News Comments: Exploring, Modeling, and Online Prediction (Abstract)[*]

Manos Tsagkias
e.tsagkias@uva.nl

Wouter Weerkamp
w.weerkamp@uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Science Park 409, 1098 XH Amsterdam

ABSTRACT

Online news agents provide commenting facilities for their readers to express their opinions or sentiments with regards to news stories. The number of user supplied comments on a news article may be indicative of its importance, interestingness, or impact. We explore the news comments space, and compare the log-normal and the negative binomial distributions for modeling comments from various news agents. These estimated models can be used to normalize raw comment counts and enable comparison across different news sites. We also examine the feasibility of online prediction of the number of comments, based on the volume observed shortly after publication. We report on solid performance for predicting news comment volume in the long run, after short observation. This prediction can be useful for identifying potentially "hot" news stories, and can be used to support front page optimization for news sites.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Comment volume, prediction, user generated content, online news

1. INTRODUCTION

As we increasingly live our lives online, huge amounts of content are being generated, and stored in new data types like blogs, discussion forums, mailing lists, commenting facilities, and wikis. In this environment of new data types, online news is an especially interesting type for mining and analysis purposes. Much of what goes on in social media is a response to, or comment on, news events, reflected by the large amount of news-related queries users ask to blog search engines [3]. Tracking news events and their impact as reflected in social media has become an important activity of media analysts [1]. We focus on online news articles plus the comments they trigger, and attempt to uncover the factors underlying the commenting behavior on these news articles. We explore the dynamics of user generated comments on news articles, and undertake the challenge to model and predict news article comment volume shortly after publication.

Let us take a step back and ask why we should be interested in commenting behavior and the factors contributing to it in the first place? We briefly mention two types of application for predicting the number of comments shortly after publication. First, in *reputation analysis* one should be able to quickly respond to "hot" stories and real-time observation and prediction of the impact of news articles is required. Second, the *lay-out decisions* of online news agents often depend on the expected impact of articles, giving more emphasis to articles that are likely to generate more comments, both in their online news papers (e.g., larger headline, picture included) and in their RSS feeds (e.g., placed on top, capitalized).

Our aim is to gain insight on the commenting behavior on online news articles, and use these insights to predict comment volume of news articles shortly after publication. To this end, we seek to answer the following questions: (i) What are the dynamics of user generated comments on news articles? Do they follow a temporal cycle? The answers provide useful features for modeling and predicting news comments. (ii) Can we fit a distribution model on the volume of news comments? Modeling the distribution allows for normalizing comment counts across diverse news sources. (iii) Does the correlation between number of responses at early time and at later time found in social media such as Digg and Youtube hold for news comments? I.e., are patterns for online responses potentially "universal"? And can we use this to predict the number of comments an article will receive, having seen an initial number?

This paper makes several contributions. First, it explores the dynamics and the temporal cycles of user generated comments in online Dutch media. Second, it provides a model for news comment distribution based on data analysis from eight news sources. And third, it tries to predict comment volume once an initial number of comments is known, using a linear model.

We explore the dataset in §2, model news comments in §3 and report on prediction results of comment volume in §4.

2. EXPLORING NEWS COMMENTS

The dataset consists of aggregated content from seven online news agents: *Algemeen Dagblad* (*AD*), *De Pers*, *Financieel Dagblad* (*FD*), *Spits*, *Telegraaf*, *Trouw*, and *WaarMaarRaar* (*WMR*), and one collaborative news platform, *NUjij*. We have chosen to include sources that provide commenting facilities for news stories, but differ in coverage, political views, subject, and type.

We turn to our first research question: What are the dynamics of user generated comments on news articles? News comments are found to follow trends similar to blog post comments as reported in [4]. The news agent commenting facilities (is it easy to comment or not) and content nature (accessible, require less understanding) show to influence the number of comments a news source receives. The time required for readers to leave a comment is on average slower for news than for blogs, although this

[*] The full version of this paper appeared in *ECIR 2010*.

DIR'11, February 4, 2011, Amsterdam.
ACM xxx-x-xxxxx-xxx-x/xx/xx.

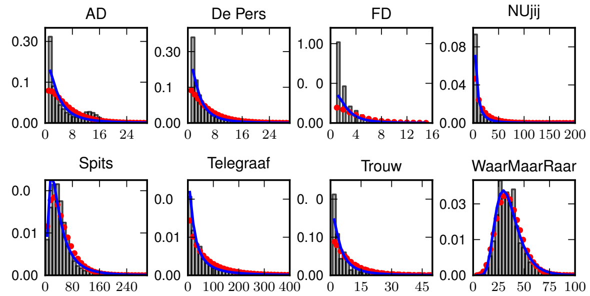

Figure 1: Modeling comment volume distribution per source using the continuous log-normal (blue line), and the discrete negative binomial distribution (red dots). Grey bars represent observed data. Probability density is on y-axis, and number of comments (binned) is on x-axis.

differs significantly per news source possibly due to differences in news readers demographics. With regards to temporal cycles, we look at monthly, weekly and daily cycles. March shows the highest comment volume across the board, and November shows the least for most sources. Weekdays receive more comments compared to weekends, with Wednesday being, on average, the most active day and Sunday the least active day across the board. The daily cycle reveals a correlation between comment volume, sleep and awake time, as well as working, lunch and dinner time: The comment volume peaks around noon, starts decreasing in the afternoon, and becomes minimal late at night. These aspects of online news seem to be inherent characteristics of each source possibly reflecting the credibility of the news organisation, the interactive features they provide on their web sites, and their readers' demographics [2].

3. MODELING NEWS COMMENTS

With regards to our second research question we seek to identify models (i.e., distributions) that underly the volume of comments per news source. We do so (1) to understand our data, and (2) to define "volume" across sources. Our approach is to express a news article's comment volume as the probability for an article from a news source to receive x many comments. We consider two types of distribution to model comment volume: log-normal and negative binomial. For evaluating the models' goodness of fit we choose the χ^2 test due to its applicability to both continuous and discrete distributions [5]. Both distributions fit our dataset well with low χ^2 scores (see also Fig. 1) leaving the final decision on which distribution to favor on the data to be modeled and the task at hand.

4. COMMENT PREDICTION

We now turn to our third research question. First, we are interested in finding out whether the correlation between early and late popularity found by [6] also holds for the news comments space. Then, assuming such a relation has been confirmed, it can be employed for predicting the comment volume of a news story. The existence of a circadian pattern implies that a story's comment volume depends on the publication time. We account for this by introducing a temporal transformation from real-time to *source-time*, a function of the comment volume entering a news site within a certain time unit.

We graph the Pearson's correlation coefficient ρ to visualize the correlation strength between comment volume at times close (early) and farther away (late) from publication. *Spits* displays a very steep comment volume curve meaning that most stories stop receiving comments short after publication. In contrast to our expectations that *NUjij* follows a fast correlation pattern similar to Digg, our findings suggest that a strong correlation is achieved much later possibly due to the different levels of effort required for digg-ing and commenting.

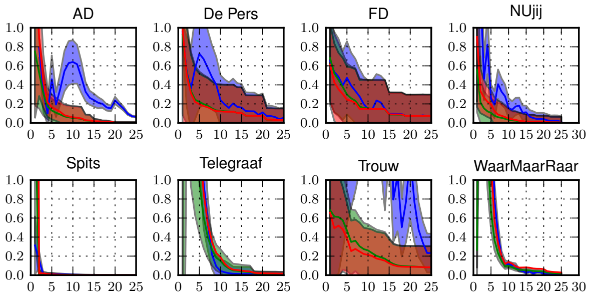

Figure 2: Relative square error using Model 1 (blue line), Model 2 (green line), and Model 3 (red line). Standard deviation is shown in the shaded areas around the lines. QRE on y-axis, observation time (hrs) on x-axis.

We follow [6] and estimate a linear model on a logarithmic scale for each source in our dataset. For evaluating our model we choose the relative squared error (QRE) metric averaged over all stories from a certain source. Fig. 2 shows that from the three models we study, the one using all stories and having the slope fixed at 1 (M2) performs the best. M2 demonstrates strong predictive performance indicated by low QRE < 0.2 for all sources, in less than 10 hours of observation. The QREs converge to 0 faster for some sources and slower for others, exposing the underlying commenting dynamics of each source as discussed earlier.

In this section we looked at natural patterns emerging from news comments, such as the possible correlation of comment counts on news stories between early and later publication time. A relation similar to the one observed for Digg and Youtube has been confirmed, allowing us to predict long term comment volume with very small error. We observed that different news sources ask for different observation times before a robust prediction can be made. QRE curves can indicate the optimum observation time per source, that balances between short observation period and low error.

Acknowledgments. This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430 (GALATEAS), by the 7th Framework Program of the European Commission, grant agreement no. 258191 (PROMISE), by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.-066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011 and by the Center for Creation, Content and Technology (CCCT).

References

[1] D. L. Altheide. *Qualitative Media Analysis (Qualitative Research Methods).* Sage Pubn Inc, 1996.

[2] D. S. Chung. Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *Journal of Computer-Mediated Communication*, 13(3):658–679, 2008.

[3] G. Mishne and M. de Rijke. A study of blog search. In *ECIR'06*, pages 289–301. Springer, April 2006.

[4] G. Mishne and N. Glance. Leave a reply. In *Third annual workshop on the Weblogging ecosystem*, 2006.

[5] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, USA, 2000.

[6] G. Szabó and B. A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.