Temporal Patterns in Memetrackers

Kartikeye Shrivastava
department of CS
Shiv Nadar University
ks731@snu.edu.in

Shashwat Tiwari
department of CS
Shiv Nadar University
st289@snu.edu.in

Tushar Nayak
department of CS
Shiv Nadar University
tn995@snu.edu.in

***Index Terms*—memetrackers, news cycle, social networks, social media**

I. Abstract

There exists growing demand to capture people's attention in the dynamics of social media and the internet, content competes to capture attention for a longer period of time and stay on the news cycle for longer. News is generated and fades in the matter of days and weeks. The dynamics on how these systems of the spread of information and the temporal patterns exhibited by the news cycle are largely unknown.

We aim to study these temporal patterns by primarily analysing the spread of certain keywords and phrases associated with a piece of news - for instance "recession" would be associated with the great 2008 financial crisis, among other phrases and keywords. Through analysis of how the frequency of the use of these terms or "memes" varies over millions of blog posts over the period of several months we hope to quantitatively measure the "news cycle".

Our analysis will offer insight into large scale human interaction with content on the internet and the nature of the temporal patterns of content on the internet.

II. Introduction

There is a diverse amount of content produced, shared, and simultaneously forgotten in the dynamics of online social media and the internet in general. Content is mainly produced by individual users in the form of blogs, videos, social media posts, as well as corporations and massive influences with huge following that influence discussion and the diffusion of memes across various internet platform, the generation of news and simultaneously the cycle of discussion and it's spread through the internet medium portrays an interesting pattern of how new information is propagated, discussed and eventually falls out of discussion - the temporal patterns of the news cycle on the internet.

This reveals interesting insight into the nature of interaction on the internet and how variations in the nature of the lifespan of content varies across different internet platforms. Blogs are a medium where individual users push new content into the system and also propagate existing content. The content is then discussed and propagated through social networks and is discussed across the web. There is little analysis into the nature of this cycle and its intricacies as the internet as a platform ages and the variations introduced by the differences in individual internet platforms such as the micro-blogging site - 'Twitter' where content is limited to 240 characters and the news cycle is slightly shorter due to the inherent nature of the website.

Here we develop a method for tracking units of information as they spread over the web. Our approach is the first to identify short distinctive phrases that travel relatively intact through on-line text as it evolves over time. We study the propagation of these distinctive phrases associated with a certain topic, for instance "election cycle" or "presidential debate" would be closely associated with the 2008 election - so these serve as distinctive markers of the topic. We refer to these distinctive phrases as "memes" as they serve as effective markers to indicate the spread of ideas over the web. The set of distinctive phrases shows significant diversity over short periods of time, even as the broader vocabulary remains relatively stable. As a result, they can be used to dissect a general topic into a large collection of threads or memes that vary from day to day. The use of these "meme" phrases is abundant and varied enough to provide us with a clear image of the logistics and distribution of a given idea - like the economy, a disaster or other forms of news.

From an algorithmic point of view, these phrases can be viewed as the analogue of genetic signatures, and similar to genetic signatures, these distinctive memes undergo changes overtime, similar to mutation in the case of genetic signatures.

As an application of our technique, we use it to produce some of the first quantitative analysis of the global news cycle. In this context, the collection of distinctive phrases that will act as tracers for memes are the set of quoted phrases and sentences that we find in articles that is, quotations attributed to individuals. This is natural for the domain of news: quotes are an integral part of journalistic practice, and even if a news story is not specifically about a particular quote, quotes are deployed in essentially all articles, and they tend to follow iterations of a story as it evolves. We perform this analysis both at a global level — under- standing the temporal variation as a whole — and at a local level — identifying recurring patterns in the growth and decay of a meme around its period of peak intensity. At a global level, we find a structure in which individual memes compete with another over short time periods, producing daily and weekly patterns of variation.

The interplay between technology, media, and social interaction has been the focus of considerable amount of research for the latter part of this decade, often however it has been

at a qualitative level exploring the context in which content and news is produced, and it's effect on public opinion. The collective behavior of the social network in response to content has consequences on how each individual interacts with said content. This has direct applications on predicting overall popularity and temporal trends of online content, it can be utilized to maximise click-through rate.

III. Our Approach

We carried out an exhaustive analysis of 2 papers:

- Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 177-186. 2011.
- Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 497-506. 2009.

We do an analysis of the approaches adopted by both papers, and the conclusions found and the implications for the web and human interaction with information on the web, as well as limitations with their research and the scope of further research.

IV. Paper 1 - Patterns of temporal variation in online media

This paper aims to establish the temporal variations exhibited by online content, and how it is shaped by different media sites, and the types of dynamics it creates. It carries out an analysis on how the popularity of content in social media rises and fades over time.

A. The Approach

The approach taken was to analyse over 170 million blogs posts over a period of one year and examine the adoption of twitter hashtags in a set of 580 million twitter posts collected in an 8 month period. The researchers measure the attention given to each topic, via tracking the mentions of "memes" over time, i.e. the volume of a particular meme. They've deveoped a K-Spectral Centroid (K-SC) clustering algorithm that to find common temporal patterns.

B. Clustering

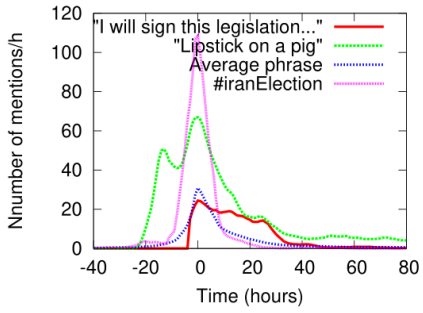

Figure 1: Source: Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media.

The goal with the K-SC clustering algorithm was to cluster similar items together. For instance, in in Figure 1, the red line, which indicates a quote from president Barack Obama on a particular bill, has distinct patterns compared to the line in green - "Lipstick on a pig", notice how the line in green has two distinct spikes.

The challenge with clustering is to group memes on the basis of the shape that they exhibit, not on volume. The clustering algorithm should cluster if two phrases exhibit similar shape irrespective of volume - so what matters here is the pattern of rise and decay in popularity - not the volume.

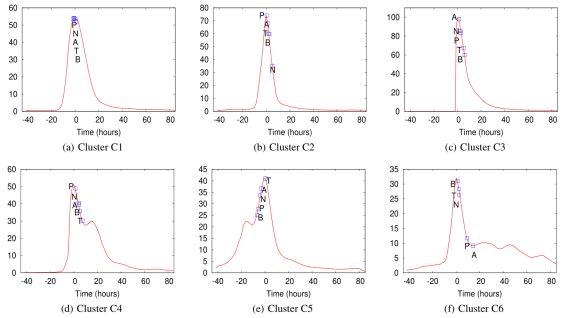

Figure 2: Clusters identified by K-SC algorithm. The letters represent when a type of website first mentions the phrases in a cluster. P: Professional Blog, N: Newspaper, A: News Agency, T: TV Station, B: Blog aggregator

The first 3 clusters exhibit normal spiky behavior, where the peak popularity lasts for less than a day. C1 looks like the average of all clusters because it has the largest volume, and it makes logical sense that most phrases follow the most average pattern. C2 and C3 have sharp peaks, but their total volume around the peak is not the same. This is due to reaction from mainstream media. C4 and C5 are mirror images of each other and the only difference is that C4 gets a rebound in popularity through blogs. The biggest outlier is C6, where a significant majority of the volume is coming from blogs - this probably corresponds to hot topics discussed primarily among blogs for several days and then caught on to by mainstream media.

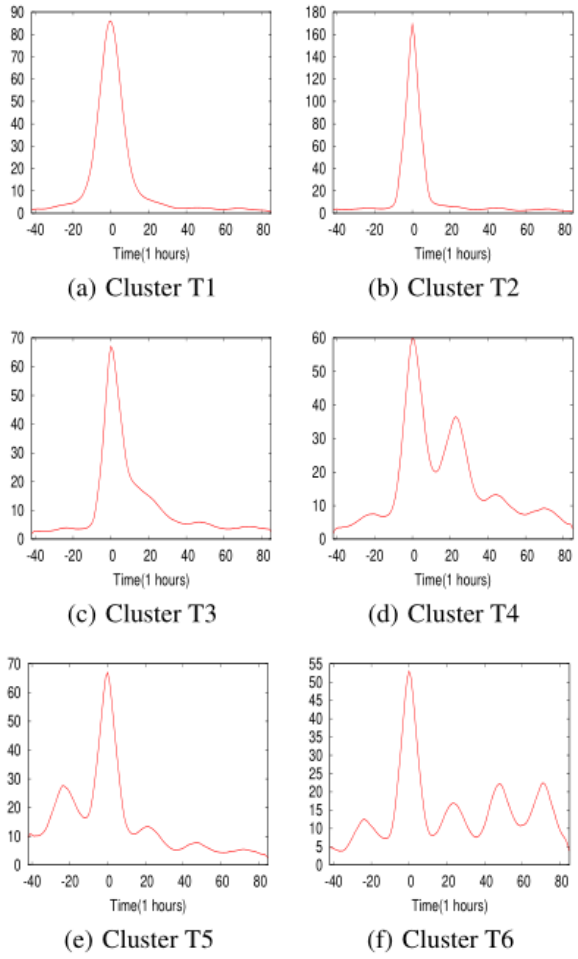

Figure 3: Clusters identified with twitter hashtags

Notably, twitter clusters are quite similar to clusters identified with other websites (compare Fig 3 to Fig 2). Twitter peaks are much more sparse, accounting for only 2-5% of peak volume. With their being significantly more rebounds as hashtags resurface because of the site being much more individual user driven. A small number of blogs has a much greater influence on adoption of news compared to even the most active users on twitter. This could be explained due to the more decentralized nature of twitter - compared to blogs.

C. Findings and Consequences

The findings of the research were that the popularity of online content can be described by a small set of time graphs. It was found that the adoption of hashtags on twitter and the propagation of memes on the web exhibit similar patterns, these patterns being heavily constrained by the type of media. Conventional press agency news shows a very steep rise in popularity followed by a slow decay. However, bloggers play an important role in extending the longevity of news, depending on when and if bloggers start participating in the discourse on a particular topic, it may experience rebounds in popularity.

The research also presents a predictive model based on a few sites or twitter users, that presents with 75% accuracy, the type of temporal pattern that a time series will follow.

Consequences. This paper gives us important insight into how users interact with online discourse, and how humans react to different content on the web. These results have important implications for directly predicting how popular content will be and how long it will last. These results can also be used to maximise online popularity and click-through rate. Most importantly, the researchers have developed scalable computational tools to understand web dynamics.

D. Limitations

This paper has been implemented on older data sets, from 2008. Potential limitations are the outdated nature of the data set, the dynamics of the web have considerably changed since 2008, with the introduction of different media sites and changes in the workings of twitter (for instance: the usage of hashtags has significantly dropped since 2008).

Furthermore, the use of blogs and the presence of conventional media has also drastically changed. Blogs have lost popularity as a medium and the presence of news outlets on the internet is much more. While the insights presented by this paper are very useful, they are limited by the time period of the data set used for their analysis.

V. Paper 2 - Meme-tracking and the Dynamics of the News Cycle

This paper aims to first identify unique distinct contextual phrases that travel over time, and study the eventual mutation of these phrases as the propagate through the web. Then, the researchers attempt to follow these phrases as threads of their own and disect a large topic into memes and threads of their own - instead of being restricted to a small set of memes fora topic. These phrases can be viewed as ”genetic signatures”, similar to genetic signatures, while these phrases remain recognizably the same over time, they still undergo significant changes. A huge computational challenge with this is to find ways of extracting these similar contextual phrases, that are on the same topic but are worded differently from the web, and cluster them together. The researchers have developed a clustering algorithm for this problem, so that memes corresponding to a particular cluster encompass all variants of a phrase.

The researchers’ aim to produce a quantitative analysis of the global news cycle. They worked with a data set of 90 million news and blogposts collected over the final 3 months of the 2008 US presidential election. Apart from a global analysis, they also aim to carry out a local analysis of a meme and it’s variation and mutation during it’s propagation and peak.

A. The Approach

The researchers’ approach is to first identify quotes, and phrases that can be identified with a particular topic. Then they develop a clustering algorithm, to cluster together different variations of the same phrase.

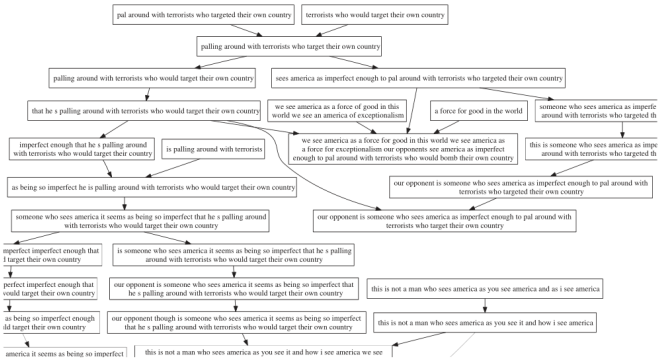

Figure 4: This phrase graph shows the evolution of a quote

1) Phrase Graph: As shown in Figure 4, an example of a phrase graph which shows the mutation and changes in a phrase over time, while the central core idea of the phrase remains the same. The researchers' goal was to build similar phrase graph where each phrase is represented by a node and directed edges connect related phrases.

2) Partitioning the phrase graph: The goal is to recognize a good phase cluster. The researchers are looking for a collection of phrases related closely enough such that they can be described as belonging to a single long phrase, or a collection of similar phrases.

B. Threads

The researchers make threads using these phrase clusters and use them to track the news cycle. The two main ingredients for tracking the news cycle are - imitation. That is, when a particular source gains significant traction, it is likely to be copied or imitated by other threads, and the source is likely to persist and grow due to adoption by other threads. The second is, recency. Newer threads are favored over older ones. The researchers come up with a model to analyse the news cycle on the basis of these two key ingredients.

C. Findings

Local Findings. These are the findings associated with the patterns of particular threads. In general, one would expect the overall volume of a thread to be very low initially; then as the mass media begins joining in the volume would rise; and then as it percolates to blogs and other media it would slowly decay - this is not true.

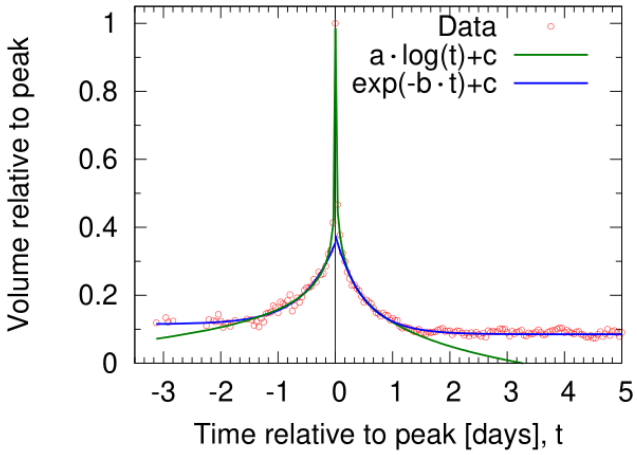

Figure 5

As can be observed from figure 5, the rise and drop of popularity is distinctly symmetric, there is no evidence for a rapid rise and a quick decay. Drop in volume is also very symmetric around the peak. The baseline popularity of the thread is also lower after the peak.

Global Findings.

1) Time lag between media and blogs: A common assertion about the news cycle is that quoted phrases first appear in the news media, and then diffuse to the blogosphere, where they dwell for some time. A quoted phrase first becomes high-volume among news sources, and is then "handed off" to blogs. The news media are slower to heavily adopt a quoted phrase and subsequently quick in dropping it, as they move on to new content. On the other hand, bloggers rather quickly adopt phrases from the news media, with a 2.5-hour lag, and then discuss them for much longer. Thus we see a pattern in which a spike and then rapid drop in news volume feeds a later and more persistent increase in blog volume for the same thread.

2) Hand-off of phrases from news media to blogs:

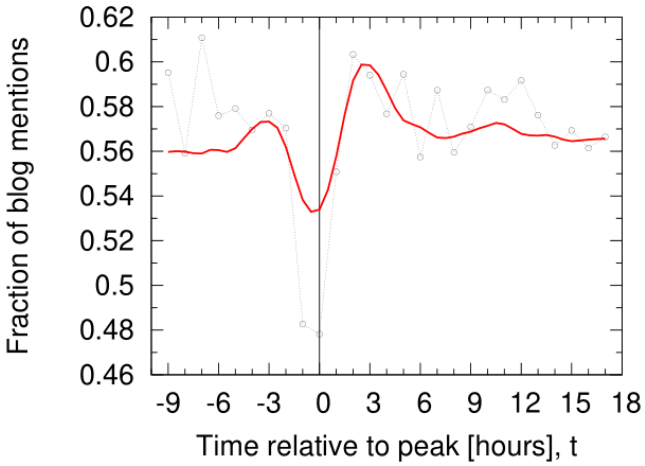

Figure 6

Figure 6 shows a "heartbeat"-like like dynamics where the phrase "oscillates" between blogs and mainstream media. The fraction of blog volume is initially constant, but it turns upward about three hours before the peak as early bloggers

mention the phrase. Once news media joins in about t=-1, the fraction of blog mentions significantly drops, however it significantly increases again after news channels drop the news and bloggers continue talking about it.

We find that the peak of news-media attention of a phrase typically comes 2.5 hours earlier than the peak attention of the blogosphere. Moreover, if we look at the proportion of phrase mentions in blogs in a few-hour window around the peak, it displays a characteristic "heartbeat"-type shape as the meme bounces between mainstream media and blogs. We further break down the analysis to the level of individual blogs and news sources, characterizing the typical amount by which each source leads or lags the overall peak.

D. Limitations

Again this paper is limited by the age of it's dataset, the nature of the web has changed since the publication if this paper in 2009, while the implementation of phrase partitioning and threads is innovative, the results from this research are largely constrained by their age.

VI. Our Conclusions

- The news cycle can be compared to an ecosytem, where each particular thread/or a meme is competing for resources (attention, clicks, occupying articles and posts).
- The primary method through which a meme gains traction is through mainstream media outlets, which mostly accounts for the peak in popularity of any given meme. However, for the longevity of any particular meme, it's discussion by bloggers and smaller users is essential for a rebound in popularity.
- This does not account for all memes however, there are memes that are first popularized by blogs and then later picked up by news outlets.
- Individual users on websites like twitter, do not have as much of an impact on online discourse as more influential players like news websites and large blogs, the discourse on twitter is still largely influenced by news outlets.

VII. Scope for further research

- Primarily, further research can be done to replicate this research on more modern datasets, to accurately use the results for the version of the web that we currently have.
- These papers present useful computational tools that can be implemented on modern datasets aswell.
- The change in the nature of the web should be accounted for as well, for instance, in the modern web, news often starts on social media sites like twitter, youtube, and then is picked up on by mainstream media - this is very different from the way the web operated in 2008. Since, a large portion of the news and discourse now focuses on events on the internet, the source is websites like twitter, and youtube - in contrast with a decade ago when discussion on the internet was largely about real life events which were reported on by mainstream media outlets before anyone else.

References

[1] Yang, Jaewon, and Jure Leskovec. "Patterns of temporal variation in online media." In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 177-186. 2011.

[2] Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 497-506. 2009.

[3] Matsubara, Yasuko, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. "Rise and fall patterns of information diffusion: model and implications." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 6-14. 2012.