News Patterns: how press interacts with social networks

Guilherme Bertini Boettcher
Instituto de Informática - UFRGS
Caixa Postal 15064
Brazil 91501-970, Porto Alegre, RS
maiona@gmail.com

João Comba
Instituto de Informática - UFRGS
Caixa Postal 15064
Brazil 91501-970, Porto Alegre, RS
comba@inf.ufrgs.br

Carla Dal Sasso Freitas
Instituto de Informática - UFRGS
Caixa Postal 15064
Brazil 91501-970, Porto Alegre, RS
carla@inf.ufrgs.br

ABSTRACT

Social media has played a big part in the adaptation process for newspapers and magazines, but innovating while going through a recession has led to a hasty evolution and automated processes for very different media. While existing social media studies and state of the art visual solutions are available for analyzing social media content and users' behaviors, no other method is optimized for finding patterns from a popularity standpoint in the specialized realm of news channels. In this paper, we propose the usage of a combination of different visualization techniques that co-relate the profile's and its reading community activities with the resulting popularity. For the period of three months, we gathered Twitter posts, the number of followers and trending topics from worldwide press profiles. We used this dataset as the seed for our bar charts, tag clouds and bubble charts to allow for multiple source comparison, so that not only the user is able to understand their own community but also the success and pitfalls faced by the competition in the same medium. We validate our analysis by interviewing a group of journalists from different established newspapers. Through interacting with our system, it was possible to reveal hidden patterns in the massive dataset of messages and comments worldwide enabling the user to have unique insight into their community's behaviors and preferences.

Keywords

information visualization; social media; temporal patterns; diffusion patterns.

1 INTRODUCTION

In the last few years, we have witnessed a dramatic change in the way newspapers and magazines communicate, as well as in the time events take to spread around the world through the Internet. The decline in sales of print media has forced the press to adapt it's business to a more current media.

Facebook and Twitter, created as a social network and a microblogging service respectively, are popular social media venues that are recognized as relevant broadcasting and influence tools. Nowadays, they are commonly used by the press to generate interest for their published material and broadcast news. However, the adaptation from an established media to a new technology was hasty and for a variety of reasons, most newspapers simply utilize automatic publications to share their content on social media, with little to no research on their readers usage of such media.

The problem of identifying behavioral patterns in social media from a popularity standpoint is applicable in varied fields other than news. It can bring interesting insight for the fields of politics, entertainment or any other popularity driven networks that can benefit from finding key behavioral patterns in a social media.

In our research, we found solutions for parts of the problem optimized for a different general group of users. Among them, the works by Guodao et al. [SWL+14], Yafeng et al. [LKT+14], Yingcai et al. [WLY+14] and Zhao et al. [ZGWZ14] [ZCW+14] are the closest to a full solution. We were not able to find a complete solution for the problem focused on the press community from a popularity standpoint.

In this paper, we propose the usage of a combination of different visualization techniques that co-relate the profile's and its reading community activities with the resulting popularity.

We organize our text as follows: in session 1 we introduce the problem and present the applicable scenarios; in session 2, we discuss the recent and state of the art techniques, their advantages and shortcomings; in session 3, we present our proposed solution in detail; in

 ISBN 978-80-86943-66-4

session 4, we analyze the experimental results obtained in three different use cases and expert reports; in session 5, we provide a review of the technique, its advantages and shortcomings and final comments; and in sessions 6 and 7, we present our acknowledgments and references.

2 RECENT SOLUTIONS

There are several works dealing with social networks analysis and visualization. As mentioned before, most are relatively simple tools to obtain statistics about users and the overall network. Literature closely related to our work can be roughly divided in two categories: analysis of information diffusion processes and visualization of diffusion patterns.

2.1 Analysis of information diffusion processes

Social networks have been studied for years [WF94], but online social networks, blogs and microblogs introduced challenges in the investigation of how people communicate using these media. The analysis of information diffusion in such networks involve measuring quantitative characteristics [KLPM10, YW10], finding relations between structure and dynamics [LG10, YC10a], predicting characteristics of the diffusion process [YC10b] and trends detection approaches [MK10, CTB$^+$12].

In addition to processing the collected data, most of these works rely on showing static plots to display the values of the metrics they are concerned about. As for trend detection, the tools must process information in real-time. For example, Twittermonitor [MK10] produces a webpage reporting recent trends in real time and provides an interface for users to rank trends according their own criteria.

2.2 Visualization of information diffusion patterns

Besides the many tools that provide graphical ways for monitoring social media activity, there are recent works that propose the visualization of information diffusion patterns [KLPM10, YC10a, CTB$^+$12, SWL$^+$14, LKT$^+$14, WLY$^+$14, ZGWZ14, ZCW$^+$14]. Among them, works of Sun et al. [SWL$^+$14], Lu et al. [LKT$^+$14], Wu et al. [WLY$^+$14] and Zhao et al. [ZGWZ14, ZCW$^+$14] are the closest ones to a full solution, and we will restrain ourselves to briefly describe them.

Sun et al. [SWL$^+$14] and Wu et al. [WLY$^+$14] aim at analyzing topics coopetition in social media (most notably, Twitter) and answer who exerts the greatest influence on a highly cooperative topic that used to be a competitive topic, what are the similarities and differences in the roles of groups of issue publics and how often they divert attention to other topics. Their tool summarizes dynamic topic competition and compares topic leaders to topics by utilizing the Theme River technique described by Havre et al. [HHN00], however is focused on the patterns of cooperation versus competition of different themes and does not contemplate the analysis of individual profiles.

Lu et al. [LKT$^+$14] describes a framework for predictive models using social media (IMDB.com, Twitter and Youtube) in an attempt to create a tool that enables non-domain experts to be competitive with experts in a given area. The tool combines line, bar, bubble and candle charts, parallel coordinates and a tag cloud with sentiment analysis as means for the user to explore the information available in the mentioned media and choose from calculated metrics to predict the popularity of movies in their opening weekends. Even though they provide a myriad of visualization techniques, their solution does not enable the user to identify patterns for the increase or decline of popularity of the subject.

Zhao et al. [ZGWZ14] describes a comprehensive tool for analysis of behavior emotion and ultimately mood of a given person in social media (most notably Twitter) over time. They provide a multi-dimensional emotion analysis tool with the ability of extracting emotional episodes and infer longer-lasting moods through an enhanced implementation of Havre et al. [HHN00] and rich interaction, however this analysis is not applicable to press profiles, since they use their microblogs to increase visibility of their online content, instead of commenting on their own day-to-day activities and is not scalable or reliable.

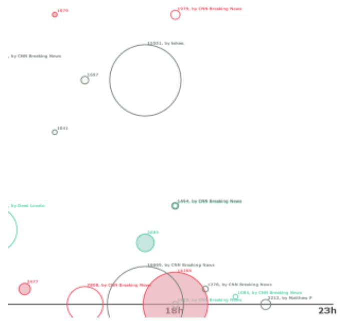
Finally, Zhao et al. [ZCW$^+$14] describes a system for the detection, exploration and interpretation of anomalous conversational threads in Twitter. This solution is applicable to the news environment, as we could consider an abnormal increase or decline in popularity of a given profile as an anomaly and apply their algorithms to further explore the available data. The problem lies in the nature of the anomalies which is the lack of a clear definition. A thread may be considered abnormal when it disseminates a message differently from the patterns of other threads in a similar topic. This is not necessarily true, especially when we take a channels popularity into account. Common subjects are covered by different profiles in unique ways that would be considered anomalous by their solution.

3 PROPOSED SOLUTION

We proposed the use of a combination of bar chart, bubble chart, tag clouds and message boards coupled with responsive interaction among the combined visualization techniques to bridge the gap left by other studies in the same area.

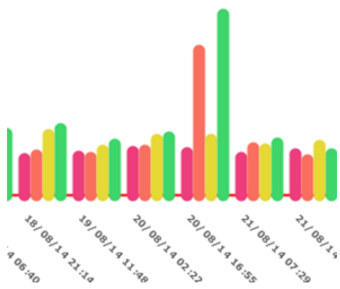
 ISBN 978-80-86943-66-4

(a) Tag cloud of words used by news source with color-coded sentiment.

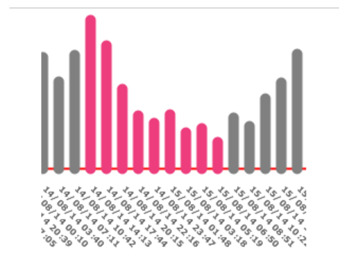


(b) Bubble chart with re-tweeting activity color-coded by sentiment.

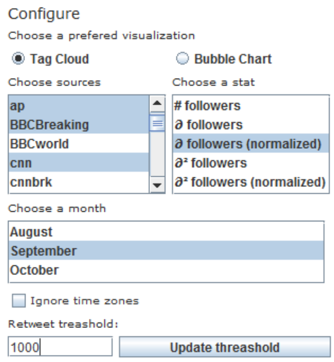
(c) Message board color-coded by sentiment with tweet and user info.



(d) Followers bar chart with multiple sources selected.



(e) Followers bar char with date and time window selected.



(f) Configuration panel.

Figure 1: Different modules of the News Patterns

We combined analytical and statistical information from original posts by news profiles and readers alike, channel popularity information and trending topics. We included configuration options that allow the user to filter out any activity that falls outside of the subject of interest and concentrate on the actions that closely relate to the popularity spikes and valleys found by the user. We took the user's expertise into account when defining which are the abnormal activities in the available dataset, since automatic search for such patterns falls into the nature of true anomalies for which there is no clear definition.

In order to gain insight into the daily work of the target users for this system and how it could benefit them in obtaining better results from their efforts in social media, we conducted interviews with journalists from different major newspapers. Based on the gathered information, we tailored our solution to answer the key questions posed by the group of journalists:

1. Why do readers stop following a profile?
2. Does the time of the post co-relate with the number of retweets?
3. How does the profile relate with it's network?
4. Is there correlation with posting subjects and the amount of followers/popularity of a profile?

After creating a working prototype, we shared our proposed solution and monitored their usage. We mention their findings and analyze their reports regarding the tool in the results chapter.

3.1 Data gathering

Data was gathered from Twitter over the period of 3 months with the use of the Firehose API, by filtering the messages by original author, retweeted status author or user mention to match the list of 19 news source Twitter IDs spread worldwide. The total numbers amount to 15 million original tweets from August 2014 to October 2014, averaging 5 million publications per month and over 10.000 posts per source per day. Each status update object is stored raw in a JSON format on plain files. Complementing this dataset, further data was gathered from each source, containing the number of followers every 30 minutes for the same time period, resulting in an average of over 1000 snapshots per source per month. Finally, a third and separate dataset gathered the trending topics from the 15 available locations in Brazil every 30 minutes, resulting in 495 unique trending topics over the three month period (an average of 165 unique trending topics per month).

 ISBN 978-80-86943-66-4

3.2 Tag Cloud

The tag cloud shows the most recurrent words and the sentiment of the messages relating to them for the selected profiles during the time period defined by the user.

The visualization is implemented as displayed in figure 1a, with either the words from selected profiles or with a localized list of trending topics. Like most implementations of this technique, occurrence of the word is mapped to it's font size, so that the most commonly used words or tags will be largest as well. Most users will be using this visualization to understand which subjects were of interest in the selected window of time, so to augment the importance of each term, the most common words are also listed first in the visualization.

The sentiment attached to each word will be calculated from the messages that contain them. We used the SentiWordNet 3.0 lexical resource for opinion mining, which assigns to each synset of WordNet three sentiment scores: positivity, negativity and objectivity. After we calculated the accumulated sentiment of all messages, we color-coded the word in a green hue if the result is positive and above a configurable threshold, red if the result is negative and below a configurable threshold or gray otherwise. Restraining the seeds to account only for the profile's original posts will create a picture of their sentiment regarding a specific topic.

3.3 Bubble Chart

The bubble chart displays the tweeting activity as shown in figure 1b for the selected source and time period. The X axis maps the hour of the day when the message was posted and the Y axis to map the popularity the selected profile(s) had during the analyzed period, ranging from the smallest to the largest amount of followers.

The size of the bubble maps the number of re-tweets for the original message in the moment of the post. As part of the nature of the fire-hose API, we do not have guaranteed access to all the messages being posted at any time, but we can compare the number of re-tweets of the original post to get an idea of how popular each message became over time.

The bubbles are color-coded with the information as discussed in 3.2, with the added information of agent mapped to the fill of the bubble. Readers can comment or simply re-tweet messages (I) directly from the source, but the same activity may come from (II) a separate profile that mentioned, commented or re-tweeted the news channel. In figure 1b, scenario (I) is represented by filled bubbles and (II) is mapped as hollow circles. In both cases, we include the raw number of re-tweets and in case (II), we displayed the screen name of the profile that generated the activity.

3.4 Message Board

Figure 1c shows original message information that belongs to the selected sources and date and time being analyzed. As in the other modules, sentiment is color coded as the background for each message. We included the original text, the date and time of the posting, the author's name, screen name and profile picture. This module allows us to get specific information in it's lowest granularity and rawest form, which can help the user identify the context from which the tag cloud and bubble charts were derived.

Clicking any message filters out every word or bubble from unrelated posts, enabling in-depth analysis of any conversation of interest.

3.5 Profile followers bar chart

The followers bar chart displays the patterns of popularity of the selected profiles according to the chosen metric over time as shown in figure 1d.

This module is very helpful in determining the relationship each news source has with their followers in terms of popularity. It allows the user to discover patterns in the growth of popularity of any channel and for the comparison between channels in a leveled playing field. This graph is a good starting point to discover points of interesting activity during an extended period of time. The user can select the appropriate window to restrain the information to the activity as shown in figure 1e in order to research the window of interest and the remaining panels will filter out any information not comprised in that period.

This implementation is fundamental to create a popularity-based solution, not found in any other work of similar proposals.

3.6 Configuration

The configuration panel is organized as in figure 1f, where users can choose the switch from each visualization type (Tag cloud and Bubble chart) and choose from the list of available profiles the ones they are interested in analyzing at each time. Choosing from different metrics of behavior, i.e. total number of followers, delta from each interval to the next, normalized delta from each interval to the next, delta squared and normalized delta squared of each interval. These two selections are necessary to display the bar chart of followers, shown in figure 1d.

To optimize the search performance, we separated the dataset into each month (August, September and October), so the user can easily switch between time windows of interest. The user also has the option of ignoring or taking the time zone into account while calculating the statistics. This separation is key to improve scalability.

The last item in the configuration panel is the re-tweet count threshold for the bubble chart. Only posts that surpass the minimum amount specified in that field will be displayed. This feature is particularly helpful when studying popular channels, such as CNN and the NY Times.

4 EXPERIMENTAL RESULTS

In order to determine the efficacy of our proposed solution, we invited the same corpus that was present during the initial phase of the project to test the tool and use their on expert knowledge, explore the existing data to find answers for the four main questions presented the introduction section. We utilize key use cases to illustrate their reported findings and our analysis.

4.1 Use case 1

Comparison between different sources is not particularly key to understand the patterns of a profile's following base, but it can be very helpful in determining what other news sources have been doing and how that affected their popularity.

According to the journalists we interviewed, most of their activity in Twitter is automated, which means that information about their following crowd is not regarded while posting. When a new piece is included in the print version, it is automatically added in the channel's online publication after the editors finalize the next morning's issue. After the issue goes live, automatic processes create the messages and posts them automatically at the same time.

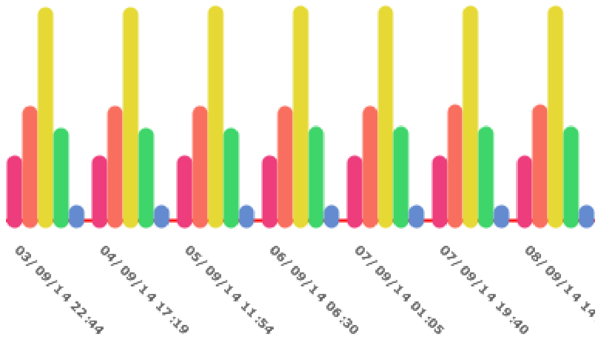
While worldwide news sources such as BBC and The Economist have a large follower count, the user is still able to compare them to local sources with a more limited reach, in order to understand if the community behavior follows a global trend or if their are driven by different motives.

As a filtering activity, the user selects the appropriate window to restrain the information to the activity as shown in figure 1e in order to research the pattern revealed and select either the tag cloud or bubble chart visualization to better understand which were the causes of that behavior.

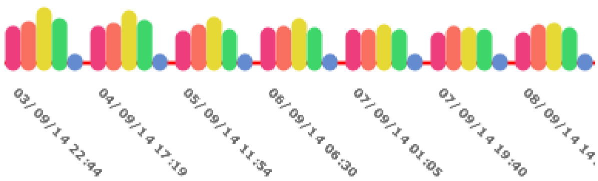
Three different views of the followers trends are shown in figure 2. There are five selected profiles (Associated Press, BBC, CNN, The Economist and The Sun) based on three different metrics (Absolute number of followers, delta and normalized delta).

It is clear that CNN has a commanding lead in popularity, while The Sun has a very limited reach on Twitter, based on figure 2a. It is also clear that the difference in popularity is sustained over time by each of the sources.

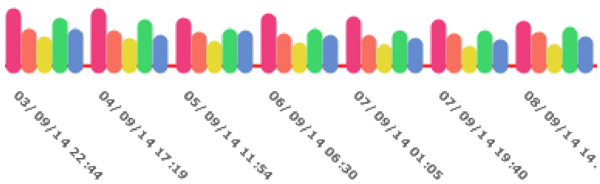
We can see in figure 2b that the number of followers increases by a similar amount for each of the profiles, The



(a) Absolute number of followers.



(b) Delta of followers.



(c) Normalized delta of followers.

Figure 2: Followers trends for Associated Press, BBC, CNN, The Economist and The Sun.

Sun being the sole exception. However, when we analyze figure 2c, it becomes clear that the effect of growth is actually greater for the Associated Press profile. This analysis means that even though it may take a long time, the Associated Press and The Economist are diminishing the gap of popularity between them and BBC.

4.2 Use case 2

To understand the relationship between a given profile and it's readers, we break their interaction into four aspects:

- Most common words used
- Time of original posts
- Time of retweets, comments and other related activity by the community

- Sentiments associated with each of the above

During the analysis of the tool, we selected Fox News as the primary source to be analyzed and noticed the same pattern of popularity repeating itself over time, so we selected that window of dates, from August 9th to 11th. Figure 3a shows the most used words for Fox News in that time.

(a) Fox news original posting tag cloud.

(b) Fox News readers comments tag cloud.

Figure 3: Comparison of original content provider and the reader community activity.

It is clear that the subjects of interest during this period for the profile were president Obama, Iraq, Robin Williams, ISIS and O'Reilly, which is not surprising since during that time, US jet fighters launched a strike on ISIS militants and Robin Williams came to pass.

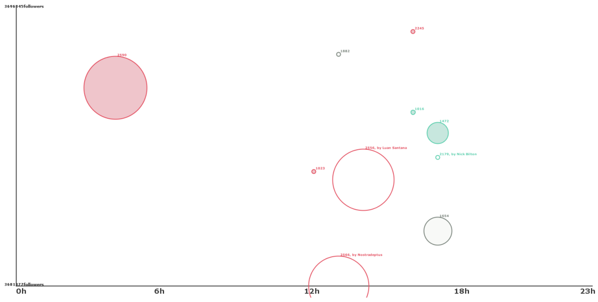
What is revealing is that even though Fox News is considered to have a republican bias, according to the analysis, they choose positive words when discussing president Obama, Iraq and the Islamics, while using negative words when mentioning their on-camera talent O'Reilly. Our users were expecting the opposite trend to be in place. Only the specific issues of "Obamacare" and Sen. Clinton are mentioned in a negative post, in confirmation to the journalists expectations.

By filtering the seeds to include the reading community's posts alone, the user is able understand their opinions in the realm of the channels messages. Figure 3b shows the Fox News' community mentions the president in a contrasting negative fashion, while confirming the position on issues like ISIS. Another interesting contrast is the related profile @foxnewspolitics, which is mentioned positively by the main account but negatively by the community. This contrast poses a very interesting insight: even though Fox News may at times be favorable to the democratic issues, they community still mentions their content in a negative scenario.

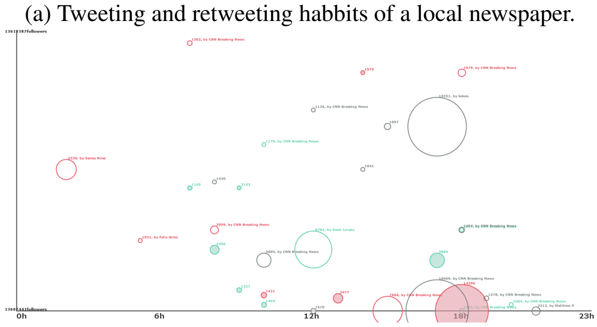
In use case 2, we will describe further cases that relate to the time patterns found for other channels.

4.3 Use case 3

The bubble chart technique allows us to understand how time and popularity influence the activity in any given message as well as to identify if when there are any key readers generating the interest instead of the original poster.



(a) Tweeting and retweeting habits of a local newspaper.



(b) Tweeting and retweeting habbist of CNN.

Figure 4: Comparison of original content provider and the reader community activity.

While the popularity of local and global news sources vary greatly, it is possible to identify similar trends in both communities: most of the activity takes place between noon and dusk. However, figure 4 shows how the local newspaper relies on separate influencer's retweeting or commenting of their post, most of the activity from the global news channel community is channeled through profiles from outside the institution. In contrast, the global source displays activity across the entire Y index, indicating that current popularity is not particularly key for generating large amounts of activity.

Figure 4a, is an interesting case for the relationship of popularity and activity. Contrary to our initial estimates, activity from the community is inversely related to the popularity of the channel. After finding this pattern in our database in repeated occasions, we investigated the relationship of the source with the community via the message board and discovered that this particular source is heavily dependent on key influencers to raise their popularity, even though this was not the case for actual activity related to their posts. Figure 4a also displays two significant influencers found in this local newspaper's community that will consistently significantly increase activity around topics of interest, generating the greatest re-tweet counts of this particular source throughout the entire three month available period. This specific discovery made by one journalist

 ISBN 978-80-86943-66-4

was deemed important to determine who are the key people that the publication can involve in order to increase its popularity.

4.4 Journalist reports

We organized a dedicated guided testing session of the tool with each of the journalists, which was comprised of a 20 minute tutorial of the visualizations and interactions followed by a 40 minute assisted free play and a final questionnaire section to determine the tools efficacy. Even though the majority of them found that performance was an issue, 75% reported that the interaction was very intuitive and provided a clear way to investigate any area of interest.

Most of the patterns found by the senior journalists were expected, based on their gathered knowledge over the years of work, however the junior journalists were more intrigued by their findings. This points to the direction that the tool may be more beneficial to unexperienced professionals as streamlined way to understand their public.

We improve on existing information diffusion pattern seeking studies by enabling (1) the discovery behavioral patterns that influence and explain increase and decline of popularity of any specific channel, (2) the discovery of key third party influencers that actively shape the community's interests and (3) the comparison of similar and distinct sources' communities in a variety of ways.

5 DISCUSSION AND FINAL COMMENTS

Our solution utilizes known visualization techniques in innovative manners, aggregating information such as sentiment and popularity to give a unique view of the behavior of the community surrounding the news source profile and their relationship with each other.

Through the use of the proposed solution, it was possible to reveal hidden patterns and gain insightful knowledge of the reading community which address real needs from the industry. The system improves the existing pool of solutions by using very clear parameters and established techniques to provide solutions to otherwise unanswered questions. While there are some gaps between the proposed motivations and our solution, based on the experts reports we understand that it shows considerable promise and believe that by enhancing the existing interaction, we will able to provide an important tool that can fine tune the way the press interacts with social networks.

The system produces convincing results for most scenarios but is not without its limitations. As mentioned in 4.4, performance remains an issue especially for live data due to the heavy calculations needed to appropriately enable the interactions between panels. Complexity analysis on our heaviest algorithm tend to $O(n^2+k)$ which, like Zhao et al. [ZCW$^+$14] poses issues with scalability. The added functionalities make the system more powerful, but also more complex. Depending on the unique patterns of a given community's activity, performance may also be affected negatively.

6 ACKNOWLEDGMENTS

We would like to thank the journalists that helped us during the development of this study. We also acknowledge the financial support from the Brazilian funding agencies CNPq, CAPES and FAPERGS.

7 REFERENCES

[CTB$^+$12] Junghoon Chae, D. Thom, H. Bosch, Yun Jang, R. Maciejewski, D.S. Ebert, and T. Ertl, *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition*, Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, 2012, pp. 143–152.

[HHN00] S. Havre, B. Hetzler, and L. Nowell, *Themeriver: visualizing theme changes over time*, Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, 2000, pp. 115–123.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, *What is twitter, a social network or a news media?*, WWW Proceedings, 2010, pp. 591–600.

[LG10] K. Lerman and R. Ghosh, *Information contagion: An empirical study of the spread of news on digg and twitter social networks*, ICWSM, 2010.

[LKT$^+$14] Yafeng Lu, R. Kruger, D. Thom, Feng Wang, S. Koch, T. Ertl, and R. Maciejewski, *Integrating predictive analytics and social media*, Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on, Oct 2014, pp. 193–202.

[MK10] Michael Mathioudakis and Nick Koudas, *Twittermonitor: trend detection over the twitter stream*, Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (New York, NY, USA), SIGMOD '10, ACM, 2010, pp. 1155–1158.

[SWL$^+$14] Guodao Sun, Yingcai Wu, Shixia Liu, Tai-Quan Peng, J.J.H. Zhu, and Ronghua Liang, *Evoriver: Visual analysis of topic coopetition on social media*, Visualization and Computer Graphics, IEEE Transactions on **20** (2014), no. 12, 1753–1762.

[WF94] Stanley Wasserman and Katherine Faust, *Social network analysis: Methods and applications*, vol. 8, Cambridge university press, 1994.

[WLY+14] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu, *Opinionflow: Visual analysis of opinion diffusion on social media*, Visualization and Computer Graphics, IEEE Transactions on **20** (2014), no. 12, 1763–1772.

[YC10a] Jiang Yang and Scott Counts, *Comparing information diffusion structure in weblogs and microblogs*, ICWSM'10, 2010.

[YC10b] ______, *Predicting the speed, scale, and range of information diffusion in twitter*, ICWSM, 2010.

[YW10] Shaozhi Ye and S. Felix Wu, *Measuring message propagation and social influence on twitter.com*, Proceedings of the Second international conference on Social informatics (Berlin, Heidelberg), SocInfo'10, Springer-Verlag, 2010, pp. 216–231.

[ZCW+14] Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and C. Collins, *Fluxflow: Visual analysis of anomalous information spreading on social media*, Visualization and Computer Graphics, IEEE Transactions on **20** (2014), no. 12, 1773–1782.

[ZGWZ14] Jian Zhao, Liang Gou, Fei Wang, and M. Zhou, *Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media*, Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on, Oct 2014, pp. 203–212.

 ISBN 978-80-86943-66-4