Survey Paper Summary On Dynamics of News Events and Social Media Reaction

Survey Paper: Dynamics of News Events and Social Media Reaction [1]

Team Name: Observer
Team Member: Hongbo Zhao, 604426609
Jing Zhao, 404426610
Qianwen Zhang, 004401414
Xiang Zhong, 204412666
Yang Pei, 304434922
Zhe Sun, 604435430

Time: Mar. 12th, 2015

1. Abstract

The increasing popularity of social media such as blogs and social networks has triggered great interest in sentiment analysis and opinion mining. With the rapid spread of user-generated contents such as reviews, ratings and recommendations, these online opinions can better serve to analysis activities, ranging from reflecting political preference to improving stock market prediction. Our survey paper tries to extract a variety of topics in news events and utilize the time sequence of the corresponding emotional relevance to reveal changes in mood and dynamic connections between events. In this survey summary, we talk about formation of the problem, the main models, some other related works and our conclusion and discussion.

2. Introduction and Background

With the advent of Information Era, the World Wide Web has changed the way we live and communicate with each other and how we manage and interact with information. We can now not merely just read the information on the screen but instead we can dynamically exchange opinions within the community and share our own views on certain topics. Due to the nature of news and opinions sharing, mediums such as blogs, forums and social networks has become a rich source of useful information. For example, a user wants to buy a book about data mining but do not know which one to choose. He then can search the keywords on Amazon and browse the reviews related to hit results about whether the specific book is worth buying or not; Apple inc. can monitor the crowds' expectation about the new product Apple Watch before they make the policies related to the pricing and product functionalities. In this way, it is crucial to mine such sentiments and opinions in order to better take advantage of the useful opinionative information hidden in these sources.

Crucial as it is, these subjective data is actually very difficult to be retrieved without the help of search engines due to the fast growing of news and information volume on the web. More importantly, it requires techniques and tools to understand the underlying mechanism, which affects the propagation of news and drives the evolution of sentiments.

In our survey paper "Dynamics of News Events and Social Media Reaction", it focuses more on exploring what causes the crowd's sentiment to change. It utilizes the current research on sentiment analysis and contradiction detection and try to mine the relationship between news events dynamics and modeling. The method is to represent publication dynamics as the result of the interplay between the original news' importance and social response.

3. Survey Summary

In this part, we give a summary of the survey paper: Dynamics of News Events and Social Media Reaction. The paper proposes a novel framework, which models the behaviour if news and social media in response to events as a convolution between event's importance and response functions, given specific media and event type. The paper reveals the connection between sentiment shifts and event dynamics. The key problem of the problem is that social

media can contribute to the shift of volumes all by themselves without any external stimuli and maintain a trending volume growth over long time periods. These effects can distract the observed event dynamics and even make them detectable. In this paper, the modeling defines the external factors from social media as a special “response” function. It represents the publication dynamics as the interplay between the inner importance of original events and social response. This modeling leads to the possibility of recovering the inner importance and its varying importance in time series.

3.1 Models

3.1.1 Framework for the model

We use the following figures to demonstrate the most important procedures of the model method proposed in the paper. Figure 1 shows the relationship between average sentiment S(t) and sentiment feature s(t), which are extracted from real life data. After using deconvolution to remove noises, the relationship is more clearer than using the original data only.

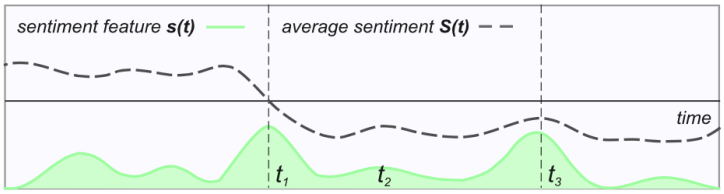

Figure 1

Figure 2 shows the news volume in time series. We can see that there is a time lag between the volume and shift of sentiments. This is due to the fact that the peak intensity of publications does not always coincide with the beginning of the events. To separate the media reaction from the importance of event, researchers follow the procedure of deconvolution.

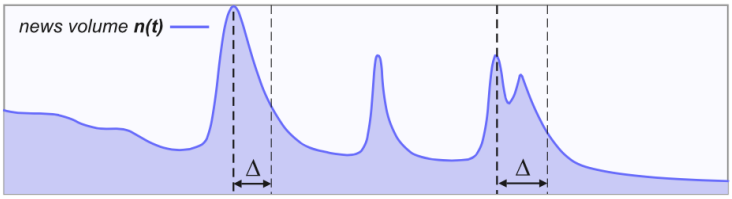

Figure 2

They model the media reaction using a media response function as is shown in the left part of the Figure 3. By doing deconvolution between news volume and media response, they get the inner importance of events.

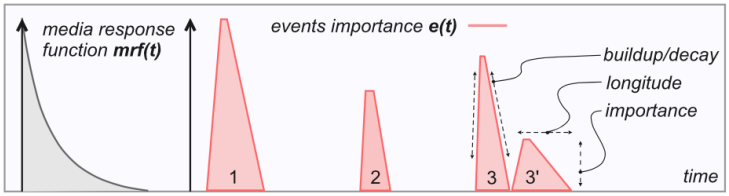

Figure 3

3.2 Proposed Method

The overall architecture of the system is divided by two main layers: News layer and Sentiment Layer. They aggregate the volume of the news for a topic in a time series in News Layer, which will be further analyzed to detect news event. To detect a news event and extract its features, we perform a deconvolution of news volume time series. The sentiments layer aggregates sentiments over time for a specific topics, and detects interesting changes. These interesting changes can refer to contradictions, outbursts of sentiments' volume, or other situations.

3.2.1 Detecting Impacting Events

Since not every kind of publications outbursts is caused by external news, the paper introduces a model for social media and news dynamics which can distinguish different kinds of events at a fine level of detail.

The paper model the observed news dynamics as a response of social media to external stimuli. The researchers represent the output as a convolution of two functions: the news events importance sequence and the media response functions. In order to recover the original event sequence, they perform a deconvolution of news frequency of news frequency time series. To model the behavior of media, they propose a family of normalized decaying functions such as linear functions, hyperbolic functions, exponential functions and symmetric and asymmetric functions. While deconvolution can be expressed in the functional form only for a limited number of response functions, the framework is designed to support any kind of finite integrable functions by taking use of frequency domain of signals. The paper models events by using piecewise linear approximation of event importance. There are rectangular models and triangular models, which can also be merged to trapezoid-like shapes. The height of the event on the event sequence indicates its importance, while the length describes its longitude.

Deconvolution is the reverse process to convolution. Following this procedure the authors are able to recreate the original event importance sequence. Moreover, deconvolution helps to detect and separate nearby, but distinct events, which may otherwise be considered as a single event due to their overlap. To estimate the parameters for the models normalizes the ascending or descending slopes of time series and then analyze them using either linear, power-law or exponential regression. Next, the method takes the average of the extracted parameters across the peaks, weighting them by regression error and level of importance.

3.2.2 Detecting shifts

The paper is interested in sentiment measures that are sensitive to particular kinds of sentiment changes and that can also be correlated with events. The particular methods which can be adopted to the problem are sentiment volume and contradiction level. First, the authors do sentiment extraction for a particular topic using the Apache Lucene index, by querying it for documents that contain a given topic's keywords. They used the methods

proposed by Thelwall et al [6]. to evaluate how twitter sentiment and its volume are changing before and after news events. The conclusions proposed by him is used to extract sentiment volume. The conclusions also reveal that external events usually lead to changes in sentiment and more importantly, can increase the level of contraction. Last but not the least, they compute sentiment contradiction level based on the first statistical moments of sentiment values. The intuition for the measure is that when the average sentiment value is close to zero, while the sentiment variance is high, then the polarization of sentiments is high, which indicates the contraction and polarization.

3.2.3 Correlating News and Sentiments

The researchers observe that different sentiment and news measures require different correlation methods, which also consider a time lag between the time series. In the case of continuous time series, which usually deviate around their average values, we can use the Pearson cross-correlation coefficient, which is defined as the normalized covariance of two time series. However, Pearson correlation is intended to determine a linear dependency between variables, which is hardly observable for bursty time series such as unexpected events or sentiment contradiction. This kind of scenarios can be measured using Jaccard coefficient to achieve better performance.

3.3 Experimental Evaluation

They analyze the news dynamics and evaluate our models on several social media datasets with different characteristics including Meme Dataset and Twitter Dataset. The Meme data set represents various events of endogenous nature. While some of the analyzed memes have a connection to real news events, the largest part of them are just sticky phrases gaining popularity from time to time. Twitter data has a distinct bias towards current events and temporal activity of users. And it shows different types of dynamics present at the same time. Using these data with different characteristics, the evaluation follows the following three steps to evaluate the result of the study:

- Compare accuracy to previous models.
- Evaluate the proposed response dynamics.
- Check the correlation of events and sentiments.

To compare the accuracy of the models for media in different medias, they perform news volume deconvolution using automatically extracted parameters, and quantity the accuracy of fitting in a way that is comparable across different time series. Errors in fitness are measured for every peak, and then normalized by its height, so guarantee that the results can be averaged and compared across peaks.

4. Related Work and Models

In this part, we research some other related papers about the several subproblems in our survey paper: Event Model, Detecting Impacting Events and Detecting Sentiment Shifts. These 3 subproblems are the most important parts in our paper. We want to know the other

approaches and models in these parts. Meanwhile, we also discuss about the motivations, advantages, disadvantages and possible improvements in each model.

4.1. Event Model

4.1.1. *Dynamical classes of collective attention in twitter* [2]

In the paper "Dynamical classes of collective attention in twitter", authors study the reason to highlight some topics of news in Twitter. The driven force can be classified as simple as 2 classes: endogenous and exogenous. The former one means the inner impact from the news itselves while the latter one represents for the externer impact from social media and other factors. The whole paper tries to figure out the relationship between the popularity of news and the news sentiment. The popularity peak from the Twitter hashtags are classified into 4 classes: expected impacting, expected non-impacting, unexpected impacting and unexpected non-impacting. For example, tsunami news reports can be classified as unexpected impacting because the catastrophe is unexpected and the click rate of the news increases in high speed when people realize the disaster and want to focus on the damage it leads to. To the contrast, the queries of Harry Potter can be classified as expected impacting. Before J. K. Rowling published the famous book, people already got news and promo about the book. After publishing, people find that the book is extremely interesting and invite their friends to read. Thus, more and more people want to know the content summary and the reviews about the book, leading to the fast increasing speed of click rate of the queries in the search engine.

They don't propose a high-level universal model. Instead, they specify all the situations and want to learn which situation is most popular among huge amount of different topics of data. I think their motivation is quite obvious. If we can learn that the external factors are more important, we can add some external stimuli by our own to produce a highly popular news event in this way. The research results can benefit some news providers. However, it is too unrealistic to some extent. The inner and external factors can be various. New factors appear every year, too. Thus, we may not get a clear answer which situation can dominate the others. In this case, the result is useless. I think they should combine the inner and external factors. They can create a new model by learning the weights of endogenous and exogenous factors using EM or some machine learning algorithms. We can score each inner and external factor. Using the model, we can calculate the correlation between the current news and our "popular" model.

4.1.2. *Robust dynamic classes revealed by measuring the response function of a social system* [3]

The authors of paper "Robust dynamic classes revealed by measuring the response function of a social system" also discussed about the news sentiment based on the user response using the time-series data from Youtube. Similar to the other event-model analysis, the paper discusses deep into the endogenous and exogenous factors from news events. The authors classified the model into 4 classes: Exogenous Subcritical, Exogenous Critical, Endogenous Critical and Endogenous Subcritical. The meaning of the 4 classes are quite the same as the classes defined in the paper "Dynamical classes of collective attention in twitter". The

difference is that the authors want to conclude the user response function model into mathematical formulas, which can be more simple, clear and precise. According to power-law distribution describing human activities, the authors extends the response function into 4 specific forms.

Their research is quite similar to the previous one. They propose a general form of the social response function figuring out the inner and external stimuli. However, they want to go deep inside the origin of the problem. Thus, they classify the inner and external factors into 4 classes. In this way, they want to learn which response function works better than others. According to their research, we can simply use the best response function. However, if we look at the 4 mathematical formulas, we may figure out that they just share slight differences. That is to say, we can relax the 4 cases into one general case. Although we may lose some precision, the final result will not change due to the slight differences among each situation. This is the solution the authors of our survey paper use. They relax the user response function into a general form. Meanwhile, they unify several other subproblems. Thus, they create a model to relax several kinds of possible situation into one. I think this is the improvement. In this case, we can handle more types of data though we may lose a little bit precision according to the relaxation.

4.2. Detecting Impacting Events

4.2.1. *Investigating query bursts in a web search engine* [4]
In the paper "Investigating query bursts in a web search engine", the authors want to track, analyze and predict the possible hot topics of some specific news in the search engine. To detect the most interesting topics, the researchers use the aggregation of event volume with both user query attentions and the news event itself to detect the query burst. When detecting the inner and external stimuli, the researchers think it will produce a burst increase of the click rate in the future. The results show that the method work very well and can classify the queries into 3 classes: bursty, random and stable. The paper want to learn a model using the previous time-series data to train some features in the 3 different classes. In this case, the model can work well for the future unknown news.

I think this paper is quite useful for the news providers to analyze the click rate and popularity of some topics of news events. If they use the system to learn about how to write some keywords or titles of some burst news can increase their page views, they can get more ads income due to the increase of page views and click rate. However, the system cannot work well if there is no external stimuli. In another word, the news event can be a hot topic by its inner factors, independently. That is to say, we should also face the condition that the news volume has noise inside, which cannot reflect its original "popularity volume" in this case. Thus, we should use some methods to cut off the negative noise inside the original data to get more precise impacting events.

4.2.2. *Inverse Problems in Physical Diagnostics* [5]

The book called "Inverse Problems in Physical Diagnostics" introduces signal model to exact original messages from complicated data including noise. This process is called "inverse problem", which is quite popular in the signal processing field. This problem is related to the media parameters, which should be solved using Fourier transformation according to the signal theorem. This can be used to do the data preprocessing to get clearer and more precise data from the original ones. The analysis tells us that it is possible to use numerically-computed Fourier transformation of any decay function. That is to say, we can use the Fourier representation of response function and other external and inner factors to filter out the noise part.

I think this solution is quite practical due to the efficiency and correctness. According to the signal theorem, Fourier transformation can be widely used in separating noise and original data. The process "convolution" and "deconvolution" are just the counterparts. If we deconvolute the data firstly, we can easily recover the data using Fourier transformation to do "convolution" with lossless data. The disadvantage is that the assumption is that the massage and news event volume can be treated as signal information, which may not fit all the situations in the real world. But according to the previous analysis and results in the physics field, the Fourier representation model can work perfectly in most cases.

4.3. Detecting Sentiment Shifts

4.3.1. *Sentiment in Twitter Events* [6]

In the paper "Sentiment in Twitter Events", the authors propose a new model detecting the impacting news sentiments base on the time-series Twitter data. They exact the top-k keywords from the Twitter messages and calculate the term frequency and inverse document frequency. To pick up the burst news, they should find out the "emerging topics" according to the frequency-matching situations. Their goal is to check whether or not the popularity has positive correlation to the increase of news sentiment. They propose the conception of sentiment strength. The results are good to find out the impacting and non-impacting news sentiment.

I think the method is quite creative referring to TF-IDF algorithm in the classification area. Because the words limitation in the Twitter messages, they can exact top-k keywords in a fast speed although the history data is huge. However, learning words is not an easy thing. People may have some typo errors or emphasis words in their tweets, which can be misclassified. For example, "happy" and "happppppppy", which emphasize the exciting mood, should be the same meaning. "Yamadata" and "Yamatada" may be the same word, which the unfamiliar japanese pronunciation leads to the typo errors. Meanwhile, they don't consider the impacts of social media reaction and user responses. When social reactions attribute to the news volume all by themselves, news sentiment is undetectable according to their model.

Meanwhile, the sentiment strength they proposed is base on the statistical popularity of sentiment volume. In some cases, it is obvious and can work very well. However, if we consider the case the average sentiment value is very low while the variance is extremely

high, we cannot detect the sentiment shifts when we use their methods. That is to say, we should also take care of the variance of sentiment values, which is really important when data is unbalanced.

4.3.2. *Scalable Detection of Sentiment-Based Contradictions* [7]
In the paper "Scalable Detection of Sentiment-Based Contradictions", the authors propose a new measure to detect the sentiment contradiction efficiently. The concept "contradiction" means the "emerging topics" we previously discussed about. Compared with the calculating sentiment shifts by using volume values and variance independently, they derive a formula combining mean and variance value given different weights. The new formula works well evaluating strong contradiction in most of cases.

I think the approach is quite useful for research of sentiment shifts. This is the first systematic model to the problem. The universal mathematical formula is derived gracefully, too. However, they don't use EM or other machine learning algorithm to train the appropriate weights. Instead, they just give a hard-code weight parameter based on the history data and experience. Although it can work well in most cases, it cannot dynamically adjust the parameter all by itself when new data is coming. The improvement is to train the parameters and update the parameters when new data is added. In this case, the formula can be more widely used into different uses.

5. Conclusion and Discussion

5.1 Contribution

The paper is the first work that proposes a principled modeling of news dynamics in various media and news interaction with sentiments. The major contribution can be summarised as follows:

- Analyze the existing models of publication dynamics, and discuss their fundamental principles.
- Model the dynamics of news volume as a convolution between events importance and media response function.
- Develop a method for news events extraction, based on deconvolution with automated parameter optimization.
- Access several sentiment features for their correlation to news events, and the possibility of predicting their changes.

The model can accommodate various response functions, suitable for different cases, which should not necessarily be expressed as a differential equation, but can be learned from the data. The results obtained by applying our methods to different real datasets confirm their robustness and universality.

5.2 Challenges and future work

While the methods enjoys many benefits, it also faces several challenges. We observe that while different media have preferences for particular response dynamics, these are often determined by event types and topics. Thus, we need to extend our method of news volume deconvolution so that it will automatically determine the best model for every particular event and process the corresponding time interval individually. This involves a refinement of the events importance model and development of a robust and precise deconvolution optimization strategy.

We also observe the existence of different parameters of response dynamics for various events even during the same topic time series. In order to enhance the accuracy of prediction of sentiment changes, we need to take into account the type of response dynamics in addition to the event's importance level, creating a more elaborate causality model. Still, predicting the types of these possible changes requires building a database of event and sentiment shift profiles, and constructing a classifier model based on these features.

Finally, we speculate that sentiment shifts may be caused by events on related topics, in addition to events on the same topic. This leads to the necessity for a more advanced correlation and causality modeling, in order to predict sentiment shifts across related topics.

6. Acknowledgement

We thank our roommates and other students who also take CS249 this quarter for their discussion and feedback about our written summary.

7. Reference

[1] Tsytsarau M, Palpanas T, Castellanos M. Dynamics of news events and social media reaction[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 901-910.
[2] Lehmann J, Gonçalves B, Ramasco J J, et al. Dynamical classes of collective attention in twitter[C]//Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 251-260.
[3] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system[J]. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653.
[4] Subašić I, Castillo C. Investigating query bursts in a web search engine[J]. Web Intelligence and Agent Systems, 2013, 11(2): 107-124.
[5] Gaikovich K P. Inverse problems in physical diagnostics[M]. Nova Publishers, 2004.
[6] Thelwall M, Buckley K, Paltoglou G. Sentiment in Twitter events[J]. Journal of the American Society for Information Science and Technology, 2011, 62(2): 406-418.
[7] Tsytsarau M, Palpanas T, Denecke K. Scalable detection of sentiment-based contradictions[J]. DiversiWeb, WWW 2011, 2011.