

Research on Model of College Network Rumor Propagation in the Era of Big Data

Yongqiang Zuo

Hebei Agricultural University, Hebei, China
Corresponding E-mail: zywelcom@126.com

Abstract

Big Data consists of large-volume, complex, growing data sets with multiple, heterogenous sources. With the tremendous development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. In the era of vigorous development network media, rumors propagation procedure becomes more complex, faster, and more dangerous. The identification and control of Internet rumors is important to related net-check departments. This paper establish an index system to evaluate the level of internet rumors and rumor credibility index system according to characteristics of rumor propagation, and analysis the advantages and disadvantages of the models through experimental verification.

Keywords: *rumor propagation audience crowd event credibility rumor harm*

1. Introduction

Big data sure involves a great variety of data forms: text, images, videos, sounds, and whatever that may come into the play, and their arbitrary combination. Big data frequently comes in the form of streams of a variety of types. Time is an integral dimension of data streams, which often implies that the data must be processed/mined in a timely or (nearly) real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in “a big data environment that can accelerate the time-to-answer critical business questions that demonstrate business values” The time dimension of big data naturally leads to yet another key characteristic of big data – speed or velocity. Gartner analysts described the dominant characteristics of big data as “three Vs” – Volume, Velocity, and Variety. Serious challenges are unfolded along each of the “V” axis. Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed scalability with demonstrated elasticity and parallelism capacities.

Rumor systematic and scientific research can be traced back to World War II. As a kind of typical social phenomenon, rumor propagation means, such as the transmission have undergone great changes in the modern society [1]. Especially in now network media rapid development period, rumors spread process becomes complicated, for the identification and control of Internet rumors become public opinion attention by the public security department.

Rumor is similar to virus in transfer methods, with reference to the model of spread of infectious disease [2]. Correspondingly, the subjects were divided into heard-rumor, non-heard rumor. Rumor of network of evaluation level evaluation index system is established based on a specific assumption.

Allport & Postman give a definition of rumors formula: rumor is equal to event importance * event fuzziness [3]. That is to say, the importance of rumor and event and

fuzzy proportional, events and more important and more blurred, rumors have more probability and effect. Importance and fuzzy which tend to zero element, will not produce a rumor, so to reveal the truth, get rid of ambiguity, it may dispel rumors, now with a confidence instead of fuzzy, with rumors of the audience instead of event of important letter, simplify the problem, according to a rumor evaluation mathematical model [4].

Especially to unexpected events, and even a variety of crisis, the role of the rumors cannot be underestimated. Under modern environment, the use of flexible and disorderly network communication, the spread of rumors faster, more powerful force. If on a general rumor push comes to shove, clearly outweigh the benefits; and for a may cause serious consequences of rumors improper disposal, it is possible to cause serious consequences. Therefore, it is very important to study the mechanism and identification of the rumor.

2. Analysis of Problem

From the infectious disease model, we establish such a model: within populations of individuals in the model is abstracted into several categories, each category are in a state, the basic state include: *s* (susceptible), easy to dye state or health state; *I* (infected) -- infection status; *R* (recovered), is to remove the status or immunity, restore state. Usually in the process of converting between these states to name different infectious model: *SIR* model (easy to dye groups were infected, and then back to health and has immunity), *SIS* model (easy to dye groups were infected after returned to easy to dye state). According to specific hypotheses to analyze rumors spread of solving heard rumors in the number of the total number of accounts for the proportion of, and to establish a network of rumor level evaluation index, according to the audience to establish the evaluation index system of the evaluation of network rumor level.

According to the fuzzy = rumor event importance * events, if direct quantitative analysis by rumor of the corresponding event importance and fuzziness, one has to take into account the factors, if will determine the importance of an event and the event fuzzy factors abstracted, be linked to the importance of events and the audience population, with the credibility of the event (the establishment of network rumor event corresponds to the occurrence of credibility level evaluation index) measure fuzzy events, so as to simplify the problem, to establish a new model: rumor = rumor network level evaluation index x events credibility index.

On the basis of above, our target system can write a report to the department of public security of opinion, put forward new measures to rumors disposal and new approaches, and for the public to provide a copy of how to identify rumor, how to correctly deal with in the era of information of all kinds of information.

Therefore, our target system should be a system with the following characteristics.

1) To establish the evaluation index system of network rumor level. (such as the scope of audience, sense of interest level, mode of transmission, consequences and so on), given the formula which can be quantified.

2) To establish the mathematical model of evaluation of rumors, including rumors of identification (for example, you can somehow establish rumors confidence limits as $[0,1]$, when a formula according to the evaluation of a message when a threshold, namely when the message is less than a value, it can be concluded that does not belong to the rumors, more than a value belongs to the rumors), rumors hazards estimation (so that the relevant departments to take corresponding measures, such as ignored, held a press conference, investigate rumor maker, *etc.*).

3) To write a report for the public security public opinion department, put forward new measures to dispose of rumors. And to provide the public with a copy of how to identify the rumors, how to correctly deal with the information age in the information age.

3. Model Hypothesis

So we establish the following mathematical model:

1) The audience were divided into two groups such as i and s , and i demonstrates the people who have heard rumor, while s demonstrates the people who have not heard rumor. And $i(t)$ demonstrates the ratio of people who have heard rumor, while $s(t)$ demonstrates the ratio of people who have not heard rumor.

2) Supposed the total number of people in the period of rumor propagation is not variable, tagged as N .

3) Ignoring the forgetting of the heard-rumor people during short period of time.

4) The average number of people per hour per person who has heard rumor is a constant tagged as a , called the daily contact rate. When heard-rumor people has broadcasted humor effectively, the non-heard-rumor people has become the heard-rumor person.

5) The average number of online information release platform for each publication is a constant B , which is called the daily rate. After the rumor has been effective, the non-heard people have become a rumor broadcaster.

4. Description of Symbols and Variables

Table 1. Variables and Symbols

Symbols	Definition
$i(t)$	ratio of heard-rumor people in total people
i_0	initial ratio of heard-rumor people
$s(t)$	ratio of non-heard-rumor people in total people
N	total people of rumor propagation in the observed region
a	average effective-propagation people per hour
b	average people who scanned rumor per hour by network information platform
A	people who has heard rumor recently through social relation
B	people who has heard rumor recently through net platform
M	Total number of network information platform
I	evaluation index of rumor
t	moment of rumor propagation period
X	credibility index
Y	evaluated index of rumor harm
K	total times of release information by platform
x	credibility of every release information through platform
m	total information through platform
n	number of false events through information platform

5. Establishment and Solution of Model

The new-added heard-rumor people can be divided into two major categories, and use differential method to solve it. And then to establish the evaluation index system of network rumors level according to the size of $i(t)$.

5.1. Design of Model Planning

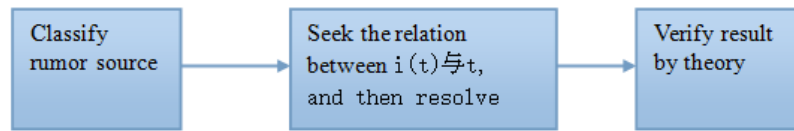


Figure 1. Model of Implement of Planning

5.2. Classify the Source of New-Added Heard-Rumor People

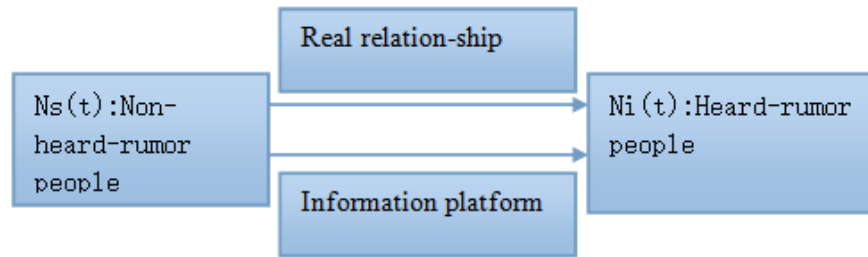


Figure 2. Rumor Broadcast Procedure Model

I) The number of people who have heard rumor through relationship is tagged as A; every heard-rumor people can make a $s(t)$ non-heard-rumor people as heard-rumor people everyday. Because the heard-rumor people number is $N_i(T)$, there will be $aN_s(t)*i(t)$ people who have heard rumor after a period of time totally, that is $A=a*N_s(t)*i(t)$. (where a for each heard rumors of the mean number of per hour contact, n is in rumor spreading period investigated the total number of $i(t)$ is heard rumors in the total number of accounts for the proportion of, $s(t)$ is heard rumors in the total number of accounts for the proportion of).

II) The number of people who have heard rumor through net information platform is tagged as B. Each information platform can make $b*s(t)$ non-heard-rumor people as new heard-rumor people. Because the total number of rumor spreading information platform is M , there will be $b*M*s(t)$ people who heard rumors, namely $B = b*M*s(t)$ (where B for each released rumors network information release platform per hour to effectively navigate the average number of, M is the total number of rumor spreading information platform).

5.3.1 Establish Model

$$N \frac{di}{dt} = A + B \tag{5-1}$$

Remarks:

$$\begin{aligned} A &= aN_s(t)i(t) \\ B &= bMs(t) \end{aligned} \tag{5-2}$$

and

$$\begin{cases} i(t) + s(t) = 1 \\ i(0) = i_0 \end{cases} \tag{5-3}$$

then

$$N \frac{di}{dt} = aN \cdot i(1-i) + bM(1-i) \tag{5-4}$$

5.3.2. Model Resolution

Therefore:

$$\frac{di}{aN \cdot i(1-i) + bM(1-i)} = \frac{dt}{N} \tag{5-5}$$

$$\int_{i_0}^i \frac{di}{aN \cdot i(1-i) + bM(1-i)} = \frac{1}{N} \int_0^t dt \tag{5-6}$$

Seek integral table and reduce it:

$$i(t) = 1 - \frac{1}{\frac{aNi_0 + bM}{(aN + bM)(1-i_0)} e^{-\frac{bM+aN}{N}t} + \frac{aN}{aN + bM}} \tag{5-7}$$

5.3.2. Analysis on Result Theoretically

$$i(t) = 1 - \frac{1}{\frac{aNi_0 + bM}{(aN + bM)(1-i_0)} e^{-\frac{bM+aN}{N}t} + \frac{aN}{aN + bM}}$$

From the formula $\frac{1}{\frac{aNi_0 + bM}{(aN + bM)(1-i_0)} e^{-\frac{bM+aN}{N}t} + \frac{aN}{aN + bM}}$, we can easily get $M \ll N$, while in the extreme cases, the conclusion can be got:

① When the total number of regions in the spread of rumors heard during the number of rumors in total number in the proportion is 0, that is: when $i_0 = 0$, $\lim_{t \rightarrow \infty} i(t) = 0$;

② When the total number of regions in the spread of rumors heard during the number of rumors in total number in the proportion is 1, that is: when $i = 1$, $\lim_{t \rightarrow \infty} i(t) = 1$;

By the results of ① and ②, the actual situation can be obtained, the results of this model have a certain desirability

5.4. The Establishment of Evaluation Index System for Evaluating the Level of Internet Rumors

For the nature of the rumor, is difficult to directly according to the data to determine, but can be by rumors in the crowd for a range of audiences from the side judge its influence, namely the size and divide the evaluation index system, due to there were rumors for a range of audiences large and small. Sometimes the difference is very big. A rumor has been produced, the audience cannot exist, it cannot be all the people, that is, or the situation cannot happen. In order to make the evaluation index of discrimination, so artificially will audiences minimal I value location 1, great as I set at 7, somewhere in between as a number of intermediate value (the Internet rumors level evaluation index). As shown in the following table:

Table 2. Classification of Model Evaluation Index

Index of range of audience crowd	$i(t)$ range	I
smallest	$0 < i(t) \leq 0.1$	1
smaller	$0.1 < i(t) < 0.26$	2
small	$0.26 \leq i(t) < 0.42$	3
medium	$0.42 \leq i(t) < 0.58$	4
large	$0.58 \leq i(t) < 0.74$	5
larger	$0.74 \leq i(t) < 0.9$	6
largest	$0.9 \leq i(t) < 1$	7

6. Establishment and Results Analysis of Rumors Credibility Model

The judgment of network rumors true and false need to set up a rumor credibility index to the authenticity of rumors which were evaluated and given a gradient judgment according to the judgment of true and false rumors, more in line with the actual, while the estimates of rumors hazards need to combine with the audience to judge.

6.1 Credibility Model Design of Rumor



Figure 3. Model Credibility Design

6.2. Establish Rumor Credibility Index

6.2.1. How to Decide Model Credibility

Establish the index system of the key is to examine the information release platform release information in false information accounted for the proportion of, if write x as a platform for information dissemination of information credibility, n for each information platform released information is the number of false information, m for each platform of information release information of the total, the total number of K as a platform for information dissemination of information,

$$x_j = 1 - \frac{n_j}{m_j}$$

6-1

$$x = 1 - \frac{\sum_{j=1}^k \frac{n_j}{m_j}}{k}$$

6-2

6.2.2. Establish Rumor Credibility Index

By on each information release platform for a credibility x , there must be the most value x_{\min} and x_{\max} , namely $x \in (x_{\min}, x_{\max})$, this interval is divided into a number of intervals, for each interval calibration reliability index (for true and false rumors of true and false judgment gradient criterion)

Table 3. Zone Classification of Rumor Credibility

Zone of x should be located	Value of X	Description of rumor
1 st Zone	1	Be rumor strictly
2 nd Zone	2	Be rumor perhaps
3 rd Zone	3	Not be rumor

With increasing of the reliability index of corresponding event, the information is more real, the more deviation from rumors of X.

6.3. Establish Rumor Index System

If event's non-credibility index is tagged as the reciprocal of confidence index, namely $\frac{1}{X}$. Then rumor = rumor network level evaluation index *events credibility index m , and gives a judgment gradient to be reference(tagged as rumors hazard assessment index).

Table 4. Index Classification of Rumor Harm

Evaluation index Y of rumor harm	rumor= rumor network level evaluation index *events credibility index	description
1	0~2.3	little
2	2.3~3.5	large
3	3.5~5	larger
4	5~7	serious

Rumors harm is related with the audience, and with the credibility of various information platform. On rumors, network public opinion department can handle it on the basis of the hazard estimates the size of the Index Y separately. If $y = 1$, ignore depending on the circumstances; if $y = 2$, need causes some attention, continue to observe the development; if $y = 3$, need to publish news to reduce rumor; if $y = 4$, in addition to the news release rumor outside, also need to close the relevant information release platform for the corresponding processing.

On the network browsing information, the public should try to choose the high credibility information platform, reduce the possibility of access rumors; secondly, as far as possible not to spread rumor or suspected rumors information, slows down the rumor propagation velocity, so as to reduce the rumors of harm.

6. Conclusion

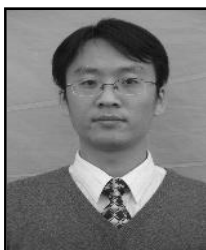
Big data sure involves a great variety of data forms: text, images, videos, sounds, and whatever that may come into the play, and their arbitrary combination. Big data frequently comes in the form of streams of a variety of types. Time is an integral dimension of data streams, which often implies that the data must be processed/mined in a timely or (nearly) real-time manner. Besides, the current major consumers of big data, corporate businesses, are especially interested in “a big data environment that can accelerate the time-to-answer critical business questions that demonstrate business values” The time dimension of big data naturally leads to yet another key characteristic of big data – speed or velocity. Gartner analysts described the dominant characteristics of big data as “three Vs” – Volume, Velocity, and Variety. Serious challenges are unfolded along each of the “V” axis. Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed scalability with demonstrated elasticity and parallelism capacities.

Through analysis on the rumor transmission way and method, and combining with the rumors spread in the Internet age, rumor spreading and reliability model is established. The model has the following advantages: model is simple and intuitive, easy to judge; the solution to every problem is linked each other, and not isolated. Model defects: the model involves many parameters, need to be conducted on a questionnaire survey of the way to be resolved.

References

- [1] N. T. J. Bailey, “The Mathematical Theory of Infectious Diseases and Its Applications”, New York: Hafner Press, (1975).
- [2] R. M. Anderson and R. M. May, “Infectious Diseases of Humans”, Oxford: Oxford University Press, (1992).
- [3] H. W. Hethcote, “The mathematics of infectious diseases”, *SIMA Review*, vol. 42, no. 4, (2008), pp. 599-653.
- [4] J. Qiyuan, X. Jinxin and Y. Jun, “Math Modeling”, Beijing: High-level Education Press, (2004).
- [5] S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks”, *Science*, vol. 337, (2012), pp. 337-341.
- [6] A. Machanavajjhala and J. P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data”, *ACM Crossroads*, vol. 19, no. 1, (2012), pp. 20-23.
- [7] S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” *Knowledge and Information Systems*, vol. 33, no. 3, (2012), pp. 523-547.
- [8] E. Birney, “The Making of ENCODE: Lessons for Big-Data Projects”, *Nature*, vol. 489, (2012), pp. 49-51.

Author



Yongqiang Zuo, male, the Han nationality, native place: Hebei. Lecturer of college students management and employment guidance, Master degree.