NewsTrust Xplorer: Content-based Rating Visualization for Exploring News Reviews and Reviewers

Hyunjong Cho  
Department of Computer Science  
University of Maryland  
cho@cs.umd.edu

Chanhyun Kang  
Department of Computer Science  
University of Maryland  
chanhyun@cs.umd.edu

Uran Oh  
Department of Computer Science  
University of Maryland  
uranoh@cs.umd.edu

Nishant S. Patel  
Department of Computer Science  
University of Maryland  
nishant@umd.edu

Abstract—NewsTrust is a non-partisan, nonprofit news service that uses a volunteer community of reviewers to rate the latest news articles on the web. The reviews consist of ratings in 18 different categories, including “facts”, “trust”, “fairness”, and “recommendation”. We obtained over 30,000 reviews spanning more than one year, covering over 15,000 unique news articles. We downloaded every full article text, and used detailed demographic data on each reviewer. Our tool, NewsTrust Xplorer, helps users to interactively discover and analyze correlations between keywords found in articles, review ratings, and the demographics of the reviewers.

Index Terms—Journalism, Information Visualization, News article rating, Trust Network, NewsTrust, news reviews

1. INTRODUCTION

In today’s world, online news media is an essential part of many people’s lives. The quality and content of online articles vary widely, especially considering the difference between traditional print media (New York Times, Washington Post), blogs, and content aggregators such as the Huffington Post. NewsTrust is a nonprofit, non-partisan project created to provide “a wide range of tools that empower citizens to access quality news and information - and learn to separate fact from fiction about important public issues”[1]. NewsTrust uses a community of 20,000 volunteer “reviewers” that read and critique online journalistic articles from a variety of sources, in order to help the public “find and share good journalism online”. The reviewers consist of general news readers, journalists, educators, and students. Each article is rated in a variety of categories, including accuracy, fairness, sourcing, context, and other journalistic criteria. In addition, the reviewers themselves are rated by other reviewers, which in turn affect the importance and power of the critiques. This data used for this study is supported by NewsTrust and further detailed below.

To visualize this enormous amount of highly specialized data, we decided to write a custom software program that would (1) Allow for dynamic navigation of tens of thousands of high dimensional records, (2) Allow users to analyze properties of the data, and discover the insights within the specific domain in depth as an valuable tool for research based on the data. Additionally, the tool should support users to import various datasets and save and the desirable results.

The rest of this paper is organized as follows: Section 2 describes past and related work. Section 3 explains the types of data we use. The content-based article visualization and the demographic-based review rating visualization are presented in section 4. Section 5 summarizes our findings from usability evaluations, which is followed by the Conclusion and Future Work.

2. RELATED WORK

The issue of news media’s credibility has interested scholars and journalists for many years. Many have focused on whether the news media possess political bias. Conservative critics argue that journalists tend to be liberal Democrats. Critics from
the left argue the opposite. To investigate such conflicting arguments, media scholars have examined partisanship among journalists and their stands on social issues.

A recent study by T. Lee [2] took an alternative approach at examining the media bias debate. Findings from that study suggest that perceptions of media bias are, in fact, caused by the audience’s own partisan or ideological biases.

Another study in media bias is the research of Chia [3], which examines the “hostile media effect” in relation to partisans’ perception of the slant of news coverage in a highly regulated press environment: Singapore. They found that partisans in Singapore perceived unbiased news to be in favor of the other side, while the nonpartisan perceived the same news to be neutral. Their findings show that hostile media effects can persist in a restricted press environment where people are aware of the government’s control of media coverage. They also found that partisans’ awareness of the government’s control of media information contributed to their perception of the article slant as well.

Andrew et al. [4] collected data from 574 participants to assess perceptions of message, website, and sponsor credibility across four genres of websites. His work was able to explore the extent and effects of verifying web-based information. He had success in measuring the relative influence of sponsor familiarity and site attributes on perceived credibility. The results show that perceptions of credibility differed - news organization websites were rated highest and personal websites lowest, in terms of message, sponsor, and overall site credibility. E-commerce and special interests sites were rated in between for the majority of cases. The results also indicated that credibility assessments appear to be primarily due to website attributes such as design features, depth of content, and site complexity rather than to familiarity with website sponsors. Finally, there was a negative relationship between self-reported and observed information verification behavior and a positive relationship between self-reported verification and internet/web experience. The findings are used to inform the theoretical development of perceived web credibility.

Gunther[5] studied the relationship between the news source and the media credibility. This research has discovered that trust in media can be better understood as a relational variable – an audience response to media content. The author insists that distrust is more likely to be a situational response, stemming from involvement with issues and groups. Using data from a national probability sample, the hypothesis was tested by analyzing the effect of numerous independent variables on respondent ratings of newspaper and television news coverage of social groups. As hypothesized, a respondent's own group identification proved to be the strongest predictor.

Several studies have been conducted about the media credibility related to audience demographics. Some researches[6-8] have identified individual differences of receivers such as age, education, gender, and knowledge about the media and the topic could contribute to the evaluation of source credibility. Ganiano and McGrath [9] conducted another type of audience-oriented studies that have developed person typologies. They used credibility of media scores to classify respondents as high, medium, or low in credibility evaluation.

There are also some works on news article visualization. For example, Tweetcatcha[10], an application that utilizes the NY Times Timeswire API and Twitter, visualizes tweets about NYT articles from the past 24 hours. There are 24 rings of tweets, where each ring represents each hour in a day. The closer a tweet is to the center of the graphic, the faster a specific article was tweeted about. It provides an interesting overview about the tweet-popularity of an article and the rapidness of its distribution.

Newsmap[11] is another online news article visualization. It collects articles from the Google News Aggregator to show which are the “hottest”. The larger the size of an article displayed, the broader the coverage among the other articles. One can also search for specific keywords or look at individual geographic regions.

3. DATA DESCRIPTION

The principal data consists of data points from NewsTrust from September 1st, 2010 to February 28th, 2011. There are three particular partitions: story information, review information, and reviewer information. In addition, we collect and process the full article texts from the news story of each review. The data was totally ammonished by NewsTrust prior to receipt, so there is no way to connect any review or reviewer to a specific individual.

3.1 Story Data

For each news article, we call it a story, we have the date of posting, the URL, the overall story rating, and the list of Review IDs for each reviewer of that story. We also have a Boolean field marking whether it is related to Baltimore, MD, used in a separate study. IDs who wrote reviews on the story. The overall story rating is the average rating that the
3.2 Review Data

Each review has the following associated information: The URL of the article (a foreign key to the story information), a unique reviewer ID, the date of the review, and the overall rating. There are also 18 different rating parameters: accuracy, balance, context, depth, enterprise, expertise, facts, fairness, information, insight, originality, recommendation, relevance, responsibility, sourcing, style, transparency, and trust. Each of these parameters is ranging from 0 to 5 and may not all be in present. Reviewers give reviews on the story by giving it numerical scores for each review parameters. Based on the context of the article, and demographics of the reviewers, the scores can easily be biased.

3.3 Reviewer Data

The reviewer data portion contains the review ID, a “member validation level” from 1 to 5, a member rating” from 1 to 5 (based on other’s option of this reviewer), an age group, primary language, journalism experience (0 to 6), education level, gender, income level, political leaning (left, right, etc), and the city and state of residence.

3.4 Full Article Texts

As described previously, one of the primary sources of data for NewsTrust Xplorer is the full text of every article in the review database. From the approximately 30,000 reviews, there were about 15,000 unique articles. Each article is referenced by its unique URL. We downloaded all the articles from the URLs and parsed them into index to retrieve keywords. Text preprocessing was then done through a Java program: the text was lowercased, and all punctuation, numbers, symbols, words with less than 3 letters were removed. We also removed approximately 200 stopwords.

After preprocessing, all the words were indexed for fast future searches. This involved building an enormous lookup table of every unique word to the complete list of articles each one could be found in, and the number of occurrences in each article. For example, querying the index for “madonna” would return a list of URLs that contain that word, along with the number of times “madonna” appears at each of those URLs. There were approximately 460,000 unique words. Any word that appeared in more than 5,000 different articles was not indexed at all.

The index was created by a separate Java
Figure 2(a). The article search results of “Obama”

Figure 2(b). Changed x-axis to the number of reviews and the y-axis to story rating. Most articles have only one review and there is one article on the right side that has 49 reviews.

Figure 2(c). Changed the color for the articles that have high story ratings.

Figure 2(d). Minimize the size for the articles

Figure 2(e). Panned and zoomed x-axis from (a)

Figure 2(f). Panned and zoomed y-axis from (a)
program and took several minutes to run. The index was then saved to disk so the primary program could load it without parsing every article text again. Loading the index takes about 30 seconds. The index is stored as a HashMap, so searches are done in constant time, and in our experience only takes a fraction of a second.

4. NEWSTRUST XPLORER

NewsTrust Xplorer is designed to explore not only trends in news article content and quality over time, but also demographic relations based on any chosen keywords. Our tool is implemented in the Java programming language and utilizes the Prefuse visualization toolkit. It’s platform independent stand-alone software.

Its user interface consists of three parts namely control panel, content-based article visualization and demographic-based review visualization as shown in Fig. 1. Each subsection describes the features of the main components.

4.1. Control Panel

This panel is mainly about filtering the articles. To search for articles that only contain certain keywords, the user can add each word of interest to the keyword list before clicking Search. For example, if one enters the words “great” and “Obama”, only articles which contain both words will appear. Searching for multiple keywords is more helpful than just a single word since it narrows the context.

A user can also filter news articles by selecting only the journalistic source, specific time period, the rating range, whether the story is from Baltimore or not, and the number of occurrences of each keyword in each article. Selecting the sources of the article is very desirable since the reviewers often tend to focus on different review parameters when they are reviewing. To be specific, selecting blogs and selecting mainstream media such as wall street journal will bring different results. Choosing the search space to only Baltimore will search the articles only from Baltimore. This feature was needed since NewsTrust wanted to see how the Baltimore market responded to news stories. Last filtering option, setting the number of occurrences of the keyword, is for searching relevant articles. For example, one can search only for articles from blogs containing the words “climate change” at least 5 times. The “number of occurrences” is especially useful, as it easily filters out articles not relevant to the search. Often, keywords will appear only once or twice in an article about an unrelated subject.

Another feature it contains is a history list. It keeps all the searched keywords as a search log for further re-searches.

4.2 Content-Based Article Visualization

This visualization is not only useful for analyzing the correlations among various attributes of stories but also very powerful for retrieving and refining the stories after filtering. The content-based article visualization is a 2 dimensional scatter plot where each rectangle shaped plot represents a single story as shown Fig 2(a). Users are able to change the size and the colors of the rectangle by selecting different variables as in Fig 2(b), (c) and (d). By plotting the searched result with different x-axis, y-axis, rectangle size, and color variables, users can discover the correlations among 4 attributes which are story rating, date, the number of occurrence and the number of reviewers of the filtered articles results. By default, the x-axis is the date and the y-axis is the rating of the story, the size of the rectangle represents the number of reviews of an article, and the color the number of appearances of the searched keyword in that article. The ability to change the color and the size of the rectangles enable users to visually emphasize the articles that are within a specific range. Fig.2(e) and (f) panning/zooming function with the scroll bars located in both x-axis and the y-axis, namely zoom sliders. With these zoom sliders, users can easily navigate and analysis the searched result dynamically. Whenever the user changes the states of the sliders, it automatically updates the Treemap of the bottom so that users are able to focus only on the demanded articles. For more detail information about the article, users can move their mouse pointers to specific rectangles. Then details such as average story rating, the number of reviewers, the number of keyword occurrence are shown upper side of this visualization. Also, if users want to view the original article then they can simply click the rectangle which directly links to the online news article.

4.3 Demographic-Based Review Visualization

As mentioned above, this visualization displays rating information about the focused articles that are shown in the content-based article visualization as a Treemap. Every time, the set of zoomed article changes, it automatically updates the Treemap. Fig. 4.3(a) shows the structure of our treemap. It shows a powerful overview of the demographic characteristics of the reviewers of the selected articles themselves. Additionally, it shows the overall ratings of the articles by the reviewers. When users want to know the details, then they can simply move their mouse pointer to the rating parameters. The details such as the age group, average rating of the rating parameter and the number of reviewers. The treemap has two-level hierarchy: The top level is grouped by age range,
and the lower level is composed of the rating metrics described in section 3.2, such as “insight” and “originality”. There are several demographic characteristics of reviewers, but we only concentrated on the age group that the reviewers belong since there were no big differences between other attributes. The size of every rectangle is proportional to the number of people in that category. As there are so many rating parameters, we added a highlighting function that brightens the location of that parameter so that users can easily find the specific parameter as shown in Fig. 4.3(b).

4.4 Additional Functions

This tool supports multiple visualizations for comparing different search results. Users are able to start a new visualization with a different keyword search list by adding tabs. Also, users can generate a new dialog to view each of the visualization with larger size, which helps users to analysis the result better. NewsTrust Xplorer also provides result saving functions. Users can save the results either the content-based article visualization or the demographic-based review visualization as an image file.

5. EVALUATION

We conducted the user experience evaluation of our visualization tool, NewsTrust Xplorer. 4 subjects from the department of Journalism were participated in this evaluation who might become our future users. Each subject was given maximum 10 minutes of training, and asked to perform 9 tasks. After experiencing main features of our tool by performing the tasks, we collected their feedbacks and opinions in written or spoken form after the experiment by filling out questionnaires. The goal of this evaluation is discovering the features of our visualization that can mislead the users rather than confirming how well it supports the users in every task.

5.1 Tasks/Questionnaires

User experience evaluation is needed to improve interactive or visual aspects of the tool. During the experiment, we had observed their uses of the tool on each task to measure the performances. We used think-aloud protocol to easily detect what each subject is trying to do to perform each task. The 9 tasks are shown below:

1. Story panel- Overview
   1.1. What is the least recent article?
   1.2. What is the most reviewed article?
   1.3. What article has the highest rating?

2. Keyword search - Search for articles with the keyword “Obama” with frequency 10 since 1/1/2011.
   2.1. What is the most relevant article?
   2.2. What is the relation between the story rating and the number of reviewers?
   2.3. If you click the square, it directly links to the article. What is the headline of that article?
3. Focus on the articles with story rating over 3-5
   3.1. What age group wrote the most number of reviews?
   3.2. What is the highest review attribute that age 50-60 gave?
   3.3. What is the average of that review attribute?

After the experiment, the subjects are asked to fill out questionnaires. Questions from 1 to 4 are in scales from 1 to 9. The lower value indicates their response is closer to “No” and the higher value means their response is closer to “Yes”.

Q1. Did you find NewsTrust Xplorer useful?
Q2. Is keyword searching with filtering easy to use?
Q3. Is content-based article visualization clear to understand?
Q4. Is demographic-based rating visualization clear to understand?
Q5. What are the features you like about NewsTrust Xplorer?
Q6. What are the features that were not clear to use?
Q7. Do you have any suggestions or comments to improve the NewsTrust Xplorer?

5.2 Results

Overall, they were able to complete all the tasks with ease. However, for the first three tasks, it seems that subjects were uncomfortable with finding the article at first since they were overwhelmed by the large number of all the articles. But then they could find the right article by changing x-axis, y-axis and using zoom sliders in content-based article visualization. They were very satisfied with the flexibility of various parameter changes. Also, the subjects were pointing out that there are some ambiguities in labels such as “frequency” in filtering panel and “search” in demographic-based rating panel. So we have changed the label “frequency” to “relevance” since it sets the minimum number of the keywords occurrence in each article for searching: if the searched keywords appeared a lot in an article, then the article is more relevant than the article that has lower number of keyword occurrence. Another change we have made is replacing the label “search” to “parameter search”. The users were confused about having two search boxes in one window and they could not expect what keyword they should search for, or what results they will get with the search box in demographic-based rating visualization.

Their responses on the third question were not that plausible than others. Since the content-based article visualization plots all the searched articles, it was hard to distinguish a single article without zooming. However, they thought our tool is useful and easy to use in general. The features they liked about are setting threshold for relevance in article searching, Treemap which is very intuitive way of representing the difference between all the age groups and saving the visualization as a result. Also, they were thrilled by the direct link on each article on content-based article visualization. The headline or the context of the article information is required to fully understand its ratings, the number of reviews and dominate age group of the reviewers.

6. CONCLUSION AND FUTURE WORK

In this work, we have implemented a visualization tool for navigating and analyzing the news articles and reviews. This tool provides visualization for enormous amount of highly specialized data of NewsTrust. The visualization allows users to explorer the numerous high dimensional review and the articles. Also, it supports the users to analyze characteristics of the reviewers and the article ratings in terms of the content of the articles. This tool will be used for generating and proving the various hypotheses.

7. ACKNOWLEDGEMENT

We would like to thank NewsTrust for sharing their valuable data for academic research, Dr. Yaros for helpful suggestions and guidelines, and Dr. Shneiderman for constant support and suggestions.

REFERENCES
[1] http://newtrust.net/about


CREDITS

Hyunjong Cho:
Implemented the demographic-based rating visualization (Treemap), re-wrote the section 2 in the final paper, recorded the first video, recorded the final video

Chanhyun Kang:
Directed overall implementation, designed data structure and user interface for the tool, implemented the content-based article visualization (zoom sliders, color selecting, changing x/y-axis, sizes), re-wrote the section 1, 4 and 5 in the paper, made presentation slides.

Uran Oh:
Implemented filtering panels, designed and conducted the Evaluation, wrote the first draft paper, recorded the first video, re-wrote evaluation part in the final paper, re-wrote the section 4 in the final, revised all the documentation.

Nishant Patel:
Programed Article download scripts for preprocessing, and indexing, set up Subversion server, wrote the abstract, introduction and section 4.4 of the first paper, revised the final paper, record audio for video, made presentation slides