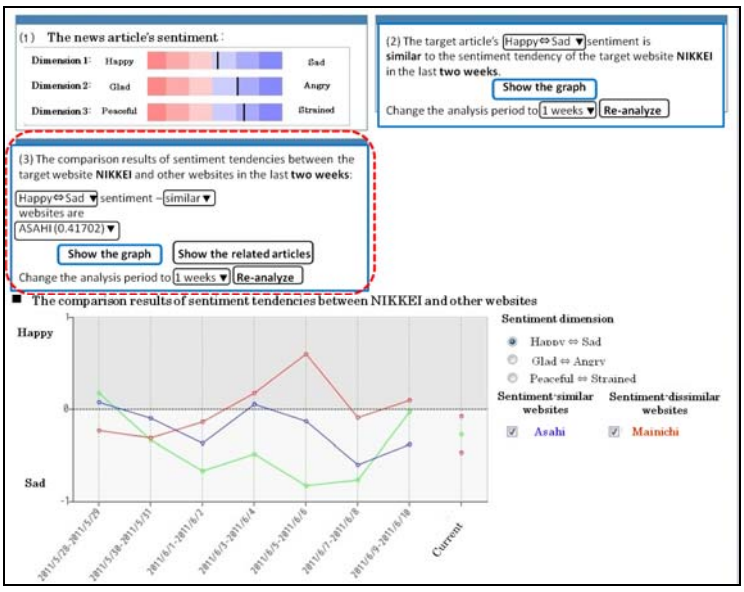

Figure 5. Snapshot of sentiment correlation between websites

6. Related work

Sentiment analysis [2, 3] is increasingly important in many research areas. Turney [4] proposed a method for classifying reviews into two categories: recommended and not recommended based on mutual information. Pang et al [5] extracted only the subjective portions of movie reviews and classified them as "thumbs up" or "thumbs down" by applying text-categorization techniques. However, these methods only consider positive-negative sentiment. Unlike these methods, our system captures more detailed sentiments of three dimensions suitable for news articles.

There also exist other several sentiment models. Plutchik [6] designed a four-dimension model: "Joy ⇔ Sadness", "Acceptance ⇔ Disgust", "Anticipation ⇔ Surprise", and "Fear ⇔ Anger". Russell [7] proposed a two-dimensional space where the horizontal dimension was pleasure-displeasure, and the vertical dimension was arousal-sleep. The remaining four variables: "excitement", "depression", "contentment", "distress", were their combination, not forming independent dimensions. We adopt three-dimension sentiments for news articles. Park et al [8] proposed an aspect-level news browsing system that aimed to mitigate news bias. Different from their works that only deal with recent news articles, our system also analyzes the past articles for extracting the sentiment tendencies of websites over time.

7. Conclusion and future work

We described a system for extracting a news article's sentiment, finding the sentiment variation inside a news website, and comparing sentiment tendencies between different websites. Our implementation enabled users to obtain visual comparison results.

We plan to construct a model for evaluating news articles' credibility based on the sentiment comparison results described in this paper. An idea is that "if news websites with different sentiment tendencies come to hold the same sentiment, the information may be credible". Developing a method for calculating credibility scores of news articles is one of our future challenge.

8. Acknowledgements

This work was supported in part by SCOPE (Ministry of International Affairs and Communications, Japan) and by the MEXT Grant-in-Aid for Young Scientists (B) (#21700120, Representative: Yukiko Kawai).

9. References

[1] T. Kumamoto, "Design of Impression Scales for Accessing Impressions of News Articles," In *SNSMW 2010*, pp. 285-295, 2010.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1-2, pp. 1–135, 2007.

[3] A. Wright, "Our Sentiments, Exactly," *Communications of the ACM*, Vol. 52, No. 4, pp. 14–15, 2009.

[4] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," In *ACL 2002*, pp. 417–424, 2002.

[5] B. Pang and L. Lee, "A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," In *ACL 2004*, pp. 271–278, 2004.

[6] R. Plutchik, "*The Emotions*," Univ Pr of Amer, 1991.

[7] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161–1178, 1980.

[8] S. Park, S. Lee and J. Song, "Aspect-level News Browsing: Understanding News Events from Multiple Viewpoints," In *IUI 2010*, pp. 41-50, 2010.