Interactive Visualization of the Polarity-based Stance of News Websites using News Genres

Masaharu Yoshioka[1], Norihiko Tatsunami[2], Masahiko Itoh[3], Noriko Kando[4] and James Allan[5]

[1]Hokkaido University, N14 W9, Kita-ku, Sapporo-shi, Hokkaido, Japan
[2]Fujitsu Japan Ltd., Shiodome City Center : Higashi Shimbashi 1-5-2, Minato-ku,Tokyo, Japan; This work was conducted during his study at Hokkaido University.
[3]Hokkaido Information University, Nishi Nopporo 59-2, Ebetsu, Hokkaido, Japan
[4]National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo Japan
[5]University of Massachusetts, Amherst, 140 Governors Drive, Amherst, Massachusetts, United States of America

Abstract

We have been working on a project that aims to characterize the stance of news websites. We are proposing a framework that can classify websites in terms of the relative proportions of positive and negative articles about particular topics. This system can represent the website's general stance for easily comparable topics such as presidential election but it is more difficult to analyze the news website from the viewpoint of its users' underlying interests. In this paper, we propose a system for the interactive visualization of a website's polarity-based stance by representing the users' interests in terms of news genres and demonstrate the system using articles refereed to Donald Trump in 2016 US presidential election periods.

Keywords

news, polarity, stance, interactive visualization, visualization of multidimensional data, news genre

1. Introduction

Nowadays, we can access a wide variety of opinions from Internet-based sources, including news articles, messages on blog websites, and other review websites. To be able to comprehend the spectrum of such opinions, it is helpful to understand the *stance* of such opinion sources [1]. Recently, there are several websites that discuss the media bias of news websites, such as Media Bias Chart [1] and Media Bias/Fact Check (MBFC) [2]. Since those pages are maintained by the human experts, it is good to have a system to analyze such bias for analyzing wide varieties of news websites. In previous work [2], we considered the polarity-based stance of articles from news websites about a particular topic. This approach could handle substantial amounts of

INRA'21: 9th International Workshop on News Recommendation and Analytics, September 25, 2021, Amsterdam, Netherlands
yoshioka@ist.hokudai.ac.jp (M. Yoshioka); tatsunami.norih@fujitsu.com (N. Tatsunami); imash@do-johodai.ac.jp (M. Itoh); kando@nii.ac.jp (N. Kando); allan@cs.umass.edu (J. Allan)
0000-0002-2096-1218 (M. Yoshioka); 0000-0002-7759-392X (M. Itoh); 0000-0002-2133-0215 (N. Kando)

[1]https://www.adfontesmedia.com
[2]https://mediabiasfactcheck.com/

relevant textual material and was therefore well suited to obtaining a macroscopic view of the stance for a particular topic in terms of a polarity-based score (the ratio of positive to negative articles for a particular topic). However, this framework was not able to discriminate between news websites that had different underlying interests.

To visualize the stance with respect to underlying interests, we propose a system for representing news websites using a *stance matrix* that combines *polarity* (positive, neutral, and negative) and *news genre* (e.g., politics or economy) for these multidimensional data. To represent users' interests interactively, several methods have been proposed that are based on modifications to a distance metric [3, 4, 5]. In this paper, we report on an application of a recently developed interactive visualization tool called "symmetric interactive representations in a unified system" (SIRIUS) [5]. This application addresses real-world problems, including the visualization of the polarity-based stance of various news websites [2].

In the remainder of this paper, Section 2 introduces our method for assessing polarity-based news stances and the interactive information visualization tool SIRIUS [5]. Section 3 describes our approach to the polarity-based stance of a news website and proposes an extension that uses news genres to represent the news-website stance in terms of multidimensional data. Section 4 presents a method for constructing news-website-stance data and demonstrates news-website visualization using SIRIUS. Section 5 concludes the paper.

2. Polarity-based Stance of News Websites and SIRIUS

2.1. Polarity-based Stance of News Websites

Currently, many users use news aggregation websites that show compiled lists of news articles from various sources. If the suggestions are based on topics related to the user's interests, those users will be exposed to news from more diverse sources, which will involve a wider array of political stances. Users must, therefore, use their judgment selectively to digest what they read, especially for controversial topics, often by assessing the source's stance.

To characterize the stance of news websites, we have already proposed a framework [2] that represents the bias of news websites toward a particular topic in terms of a vector. This *Polarity-based Stance* vector expresses the proportions of positive, neutral, and negative articles for a particular topic. Comparing the *Polarity-based Stance* for various topics is useful for characterizing the stance of a news website (e.g., being for or against a particular presidential candidate).

2.2. SIRIUS: Interactive Visualization Tool for Multidimensional Data

SIRIUS [5] is a recently developed interactive visualization tool for multidimensional data. SIRIUS uses multidimensional scaling (MDS) as the basic tool for visualization and proposes a framework for visualizing data using symmetric duality in the multidimensional data. MDS characterizes the data by calculating the *distances* between data items. In addition to giving the MDS results for the data, SIRIUS can visualize the attributes (conceptual names for the values in each dimension) used by MDS and the transposed matrix used to calculate distances among the data items. SIRIUS implements interactive visualization in terms of weighted MDS (WMDS) [6].

WMDS uses the weighted Euclidean distance between the i-th and j-th data item (wd_{ij}) instead of the standard Euclidean distance. The weighted Euclidean distance is calculated as

$$wd_{ij} = \sum_{k=1}^{n} w_k(d_{ik} - d_{jk}), \tag{1}$$

where n, w_k, and d_{ik} denote the dimensionality, the weight value for the k-th dimension of weight vector $\boldsymbol{w}$, and the k-th attribute value of the i-th data item, respectively.

WMDS can represent the multidimensional data based on users' interests ($\boldsymbol{w}$) by using the weighted Euclidean distance instead of the standard Euclidean distance.

SIRIUS offers a framework for estimating the value of the weight vector $\boldsymbol{w}$ based on interaction with the visualization results. The user can reorganize the data to represent better the user's interests. SIRIUS updates the weight vectors using WMDS, aiming to minimize the sum of the differences between distances in the reorganized data and those in the new visualization data.

Feedback from the system is given via visualization of the attributes. Users can understand their interests through the weight values assigned for each dimension and the scatterplots of the data given by WMDS.

3. Visualization of the Polarity-based Stance Matrix using News Genres

3.1. News-genre Polarity-based Stance Matrix

To understand better the characteristics of a news website, it is useful to know its degree of interest in various topics. We assume that its interests can be characterized in terms of news genres. For example, a news website could be positive toward a topic with respect to its economics aspects, but negative with respect to its political aspects. Therefore, we propose a novel representation for the stance of news websites using both polarity-based scores and news genres in terms of a news-genre polarity-based stance matrix (NPSM):

$$\mathrm{NPSM} = \begin{pmatrix} p_1 & p_2 & \cdots & p_m \\ \mathrm{neu}_1 & \mathrm{neu}_2 & \cdots & \mathrm{neu}_m \\ n_1 & n_2 & \cdots & n_m \end{pmatrix}$$

.

Here, m is the number of news genres considered, and p_i, neu_i, and n_i correspond to the proportions of positive, neutral, and negative articles, respectively, for news genre i among all available articles on the topic. The reason for including the proportion of neutral articles, which was not used in the previous definition, is that this value is useful in characterizing the news websites' degree of interest in the topic. This matrix for each news website is converted to $3 \times m$ dimension vector for calculating distance between other news website.

3.2. Data Construction

3.2.1. Assigning a news genre to each article

To analyze news websites using *NPSM*, it is necessary to select a list of news genres. For this purpose, we analyzed several well-known news websites, namely `Los Angeles Times`, `Washington Post`, `Business Insider`, and `Yahoo!News`, that have news genres. Based on this analysis, we selected the following seven top-level genres as a candidate list of news genres for this analysis (i.e., $m = 7$): politics (pol), domestic (dom), international (int), economy (eco), science (sci), lifestyle (lif), and entertainment and sports (e/s) for analyzing the political stance of a news website. In the real data, there are articles that are assigned two or more news genres. However, in this research, we assign one news genre for all articles to make this assignment task simpler. It is necessary to analyze the effect of this setting for the future works of our system.

A possible problem in collecting data for each news website is that insufficient information may be available to estimate the genres for its news articles. If a news website organizes its articles using the genres proposed in this paper, there is no problem. However, there are news websites that use different names for news genres and others that omit genre information. For the former case, we manually construct rules that map the news genres used by the news website into our seven news genres. For the latter case, it is necessary to develop a system for assigning news genre information about the news articles.

In this work, we planned to use the GDELT[3] database, which contains information about the list of named entities mentioned in the paper, polarity information, and universal resource locators (URLs). We use the URLs to estimate the news genre because they often contain information related to the news genre or title of the article and propose a news genre classification tool that uses the deep-learning-based natural-language-processing tool BERT [7]. BERT is an infrastructure for creating contextual text-analysis tools via training with a big-data corpus for general contextual tasks. For the training data, we used URLs for the news websites that were properly categorized in terms of our seven news genres and URLs for news websites whose news genre was assigned manually ("gold standard" news genre data).

3.2.2. Construction of news genre polarity-based stance data

For each news article, a polarity type (positive, neutral, or negative) is assigned via the method proposed in [2]. In this method, the "tone score" t, which is a score for representing the positive contents minus the negative contents and provided by GDELT, is used for this classification. That is, positive is represented by $t > 1$, negative by $t < -1$, and neutral otherwise).

As a result, all news articles can be classified in terms of news genre and polarity. The system then constructs *NPSM* by calculating the ratios (seven news genres × three polarities) for each news website. After constructing the 7×3 matrix, we translate this matrix to 21-dimensional vectors for input to SIRIUS.

[3] https://www.gdeltproject.org/

4. Experiments

4.1. Data Construction using GDELT

Following [2], we collected article data from GDELT for news websites listed in the Media Bias Chart 5.1[4], which provided information about bias and quality aspects of news, mainly for US news websites. Examples of the type of information about an article included whether the content of the article related to presidential candidates Donald Trump or Hillary Clinton, the date of distribution of the article, its URL, and the value of its polarity (calculated by GDELT from words used in the body of the article). The data collected were for 77 news websites during the period from September 1, 2016 to November 31, 2016.

4.2. News Genre prediction using BERT

We constructed a news genre classification system, as discussed in Section 3.2.1, using the collected data. For URL data, we used their final element (e.g., `aaa.html` in `https://www.aa.com/bb/aaa.html`) for input into the system. The pretrained model for BERT was the 12-head uncased model released by Google in 2018 [5] and we used a 12-layer, 768-dimensional hidden layer for fine tuning. We split the data collected as "gold standard" news genre data into a training set, a development set, and a test set at ratios of 6:2:2, respectively, and the task was to estimate the news genre based on the probabilities predicted for each news genre. In this system, the highest-probability news genre was assigned as the news genre of the article. The results of the predictions for the test data are shown in Table1.

Table 1
Performance in Predicting Article News Genre for Test Data

News Genre	Precision	Recall	F1-score	Support
pol	0.96	0.86	0.91	8,304
dom	0.74	0.79	0.76	2,970
int	0.90	0.87	0.88	3,286
eco	0.76	0.86	0.81	2,833
sci	0.86	0.87	0.87	3,147
lif	0.84	0.93	0.88	2,882
e/s	0.88	0.89	0.88	3,169
Accuracy			0.87	26,591

This system was used to assign a news genre to news articles whose URLs did not contain news genre information. Table 2 gives the numbers of articles related to Donald Trump used for the data analysis.

Table 2
Number of Articles by Classification Method

Classified manually	24,564
Classified by BERT	120,440
Total	145,004

[4]https://www.adfontesmedia.com
[5]https://github.com/google-research/bert

4.3. Data Analysis using SIRIUS

Using our collected news-website information, we aimed to analyze the polarity-based stance of news websites toward Donald Trump during the presidential election period, September 1, 2016 to November 31, 2016. Right-upper side of Figure 1 shows an initial plot of the news websites (gray background) and attributes (blue background) using SIRIUS. For the news websites plot, initial plot is calculated by ordinal MDS (weights for all dimensions are same). Red and blue colored nodes shows the right-wing and the left-wing news websites defined by the information in Media Bias Chart 5.1 respectively. Attributes plot shows the results similarity among parameters. The radius of the node shows weight of the attribute. Since distance between two attributes are calculated by using attribute-news website matrix by transposing news website-attributes matrix, attributes whose distribution are different from others are plotted far from the middle.

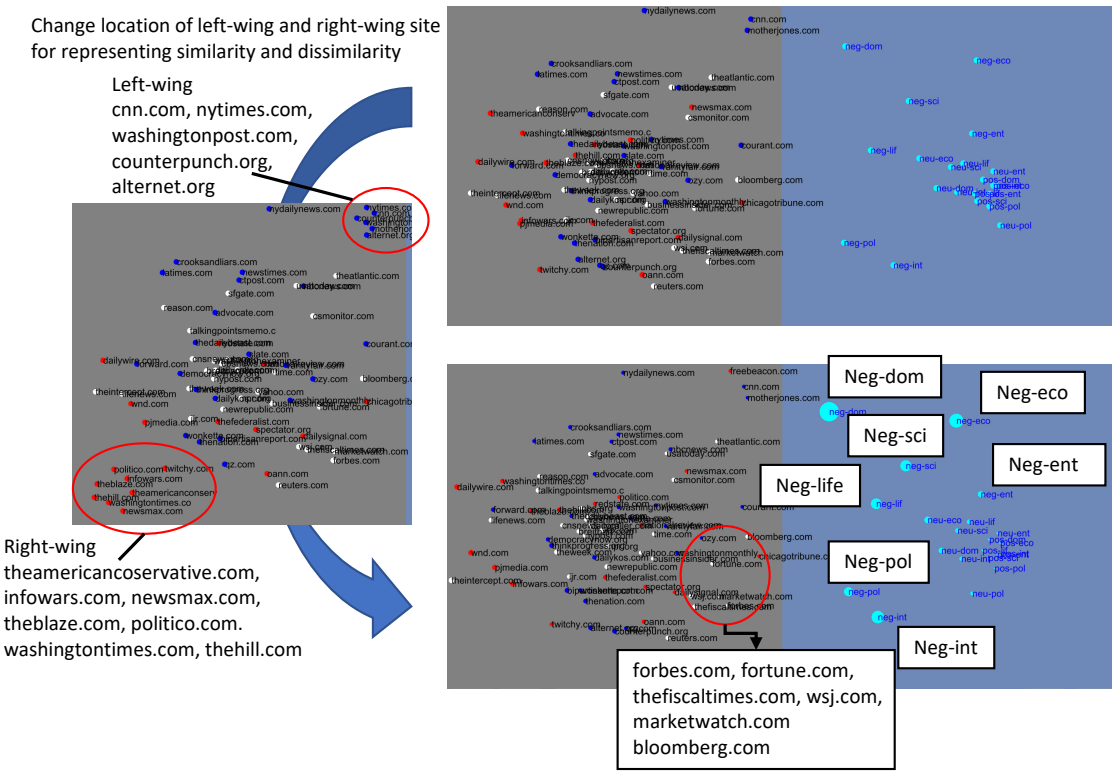

Figure 1: Interactive updating by moving news-website nodes

From this initial plots, the user change the position of some right-wing news websites (`theamericanconservative`, `newsmax`, `theblaze`, `washingtontimes`, `infowars`, `politico`, `thehill`) and left-wing news websites (`cnn`, `nytimes`, `washingtonpost`, `alternet`, `counterpunch`) to the corner for reprsenting similarity and dissimilarity (left side of Figure 1). SIRIUS update the weight of attribute based on this node location and update the results (right bottom of Figure 1). In this case, most of the negative attributes are plotted far from the middle. The interpretation is

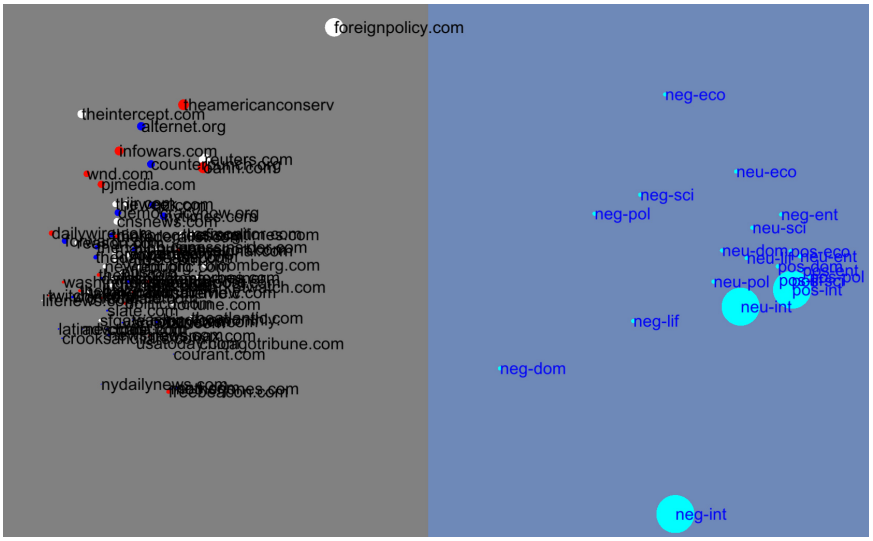

Figure 2: Result of replotting with the weight of the international (int) topic increased

that ratio of negative articles of various news genres can be used to characterize the differences among news websites using wing specific information. Note that neutral financial news websites such as `Forbes` and `Fortune` are gathered together at the bottom.

Another possible operation is to control the weight. Figure 2 illustrates the result of updating the plot by putting an emphasis on "international" (int) (the positive, negative, and neutral weights for int were 0.2133 and were 0.02 for others). In this figure, note that the node for `foreignpolicy.com` is large and plotted at a position distant from other objects. This is because this news website focuses on issues related to foreign policy (i.e., related to "international"), whereas news websites that are less interested in international issues are gathered in the middle.

4.4. Discussion

Based on these examples, we can confirm that the interactive visualization of news websites using SIRIUS promises to represent the characteristics of news-stance differences based on the news genres. This reflects those news websites may have different interests on issues classified by news genre. Interactive operation to such multiple dimension data may be useful to analyze similarity with multiple aspects. For example, feedback from the system as attribute information is helpful in understanding better the replotted results (Figure 1) and direct operation to the weight change can generate a visualization result that focused on particular news genre (Figure 2). However, it is not always easy to interpret the final results in depth.

5. Conclusions and Future Work

In this paper, we propose a system to evaluate the stance of news websites using interactive multidimensional data-visualization tools. In particular, we discuss the characteristics of the interactive visualization tool SIRIUS. Because this is the first attempt at visualizing political

stances using news genres, it will be necessary to apply this system to other examples and obtain feedback from real users in its evaluation.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 18H0333808.

References

[1] D. Küçük, F. Can, Stance detection: Concepts, approaches, resources, and outstanding issues, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2673–2676. URL: https://doi.org/10.1145/3404835.3462815. doi:10.1145/3404835.3462815.

[2] M. Yoshioka, M. Jang, J. Allan, N. Kando, Visualizing polarity-based stances of news websites, in: Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018., 2018, pp. 6–8. URL: http://ceur-ws.org/Vol-2079/paper2.pdf.

[3] E. T. Brown, J. Liu, C. E. Brodley, R. Chang, Dis-function: Learning distance functions interactively, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, IEEE, 2012, pp. 83–92.

[4] M. Yoshioka, M. Itoh, M. Sebag, Interactive metric learning-based visual data exploration: Application to the visualization of a scientific social network, in: E. Grant, D. Kotzinos, D. Laurent, N. Spyratos, Y. Tanaka (Eds.), Information Search, Integration, and Personalization, Springer International Publishing, Cham, 2016, pp. 142–156.

[5] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, C. North, Sirius: Dual, symmetric, interactive dimension reductions, IEEE Transactions on Visualization and Computer Graphics 25 (2019) 172–182. doi:10.1109/TVCG.2018.2865047.

[6] A. Endert, C. Han, D. Maiti, L. House, S. Leman, C. North, Observation-level interaction with statistical models for visual analytics, in: 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), 2011, pp. 121–130.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: http://arxiv.org/abs/1810.04805, cite arxiv:1810.04805Comment: 13 pages.