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

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

Chanhyun Kang
Department of Computer Science
University of Maryland
kang.chanhyun@gmail.com

Uran Oh
Department of Computer Science
University of Maryland
oh.uran@gmail.com

Nishant S. Patel
Department of Computer Science
University of Maryland
nishant.s.patel@gmail.com

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 19 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.

The rest of this paper is organized as follows. Section 2 describes the relate works at present, section 3 explains the four types of data we use, the content-based article visualization and the demographic-based review rating visualization are presented in section 4. The section 5 contains the result of evaluations of NewsTrust Xplorer, and the conclusion and the future work of our study are shown in the last section.

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 [1] took an alternative approach to examining the media bias debate. Findings from that study suggest that perceptions of

\(^1\) http://newstrust.net/about
media bias are, in fact, caused by the audience’s own partisan or ideological biases.

Another study in media bias is the study of Chia et al. [2] 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 nonpartisans 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. [3] collected data from 574 participants were used to assess perceptions of message, site, and sponsor credibility across four genres of websites; to explore the extent and effects of verifying web-based information; and to measure the relative influence of sponsor familiarity and site attributes on perceived credibility. The results show that perceptions of credibility differed, such that news organization websites were rated highest and personal websites lowest, in terms of message, sponsor, and overall site credibility, with e-commerce and special interest sites rated between these, for the most part. The results also indicated that credibility assessments appear to be primarily due to website attributes such as design features, depth of content, 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.

There are various related studies on news article visualization. For example, Tweetcatcha², which is an application developed with NY Times Timeswire API and Twitter, visualizes the tweets from the online news articles on New York Times website that are appeared within latest 24 hours. There are 24 rings where each ring represents every hour in a day, and the tweets are located any of these rings. If the tweet is located closer to the center, then it indicates that it took shorter period of time to tweet the article from the time the article is posted on NY Times. So, it displays the number of tweets of a specific article and the time that each tweet took since that article is posted.

Another application that visualizes news articles is called Newsmap ³, which uses treemap visualization format. It collects numerous news articles from the Google News Aggregator to show which news is the hottest. The larger size of an article indicates its coverage among all the articles. One can also choose region for the source of the articles and search for news article that contain a specific keyword. This news visualization tool is very intuitive so that it’s easy to learn and use.

3. DATA DESCRIPTION

The data are extracted from NewsTrust where it contains every news article reviews from September 1st, 2010 to February 28th, 2011 and it can be partitioned into three types: story, review and reviewer where story indicates the online news article or blog that are reviewed.

3.1 Story Data

For each story, the data have the information of whether the articles is from Baltimore NewsTrust or not, date of the story and the URL, the rating and the list of reviewers’ IDs who wrote reviews on the story. The story rating is the average rating that the story received from various reviewers, and its value is in between 1 and 5.

3.2 Review Data

For every review, the information on reviewer ID and the story URL that the reviewer reviewed, the date of the review was written and its rating. Also, there are up to 19 different metrics that a reviewer gives score on such as accuracy, balance, context, depth, enterprise, expertise, facts, fairness, information, insight, knowledge, originality, recommendation, relevance, responsibility, sourcing, style, transparency, and trust. Each metric is scale from 1 to 5 and could be biased.

3.3 Reviewer Data

The reviewer data contain the reviewer ID, member validation level values from 1 to 5, member rating in between 1 and 5, age group of the member where there the reviewer belongs, the mother language of the reviewer, journalism experience scales from 0 to 6, education level scales from 0 to 5, gender, income, politic of the reviewer, city and state the reviewer lives.

3.4 Full Article Texts

As described previously, one of the primary

² http://a.parsons.edu/~drumb588/tweetcatcha/TweetCatcha.swf

³ http://newsmap.jp/
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 note that the URLs are consistent, meaning the exact URL string is used for every review of that same story, despite non-unique URLs in general use. For example, http://abclocal.go.com/kgo/story?section=7live&id=7772021 and http://abclocal.go.com/kgo/story?section=6live&id=7772021 refer to the same story.

First, a list of every URL in the database received from NewsTrust was created. Then, the GNU utilities “uniq” and “sort” were used to sort the list and remove all duplicates. A bash script was then used to download every article and render the text. The GNU Wget program was used to get the HTML pages, and Martin Bayer’s “html2text” program was used to render the text in the HTML. This would be roughly equivalent to copy-and-pasting the text of the web page after loading it in a browser (not copying the HTML source).

Download and parsing every article took several hours, although this could be done much faster in parallel.

After each article text was downloaded, it was preprocessed through a custom Java program. All the text was lowercased, and all punctuation, numbers, symbols, etc. were removed. All words with less than 3 letters were removed. We also removed approximately 200 “stopwords” (extremely common or useless words like “I”, “if”, “in”, “no”, etc., plus words like “subscribe”, “log”, “help”, and “terms”).

After preprocessing the text, all of 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, plus 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 articles was not indexed at all.

The index was created by a separate Java 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 take a fraction of a second.

4. NEWSTRUST XPLORER

NewsTrust Xplorer is designed to explore not only the trend of contents of news articles over time but also the demographic information when a set of keywords is searched. Our tool is implemented with Java programming language using the Prefuse visualization toolkit as a desktop application. It imports two types of data files, one with the story, review and reviewer information and the other one with the story articles. Then users can search specific keywords and explore the search results such as the articles that has the keyword and the demographic information of the reviewers and the rating they gave to the articles. While navigating the results, a user can store the filtered results as an captured image whenever they are willing to, which is very useful when one discovered interesting insight.

4.1 Content-Based Article Visualization

To get the information about the articles which contain a single keywords or a set of certain keywords, a user can type in and add the desirable keywords to the keyword list box and then click the search button. For example, if you enter the keywords “greet” and “Obama”, then NewsTrust Xplorer will find all the articles that have both two keywords at least once. Searching for a set of keywords is much more helpful for users to get the context of the keyword that is written in articles that are filtered than searching for a single keyword.

A user can also filter out the news articles by selecting only the wanted source, specific time period, the rating range, whether the story is from Baltimore or not, and the number of occurrence that the keywords in each article.

The scatter plot on the top left represents the search result where each rectangle represents an article. A user can choose the x-axis and y-axis, size, color. By default, the x-axis is the date and the y-axis is the rating of the story, the size of the rectangle is proportion to the number of reviews of an article and the color by the number of occurrence of the searched keyword in an article. It is possible to zoom in the results with the scrollbars on each axis.

4.2 Demographic-Based Review Visualization

After filtering out articles by using keywords, we visualize the reviews on them with a treemap visualization. Fig. 4.2.1 shows the structure of our treemap. It is designed to display the demographic information of the reviewers who wrote reviews of the articles that contain the specific searched keywords. Also, it shows the overall ratings of the articles that the reviewers gave to the one of those articles. The treemap we use has two-level hierarchy. The top hierarchy is age group by default-a user can change this to any attribute of reviewers such as gender, income or education. On the other hand, the lower hierarchy is fixed to the rating metrics that are mentioned in section 3.2. The size of the first node of the tree is proportion to the number of people each depth of the tree, which makes the size of each leaf node too small. Thus we added function that

![Figure 2. Content-Based Article Visualization. Each of the rectangles indicates one article. The size represents the number of reviewers, and the color depends on the number of occurrences of the searched keyword in each article.](image-url)
Figure 3. Demographic-based review visualization. The size of each node represents the number of people who wrote reviews on, the color indicates the rating of each category. Trust metric is highlighted by typing the name of the metric.

5. EVALUATION

We are going to conduct two types of evaluations of our visualization tool, NewsTrust Xplorer, in terms of User Experience and Visual Data Analysis and Reasoning.

5.1 User Experience

To improve interactive or visual aspects of the tool, we will have 7 subjects train no more than 2 minutes, and then perform 7 tasks. We are going to collect not only the accuracy and efficiency of their task performance but also each their feedback and opinions in written or spoken form during and after the experiment. The goal of this evaluation is to understand to what features of our visualization supports the specific tasks in users’ point of view. By collecting the participants’ reactions to the visualization, we can refine our visualization or interface design. Every subject will be asked to answer his/her satisfaction, trust and liked/disliked. Also, every task they perform will be measured in terms of perceived effectiveness, perceived efficiency, perceived correctness. Subjects will be recruited among the students who are familiar with using visualization tools and have basic knowledge on information visualization and they will be given a maximum of 2 minutes of training of our tool. Each subject will be given 10 minutes to fulfill the 7 tasks below. Think-aloud protocol will be used.

- Search for the articles that contain the keyword ‘Japan’ with time period from 11/30/2010 to 1/30/2011.
- Search the keyword again with the minimum frequency of 3 and the story article rate between 3 and 5.
- Add one more keyword ‘Korea’ with all time period and then search again.
- When is the time that people’s interest in ‘Japan’ and ‘Korea’ reached its pick?
- Which age group has the most interest in the set of keywords?
- What is the rating value that people in age group from 50 to 64 most cared?
- Start another Search on ‘Japan’ and ‘China’ and compare the results with the previous result.

After the experiment, each subject will be asked to fill out the questionnaires below.

- What are the features you like?
- What are the features that were not clear to use?
- What are the features that might be useful to be added?

5.2 Visual Data Analysis and Reasoning

While evaluating user experience only focus reasoning evaluation is to see how our information
visualization tool supports users in analytic process such as reasoning the data, generating knowledge, insight or hypothesis. For this evaluation, subjects are required to be experts in our domain, which is the tendency of reviewers of online news articles. We will conduct interviews with 3 subjects after the 2 minute training and direct observation of each subjects’ 10 minute self-exploring. Again, we will use think aloud protocol and have it recorded for further reference. The goal of this evaluation is to get the feedback of how useful the tool is to find an insight in the user’s domain. The questions that will be given during the interviews are written below:

- How well/poor it support data exploration such as searching, filtering and reading and extracting information?
- How well/poor it support the knowledge discovery process?
- How well/poor it support hypothesis generation? and interactive examination?
- How well/poor it support the communication and application of analysis results?

The output of this evaluation such as responses and observed the reactions of subjects along with the feedbacks from the user experience experiment will give us the idea of improving our visualization design in terms of not only its usability but also the utility.

6. CONCLUSION AND FUTURE WORK
Since, News Trust Xplorer is optimized in navigating news articles with given keyword,

REFERENCES