What Memes Say about the News Cycle?

CS322 Final Project Report

Shayan Oveis Gharan
shayan@stanford.edu

Farnaz Ronaghi Khameneh
farnaaz@stanford.edu

Ying Wang
yw1984@stanford.edu

ABSTRACT

In this project we investigate different properties of the news cycle. We address the correlation between news topic and the media coverage distribution. We present a simple method for finding pioneer bloggers in the blog space. Results of our work show that pioneers mostly write about time-sensitive topics such as: politics and finance. We also study differences and similarities between news agencies in their attitude towards news. We show that they share same properties at large-scale, but they have many structural differences. Although, we believe more detailed data will help us in recognizing structural similarities in a better way.

Keywords

Data Mining, Topic tracking, News cycle

1. INTRODUCTION

The meme-tracker is an application developed by Leskovec et al. [2] to track memes across the Web. As described in [2], "The application builds maps of the daily news cycle by analyzing around 900,000 news stories and blog posts per day from 1 million online sources, ranging from mass media to personal blogs." Leskovec et al. in [2], track the quotes and phrases that appear most frequently over time across this entire spectrum. This makes it possible to see how different stories compete for news and blog coverage each day, and how certain stories persist while others fade quickly.

Output of the meme-tracker is organized into clusters of stories, where each story contains urls and frequency information for all articles which have mentioned it. We use this data to study various properties of blogs and mainstream media. Characteristics of the news cycle has always been of particular interest for politicians, sociologists and economists. One of the interesting directions is to look for possible follow-ups for a story which happen in distant future. Different news agencies have different attitude towards spreading news stories, they react to various topics in a different way. Comparing two news agencies is another problem we investigated. In [3], researchers show that there are memes that are migrating from the blog space towards the news media. We defined a simple measure for finding the pioneers of the blog space. The people with high score can be considered as the influential bloggers. We looked at the dominant topics of pioneers posts and studied the time dependency of those topics.

Our study shows that topics such as "politics" and "finance" are very time-dependent, topics such as "sports" and "living" are almost the opposite. Another finding is that the gap between peak points follow approximately an exponential distribution, suggesting that follow-ups may arrive as a Poisson process. Though large news agencies have similar statistics like degree distribution, our experiment on network alignment shows that they actually have very different structure. This part of the work needs more validation though as we address later.

2. COVERAGE DISTRIBUTION AND STORY FOLLOW-UPS

For each story, we can define a coverage distribution showing the volume of media coverage for that story over time. For many stories in the meme-tracker data, there are multiple peak points in the coverage distribution. These peak points show multiple points of media attention to a news story. We studied how often and how far away these multiple peaks are.

First of all, we need to identify peak points. Finding a local maxima is itself challenging. We recognize a peak point as a point in time with enough media coverage. There should also be a steep change in the neighborhood of the peak point. To capture these properties, we calculated a weighted sum of frequency differences around the point. Weights reduce exponentially suggesting the fact that the bigger the difference, between the local maxima and its' close neighbors, the better. Let's call this weighted sum as the quality of peak point. We then select the two highest-quality peak points for each of the stories and calculate the time gap between them. We have plotted the distribution of time gap between peaks in figure 2. The plot suggests that gap between high quality peaks follows an exponential distribution. It can be inferred that there are lot's of stories with time gaps smaller than 10 days, but there also are a remarkable number of them with far apart points of media attention. Intuitively, far apart peaks of media coverage can be considered as pos-

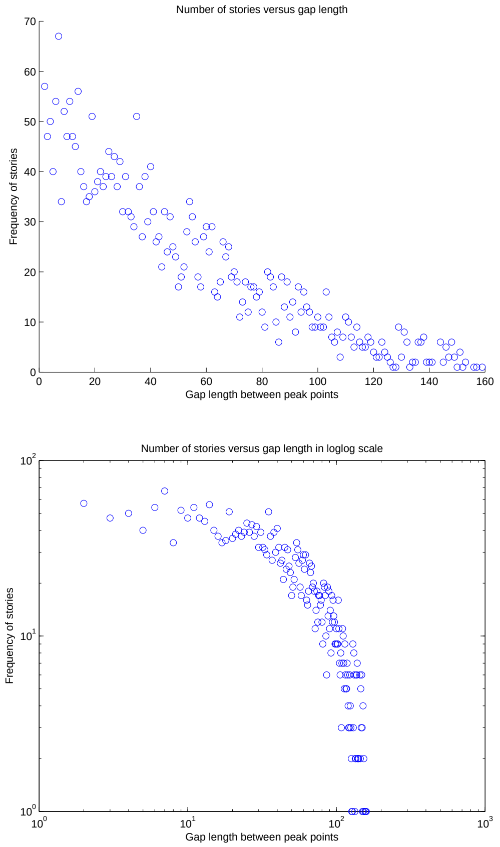

Figure 1: Number of stories versus gap length in log-log scale.

sible follow-ups to stories. Such follow-ups are regular for sports, living stories, music and similar topics. The exponential inter-arrival time suggest that the follow-ups may come following the Poisson process, i.e. follow-ups happen with a constant rate at any time.

3. A STUDY OF THE TIME DEPENDENCY FOR DIFFERENT TOPICS

In this section, we study the time dependency of different topics based on available data. Suppose that we have a story s. The main question is whether the topic of a story s has any impact on its density function.

We will start this study by finding topics for all articles in data. We did not expect to have accurate topics via clustering or dense subgraphs of the data. As we explored the url structure for news agencies, we found that many of them have particular structures and include the category of articles inside that structure. We found these patterns one by one for different agencies. For each cluster of stories, a major proportion of articles are categorized using this method. We can use the dominant category of each cluster as the topic of all articles in that cluster.

In order to compare the density functions of stories with different topics, we use the standard deviation. Since wider density functions tend to have larger standard deviations and vice versa.

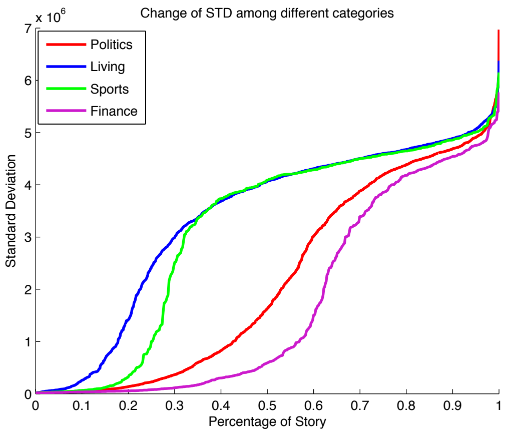

Figure 2: Comparing the time dependency of various topics using the standard deviation of the posts of different stories

We compare the standard deviation of coverage distribution for stories with different topics. The result is shown in Figure 3. In an overall view, ”Finance” and ”Politics” seems to be more time-dependent compared to ”Living” and ”Sports”. In particular, about half of the stories in ”Finance” and ”Politics” have very small standard deviation. The result is reasonable. Indeed, we expect stories about ”TV series” or ”Football tournaments” to be less dependent on time, as they occur during a long period and quotations often recur. On the other hand, political or financial events such as presidency election or financial crisis usually occurs at a specific time and the interest decreases rapidly.

4. PIONEERS OF THE BLOG SPACE

Leskovec et al show that the peak of blog attention to a news story in web blogs happens one hour and a half later than the mainstream media [3]. It is interesting to find pioneers of the blog space. From our point of view, a pioneer blogger is someone who writes ahead of time about different stories. To find pioneer bloggers, we consider two popular blogging hosts and we identify the influential bloggers in each of them. Since we do not have the citation graph, it is a hard problem to find the most influential bloggers by only considering a set of quotations. Indeed, when two bloggers write a series of posts about a particular context, even if b_1 starts to write ahead of b_2, we can not say b_2 is necessarily copying from b_1; However, we can say that b_1 is not copying from b_2. Therefore, those who start to write mostly ahead of time can be considered as influential bloggers.

Finding the pioneers requires defining the *ahead of time* factor explicitly. We define a blogger b, who writes ahead of time, to be a blogger that starts writing much earlier than the average. We quantify this factor by considering the difference between the start time of writing a story, and the time of average media coverage for that story, $E[s]$. Moreover, because we want to compute an aggregated ahead of time factor for a blogger, we need to normalize the factor for each story to remove impacts of difference in their distributions. Therefore, we divide the ahead of time factor by the standard deviation of coverage distribution for each particular story. In other words, ahead of time factor for a specific blogger b and a specific story s is the difference of $t(p)$ and $E(s)$ in standard deviation units. Intuitively it shows the extent to which b has written about s ahead of the others. Thus we have:

$$f(b,s) = \min_{p \in Posts(b,s)} \frac{t(p) - E[s]}{\sigma_s},$$

where $Posts(b,s)$ is the set of all posts of the blogger b for story s, and $t(p)$ is the time of post p. The total ahead of time factor for the blogger b is defined as

$$f(b) = \sum_{s \in Story} f(b,s).$$

We computed this factor for bloggers in `www.blogspot.com` and `www.wordpress.com` who have sufficiently long threads of posts on a particular story. Figure 4 shows the *ahead factor* versus the number of stories for which the blogger has enough posts. The blogger with larger number of stories and smaller *ahead factor* is a pioneer because he writes follow-ups about many stories and he is ahead of time on average.

The colors in the graph show the dominant topic of blogger's posts. For pioneer bloggers, dominant topic turns out to be an accurate measure in practice as we investigated the content of the pioneer weblogs. For example, the blog `http://cnnpoliticalticker.wordpress.com` (lower right corner on the plot) turns out to be one of the political blogs by CNN and probably it is one of the most influential blogs. Interestingly, the influential blogs are mostly in "Politics" and "Living" and "Finance". Due to the results of previous section, finance and politics are very time-dependent and mostly have narrow distributions so being ahead of time in these categories is a challenging process.

5. NETWORK ALIGNMENT

Different news agencies may have preferences on different regions or topics, but do they behave similarly in spreading important news stories? In particular, if we look at the evolution of stories in different news agencies, do they share similar patterns? This is a relatively unexplored area. In this section, we address these questions using the network alignment approach [1]. Our result shows that the evolution graphs in major news agencies do not align well. Assuming our definition of evolution graph truly reflects the evolution of stories, the major news agencies must behave very differently in spreading news.

The network alignment is a technique for computing similarity of different graphs, and it finds a good mapping between two graphs when it exists. Formally, given graphs A, B, and a set of potential matches L between A and B, the objective is to find a matching using edges in L, such that it maximizes the overlap of A and B. This is demonstrated graphically in Figure 5. In our study, we compute and compare evolution graphs from three major news agencies: CNN, BBC, and Reuters. Each vertex in the evolution graph corresponds to a url. Ideally, the evolution graph should reflect the influences between different urls. For instance, if one url is a follow-up of another url, we should put a link between them. Such relationship can be identified by the content and timestamp of the urls, as well as the hyperlinks. In the meme-tracker clusters data, however, we only have the timestamp and the quotations for each url. Therefore we build the evolution graph in the following way: for each pair of urls that share a quotation in common, we put an undirected edge between them. We also tried the directed version but the result was not much different.

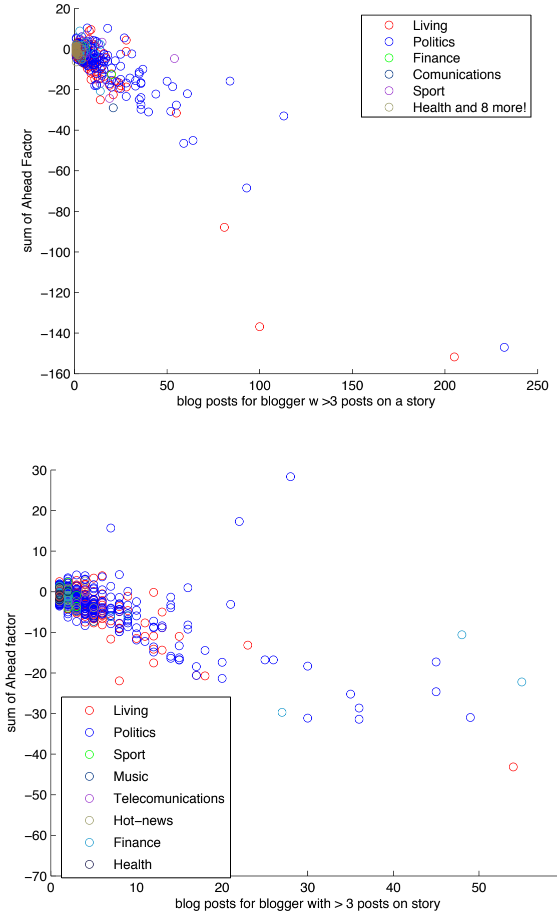

Figure 3: Ahead of time factor for bloggers with sufficiently large length thread of posts. Bloggers are colored according to their interested topics. The bottom diagram represents the influential bloggers of `www.wordpress.com` and the top one represents those of `www.blogspot.com`

To compute the potential matchings, we compare the quotations for each pair of urls from different news agencies. We define the similarity of two quotations as an exponential function of the inverse of their string edit distance. The idea is to let identical quotations score high, and at the same time tolerate small variations caused by typo or punctuation. The score is further adjusted by the difference in volumes of the two quotations to avoid matching a very popular story to an unpopular one. The scores are added for each pair of

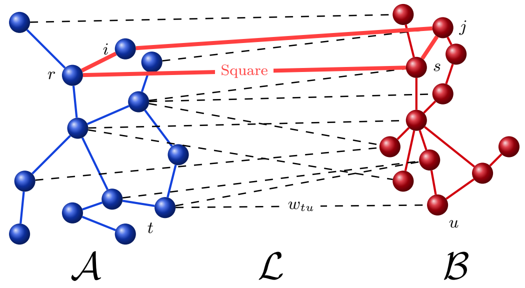

Figure 4: Graphical demo of network alignment

quotations in the urls, and potential matches are selected by setting a threshold on the total score.

After constructing the graphs we computed their alignment. For BBC and CNN, their evolution graphs have approximately 20,000 vertices each, and around 50,000 edges. The alignment produces about 2,500 overlapped edges, which is only a small portion of the total edges. We tried various parameters for generating L, but the result was not improved much.

We thought the poor alignment may indicate the two graphs do not align well globally, so our next attempt was to align them in a narrower space. We extracted urls about a specific topic (like politics or sports) and tried to align a much smaller graph. The alignment was still poor. For the topic ”Politics”, the graphs had about $2,000$ edges but the overlap was around 100. We solved an LP of the network alignment problem and found that the upper bound on overlap was only slightly higher than what we got, which suggest the two graphs really does not align well.

The result of this experiment suggest that the evolution graphs of different news agencies do not align well, but the result should be interpreted carefully. The poor alignment may come from several reasons.

1. The evolution graph we constructed does not necessarily reflect the true relationship between urls. Sharing a quotation may not imply a strong tie between the two urls. The evolution graphs we generated are like collections of cliques, where the true evolution graph should be much more tree-like.
2. Our distance measure for urls needs improvement. Computing only the string edit distance between quotations can go very wrong because string distance does not truly reveal conceptual relevance in most cases. However, based on the data we have it is hard to come up with something smarter. Having the full text and hyperlinks will definitely provide more possibilities.
3. The popularity of a news usually differs between various news agencies. For example, BBC mostly concerns the news in all over the world, while CNN main concern is on the news in United States.

We view these results as preliminary. There are several things we can do that may improve the result. As mentioned above, having more data will enable us to compute more accurate evolution graphs. Moreover, since it is hard to align at a large scope, we can study the evolution of a single story. For example, we can compare the evolution of ”Joe the plumber” on CNN and BBC, which may bring us more insight. Finally, this experiment shows a problem of network alignment: when it works, it is great; but when the alignment is not good, it is hard to argue whether the two entities are not similar, or the graphs are not set up correctly. This is bad because it is actually not easy to find similar graphs in practice and we expect most graphs do not align well. How to use network alignment to give a convincing negative result, this is a problem we need to think about.

6. CONCLUSION

Our study reveals several interesting properties about the evolution of memes. First, the exponential distribution of story follow-up gap time indicates the follow-up happens at a constant rate that is not dependent on time. This is somewhat counter-intuitive because people may expect follow-ups to happen in a relatively short time. Our result shows that except for some highly time-dependent stories, other memes do have a good spread and recur from time to time.

The experiment on finding pioneer bloggers is interesting in both theory and practice. The model helps us identifying influential bloggers. When we investigate those influential blogs, we see good reasons why they are influential. Based on the model, we can build an online tool, such that each blogger can submit the link to their blog and the tool will tell them whether they are the creators of memes, or the copy cats.

Our study on aligning evolution graphs suggests that even though large news agencies may look similar from high level, they have intrinsically different structures, though we feel that the evidences are not strong enough. Nevertheless, it is a new and unexplored idea, and we believe that the result can be improved in a few directions using the links in raw data. This is definitely a promising line of research, as the network alignment reveals more detailed structural similarity between graphs, rather than the high-level statics like degree distribution, which are very similar in most cases. Our current method for making graphs from the current data results in small cliques being connected together, this graph is very different from the real citation graph. We believe that network alignment fails due to the incorrect definition of graph, we hope to continue in this direction with a better data.

7. REFERENCES

[1] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang. Algorithms for large, sparse network alignment problems. 2009.

[2] J. Leskovec, L. Backstrom, and J. Kleinberg. Memetracker: tracking news phrases over the web. available at `http://memetracker.org`.

[3] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and*

data mining, pages 497–506, New York, NY, USA, 2009. ACM.