Time-Series Analysis of Blog and Metaphor Dynamics for Event Detection

Brian J. Goode, Juan Ignacio M. Reyes, Daniela R. Pardo-Yepez, Gabriel L. Canale, Richard M. Tong, David Mares, Michael Roan and Naren Ramakrishnan

Abstract Open source indicators (OSI) like social media are useful for detecting and forecasting the onset and progression of political events and mass movements such as elections and civil unrest. Recent work has led us to analyze metaphor usage in Latin American blogs to model such events. In addition to being rich in metaphorical usage, these data sources are heterogeneous with respect to their time-series behavior in terms of publication frequency and metaphor occurrence that make relative comparisons across sources difficult. We hypothesize that understanding these non-normal behaviors is a compulsory step toward improving

B.J. Goode (✉) · N. Ramakrishnan
Discovery Analytics Center, Virginia Tech, 900 N. Glebe Rd., Arlington, VA, USA
e-mail: bjgoode@vt.edu

N. Ramakrishnan
e-mail: naren@vt.edu

J.I.M. Reyes · D.R. Pardo-Yepez · G.L. Canale · D. Mares
Center for Iberian and Latin American Studies (CILAS), UC San Diego,
Gilman Dr., La Jolla, CA 9500, USA
e-mail: jireyes@ucsd.edu

D.R. Pardo-Yepez
e-mail: dpardoye@ucsd.edu

G.L. Canale
e-mail: gcanale@ucsd.edu

D. Mares
e-mail: dmares@ucsd.edu

R.M. Tong
Tarragon Consulting Corporation, 1563 Solano Avenue, Berkeley, CA #350, USA
e-mail: rtong@tgncorp.com

M. Roan
Department of Mechanical Engineering, Virginia Tech, 445 Goodwin Hall,
Blacksburg, VA, USA
e-mail: mroan@vt.edu

analysis and forecasting ability. In this work, we discuss our blog data set in detail, and dissect the data along several key characteristics such as blog publication frequency, length, and metaphor usage. In particular, we focus on occurrence clustering: modeling variations in the incidence of both metaphors and blogs over time. We describe these variations in terms of the shape parameters of distributions estimated using maximum likelihood methods. We conclude that although there may be no "characteristic" behavior in the heterogeneity of the sources, we can form groups of blogs with similar behaviors to improve detection ability.

Keywords Open source indicators · Metaphor · Temporal clustering

1 Introduction and Related Work

Open source indicators are a useful tool for identifying precursor signals in social media for large scale events. Past work has focused on forecasting protest events and disease outbreaks [1], but the methods employed for mining social media apply to a number of useful event forecasting scenarios. The open source indicators used for forecasting are numerous and include well-known outlets such as Facebook, Twitter, and RSS feeds of news sources and blogs. Success in finding documents within these sources that are of interest is typically subject to various algorithms using keywords [2, 3]. However, our recent work has suggested that more complex keys, such as linguistic metaphors, can be used to identify signals of interest.

The focus of this paper is to understand blog and metaphor content behavior along three dimensions in order to generate signals for event detection. Prior research suggests that blog behaviors can be classified as highly clustered (i.e., "bursty") in terms of temporal rates of appearance and diffusion across a blog network [4]. Such behavior may be intrinsic to human prioritizing behaviors [5, 6], and manifest in blog and metaphor usage.

In this paper, we discuss the methodology of recent work wherein blogs in Latin America are filtered for political metaphors to create signals for event detection. In particular, we address the issue of combining heterogeneous blog sources with sparse key metaphor signals. There are many different methods for addressing this problem, and the approach we take here is one of analyzing occurrence clustering by not assuming exponential rates of blog publication and metaphor appearance using a discrete Weibull distribution. Similarly, we analyze word counts (blog lengths) through the lens of a log-normal expectation. The conclusions we draw are that parsing similar blogs along these dimensions, with an eye toward complex system analysis, results in a set of signals that can be combined to yield an event signal without there being a "characteristic" blog to normalize. We conclude the paper with an example of how this information is aggregated to produce an event signal.

2   Characteristics and Analysis of Blogs and Metaphors

Our data set consists of blog documents that were extracted from four Latin American countries for this study: Argentina (89 blogs), Ecuador (59 blogs), Mexico (97 blogs), and Venezuela (82 blogs). The number of documents in each blog ranges from less than 100 to more than 10,000. In total, 589,089 documents are analyzed. We study three properties of these blog documents grouped by blog source: length (word count), publication frequency, and frequency of political metaphor usage. The distributions that we present of these features reveal interesting properties of the blog sources and political metaphor content. We use maximum likelihood estimation to model the word count as a log-normal distribution. Publication frequency and frequency of metaphorical usage features tend to follow a discrete Weibull distribution [7] and show evidence of varying degrees of temporal clustering. The shape parameter of the discrete Weibull distribution is a measure of related blog behavior for temporal clustering.

Word Counts. Word count distributions are formed for each blog in each country. There are two main types of word count distributions that emerge from the data. The first, shown in Fig. 1a, is one where the expected document length increases with length up to the peak of the distribution, and then starts to decay. One way to think of the distribution is not just a measure of the length of the document viewed as a whole. Rather, the word count distribution is the likelihood that if words were being displayed to an observer, what is the likelihood that a thought will be complete in some finite number of words? In the first case in Fig. 1a, the documents composing this blog series typically require a lead-up to complete the document. This assumes that these documents are full text and prose and not just lists. These types of documents are documents corresponding to authors that are more verbose in their prose. The second type of distribution, shown in Fig. 1b, is a very heavy tailed distribution where the likelihood of a continued document length decreases with document length. These are more "twitter" style blogs. Assuming full text and prose, then these documents tend to get their message out earlier than later in the document. Such documents can correspond to comments about videos or take the format of alerts where one tries to make the point as soon as possible.

The distributions here do not reveal anything about the content of the blogs – that would be a representation error due to speculation. However, if we assume that the documents are written for a purpose, then the distributions do reveal information about the relative nature of where the purposeful information appears in the document. A closer inspection of the distributions (a) and (b) in Fig. 1 show that they are not very different in terms of frequency of length leading toward the right tail. In fact, the Fig. 1b tail appears to be slighyly longer. However, the short documents in Fig. 1b suggests that there is a prevalence for short and concise documents given the increased probability of occurrence.
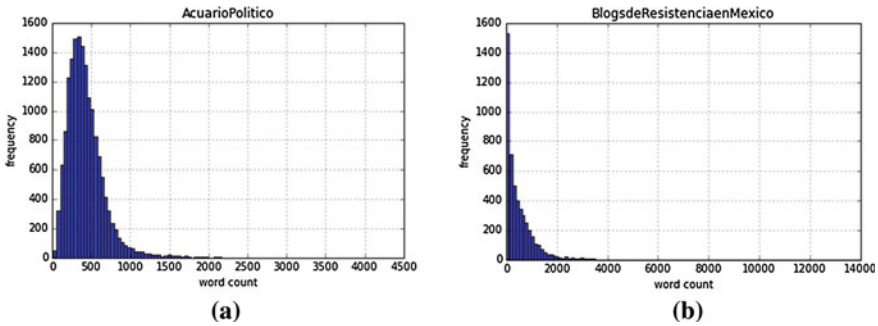
Fig. 1 Distribution of blog length (by word count) for blogs "Acuario Politico" and "Blogs de Resistencia en Mexico". Cast as a log-normal distribution, this represents the number of words until a document ends. In this particular case, such a distribution describes the probability that a "thought" will end given the amount of words at present in the document

For each country, a maximum likelihood estimate (MLE)[1] of the log-normal distribution was fit to the documents in each blog. Box plots of the log-mean and log-standard deviation are shown in Fig. 2 for each country. For each country, the quartiles for both log-mean and log-standard deviation tend to be comparable on the log scale. The majority of the mass of the distributions tend to be centered around a document length of 400 with a confidence interval of spanning [500, 800] words (Cox method). On the log scale, the differences in the mean quartiles can be quite substantial. Argentina and Ecuador, for instance, show more diversity in the expected blog length values than Mexico and Venezuela. The larger the spread of the quartiles indicates more diversity in the lengths of blogs, and subsequently, the prevalence for how long information is communicated in a particular blog.

Publication Frequency. As discussed above, the temporal behavior of blogs is subject to clustering (i.e., "bursty") behavior along several dimensions including temporal publication frequency and references to other blogs. We hypothesize that the temporal clustering (publication frequency) behavior of the blogs is indicative of the type or quality of the content of the documents published. An example is shown in Fig. 3 where two very different publication behaviors are seen. In Fig. 3a the blog "Carpe Diem" is seen to exhibit numerous short-order occurrences with fewer long-term occurrences decreasing in probability as the temporal duration increases. This trend is indicative of temporal clustering, and increases in publication activity could serve as an event indicator. Many blogs follow this format. There are also cases seen in Fig. 3b where there is a definitive publication frequency. For this blog, publication appears to occur regularly around every two weeks with few deviations from the trend. The issue with this sort of blog is that

---

[1]MLE for all sample distributions analyzed with Q-Q plots. Results hold well for most distributions. Distributions with larger tails show deviations in the higher quartiles, but with no particular trend. Errors in tail estimation are attributed to non-aggregate and micro-level events beyond the scope of this paper.
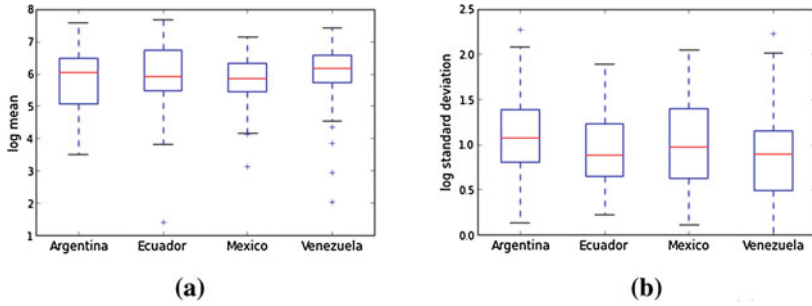
Fig. 2 Box plot of length log-mean and log-standard deviation parameters by country

there is an increased chance that the text of the recorded date does not coincide well with actual events. In this case, event detection will have to take into account the temporal references in language present in the document. However, it may also be indicative of a more reputable news source that publishes regularly, and the content may have higher quality content. Removing speculation as to the cause, simply aggregating blogs of different temporal clustering characteristics may lead to inconsistencies with respect to both temporal alignment and content type.

We propose that the temporal clustering in our data can be measured by the shape parameter, β, of the discrete Weibull distribution. Specifically, if we measure deviation from an exponential temporal duration of publication, then $\beta = 1$ indicates exponentially decreasing occurrence rates with time. $\beta < 1$ indicates temporal clustering behavior like that seen in Fig. 3a, and $\beta > 1$ indicates behavior of increasing likelihood of publication with greater temporal duration. The box plots shown in Fig. 4 show how the proportion of blog temporal clustering in our dataset in every country. For Argentina, Ecuador, and Venezuela three quartiles exhibit
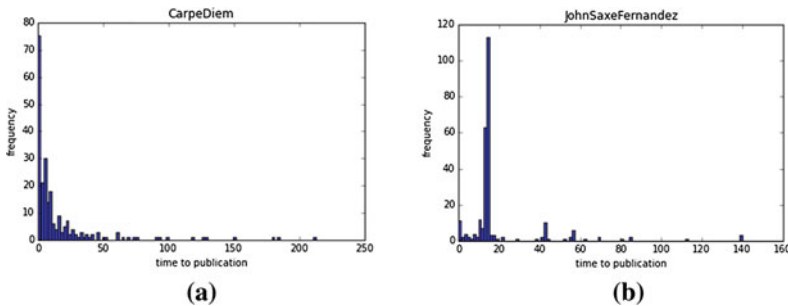

Fig. 3 Distributions of publication frequency. The distribution on the *left* **a** shows temporal clustering where the distribution on the *right* **b** shows a regular bi-weekly publication rate. When combining sources, different clustering behaviors in aggregate can change the perceived dynamics of the underlying content. Grouping by similar clustering behaviors can help mitigate effects of different time stamps
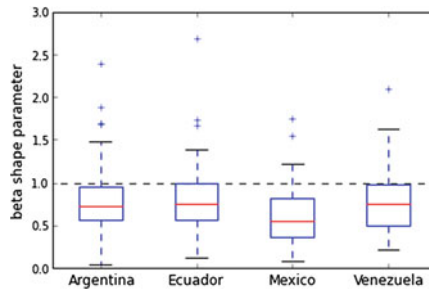
Fig. 4 Box plot of distribution of publication frequency. The *dashed line* represents a uniform rate of publication. All of the datasets in each country tend to exhibit mostly clustered publication rates, with one quartile of the blogs appearing to be a more uniform publication rate. However, the political blogs selected in Mexico tend to display more temporal clustering than the rest of the dataset

temporal clustering where one exhibits less temporal clustering. Our Mexico blog dataset exhibits more temporal clustering than the other countries.

Metaphor Usage. In each blog document, linguistic metaphors were discovered using a metaphor extraction algorithm [8]. The metaphors are mapped to a set of political target and source concepts (see Methods section). Instead of key words, these concept mentions serve as the basis for forming the event signal. The frequency of occurrence of specific source and target concepts in the blog documents is recorded over time. Similar to publication frequency, if linguistic metaphors appear regularly in particular blogs, then it may be more difficult to distinguish the signal from the noise, or ordinary use of language. For this, we use MLE to estimate the shape parameter of the discrete Weibull distribution for metaphor time series data, for each blog. The resulting box-plots separated by country are shown in Fig. 5. These results show that all of the blogs tend to exhibit a degree of temporal clustering with respect to metaphor usage, with some blogs being very cluster-centric in their use of political metaphor. Across countries, similar results are seen, with Ecuador exhibiting less clustering and Venezuela having a larger range at the center of the shape parameter distribution.

Fig. 5 Box plot of shape parameter distributions for metaphor time series data
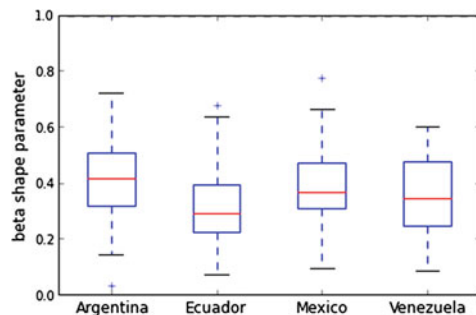
These results are encouraging in that metaphor usage within blogs appears to be clustered and more likely to surround an event of interest. However, this does not preclude that metaphors are not persistent throughout a population and only referenced on occasions of interest. To the contrary, some metaphors like those that compare elections to a game tend to be quite persistent when aggregated across our whole blog corpus. To compensate for this, we use the results of publication frequency and blog length to aggregate across groups of similar blogs to effectively normalize for different blog dynamics and reduce the temporal complexity present within each aggregated unit of blogs.

3   Event Detection

We now provide an example of a use case where we cluster metaphor time series based on correlated metaphor behavior and temporal clustering of blog publication frequency. For this example, we use our Argentinian dataset and plot these metrics on the set of axes shown in Fig. 6. Information about the blog dynamics are shown on the axes themselves. The y-axis is the occurrence measure (k = β) for the blog publication frequency, and the frequency of document publication are shown on the x-axis. Each marker represents a subset of the Argentinean blogs displaying similar trends in specific metaphor activity based on correlated metaphor time series. The opaqueness of the marker shows the metaphor time series trending strength of
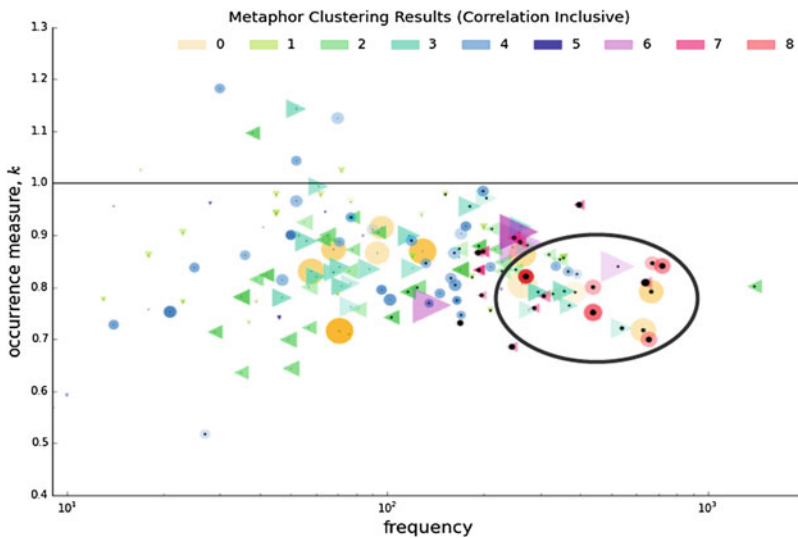

Fig. 6 Plots of correlated groups of blogs and metaphors in terms of frequency of appearance and occurrence measure. We look at groups of blogs and metaphors that have low occurrence measures (high clustering) and high frequency of appearance

groups of metaphors separated by co-clustering on signal correlation with a 30-day rolling Hamming window. The size of the marker indicates the number of similar blogs, and the size of the black marker represents the number of similar metaphor clusters within each subset of blogs. The end result of this procedure are groups 1–8 that represent aggregated collections of *both* blogs and metaphors displaying similar behaviors as defined in the previous sections.

For this example, we target the lower right region of the plot which contains blogs that show large numbers of entries with a propensity to show temporal clustering (low occurrence measure). We also target highly correlated metaphors which give more signal weight to the temporally clustered metaphor time series. Of the available groups, Group 8 is well represented in this region of the chart. This group is represented by the following metaphor source-target concept pairs: GOVERNMENT:{MACHINE, MOVEMENT, STRUGGLE} and WEALTH: {MOVEMENT, RESOURCE}. This indicates that there were instances of linguistic metaphors referencing concepts that GOVERNMENT is like a MACHINE or WEALTH is like a RESOURCE.

We plot the aggregated time series for Group 8 in Fig. 7. We investigate the signal for peaks in metaphor usage around particular dates. For Group 8, the concept pair of GOVERNMENT and MACHINE shows several peaks in the years spanning
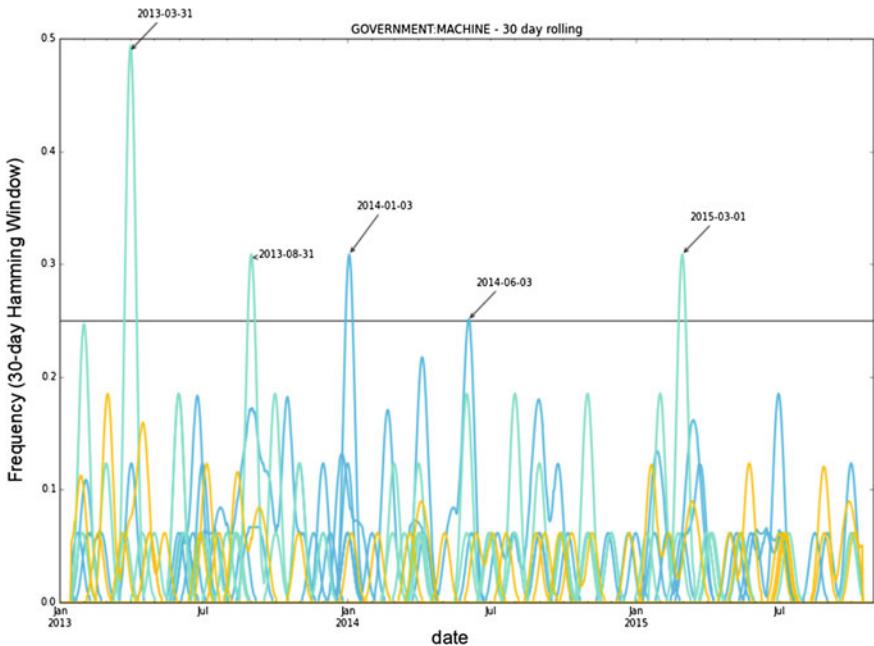

Fig. 7 The time series are shown for the GOVERMENT:MACHINE metaphor concept. The time series displayed here are all from blog and metaphor combinations that show similar temporal clustering, document length, and correlated metaphor usage

2013–2015. The highest peak, which occurs on 3/31/2013 on a centered monthly rolling average corresponds to the dates following when Pope Francis, from Argentina, was elected Pope by the Catholic Church. Also prevalent during this time was the Kirchner money laundering allegations appearing on an Argentinean news show [9]. The documents composing these time series make references to both Pope Francis and government finance, through the lens of political discontent.

4 Conclusions

Blog sources can be widely varied in terms of characteristics governing document length, publication frequency, and content (e.g., metaphor usage). For event detection, we use the key linguistic metaphor concepts to form time series for event detection. Our analysis of blog dynamics from our set of Latin American blogs shows that the document length tends to follow a log-normal distribution. This is indicative of the preference for communicating ideas, either in short bursts or more lengthy discussions. The publication frequency of the blogs were measured using maximum likelihood parameter estimation from a discrete Weibull distribution. The shape parameter of this distribution serves as a measure of temporal clustering of document publications. Blogs with high clustering behavior are more likely to coincide with events of interest than those with a constant publication rate. Our content analysis revealed that no one blog was likely to have persistent metaphor usage, contributing to the strength of the time series signal. Finally, we demonstrate a grouping relation among both blogs and metaphor content in order to generate a signal for event detection. For this, we targeted blogs with high temporal clustering, large document lengths, and similar trends in metaphors. We demonstrate that we can extract peaks by aggregating blogs of similar behaviors.

5 Methods

The blogs in this study were read and selected for political slants and references to political topics. Blog searches were conducted either using a directory (e.g., http://blogsdemexico.com.mx) or through Google keyword searches. Political blogs were identified by references to political entities, events, and people (e.g. "Maduro blogspot"). Also included in the corpus are blogs discussing issues such as the economy. Our blog dataset is biased toward political blogs, and specifically those with high metaphor content. Over all countries included in the analysis, the mean linguistic metaphor to document ratio was 0.49, meaning 1 out of every 2 documents on average contains a metaphor reference. All blogs were extracted using an automated tool to search, pull each document present within the blog structure, and identify meta information such as publication date, author, and title. Blogs were excluded if they were coded using JS Widgets.

The metaphors referenced in the text were identified using software, called the Metaphor Detection System (MDS), developed on the IARPA Metaphor program.[2] The MDS detects linguistic metaphors in Spanish, and has a detection F-score of 0.74 (precision 0.82, recall 0.68). Each linguistic metaphor is mapped to a target concept: BUREAUCRACY, DEMOCRACY, ELECTIONS, GOVERNMENT, POVERTY, TAXATION, and WEALTH. Each of the target concepts are linked to a source concept describing the target as mapped in the linguistic metaphor generating a source-target concept pair. The frequency of appearance of these concept pairs over time constitutes the time-series. These pairs are established a priori in the MDS architecture and can serve as a bias.

All MLE parameter estimates for distributions were found using R statistical software. All co-clustering operations were conducted using the scikit-learn package in Python. Co-clustering was performed over 20 randomized trials to determine the optimal number of clusters. The metaphor time series used to generate the correlation matrix were filtered with a central Hamming window of length 30 days to smooth out sparse signal irregularities.

References

1. Ramakrishnan, N., et al.: 'Beating the news' with EMBERS: forecasting civil un-rest using open source indicators. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1799–1808. KDD '14, ACM, New York, NY, USA (2014)
2. Muthiah, S., et al.: Planned protest modeling in news and social media. Innovative Appl. Artif. Intell. (2015)
3. Chakraborty, P., et al.: Forecasting a moving target: ensemble models for ili case count predictions. In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 262–270 (2014)
4. Goetz, M., Leskovec, J., McGlohon, M., Faloutsos, C.: In: International AAAI Conference on Web and Social Media (2009)
5. Barabasi, A.: The origin of bursts and heavy tails in human dynamics. Nature **435**, 207–211 (2005)
6. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley Press Inc, Cambridge (1949)
7. Nakagawa, T., Osaki, S.: The discrete Weibull distribution. *IEEE Trans. Reliab.* **R-24**(5), 300–301 (1975)

[2]See: http://www.iarpa.gov/index.php/research-programs/metaphor.

8. Mohler, M., Tomlinson, M., Rink, B.: Cross-lingual semantic generalization for the detection of metaphor. Int. J. Comput. Linguist. Appl. **6**(2), 115–136 (2015)
9. Pérez, S., Turner, T.: In Argentina, mix of money and politics stirs intrigue around kirchner. Wall Street J. July 28, 2014. [online] Accessed March 9 2016

Springer