RESEARCH ARTICLE

Words Analysis of Online Chinese News Headlines about Trending Events: A Complex Network Perspective

Huajiao Li[1,2,3,4], **Wei Fang**[1,2,4]*, **Haizhong An**[1,2,4], **Xuan Huang**[1,2,4]

1 School of Humanities and Economic Management, China University of Geosciences, Beijing, China, **2** Key Laboratory of Carrying Capacity Assessment for Resource and Environment, Ministry of Land and Resources, Beijing, China, **3** Department of Energy and Mineral Engineering in the College of Earth and Mineral Sciences, The Pennsylvania State University, State College, Pennsylvania, United States of America, **4** Lab of Resources and Environmental Management, China University of Geosciences, Beijing, China

* itasstudio@126.com

OPEN ACCESS

Citation: Li H, Fang W, An H, Huang X (2015) Words Analysis of Online Chinese News Headlines about Trending Events: A Complex Network Perspective. PLoS ONE 10(3): e0122174. doi:10.1371/journal.pone.0122174

Academic Editor: Zi-Ke Zhang, Hangzhou Normal University, CHINA

Received: June 20, 2014

Accepted: February 8, 2015

Published: March 25, 2015

Data Availability Statement: The data used in this paper are mainly extracted from the Baidu (http://www.baidu.com) search engine, which is generally acknowledged as the most widely used free and public search engine in China. We obtained the URL links for news regarding the two trending events from Baidu"2010 Mexico oil spill. http://news.baidu.com/ns?word=%E5%A2%A8%E8%A5%BF%E5%93%A5%E6%B9%BE%E6%BC%8F%E6%B2%B9&pn=[*]&cl=2&ct=1&tn=news&rn=20&ie=utf-8&bt=0&et=0" "2011 Bohai Bay Oil Spill http://news.baidu.com/ns?word=%E6%B8%A4%E6%B5%B7%E6%B9%BE%E6%BC%8F%E6%B2%B9&pn=[*]&cl=2&ct=1&tn=

Abstract

Because the volume of information available online is growing at breakneck speed, keeping up with meaning and information communicated by the media and netizens is a new challenge both for scholars and for companies who must address public relations crises. Most current theories and tools are directed at identifying one website or one piece of online news and do not attempt to develop a rapid understanding of all websites and all news covering one topic. This paper represents an effort to integrate statistics, word segmentation, complex networks and visualization to analyze headlines' keywords and words relationships in online Chinese news using two samples: the 2011 Bohai Bay oil spill and the 2010 Gulf of Mexico oil spill. We gathered all the news headlines concerning the two trending events in the search results from Baidu, the most popular Chinese search engine. We used Simple Chinese Word Segmentation to segment all the headlines into words and then took words as nodes and considered adjacent relations as edges to construct word networks both using the whole sample and at the monthly level. Finally, we develop an integrated mechanism to analyze the features of words' networks based on news headlines that can account for all the keywords in the news about a particular event and therefore track the evolution of news deeply and rapidly.

Introduction

With the development and popularization of information and network technology, the Internet has become the main medium from which people obtain information and news. Helping solve a serious information overload problem [1], search engines are recognized as one of the most useful and popular services on the web [2, 3]. Generally, the web (and a search engine) is the first source a person turns to for information or news [4]. People have grown accustomed to inputting a few keywords into search engines and then clicking on one or more headlines

news&rn=20&ie=utf-8&bt=0&et=0", and the relationship of the URL links and the pages can be expressed as the following: [*]=2p-10 p=1,2,3,4,```,N (1) where p is the online page number of Baidu News about the trending events, and N is the total number of pages of search results about the given hot event. In this research, there are a total of 38 pages regarding the 2010 Gulf of Mexico oil spill and 37 pages regarding the 2011 Bohai Bay oil spill. In the pages, we captured the title, media source, date and time by two different labels, and we automatically gathered all 1,487 pieces of Chinese news on 29 October 2014 about the 2011 Bohai Bay oil spill and the 2010 Gulf of Mexico oil spill. In the initial gathered data, there were 49 pieces of duplicate news form the same media at the same time and four news duplicate pieces of news before the event occurred in the 748 news stories about the 2010 Gulf of Mexico oil spill and 29 pieces of duplicate news from the same media at the same time and eight duplicate news pieces from before the event occurred in the 739 news about the 2011 Bohai Bay oil spill. Thus, after data cleaning, we obtained 695 pieces of news about the 2010 Gulf of Mexico oil spill and 702 pieces of news about the 2011 Bohai Bay oil spill.

Funding: This research is supported by grants from the National Natural Science Foundation of China (Grant No. 71173199), the China Scholarship Council (File No. 201406400004), the Humanities and Social Sciences planning funds project under the Ministry of Education of the PRC (Grant No.10YJA630001), Fundamental Research Funds for the Central Universities (Grant No. 2-9-2014-104), the Science and Technology Innovation Fund of the China University of Geosciences (Beijing), and the Key Laboratory of Carrying Capacity Assessment for Resource and the Environment (No. CCA2015.05). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

out of the voluminous search results. Users can choose based on the closeness of the match and the users' desire to obtain a detailed description of the news, and some scholars have successfully researched recommendation algorithms regarding the news and social networks based on users' behavior and their similarities [5, 6, 7, 8]. Furthermore, more and more people realize that online news plays an important role in the spread of public opinion; thus, it is of great importance to know what and how different news sources present information. A headline is a significant component of the news and not only presents or relates the main points of news content but also must attract and hold the reader's attention [9]. Some scholars have provided evidence that there are connections among public relations, public awareness and news [10]. As networks develop, crisis communication theories require further modification and perfecting [11].

To analyze the information contained in news headlines, we should begin with information extraction technology. Information extraction can be traced back to 1960, when scholars first attempted to extract structured information from natural language text. In the news field, previous studies have mainly focused on text mining techniques and tools [12, 13], semantic analysis [14], analysis of sentiment [15, 16], etc. Some scholars have observed that news has value to an extent. Yoon (2012) observed that it is useful to detect weak signals for long-term business opportunities using text mining of web news [17]. Huang, Liao, Yang, & Chang (2012) proposed a financial news headline agent to assist with investment decisions in the Taiwanese stock market after receiving essential real-time news headlines disseminated by the agent [18]. Regarding text mining methods, scholars have studied data pre-processing [19], text mining [20, 21, 22] and visualization [23, 24]. Chen and Hsieh (2006) observed web page classification based on a support vector machine using a weighted vote schema [25]. Magerman, Van and Song (2010) explored the feasibility and accuracy of latent semantic analysis performed by text mining techniques and detected similarities between patent documents and scientific publications [26]. However, the majority of current theories and tools are directed at identifying one website or one piece of online news, although the information from one website or one piece of online news may be biased and insufficient to rapidly develop an understanding of a topic. As the trending event, there are hundreds of pieces of news reporting it, and there are thousands of words included in the headlines, the words can be linked to each other to form a big words network. Complex network theory can provide an improved approach to analyzing the evolution of the words network of news headlines.

Complex network theory has attracted great interest with respect to solving complex issues in recent years. A theory derived from the study of physics, complex network theory has been applied to many empirical studies, particularly in management [27], sociology [28, 29], and economics [30, 31, 32]. The fundamental principle of complex network theory is to identify units and relations between the units, which enables the construction of a network utilizing the units as nodes and the relations as edges to analyze and solve problems holistically.

In this paper, we chose the Deepwater Horizon oil spill (http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill, which is also known as the 2010 Gulf of Mexico oil spill) and the 2011 Bohai Bay oil spill (http://en.wikipedia.org/wiki/2011_Bohai_Bay_oil_spill) as our empirical subjects or themes, and we gathered and pretreated all the Chinese news headlines in the search results on the two trending events from Baidu (http://www.baidu.com), the most popular and well-known Chinese search engine. We used Simple Chinese Word Segmentation to segment all the headlines into words and then used the words as nodes and the adjacent relations as edges to construct the words networks across the whole sample as well as at the monthly level. We integrated statistics, word segmentation, complex network and visualization to analyze all the headlines' keywords and the evolution of online news about the two different trending events.

Main Process and Data

Main process

As shown in Fig. 1, the main process involved in this research occurs over nine steps. First, we chose the theme of the analysis and determined the search terms we would use. For this paper, as described above, we chose two trending events involving oil spills, the Gulf of Mexico oil spill, which occurred in May, 2010, and the Bohai Bay oil spill, which occurred in June, 2011. Second, we chose the search engines and input search terms to obtain the search results. In the third step, we analyzed and came to understand the rule of the search results and obtained the data structure to provide the foundation for Step 4. In Step 4, we developed tools to capture the search results automatically by inputting the data structure, which allowed us to export and clean the data based on the research object. In Step 5, we chose a suitable word segmentation tool to divide the Chinese headlines into different words. Steps 2 to 5 will be explained in more detail in section 2.2 below. After obtaining the words, we can construct different words networks according to the complex network methodology and calculate different features and analyze all the headlines' keywords and the evolution of online news about the given theme. Steps 6 to Step 9 will be explained in detail in the remaining chapters of this paper.

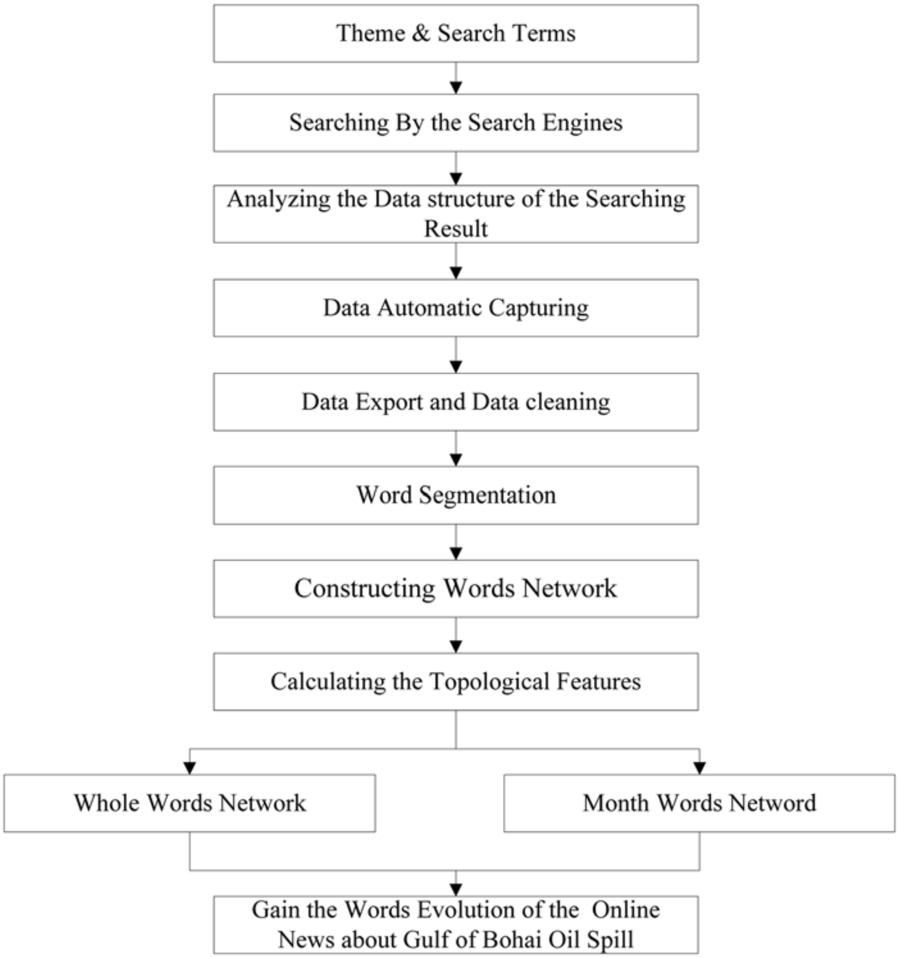
Theme & Search Terms
Searching By the Search Engines
Analyzing the Data structure of the Searching Result
Data Automatic Capturing
Data Export and Data cleaning
Word Segmentation
Constructing Words Network
Calculating the Topological Features
Whole Words Network
Month Words Netword
Gain the Words Evolution of the Online News about Gulf of Bohai Oil Spill

Fig 1. The main process of the research.

doi:10.1371/journal.pone.0122174.g001

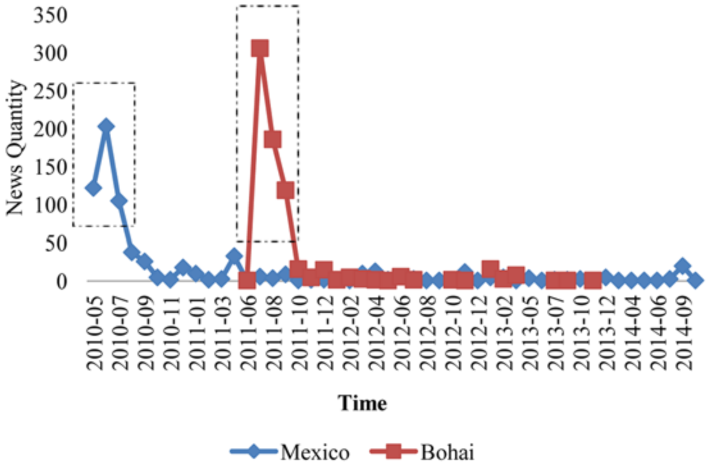

Fig 2. The evolution of the quantity of news after the events occurred.

doi:10.1371/journal.pone.0122174.g002

Data

The data used in this paper are mainly extracted from the Baidu (http://www.baidu.com) search engine, which is generally acknowledged as the most widely used search engine in China. We obtained the URL links for news regarding the two trending events from Baidu:

2010 Mexico oil spill: http://news.baidu.com/ns?word=%E5%A2%A8%E8%A5%BF%E5%93%A5%E6%B9%BE%E6%BC%8F%E6%B2%B9&pn=[*]&cl=2&ct=1&tn = news&rn=20&ie = utf-8&bt=0&et=0

2011 Bohai Bay Oil Spill: http://news.baidu.com/ns?word=%E6%B8%A4%E6%B5%B7%E6%B9%BE%E6%BC%8F%E6%B2%B9&pn=[*]&cl=2&ct=1&tn = news&rn=20&ie = utf-8&bt=0&et=0)

The relationship of the URL links and the pages can be expressed as following:

$$[*] = 2p - 10p = 1, 2, 3, 4, \text{```}, N \quad (1)$$

where p is the online page number of Baidu News about the trending events, and N is the total number of pages of search results about the given hot event. In this research, there are a total of 38 pages regarding the 2010 Gulf of Mexico oil spill and 37 pages regarding the 2011 Bohai Bay oil spill.

In the pages, we captured the title, media source, date and time by two different labels: "<h3 class = "c-title"> </h3>" and "<p class = "c-author"> </p>", and we automatically gathered all 1,487 pieces of Chinese news on 29 October 2014 about the 2011 Bohai Bay oil spill and the 2010 Gulf of Mexico oil spill. In the initial gathered data, there were 49 pieces of duplicate news form the same media at the same time and four news duplicate pieces of news before the event occurred in the 748 news stories about the 2010 Gulf of Mexico oil spill and 29 pieces of duplicate news from the same media at the same time and eight duplicate news pieces from before the event occurred in the 739 news about the 2011 Bohai Bay oil spill. Thus, after data cleaning, we obtained 695 pieces of news about the 2010 Gulf of Mexico oil spill and 702 pieces of news about the 2011 Bohai Bay oil spill.

Fig. 2 shows the time distribution of the news regarding the two trending events and enables us to find that both trending events were well covered in the media in the three or four months

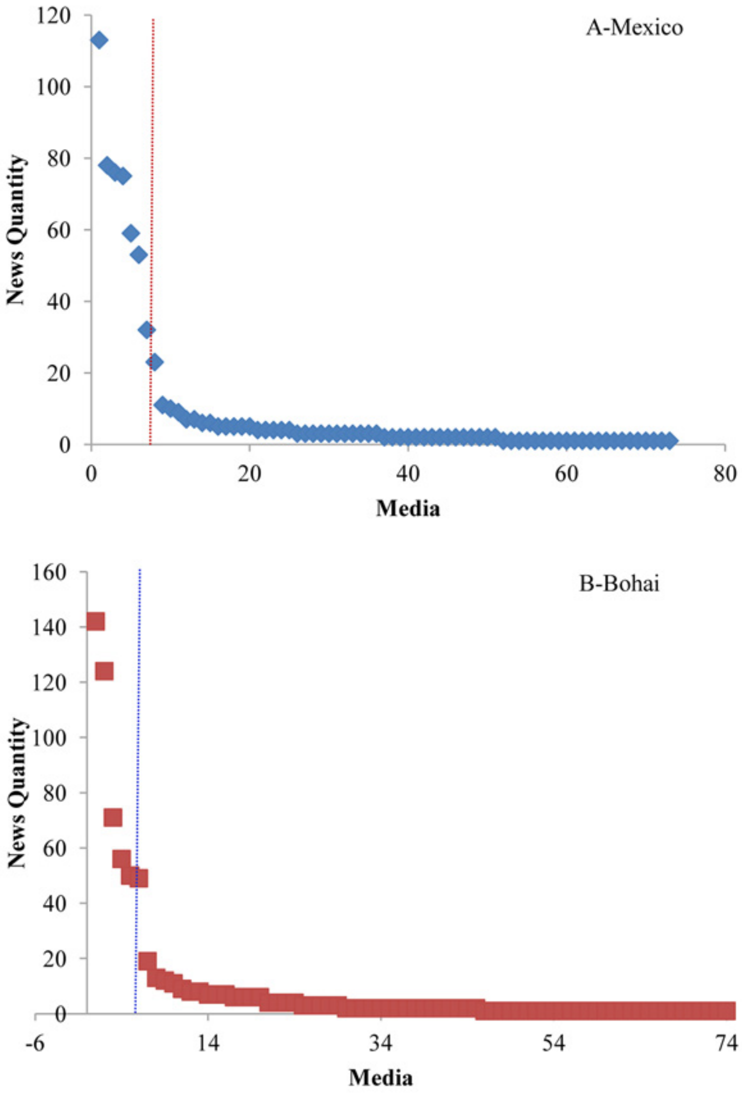

Fig 3. News Media Distribution of the two trending events.

doi:10.1371/journal.pone.0122174.g003

after each occurred, and then faded away to be talked about in the media only occasionally thereafter. Meanwhile, there is one notable difference between the news about the two trending events: the media first reported the 2010 Gulf of Mexico oil spill accident immediately after it occurred but first reported the 2011 Bohai Bay oil spill one month after it had occurred.

Fig. 3 shows the media distribution of the two trending events. According to Fig. 3, only a few media outlets contributed the majority quantity of the news, and most media outlets reported no more than 10 pieces of news. The top six media outlets that reported the 2010 Gulf of Mexico oil spill are Netease, Sina, Xinhuanet, Sohu, Ifeng, Tencent, and the top six media that reported the 2011 Bohai Bay oil spill are Sina, Ifeng, Sohu, Hexun, Tencent, and Netease, which comprise all the mainstream media outlets in China.

Method

Method of headlines' word segmentation

We used the open source word segmentation software called Simple Chinese Word Segmentation (http://www.xunsearch.com) based on the scripting language PHP. Simple Chinese Word Segmentation employs a dictionary containing more than 260 thousand Chinese words. The part-of-speech tagging used in this software is Peking University annotation, which contains 47 parts of speech. The input information is the headlines and the serial numbers of the headlines, whereas the output information consists of the serial numbers of the words, the words, the words' part of speech, and the serial numbers of the headlines.

Method of constructing words network

As described above, the main job of constructing the word network is to determine the nodes and edges as well as the weights of the edges. There are different ways of constructing networks, such as equivalence relationships (complete graph) [30], affiliation relationships (bipartite graph) [33, 42], and so on. In this paper, in order to show the words contextual relationships in the title, we gleaned the segmented words from the news headlines according to the features of the study subject (theme), and then we took each word as a node and connected nodes with edges based on the sequence of the words in the headlines, i.e., the former node as the start node and the node following the former node as the end node. This process was conducted repeatedly and sequentially for all the words in the titles.

Fig. 4 shows the linear network for one title. Next, the linear networks of different headlines were superimposed; the weights of the edges are the times of the appearance of the edges between two nodes in different linear networks. Let graph G = (V,E,W) represent the directed weighted network in which V and E are the set of nodes and edges, and W represents the

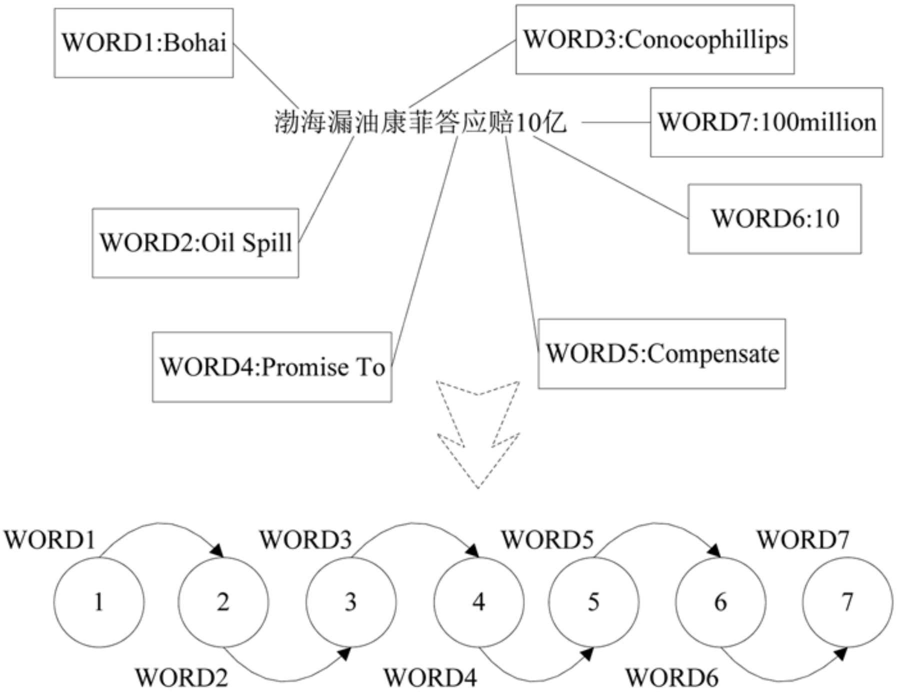

Fig 4. Construction of the Network (based on one title).

doi:10.1371/journal.pone.0122174.g004

weight of the edges. Formula (2) shows the definition of the edges of the words in one title. In addition, the weight of the edge between two different nodes is the sum of e_{ij}.

$$\begin{cases} e_{ij}(k) = 1 & V_i\text{is the nearby former node of } V_j\text{in title k} \\ e_{ij}(k) = 0 & V_i\text{is not the nearby former node of } V_j\text{in title k} \end{cases} \tag{2}$$

The calculation methods of topological features

There are numerous topological features of the nodes, edges and networks in complex network theory. In this paper, we mainly analyzed the two different levels of networks using the following seven different topological features: degree, degree assortativity, weighted degree, average shortest path length, clustering coefficient, community structure, and stability coefficient. Meanwhile, on the basis of the distribution of the degree and the weighted degree of the words network, we analyzed the scale-free characteristics of the whole-sample words network, and we also analyzed the small-world properties of the words networks based on the average shortest path length and clustering coefficient between the words networks and the two different random networks, one with same average degree and another with same degree sequence as the words networks by network reshuffling [33].

The node's degree indicates how many nodes connect. The more connections a node makes, the more importance that node has.

$\mathcal{R}_i^{in}(t)$ represents the in-degree of node i:

$$\mathcal{R}_i^{in}(t) = \Sigma_{j=1}^{n} e_{ji} \tag{3}$$

$\mathcal{R}_i^{out}(t)$ represents the out-degree of node i:

$$\mathcal{R}_i^{out}(t) = \sum_{j=1}^{n} e_{ij} \tag{4}$$

$\mathcal{R}_i(t)$ represents the sum of the in-degree and the out-degree of node i:

$$\mathcal{R}_i(t) = \mathcal{R}_i^{in} + \mathcal{R}_i^{out} \tag{5}$$

In order to analyze the degree assortativity of the network, we use Pearson correlation coefficient of the degrees of any of the two nodes connected by a link to calculate it [34, 35].

$$r^{in} = \langle r_i^{in} \rangle = \langle \frac{(\mathcal{R}_i^{in} - \overline{\mathcal{R}_i^{in}}) * (\mathcal{R}_j^{out} - \overline{\mathcal{R}_j^{out}})}{s(\mathcal{R}_i^{in}) s(\mathcal{R}_j^{out})} \rangle \tag{6}$$

$$r^{out} = \langle r_i^{out} \rangle = \langle \frac{(\mathcal{R}_i^{out} - \overline{\mathcal{R}_i^{out}}) * (\mathcal{R}_j^{in} - \overline{\mathcal{R}_j^{in}})}{s(\mathcal{R}_i^{out}) s(\mathcal{R}_j^{in})} \rangle \tag{7}$$

where r^{in} is the in-degree assortativity of the network, r^{out} is the out-degree assortativity of the network, $\bar{\mathcal{R}}$ is the average degree and σ is the standard deviation of different degrees.

For a weighted network, the importance of a node is determined not only by the number of nodes it connects but also by the weight between the node and other nodes. The higher the weight, the more frequently the two words will appear together.

$W\mathcal{R}_i^{in}$ represents the in-weighted degree of node i:

$$W\mathcal{R}_i^{in} = \sum\nolimits_{j=1}^{n} w_{ji} \tag{8}$$

$W\mathcal{R}_i^{out}(t)$ represents the out-weighted degree of node i at time t:

$$W\mathcal{R}_i^{out} = \sum\nolimits_{j=1}^{n} w_{ji} \tag{9}$$

$W\mathcal{R}_i(t)$ represents the sum of the in-weighted degree and the out-weighted degree of node i at time t:

$$W\mathcal{R}_i = W\mathcal{R}_i^{in} + W\mathcal{R}_i^{out} \tag{10}$$

Most real-world network distributions have long right tails of values that are far above the mean, and the degree distribution of the nodes obeys a power law according to M. E. J. Newman [36, 37]; thus, we say that the network has scale-free characteristics. In a scale-free network, the degree (weighted degree) distribution follows a power law:

$$P_R \propto R^{-\lambda} \tag{11}$$

while λ can be calculated by:

$$\ln P_R \propto -\lambda \ln R \tag{12}$$

The shortest path length of two words means the least quantity of edges between them. The average shortest path length represents the connectivity of different words as well as the words network, and it can be calculated by [38]:

$$d_{ij} = \frac{1}{N(N-1)} \sum\nolimits_{i,j \in N, i \neq j} e_{ij} \tag{13}$$

The clustering coefficient means the connectivity of the neighbor nodes of a given node and is given by the ratio of existing edges (E_i) between its first neighbors ($\mathcal{R}_i$) to the potential number of such ties ($\frac{1}{2}\mathcal{R}_i(\mathcal{R}_i - 1)$). In addition, we can obtain the clustering coefficient of the network by averaging the clustering coefficient of all nodes in the network. [39]

$$C = < C_i >_i = < \frac{2Ei}{\mathcal{R}_i(\mathcal{R}_i - 1)} > \tag{14}$$

Moreover, if the network presents a high probability that two neighbors of one given node are also connected themselves with a small average shortest path length between two nodes, we call the network a "small-world" network.

Meanwhile, the community structure and stability coefficient are two useful features that are often used in evolution analysis. In this paper, we used the heuristic method [40] and "auto-correlation function" (or "similarity coefficient function") [41, 42] to analyze the evolution of the monthly-words networks. Formula (15) and Formula (16) shows the main step of the two different methods.

$$\Delta M = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{15}$$

where $\sum_{in}$ represents the degrees of all the links inside the community K, $\sum_{tot}$ represents the total degrees of the nodes in K, k_i represents the total degrees of node i, $k_{i,in}$ represents the total degrees of the links from i to all the nodes in K, and m represents the total degrees of the network. The community will be combined repeatedly until ΔM is negative while combining the communities.

$$S(t) = \frac{\mathcal{N}_t \cap \mathcal{N}_{t-1}}{\mathcal{N}_t \cup \mathcal{N}_{t-1}} \tag{16}$$

where S(t) is the stability coefficient (similarity coefficient) of the words network. $\mathcal{N}_t$ represents the set of nodes in the words network in month t, and $\mathcal{N}_{t-1}$ represents the set of nodes in the network in month t-1. $\mathcal{N}_t \cap \mathcal{N}_{t-1}$ is the number of common nodes (words) at $\mathcal{N}_{t-1}$ and $\mathcal{N}_t$, and $\mathcal{N}_t \cup \mathcal{N}_{t-1}$ is the number of nodes at the union of $\mathcal{N}_{t-1}$ and $\mathcal{N}_t$.

Results and Analysis

The topological features of the whole-sample words network

The visualization of the whole-sample words network. After application of the Simple Chinese Word Segmentation software, we obtained 5,661 words regarding the 2010 Gulf of Mexico oil spill and 6,821 words regarding the 2011 Bohai Bay oil spill (after eliminating punctuations). After cleaning duplicate words, there were 1,288 different words in all the online Chinese news headlines regarding the 2010 Gulf of Mexico oil spill and 1,572 different words in all the online Chinese news headlines regarding the 2011 Bohai Bay oil spill, which means there are 1,288 nodes in the whole-sample words network about Mexico and 1,572 nodes in the whole-sample words network about Bohai. Fig. 5 presents the visualization results of the two whole-sample words networks regarding Mexico and Bohai (the color of the node is determined by the community Id which the node belongs to).

The scale-free characteristics and degree assortativity of the whole-sample words network. According to Formula (5) and Formula (10), we can calculate the degree and weighted degree of each node in the two whole-sample words networks and obtain the keywords of the two trending events in the whole-sample words network perspective. Tables 1 and 2 show the keywords of the two trending events (CNOOC represents "China National Offshore Oil Corporation").

According to Fig. 6, both the degree distribution and the weighted degree distribution of the two whole-sample words networks can be approximated by the power-law : $\ln P_R \propto -\lambda \ln R,$, with good R^2 (goodness of fit). Thus, we can conclude that the two networks are scale-free. Meanwhile, according to Formula (6) and Formula (7), we can get that the in-degree assortativity of two whole-sample words networks is 1.38671E-05 and -9.69644E-06, respectively, and the out-degree assortativity of two whole-sample words networks is-4.06877E-05 and -2.39787E-05, respectively, which are close to zero and much lower than the degree assortativity of real world networks [43]. So we can conclude that the words networks constructed in this paper have no significant assortative or disassortative mixing features.

The small world properties of the whole-sample words network. A small world network means that the neighbors of a given node have a high probability of contact with one another with a short average length. In the words networks, if it is small-world, it indicates that the

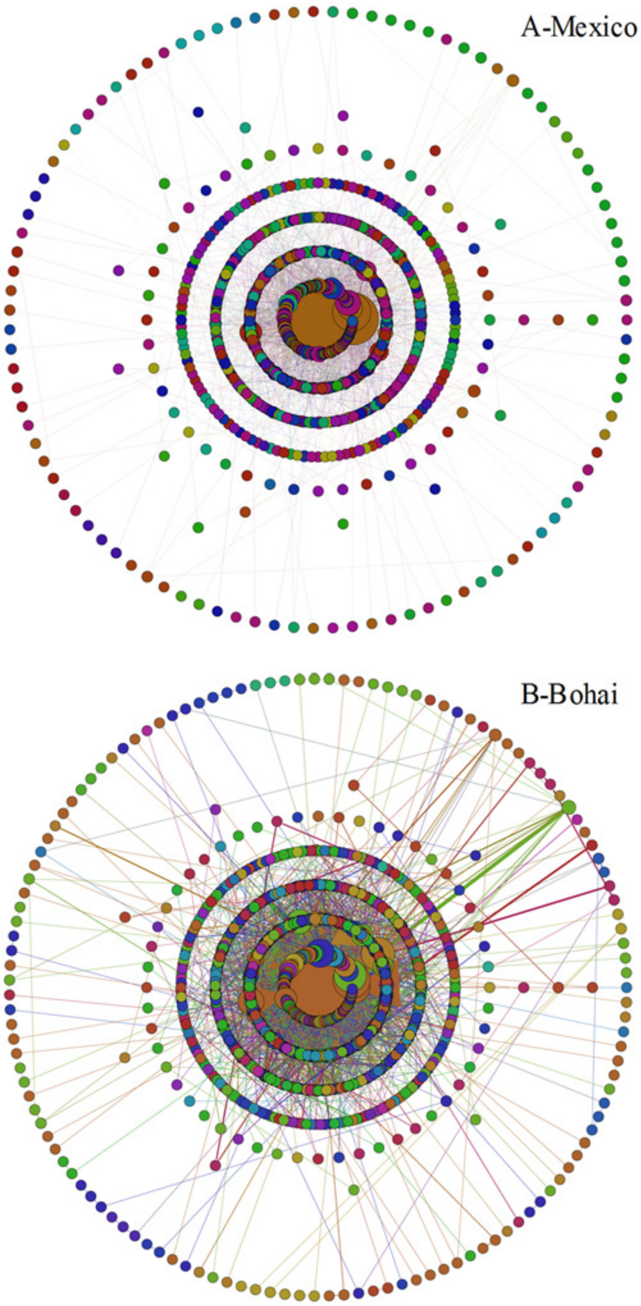

Fig 5. Visualization results of the whole-sample words networks of the two trending events.

doi:10.1371/journal.pone.0122174.g005

words of the headlines contact very well with one another, and most of the points of the news are well connected. According to Formula (13) and Formula (14), we can gain both the average shortest path and the average clustering coefficient of the two whole-sample words networks. The average clustering coefficients of the two whole-sample words networks about the 2010 Gulf of Mexico oil spill and the 2011 Bohai Bay oil spill are 0.042 and 0.054, respectively. They are much larger than the clustering coefficients of the random networks of the identical size, which are both 0.001, as well as the random networks with the same degree sequence, which are 0.004 and 0.005, respectively. The average shortest path lengths of the two whole-sample

Table 1. Keywords of the 2010 Gulf of Mexico oil spill by degree and weighted degree.

Keywords By Degree		Keywords By Weighted Degree	
ID	Words	ID	Words
W00010	Oil Spill	W00010	Oil Spill
W00595	Gulf of Mexico	W00595	Gulf of Mexico
W00011	Event	W00011	Event
W00076	America	W00049	Oil
W00020	Accident	W00020	Accident
W00965	BP	W00077	Company
W00219	Already	W00076	America
W00049	Oil	W00023	100 Million
W00062	Probably	W00219	Already
W00012	Impact	W01001	Britain

doi:10.1371/journal.pone.0122174.t001

words networks are 4.01 and 3.931. They are shorter than the random network with the same mean degree (1.969 and 2.27), which are 5.316 and 5.026, and the random network with the same degree sequence, which are 11.525 and 11.211. Thus, we can conclude that the two networks have small-world properties.

The results of scale-free characteristics and small world properties of the whole-sample words networks indicate that, the networks constructed in this paper are nonrandom and well-connected than the random networks with the same mean degree as well as the random networks with the same degree sequence as the words networks by network reshuffling. So, words in the online news titles are well connected by the regular grammatical rules and media preference of the words related to the topic of the trending events. However, by analyzing the degree assortativity, we can find that, most of them show very weak disassortative mixing, which is similar as model of Barab´asi and Albert and random networks [43].

The evolution of the monthly-words network

The visualization of the monthly-words network. To analyze the evolution of words in the headlines about the two trending events, we constructed the words networks for different months and analyzed the evolution of different topological features regarding the monthly-words

Table 2. Keywords of the 2011 Bohai Bay oil spill by the degree and weighted degree.

Keywords By Degree		Keywords By Weighted Degree	
ID	Words	ID	Words
W00010	Oil Spill	W00010	Oil Spill
W00020	Accident	W00009	Bohai Bay
W00009	Bohai Bay	W00020	Accident
W00026	Conocophillips	W00069	CNOOC
W00011	Event	W00295	Bohai
W00025	Claims	W00026	Conocophillips
W00295	Bohai	W00011	Event
W00071	Declare	W00259	Oil Field
W00440	Pollution	W00071	Declare
W00069	CNOOC	W00025	Claims

doi:10.1371/journal.pone.0122174.t002

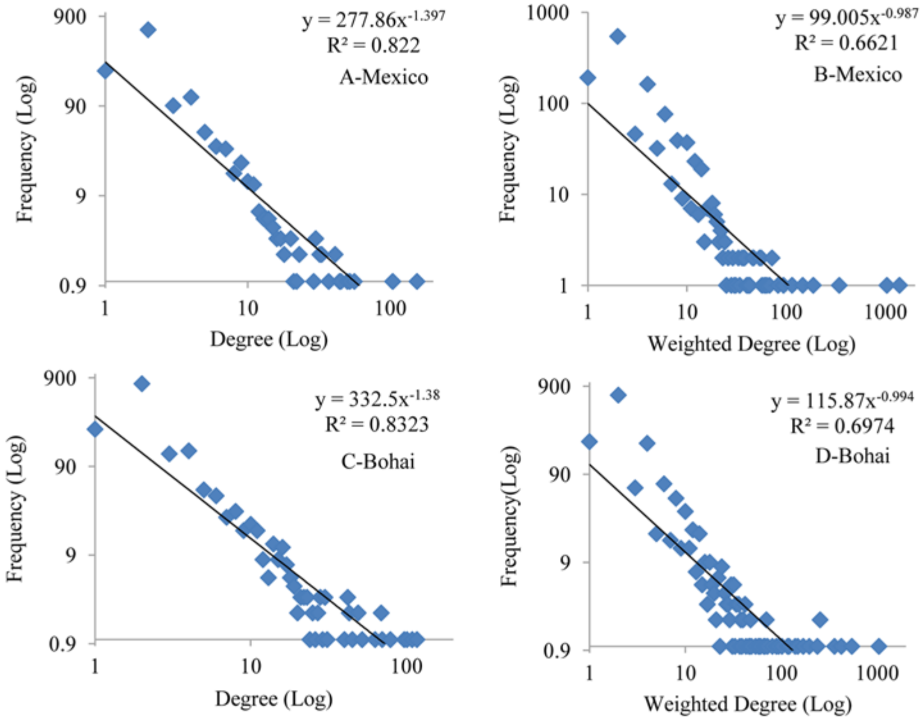

Fig 6. Degree and weighted degree distribution of the whole-sample words networks.

doi:10.1371/journal.pone.0122174.g006

networks. According to Fig. 2, the 2010 Gulf of Mexico oil spill was widely covered by the Chinese media immediately after it occurred, whereas the 2011 Bohai Bay oil spill was not widely covered by the Chinese media until one month after it occurred. Both events were hotly debated for approximately three months in the media; for the Gulf of Mexico oil spill, that time period was May-July 2010, whereas for the Bohai Bay oil spill, that time period was July-September 2011. Thus, in this paper, we construct three different monthly-words networks for each event and analyze their evolution. Fig. 7 shows the visualization results of the monthly-words networks of the two events (the color of the node is determined by the community Id which the node belongs to.), whereas Fig. 8 shows the evolution of nodes and the average degree and the weighted degree of monthly-words networks.

In Fig. 8, the circles represents the different distances between the words and the core keywords, we can discover that the most of the keywords (the big nodes) in different period are closely connected in the networks. Fig. 8 shows that both of the two trending events were most highly concerned by the media in the next month after they occurred. However, since 2010 Mexico oil spill were reported immediately by the parties responsible for the accident, and 2011 Bohai Bay Oil Spill were reported delayed by the the parties responsible for the accident, the online news about the 2010 Mexico oil spill were well published in the month it occurred, and reached the top one month later, then it declined slightly in the third month. Meanwhile, most of the media reported the 2011 Bohai Bay Oil Spill one month later after it occurred, and then it declined slightly in the next two months.

The keywords evolution of the monthly-words network. According to Formula (5) and Formula (10), we obtained the degree and weighted degree of each node in the six different monthly-words networks of the two trending events. Table 3 and Table 4 show the evolution of the Top 10 keywords regarding the two trending events when they were well covered by the

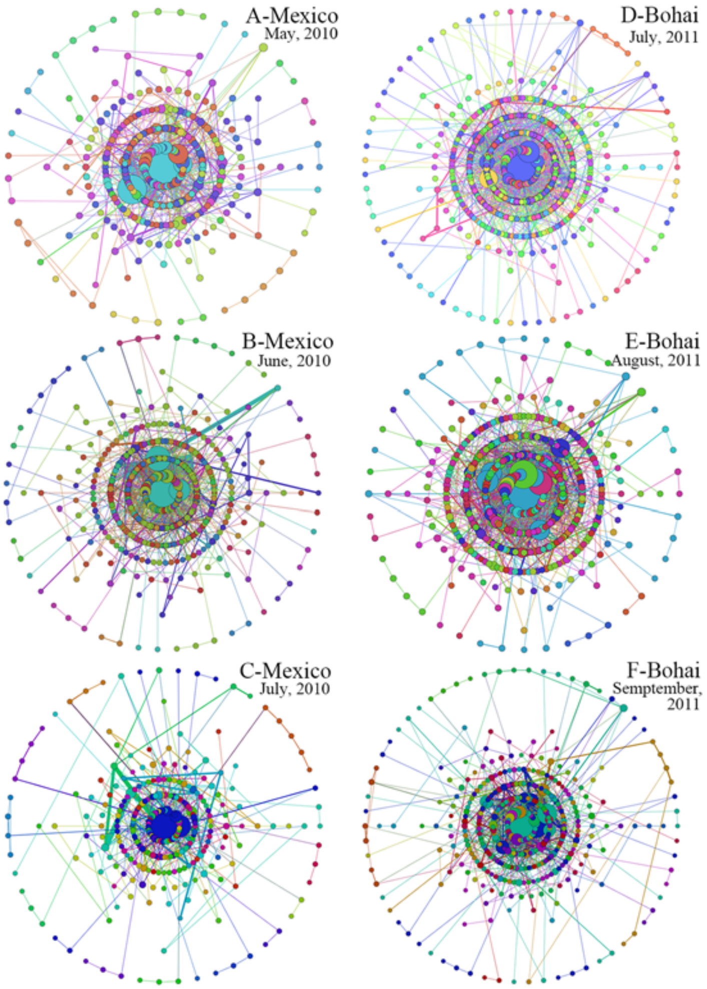

Fig 7. Visualization results of the monthly-words networks of the two trending events.

doi:10.1371/journal.pone.0122174.g007

Chinese media (D represents "Degree", and WD represents "Weighted Degree"). According to the two tables, it is clear that both the trending events have 21 different keywords and that each of the trending events lasted three months. Comparing the Top 10 keywords between the two trending events reveals that the keywords similarity coefficient (Formula (16)) between the two events is only 20%, which means that most of the keywords of the two trending events are different. For a single hot event, there are clear features of evolution; for the 2010 Gulf of Mexico oil spill more of the media outlets were concerned with topics such as "disaster" and "control" as time went by, whereas for the 2011 Bohai Bay oil spill, the media became increasingly concerned with "claims" and "compensation" as time passed. Meanwhile, in the beginning, the media focused more on "CNNOC", and later, more media attention focused on "Conocophillips".

The community evolution of the monthly-words network. According to Formula (15), the modularity (MC) means the independence between the communities, and the members in one community indicate that they have strong connections between one another. For the words networks, the nodes (words) in the same community indicate that they are well connected and more frequently appearing in same titles of the online news about the trending event. Fig. 9 and Fig. 10 show the members' quantity and the total degrees and weighted

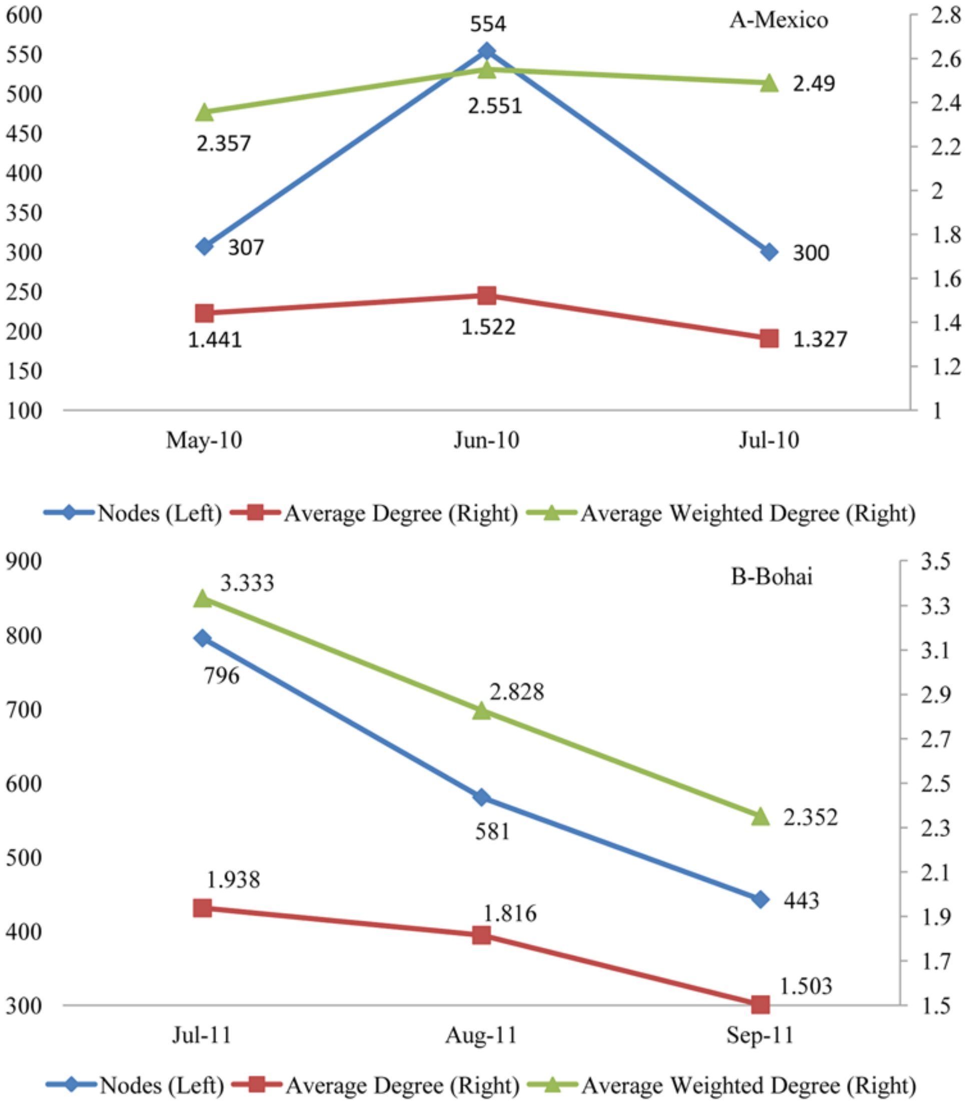

Fig 8. Evolution of nodes and the average degree and the average weighted degree of monthly-words networks about the two trending events.

doi:10.1371/journal.pone.0122174.g008

degrees of each community in different monthly-words networks. Meanwhile, the two figures also show the evolution of the community quantity and modularity of each monthly-words network.

Clearly, in monthly-words networks, the community with the most members does not have the highest total degrees and weighted degrees. For example, in Fig. 9(a), the community with the most members is community 1, but the community with the highest total degrees and weighted degrees is community 5; as we analyzed the members of communities in detail, we found that community 5 contains five of the Top 10 keywords as its members, i.e., W00010, W00595, W00011, W00020, and W00071. Fig. 9 demonstrates that the modularity (MC) decreased from 0.607 to 0.579 and then increased from 0.579 to 0.64, which means the independence of the communities became weaker from May, 2010 to June, 2010, and then, it became stronger from June, 2010 to July, 2010. By contrast, the modularity (MC) of monthly-words

Table 3. Keywords evolution of the 2010 Gulf of Mexico oil spill.

Keywords		May, 2010		June, 2010		July, 2010	
		D	WD	D	WD	D	WD
W00010	Oil Spill	1	1	1	1	1	1
W00595	Gulf of Mexico	2	2	2	2	2	2
W00076	America	3	5	4	5		
W00011	Event	4	3	3	3	3	7
W00077	Company	5	7			6	4
W00020	Accident	6	4	8	4	8	
W00071	Declare	7					
W00012	Impact	8	10	9	9		
W00062	Probably	9		5	7		
W00013	Severe	10					
W00049	Oil		6	10	8	4	3
W00177	Quantity		8				
W00121	Over		9				
W00873	Disaster			6	6	7	
W00219	Already			7		5	10
W00682	sustain				10		
W01410	Success					9	5
W00263	Control					10	
W01895	Sealed						6
W00541	Announce						8
W01001	Britain						9

doi:10.1371/journal.pone.0122174.t003

networks regarding the 2011 Bohai Bay oil spill increased continuously from July, 2011 to September, 2011 according to Fig. 10.

By comparison of Fig. 9 and Fig. 10, we can find that, although the nodes in monthly-words networks about the 2011 Bohai Bay oil spill are larger than the 2010 Gulf of Mexico oil spill, the community quantity of monthly-words networks about the 2011 Bohai Bay oil spill is much smaller than the 2010 Gulf of Mexico oil spill, which means that the words about the 2011 Bohai Bay oil spill are more focused on a few topics and that the words have stronger connections among one another. Fig. 11 shows the links between the communities of the two monthly-words networks, the 2010 Gulf of Mexico oil spill in May, 2010 and the 2011 Bohai Bay oil spill in July, 2011 (the color and size of the node is determined by the out-degree of community). It is obvious that, the communities of the monthly-words network about the 2010 Gulf of Mexico oil spill in May, 2010 is less linked than the communities of the monthly-words network about the 2011 Bohai Bay oil spill in July, 2011. It provides further evidence about why the modularity (MC) of the former monthly-words network is larger than the later one.

The evolution stability of the monthly-words network. To analyze the stability and similarity of the words in different monthly-words networks, we used Formula (16) to calculate the stability coefficient of the monthly-words networks. According to Fig. 12, we can find that the stability coefficient of both the two trending events decreased gradually. In addition, the stability coefficients of the monthly-words networks regarding the 2011 Bohai Bay oil spill are larger than for the 2010 Gulf of Mexico oil spill, which means that the words regarding the 2011 Bohai Bay oil spill in different months are more similar. However, all four stability coefficients

Table 4. Keywords evolution of the 2011 Bohai Bay oil spill.

Keywords		July, 2011		August, 2011		September, 2011	
		D	WD	D	WD	D	WD
W00010	Oil Spill	1	1	1	1	1	1
W00020	Accident	2	2	7	4	7	5
W00009	Bohai Bay	3	3	5	3	2	2
W00011	Event	4	6				
W00069	CNOOC	5	4	8			
W00295	Bohai	6	5	6	5		6(1)
W00440	Pollution	7	10(1)				
W00071	Declare	8	8	4	7	4	
W00203	Conceal	9	9				
W00219	Already	10	10(2)				
W00259	Oil Field		7			10	6(2)
W00026	Conocophillips			2	2	3	3
W00025	Claims			3	6		
W00077	Company			9(1)	8	5(2)	8
W00110	Compensation			9(2)		8	
W00012	Impact			9(3)			
W00023	100 Million				9		
W00187	Apologize				10		
W00053	Fund					5(1)	4
W00043	Still Not					9	10
W00315	To Set Up						9

Note: 100 million is a Chinese unit of quantity, Yi

doi:10.1371/journal.pone.0122174.t004

are less than 0.22, which indicates that most words appearing in the news headlines regarding the two trending events are new.

Discussion and Conclusion

Complex network method has been well used in different empirical areas [44–48]. In this paper, we studied an infrequently considered but quite important method for developing a rapid and deep understanding of all the websites and all the news regarding one topic which integrates statistics, word segmentation, complex network theory and visualization to analyze all the online news headlines' keywords and their evolution regarding two trending events, the 2010 Gulf of Mexico oil spill and the 2011 Bohai Bay oil spill.

We presented an integrated method to analyze both the whole-sample words network and monthly-words network regarding the online news headlines of the two trending events. Through our research, we found that, as with most empirical complex networks, the words networks of online news headlines regarding the two trending events have scale-free characteristics and small-world properties, and the degree assortativity coefficients of the two whole-sample words networks are very low. By calculating the topological features of the nodes, we obtain both the keywords of the whole-sample words network and the keywords of the monthly-words network. Meanwhile, we also obtained the inner relationship and evolution of the words. Compared with the 2010 Gulf of Mexico oil spill, we found that the words regarding the 2011 Bohai Bay oil spill are more focused on a few topics, and the connections between the

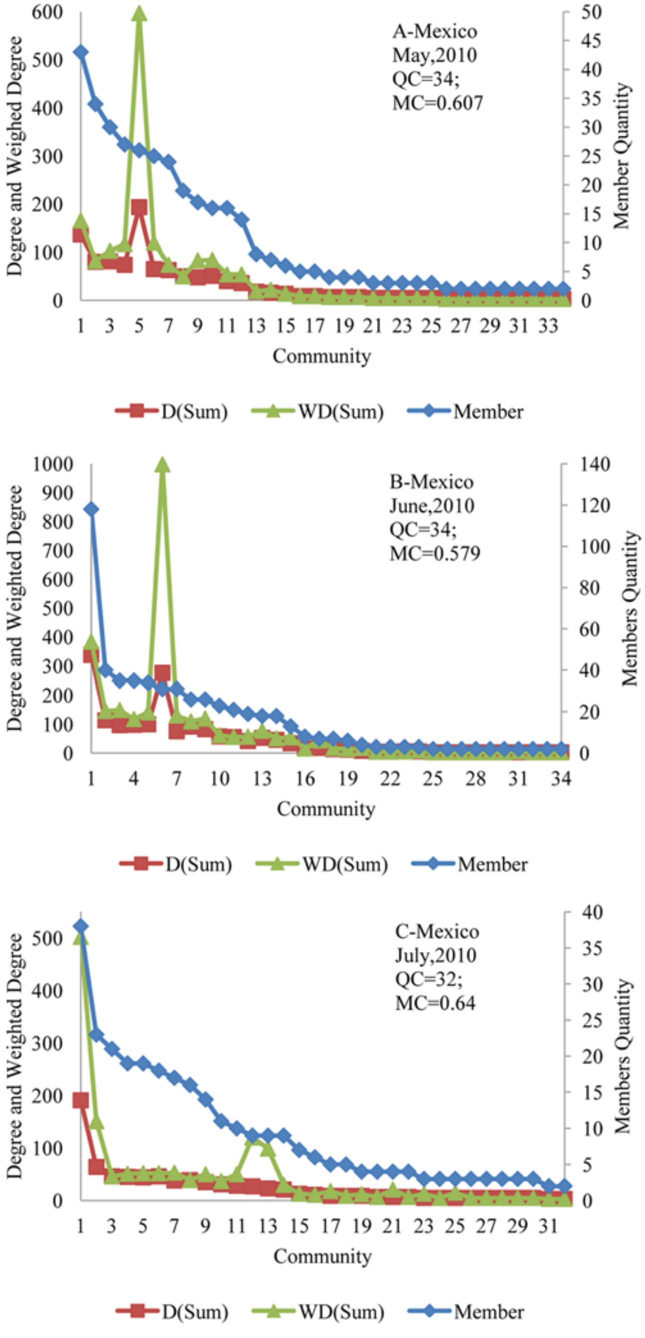

Fig 9. Evolution of monthly-words network community regarding the 2010 Gulf of Mexico oil spill.

doi:10.1371/journal.pone.0122174.g009

words as well as the communities are stronger. We also found that both the words in the online news headlines regarding the 2010 Gulf of Mexico oil spill and the 2011 Bohai Bay oil spill changed obviously as time passed.

Such word-network analysis is a helpful tool with which scholars and companies may analyze and address the public concern regarding an event in a given theme. However, many problems remain to be studied. For example, some of the online news cannot be indexed by existing

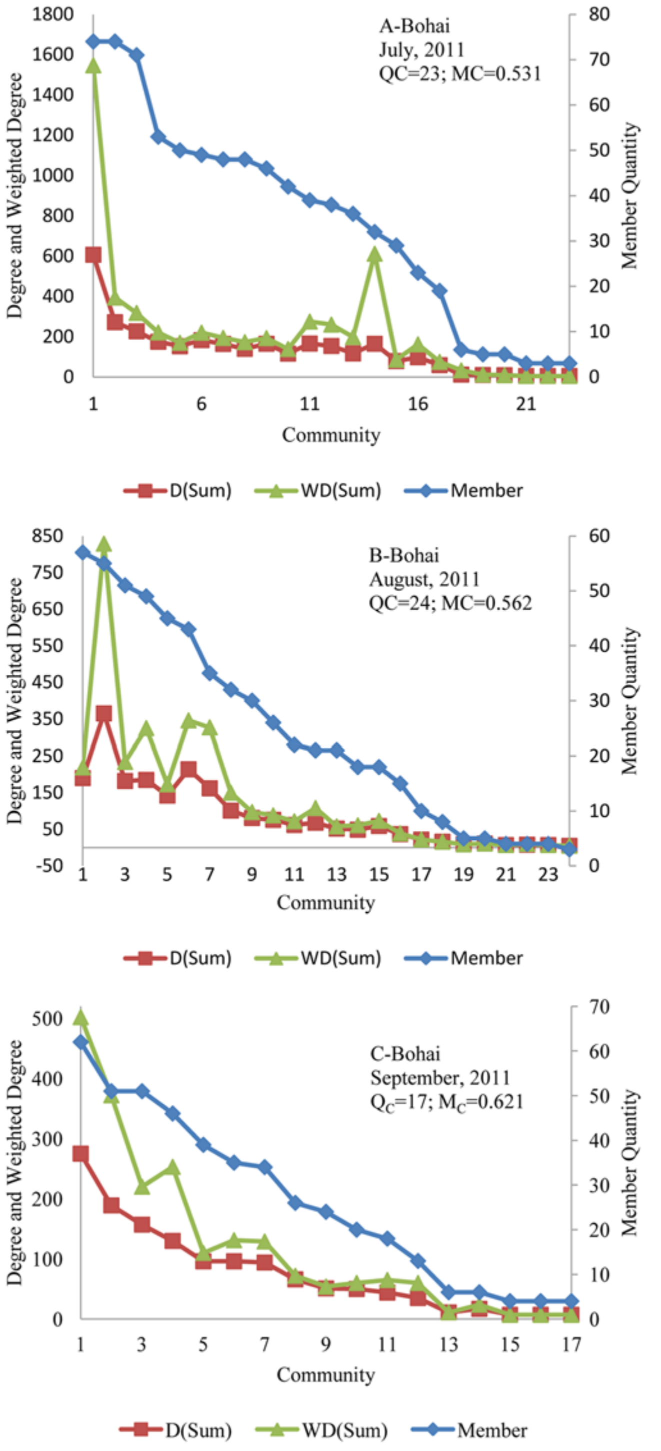

Fig 10. Evolution of monthly-words network community about the 2011 Bohai Bay oil spill.

doi:10.1371/journal.pone.0122174.g010

search engines. If we want to gather information regarding word networks more precisely, we must explore more methods to search the news. Therefore, in the future, we could extend the methods of data searching and try to construct the word networks of the headlines according to reality. Certainly, some of the titles are sensationalized or misleading, which does not reflect

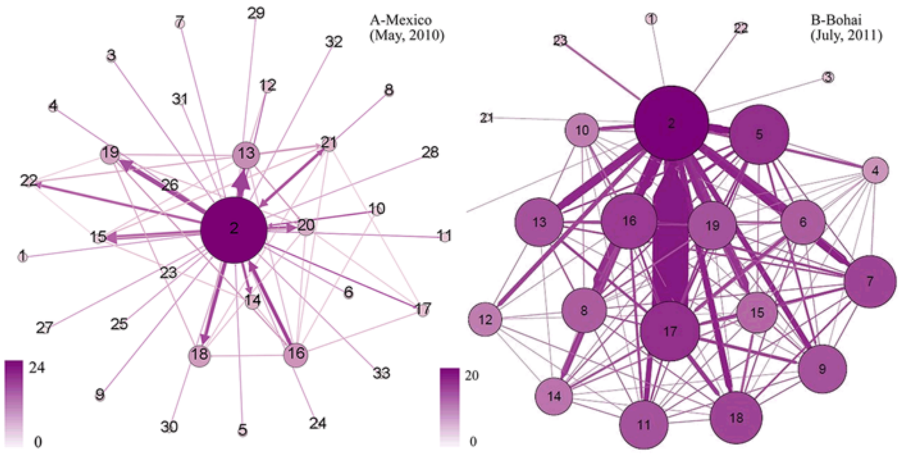

Fig 11. Community networks about the two trending events.

doi:10.1371/journal.pone.0122174.g011

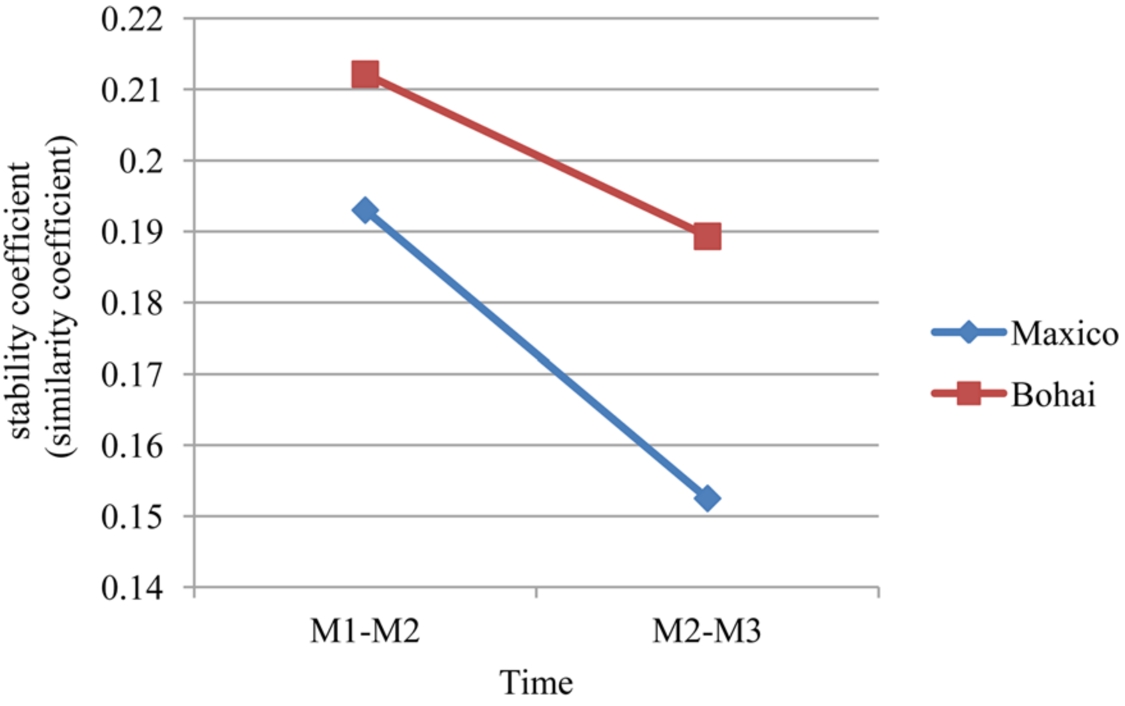

Fig 12. Stability coefficient of the monthly-words networks.

doi:10.1371/journal.pone.0122174.g012

the real meaning of the contents of the news; thus, as a next step, we can identify a new method to judge the degree of correlation between the titles and the contents of the news.

Acknowledgments

The authors would like to express their gratitude to the anonymous reviewers who gave them many helpful and insightful suggestions to modify and improve this research. Meanwhile, the authors would like to thank Songtao Mou, Jiachen Huang, Jing Chen, Xiangyun Gao, Xiaojia

Liu who provided valuable suggestions while writing and revising this paper, as well as AJE for their professional help regarding language usage, spelling, and the grammar in this paper.

Author Contributions

Conceived and designed the experiments: HL HA. Performed the experiments: HL WF. Analyzed the data: HL. Contributed reagents/materials/analysis tools: WF HL. Wrote the paper: HL HA XH.

References

1. Chen DB, Wang GN, Zeng A, Fu Y, Zhang YC. Optimizing Online Social Networks for Information Propagation. PloS one 2014; 9: e96614. doi: 10.1371/journal.pone.0096614 PMID: 24816894
2. Bharat K, Broder A. A technique for measuring the relative size and overlap of public web search engines. Computer Networks and ISDN Systems 1998; 30: 379–388.
3. Risvik KM, Michelsen R. Search engines and web dynamics. Computer Networks 2002; 39: 289–302.
4. Morris MR, Teevan J, Panovich K. A Comparison of Information Seeking Using Search Engines and Social Networks.ICWSM 2010; 10: 23–26.
5. Qiu T, Zhang ZK, Chen G. Information filtering via a scaling-based function. PloS one 2013; 8: e63531. doi: 10.1371/journal.pone.0063531 PMID: 23696829
6. Medo M, Zhang YC, Zhou T. Adaptive model for recommendation of news. EPL (Europhysics Letters) 2009; 88: 38005.
7. Zhang ZK, Liu C. Hybrid recommendation algorithm based on two roles of social tags. International Journal of Bifurcation and Chaos 2012; 22:1250166
8. Chen D, Zeng A, Cimini G, Zhang YC. Adaptive social recommendation in a multiple category landscape. arXiv preprint 2012; arXiv:1210.1441.
9. Shie JS. Conceptual metaphor as a news-story promoter: The cases of ENL and EIL headlines. Intercultural Pragmatics 2012; 9:1–21.
10. Kleinnijenhuis J, Schultz F, Utz S, Oegema D. The mediating role of the news in the BP oil spill crisis 2010: How US news is influenced by public relations and in turn influences public awareness, foreign news, and the share price. Communication Research 2013; 0093650213510940.
11. Utz S, Schultz F, Glocka S. Crisis communication online: How medium, crisis type and emotions affected public reactions in the Fukushima Daiichi nuclear disaster. Public Relations Review 2013; 39:40–46.
12. Mahgoub H, Rösner D, Ismail N, Torkey F. A Text Mining Technique Using Association Rules Extraction. International journal of computational intelligence 2008; 4: 21–28.
13. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999; 27: 1210–4. PMID: 10631500
14. Choi Y, Jung Y, Myaeng SH. Identifying controversial issues and their sub-topics in news articles, In Intelligence and Security Informatics. Springer Berlin Heidelberg 2010;140–153.
15. Balahur A, Steinberger R. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceeding of WOMSA 2009; 9
16. Bhowmick PK. Reader perspective emotion analysis in text through ensemble based multi-label classification framework. Computer and Information Science 2009; 2: 64–74.
17. Yoon J. Detecting weak signals for long-term business opportunities using text mining of web news. Expert Systems with Applications 2012; 39: 12543–12550.
18. Huang CJ, Liao JJ, Yang DX, Chang TY, Luo YC. Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications 2010; 37: 6409–6413.
19. Tanasa D, Trousse B. Advanced data preprocessing for intersites web usage mining. Intelligent Systems, IEEE 2004; 19: 59–65.
20. Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems 2010; 48: 354–368.
21. Lin C, Xie R, Guan X, Li L, Li T. Personalized news recommendation via implicit social experts. Information Sciences 2014; 254:1–18.
22. Wagner H, Dłotko P, Mrozek M. Computational topology in text mining, In Computational Topology in Image Context. Springer Berlin Heidelberg 2012; 68–78.

23. Afzal S, Maciejewski R, Jang Y, Elmqvist N, Ebert DS. Spatial text visualization using automatic typographic maps. IEEE Transactions on Visualization & Computer Graphics 2012; 18: 2556–2564. PMID: 24783264

24. Gürkan A, Iandoli L, Klein M, Zollo G. Mediating debate through on-line large-scale argumentation: Evidence from the field. Information Sciences 2010; 180: 3686–3702.

25. Chen RC, Hsieh CH. Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications 2006; 31: 427–435.

26. Magerman T, Looy BV, Song X. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics 2010; 82: 289–306.

27. Dodds PS, Watts DJ, Sabel CF. Information exchange and the robustness of organizational networks. Proceedings of the National Academy of Sciences 2003; 100: 12516–12521. PMID: 14528009

28. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 2002; 311: 590–614.

29. Hanaki N, Peterhansl A, Dodds PS, Watts DJ. Cooperation in evolving social networks. Management Science 2007; 53: 1036–1050.

30. Li HJ, An HZ, Huang JC, Gao XY, Sh YL. Correlation of the holding behaviour of the holding-based network of Chinese fund management companies based on the node topological characteristics. Acta Phys. Sin. 2014; 63: 48901–048901.

31. Gao X, An H, Zhong W. Features of the Correlation Structure of Price Indices. PLoS one 2013; 8: e61091. doi: 10.1371/journal.pone.0061091 PMID: 23593399

32. Serrano MA, Boguná M. Topology of the world trade web. Physical Review E 2003; 68: 015101. PMID: 12935184

33. Zhang CJ, Zeng A. Behavior patterns of online users and the effect on information filtering. Physica A: Statistical Mechanics and its Applications 2012; 391: 1822–1830.

34. Hu H., Wang X. Evolution of a large online social network. Physics Letters A 2009; 373:1105–1110.

35. Piraveenan M, Prokopenko M, Zomaya A. Assortative mixing in directed biological networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2012; 9: 66–78.

36. Newman MEJ. The structure and function of complex networks. SIAM review 2003; 45: 167–256.

37. Li H, An H, Gao X, Huang J, Xu Q. On the topological properties of the cross-shareholding networks of listed companies in China: Taking shareholders' cross-shareholding relationships into account. Physica A: Statistical Mechanics and its Applications 2014; 406: 80–88.

38. Brandes U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 2001; 25:163–177.

39. Ebel H, Mielsch LI, Bornholdt S. Scale-free topology of e-mail networks. Physical Review E 2002; 66: 035103. PMID: 12366171

40. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008; 10: 10008.

41. Palla G, Barabási A L, Vicsek T. Quantifying social group evolution. Nature 2007; 446: 664–667. PMID: 17410175

42. Li H, Fang W, An H, Yan L. The shareholding similarity of the shareholders of the worldwide listed energy companies based on a two-mode primitive network and a one-mode derivative holding-based network. Physica A: Statistical Mechanics and its Applications 2014; 415: 525–532.

43. Newman MEJ, Assortative mixing in networks. Physical review letters 2002; 89: 208701. PMID: 12443515

44. Qi H, An H, Hao X, Zhong W, Zhang Y. Analyzing the International Exergy Flow Network of Ferrous Metal Ores. PloS one 2014; 9: e106617. doi: 10.1371/journal.pone.0106617 PMID: 25188407

45. Hao X, An H, Liu X, Gao X, Cong L. Analysis on main mineral products in international trade. Resources & Industries 2013; 15: 35–43

46. An H, Gao X, Fang W, Huang X, Ding Y. The role of fluctuating modes of autocorrelation in crude oil prices. Physica A: Statistical Mechanics and its Applications 2014; 393: 382–90.

47. An H, Zhong W, Chen Y, Li H, Gao X. Features and evolution of international crude oil trade relationships: A trading-based network analysis. Energy 2014; 74: 254–259.

48. An J, An H, Yang G. Relation of financeial institutions and listed mining entities in equity financing based on complex network. Resources & Industries 2014; 16: 124–1