Impact Analysis of the Person in Topic Event Mining

Fenghuan Li, Dequan Zheng*, Tiejun Zhao

MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001, P. R. China
{finelee2012, dqzheng2007}@gmail.com, tjzhao@hit.edu.cn

Abstract. Impact analysis of the person can help us to understand the role that this individual plays and the development of the event. This paper focuses on a study of persons’ impact on events. The problem is addressed by exploiting concepts coming from physics and using indicators typically exploited in stock market analysis. We assess person’s impact on a specific topic event by expressing ‘momentum’ and referring to a key period as the interval when this impact is profound and is kept relatively stronger. Key features of the model (viz. the impacts and key periods) are estimated by the combination of Moving Average Convergence Divergence (MACD) and MACD histogram drawn from stock analysis. Experimental results reported for four well-known topic events in China show that the model could achieve desirable result in analyzing person’s impact on the topic event.

Keywords: Impact analysis; Topic event; Momentum; MACD; Histogram

1 Introduction

The analysis focused on ‘person’ becomes an important task in data mining area, such as information extraction [1,2], information retrieval[3,4], machine vision and application[5,6], public opinion analysis[7] and social networks[8]. But, there is few studies on ‘person’ in topic event. We study a problem about a person’s impact on topic event, which is defined as the role this individual plays. Topic event is a set of news stories, such as “Wenchuan earthquake in 2008”. News on a topic event are posted chronologically, which are viewed as a kind of data stream, the same as one person’s information. Therefore, the impact analysis can be viewed as a problem of time series analysis.

It is generally the case that the more times a name is mentioned in a newspaper, the stronger impact this particular individual has. That is to say, a pure frequency of one person’s name may be used to evaluate his impact. But there are some limitations with this regard. First, thresholds must be used to determine the impact and identify the key period. If the threshold is too high, the period when the impact is strong may be neglected. On the other hand, the period with weak impact may be detected. It has to be noted the high person’s occurrence frequency does not directly imply a strong impact. For example, if one person is the journalist following the event, his name, which is served as a stopword, will be mentioned frequently in news reports. Therefore, the impact of this individual on the event is extremely weak.

To address this new and interesting problem, this study borrows the idea [9], which is used to detect topic burst. Though ideas are something similar, two tasks are extremely different. Essential differences are that impacts and key periods are quantified and a combination of two indicators is used in our model, while only bursts are identified and merely one indicator is used to estimate bursts. The study exhibits following original facets. (1) A novel study is proposed. The problem includes impact tracking and key period detection. (2) The problem is addressed by concepts from physics and indicators in stock analysis. (3) Person’s social attributes are considered and some disadvantages in other algorithms are avoided. (4) Parameters can be adjusted so that the model can be utilized in other applications.

The paper is organized as follows. The second section discusses some related work and baseline systems. The third section details the relevant terms and impact model. The fourth section elaborates on results of experiments for four topic events. The fifth section concludes this work with a summary.

2 Related Work

There has been increasing interest in time series [10,11], person’ action[4,6], topic event[9,12,13]. Cokol et al. [11] captured variations in scientific impact over time to compare relative significance and evolution of fields similar to a financial market scorecard. Zhang et al. [4] analyzed researchers’ behaviors over time and investigated if time series analysis is a valid method for predicting relationships between searcher actions. Kleinberg[12] modelled an infinite state automaton in which each state represented a message arrival rate to detect topic burst. Arrival rate of the word can be defined as the number of articles containing the word, frequency of the word and alike. Suppose there are \( n \) batches of documents and the \( r^t \) batch contains \( r_t \) relevant documents out of a total of \( d_t \). Let \( R = \sum_{t=1}^{n} r_t \) and \( D = \sum_{t=1}^{n} d_t \). For each state \( q_t \), there is an expected fraction of relevant documents \( p_t \). Set \( p_0 = R/D \), and \( p_t = p_0s^t \), where \( s>1 \) is a scaling parameter. The weight of one topic in an interval \([t_0, t_1]\) is defined as:
The weight is equal to the improvement in cost incurred by using state $q_1$ over the interval rather than state $q_0$. That is to say, topics with higher weights correspond to more prominent periods of elevated activity. This method could be quite limited, since more significant information could be ignored. For example, a person’s social attributes may be more important than the arrival rate of his name mention.

Many systems\cite{9,14} were based on MACD histogram, which was used to track trends. The model\cite{9} reconstructed bursts as a dynamic phenomenon using concepts originating from physics instead of focusing on arrival rates. They viewed topic bursts as intervals of increasing ‘momentum’, which is estimated by MACD. The MACD histogram reflects the difference between MACD and MACD’s average. If MACD histogram is positive, it indicates recent upward momentum is more than the previous and the uptrend is getting stronger, therefore, a topic burst can be anticipated. As an example, let us refer to Fig. 1, in which data were based on our dataset and analyzed with the use of the model\cite{9}. MACD histogram values are positive during the interval 2011/7/25-2011/7/28, therefore, there is a burst. This model can be investigated to solve the impact analysis presented in this study, however one has to proceed with caution. There may be no topic burst when histogram values were positive in the period of August 7 - August 16 (in red). The reason is that $X$ decreased from July 28, so MACD became negative (from August 1), meanwhile $X$ became equal to 0 from August 4 to 16, so MACD approached 0 gradually. This phenomenon has resulted in the positive histogram (August 7-16), but without any topic burst. Otherwise, if it is merely determined by MACD, the impact may be very weak and decrease gradually, such as the data (in blue) in Fig. 1.

3 Impact Analysis Based on Momentum and MACD

The person’s impact is the role this individual plays in the event. ‘momentum’ is an object’s movement trend in the direction of movement. It also quantifies an object’s energy when it is moving. Therefore, we can regard the impact as ‘momentum’ of the person. The impact changes with the change rate of the momentum. When the change rate of momentum is positive, the impact increases. This change rate has an interpretation of ‘acceleration’ or ‘force’, which is a basic notion present in physics. To measure the change of momentum, analysis of the stock market is incorporated here, such as Exponential Moving Average (EMA), MACD, MACD signal and MACD histogram. These indicators are used to spot changes in the strength, direction, momentum and duration of a trend in a stock’s price. ‘Impact’ is viewed as ‘momentum’ in our model. The impact of the person on the topic event can have several descriptors which are used to represent ‘momentum’ as follows:

1. Position $x(t)$: A measure of arrival rate at time $t$. It may be the number of documents, the number of documents containing mentions of the person, the occurrence frequency of this person and alike.
2. Velocity $v(t)=dx/dt$: It can be computed as the derivative of position. Velocity is interpreted as the change rate of arrival rate over time.
3. Acceleration $a(t)=dv(t)/dt$: It is the rate at which the velocity changes with time. Changes in momentum are measured with acceleration if mass is a constant value. If acceleration is positive, velocity increases and then momentum increases.
4. Mass $m(t)$: A measure of importance at time $t$. It is the person’s social attribute, which may be the grade of professional title, the grade of position, educational attainment, the age etc. It could be constant or a variable associated with time $t$. $m(t)$ is set to $\varepsilon$, which is a small positive number, for those who don’t possess any professional title or position.

5. Momentum $p(t)=m(t)\cdot v(t)$: It is the product of mass and velocity. It measures the impact of the person on the topic event. When momentum increases, impact increases, otherwise, decreases.

6. Force $F(t)=m(t)\cdot a(t)$: It is any influence that causes momentum to undergo a certain change. So it measures changes in momentum. When the force is positive, there is an increasing trend in impact.

The MACD indicator is a collection of three signals, calculated from historical price data, showing the difference between a fast and slow EMA of closing prices. Three signals are: MACD, MACD signal and histogram. MACD is based on EMA which highlights recent changes in stock’s prices. By comparing EMA values in different periods, MACD line can indicate changes in the trend of a stock. By comparing that difference with an average, histogram line can indicate subtle shifts in the stock’s trend. The time period for EMA on which an MACD is based, is flexible. But the most commonly used parameters involve a short period EMA of 12 days, a longer period EMA of 26 days, and a 9-day EMA of the difference between these two. As shown in Fig. 2, if EMA of previous 12 periods is going over EMA of 26 periods, it would indicate in recent time the market started moving upward, otherwise, going down. Signal line is the EMA of MACD for previous 9 days. If MACD line moves over signal line, it indicates the recent upward momentum is higher than the previous and the uptrend is getting stronger, otherwise, downtrend becomes stronger. The larger the histogram value is, the stronger this trend is. The above indicators are concepts in stock analysis. Since topic events are almost emergencies, reports appear in large batches. Essentially, dozens of news reports appear together every day and $m(t)$ can be a measure of one person’s impact on the topic event. Instead, we need to model this kind of phenomenon. Suppose there are $r_t$ reports in the $t^{th}$ day. In this case, similar indicators are defined as follows.

![MACD indicators of example stock during 2011/08-2011/12](image)

1. MEMA: It is a smoothing model whose weighting factors decrease exponentially. The $N$-day MEMA for a $n$-day series $X=(x_1,\ldots,x_n)$ may be calculated recursively:

$$MEMA(N)[X]_t = \begin{cases} \sum_{i=1}^{t} m(t_i)x_{it} & t = 1 \\ \alpha\sum_{i=1}^{t} m(t_i)x_{it} + (1-\alpha)MEMA(N)[X]_{t-1} & t > 1 \end{cases}$$

(2)

Where $\alpha$ represents a degree of weighting decrease, a constant factor between 0 and 1. MEMA is used to smooth out short-term fluctuations and highlight longer-term trends. $m(t_i)$ could be constant or a variable which describes person’s social attribute.

2. MMACD line: It expresses the difference between a short period MEMA, and a longer period MEMA. A short period MEMA responds more quickly than a longer period MEMA to recent changes in position $x_t$. Compared with MEMAs of different periods, MMACD can indicate changes in the trend of $x_t$. So, MMACD is a filtered descriptor of velocity $v(t)$ or $p(t)$. MMACD line crossing zero suggests the direction of average velocity is changing. MMACD of $X$ can be defined as follows, where $N_1<N_2$.

---

1. [http://zh.wikipedia.org/wiki/%E8%81%8C%E7%A7%B0](http://zh.wikipedia.org/wiki/%E8%81%8C%E7%A7%B0)
2. [http://zh.wikipedia.org/wiki/%E4%B8%AD%E5%8D%8E%E4%BA%BA%E6%B0%91%E5%85%B1%E5%92%8C%E5%9B%BD%E5%85%AC%E5%8A%A1%E5%91%98](http://zh.wikipedia.org/wiki/%E4%B8%AD%E5%8D%8E%E4%BA%BA%E6%B0%91%E5%85%B1%E5%92%8C%E5%9B%BD%E5%85%AC%E5%8A%A1%E5%91%98)
$MMACD(N_1, N_2) = MEMA(N_1) - MEMA(N_2)$  \hspace{1cm} (3)

3. Msignal line: Msignal line is a MEMA of MMACD line. It results average velocity, which is filtered again. It is defined as follows:

$Msignal(N_1, N_2, N_3) = MEMA(N_3)[MMACD(N_1, N_2)]$  \hspace{1cm} (4)

4. Mhistogram line: It is the difference between MMACD line and Msignal line. It is a measure of acceleration $a(t)$ or force $f(t)$. MMACD line crossing Msignal line suggests that the acceleration is changing direction. Mhistogram is described as:

$Mhistogram(N_1, N_2, N_3) = MMACD(N_1, N_2) - Msignal(N_1, N_2, N_3)$  \hspace{1cm} (5)

Changes in momentum are described by force, therefore, Mhistogram suggests changes in the momentum and changes in the impact. When Mhistogram is positive, the impact of one person increases, otherwise, decreases. Different from the model[9] which is merely determined by histogram, we detect the key period when impact is relatively strong by the combination of MMACD and Mhistogram. If it is merely determined by Mhistogram, there may be a problem mentioned in section 2 that there is positive Mhistogram, but no impact, such as the data (in red) in Fig.1. Otherwise, if it is merely determined by MMACD, the impact may be very weak and decrease gradually, such as the data (in blue) in Fig.1. Therefore, if $\forall t \in [t_1, t_2]$ follows that $MMACD \geq 0$ and $Mhistogram \geq 0$, then the interval $[t_1, t_2]$ is one key period.

4 Performance Analysis

The proposed model is applied to four well-known topic events in China. We select two persons in every event, including individuals and groups. Detailed information about events is shown in Table 1. Our model is compared with Kleinberg’s model[12] and He and Parker’s model[9]. $m(t)$ is set to 0.001 for those who do not possess any professional title. Experimental data were collected from China People’s Daily in 1998 and online\(^3\). Because of space limitation, we show results of only one person for each topic event.

<table>
<thead>
<tr>
<th>Topic Event</th>
<th>Persons</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Shanghai subway accident”</td>
<td>Jianguang Xu, ‘injured’</td>
<td>2011.09.13-2011.10.10</td>
</tr>
</tbody>
</table>

4.1 Experimental Results

Results are shown in Fig. 3. MMACD(12, 26) and Mhistogram(12, 26, 9) are used to predict the impact and detect relatively significant period. If MMACD and Mhistogram are positive, impact increases when one person has a significant impact. MEMA1 is 1-day MEMA, which relates to public attention to given person. For illustration, MEMA1 is normalized to fit into Mhistogram. As we can see, results generated by the constructed model are consistent with the development of every event.

Shanghai subway accident happened on September 27, 2011. Jianguang Xu, director of health bureau, kept a close eye on it. Therefore, his impact peaked on September 27. Subway operation returned to normal on September 28. A committee was formed to investigate the accident. Results were disclosed on October 6, which elicited an unexpectedly strong response. Therefore, in Fig. 3(a), his role was diminished from then on.

Wenzhou high-speed train crash happened on July 23, 2011. Last survivor was Yiyi, a 3 years old girl, whose parents died in the accident. She was seriously injured and became the focus of the society. Her hospital course and living quality were attracting more and more attention. She had undergone at least 5 operations. Each treatment had been followed, so her impact changed with every treatment as shown in Fig. 3(b).

Wenchuan earthquake happened on May 12, 2008. Its magnitude was 8 and there were frequent aftershocks. An emergency rescue was carried out from May 13 to May 20. It is shown in Fig. 3(c), MMACD of ‘Army’ increased sharply during May 13-16 and reached a maximum. As time went by, there was a falling off in rescue operations. It indicated

\(^3\) http://news.ifeng.com/mainland/special/shanghaiditixiangzhuang/content-4/list_0/0.shtml
http://news.ifeng.com/mainland/special/wzdongchetuogui/content-3/list_0/0.shtml
http://news.sina.com.cn/z/08earthquake/scroll/1.shtml
rescue operation had withstood a critical point when MMACD of ‘Army’ reached a minimum on June 5. Impact of ‘Army’ was absent after rescue.

A flood disaster took place in China from early July to late August, 1998. Jiancheng Gao, one troop’s instructor, sacrificed on August 1. Therefore, in Fig. 3(d), his impact increased greatly in the following days and MMACD reached its local maximum. After rescue, he was awarded, so his impact increased again and peaked. After the disaster relief work, his impact gradually declined and MMACD became 0.

4.2 Parameters Selection

As we can see from subsection 4.1, the impact of one person on the event can be interpreted by our model, and results of impact analysis are consistent with event developments. However, as shown in Fig. 3, there are some differences between experiment results and the reality. For example, in Fig. 3(c) and Fig. 3(d), the detected period is so long that the interval with no public attention is recognized, what’s more, in other subfigures in Fig. 3, some important intervals with much public attention are not recognized.

It is found that since the duration of one event’s critical period was relatively short and parameters of Mhistogram(12,26,9) were relatively large, some farther historical data affected the analysis. A study finds identified periods are
shorter when parameters are smaller. If parameters are too small, identified periods are so short that some key periods when the impact is strong enough are ignored. Therefore, parameters are tuned according to the duration of one event’s critical period, instead of default parameters being used in stock market. Generally, there is a one-week critical period during every topic event, so Mhistogram(6,9,3) and MMACD(6,9) are adopted. The results are shown in Fig. 4 after parameters are tuned. When values of MMACD are positive, they can illustrate one person’s impact and its trend better. It is shown in Fig. 4 that MMACD can be more consistent with the public attention (MEMA1) and key periods when impacts are relatively stronger are more accurately identified according to Mhistogram and MMACD. Problems in Fig. 1 and Fig. 3 have been addressed.

4.3 Performance Comparison

Kleinberg’s automaton model is to explore hierarchical structures and topics of document streams based on the role of time and document arrival rate in e-mail. It is used to identify topic bursts in the stream of titles of all papers from database conferences SIGMOD and VLDB for the years 1975-2001 and titles of all papers from the theory conferences STOC and FOCS for the years 1969-2001. He and Parker reconstructed bursts as a dynamic phenomenon to detect topic bursts which were merely determined by histogram for the large PubMed/MEDLINE database of biomedical publications. We use Kleinberg’s model and He and Parker’s model to carry out a similar analysis to identify the relatively important periods when one person’s impacts are relatively strong. The results are shown in Fig. 5. There are significant differences between the periods detected by the three different methods.

Our model is more appropriate, accurate and efficient. Kleinberg’s model is to identify topic bursts based on arrival rate. It is content-independent and character-independent. Our model utilizes more information about time series and persons’ character information. The comparison is shown in Fig.5. Only one period can be identified when the impact is the strongest in Kleinberg’s model, whereas periods when any significant and subtle increase is detected are identified in our model. Therefore, our model can identify more significant periods for the person. It is shown periods produced by Kleinberg’s model are usually longer, so some important changes cannot be detected from the subtle. Furthermore, our model exhibits another advantage that its parameters can be tuned to a given application. He and Parker reconstructed bursts as a dynamic phenomenon which were merely determined by histogram.

Main difference between our model and He and Parker’s model is MMACD and Mhistogram are used in mine while Mhistogram is just used in He and Parker’s model. As shown in Fig.5, more periods are detected using He and Parker’s model which contains many irrelevant intervals. For example, in Fig.5(c), rescue works peaked on May 15, then its intensity declined. MMACD of ‘Army’ gradually became negative and reached the minimum on May 27. After June 9, MEMA1 and MMACD gradually became 0 when rescue works were completed. MMACD (0) during June 21- June 30 were higher than those (<0) during May 27- June 20. Therefore Mhistogram was positive, but no ‘Army’ participated during June 21- June 30. It is the same situation in intervals July 11- July 19 and August 5- September 4. This led to a problem where there was neither the participant nor attention, but the period was detected. The problem of this nature can be avoided in our model.

![Fig. 5.](image)
5 Conclusions

This paper presents a model for impact analysis of several persons on topic events. Compared to some existing methods, this framework defines the impact by using concepts stemming from physics. A key period when one person’s impact is profound in the course of one event is modelled as the interval of positive and increasing momentum. Impacts and key periods are estimated by a combination of MMACD and Mhistogram drawn from technical analysis of stocks. Our experiments show that our model works well for analyzing impact qualitatively and quantitatively. The model not only can monitor the impact and key periods, but also can tune parameters of indicators to the application. The accuracy and effectiveness of the method have been demonstrated through a series of experiments completed for four well-known topic events present in China.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant No. 61173073 and International S&T Cooperation Program of China under Grant No. 2014DFA11350.

References
