A Computational Analysis of Agenda Setting

Yeooul Kim, Suin Kim
KAIST
Republic of Korea
suin.kim@kaist.ac.kr

Alejandro Jaimes[*]
Yahoo Research
New York, USA
ajaimes@yahoo-inc.com

Alice Oh
KAIST
Republic of Korea
alice.oh@kaist.edu

ABSTRACT

Agenda setting theory explains how media affects its audience. While traditional media studies have done extensive research on agenda setting, there are important limitations in those studies, including using a small set of issues, running costly surveys of public interest, and manually categorizing the articles into positive and negative frames. In this paper, we propose to tackle these limitations with a computational approach and a large dataset of online news. Overall, we demonstrate how to carry out a large-scale computational research of agenda setting with online news data using machine learning.

1. INTRODUCTION

In communications and media studies, *agenda setting* [3] is an important theory for explaining the effect of media on public opinion. We propose that the vast and ever-increasing amounts of user comments and social sharing data on online news sites, combined with recent advances in text mining techniques, enable a deeper and broader analysis of media effects. In this paper, we apply machine learning to extract meaningful patterns from online news and offer an effective alternative to traditional agenda setting research.

Specifically, we address the following. First, because of limited resources in terms of news data and public input, usually measured by surveys about the "most important problems" (MIP), each agenda setting study only considers a few issues, hand picked by the researchers [5]. The validity of the survey-based methods are questioned, and researchers discuss how social media may replace or supplement the surveys [6]. Also, content analysis for positive and negative tone is done by manual coding which is time-consuming [2].

We tackle these limitations with a computational approach to agenda setting research. The first step of our approach is to obtain data from a news website and collect user comments and sharing counts for observing the effect of media. The second step is to automatically discover issues using machine learning and analyze the media effects for the issues. The third step is to replace manual coding by an algorithm to automatically discover the polarity of the articles and the corresponding comments. With these, we define a new perspective of media effects, we show the advantages of computational media effects research, and we reveal the potential of applying machine learning to discover issues and sentiment patterns in articles and comments. While recent research [4] has modeled agenda setting with a machine learning algorithm, our focus is on the effect on the public, measured by their responses in the form of comments and shares.

2. DATA AND METHOD

We use `npr.org` (National Public Radio), and we collect news articles, all comments on those articles, and the commenters' user IDs from eight news sections from Jan 2011 to Apr 2013. We collected 17,674 articles and 763,721 comments from 117,111 commenters. In addition to the comments, users share the articles using social network services such as Facebook, and we collected the number of shares for each article, with a total of 27,886,921 shares for all articles.

To identify important issues, we use a nonparametric topic model, the hierarchical Dirichlet process (HDP) [7]. With the goal of assigning a single issue per each article, we set the hyperparameter α to a small value such that each article is constrained to have a very few number of issues. We control the granularity of the issues by adjusting the η prior.To analyze the degree of agenda setting, we look at users' news sharing and commenting behaviors separately, and analyze the degree of media effect by looking at the correlations of those behaviors with the amount of news coverage. We look at the correlation patterns of issue coverage and the three factors, number of comments, number of unique commenters, and sharing counts, for each issue.

3. RESULTS AND DISCUSSIONS

Using HDP, we found seven issues from the Sports section, ten issues from Technology, eleven issues from Business and World, twelve issues from Politics and Opinion, thirteen issues from Health, and fifteen issues from Science. The number of articles of each issue varies, and we select and use issues with more than one hundred articles. Table 1 shows a few example issues from the Politics section.

The results show large and significant correlations for almost all issues and comments, and most of the issues for shares. Among the eight sections, Sports and Technology

[*]This work was performed while the author was a visiting professor at KAIST, Division of Web Science and Technology, under the WCU (World Class University) program.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577379.

keywords	#articles
romney gingrich republican santorum	575
government court gun voter rights	195
state republican election party walker	167

Table 1: Keywords and no. of articles from Politics.

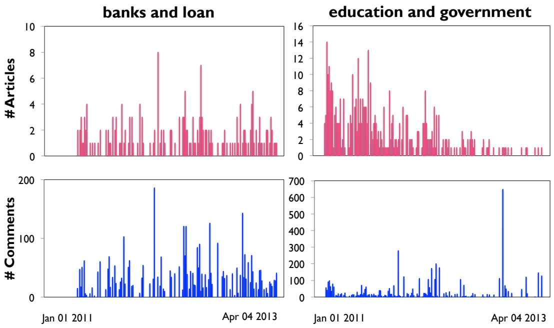

Figure 1: (a) *banks and loan* and (b) *education and government*. Red bars (upper) show the no. of articles, and blue bars (lower) the no. of comments.

show the largest correlations between issue coverage and sharing, and World shows the smallest. A closer look at two examples showing different degree of agenda setting are shown in Figure 1. The Business issue on *banks and loan* shows a large agenda setting effect, whereas the Opinion issue on *education and government* shows a small effect.

Our sentiment analysis works by taking a set of seed words and using topic modeling to expand the positive and negative words [1]. Sentences in the article (or comments) are assigned sentiment scores according to the number of positive and negative sentiment words in the sentence. This sentiment lexicon expansion shows issue-dependent sentiment words, such as “cutthroat” (negative in Business), “victory” (positive in Sports), and “benign” (positive in Health).

We look at the sharing and commenting behaviors for articles with positive and negative sentiments and find interesting patterns. Figure 2 shows the average number of comments and shares in positive and negative articles, as well as the entire set of articles. An interesting pattern is that users’ shares show high interest in Science and Health, but their comments show high interests in Politics and Business. In general, users share more positive articles than negative, except in Politics and Sports. In Sports, negative articles are about losses, so there is no difference between negative and positive articles for sharing. Positive articles in Science are about new findings, so there is a significant increase in sharing. For commenting, users tend to leave more comments on negative articles than on positive articles, especially for Politics. This reflects frequent heated discussions on criticisms about the current politics. The reverse is true for Health, where users discuss positive articles about improving health.

These results illustrate that using a computational approach to media effects analysis alleviates several limitations of the traditional media effects studies. The most important advantage of a computational approach is the scale of the study. By using machine learning to automatically discover the issues and sentiments, we were able to get results from over 17,000 articles, 763,000 comments, 28 million shares.

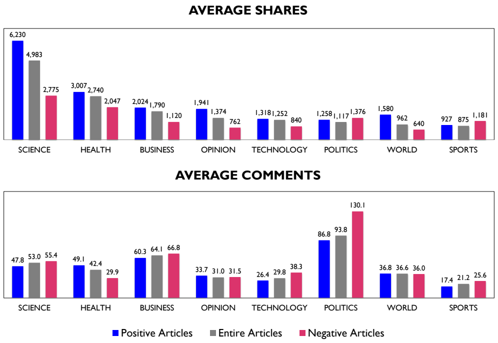

Figure 2: Average number of shares and comments in pos/neg articles in the eight sections.

There are several questions we have not addressed in this paper. We only used NPR as our dataset to identify media effects. Although NPR is one of the major online news outlets, for more comprehensive analysis of media effects on general public, we would need to collect data from a variety of different sources. Many questions remain, such as the accuracy of all the steps of the computational analysis. While the individual methods have been tested and used extensively, the validity and effectiveness for analyzing agenda setting effects are not as clear. Finally, agenda setting research has many deep questions. As this is a first study in applying computational methods to agenda setting theory, we plan to delve into some of the more detailed questions in agenda setting research with computational tools.

4. ACKNOWLEDGMENTS

This research is funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R & D Program 2013. Part of this work has been funded by the European Union 7th Framework Programme Social Sensor projects.

5. REFERENCES

[1] Y. Jo and A. Oh. Aspect and sentiment unification model for online review analysis. In *Proc. WSDM*, 2011.

[2] S. Kiousis. Explicating media salience: A factor analysis of new york times issue coverage during the 2000 us presidential election. *Journal of Communication*, 54(1):71–87, 2004.

[3] M. McCombs and D. Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.

[4] V.-A. Nguyen, J. Boyd-Graber, and P. Resnik. Lexical and hierarchical topic regression. In *NIPS*, 2013.

[5] M. Roberts, W. Wanta, and T.-H. D. Dzwo. Agenda setting and issue salience online. *Communication Research*, 29(4):452–465, 2002.

[6] T. W. Smith. Survey-research paradigms old and new. *International Journal of Public Opinion Research*, 25(2):218–229, 2013.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.