Sustainability **2014**, *6*, 1896-1912; doi:10.3390/su6041896

OPEN ACCESS

sustainability

ISSN 2071-1050

www.mdpi.com/journal/sustainability

Article

Using Web Crawler Technology for Geo-Events Analysis: A Case Study of the Huangyan Island Incident

Hao Hu [1], Yuejing Ge [1,*] and Dongyang Hou [2]

[1] School of Geography, Beijing Normal University, 100875 Beijing, China; E-Mail: bsdhao@gmail.com

[2] School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; E-Mail: houdongyang1986@gmail.com

* Author to whom correspondence should be addressed; E-Mail: geyj@bnu.edu.cn; Tel.: +86-139-106-26308.

Received: 21 February 2014; in revised form: 26 March 2014 / Accepted: 26 March 2014 / Published: 9 April 2014

Abstract: Social networking and network socialization provide abundant text information and social relationships into our daily lives. Making full use of these data in the big data era is of great significance for us to better understand the changing world and the information-based society. Though politics have been integrally involved in the hyperlinked world issues since the 1990s, the text analysis and data visualization of geo-events faced the bottleneck of traditional manual analysis. Though automatic assembly of different geospatial web and distributed geospatial information systems utilizing service chaining have been explored and built recently, the data mining and information collection are not comprehensive enough because of the sensibility, complexity, relativity, timeliness, and unexpected characteristics of political events. Based on the framework of Heritrix and the analysis of web-based text, word frequency, sentiment tendency, and dissemination path of the Huangyan Island incident were studied by using web crawler technology and the text analysis. The results indicate that tag cloud, frequency map, attitudes pie, individual mention ratios, and dissemination flow graph, based on the crawled information and data processing not only highlight the characteristics of geo-event itself, but also implicate many interesting phenomenon and deep-seated problems behind it, such as related topics, theme vocabularies, subject contents, hot countries, event bodies, opinion leaders, high-frequency vocabularies, information sources, semantic structure, propagation paths, distribution of different attitudes, and regional difference of net citizens' response in the Huangyan Island incident. Furthermore, the text analysis of network information with the help of focused web crawler

is able to express the time-space relationship of crawled information and the information characteristic of semantic network to the geo-events. Therefore, it is a useful tool to collect information for understanding the formation and diffusion of web-based public opinions in political events.

Keywords: web crawler technology; text information; sentiment analysis; Huangyan Island Incident

1. Introduction

With the increasing influence of social networking in the era of big data, especially in economic, social, and political fields, there are abundant text information and social relationships being brought into our daily lives. Researchers are becoming more and more interested in using these big data to study the phenomenon and discipline of our changeable society. The emergence of new search engine technology is making the information grabbing and data collecting more easily and efficiently. The web crawlers as one of the tools have been widely used in agricultural exploration [1], information searching and data mining [2,3], toponym database updating [4], infection disease monitoring [5], network monitoring and management of cultural content [3]. Especially with regard to computers, there are many more research topics on the development of technology. Meanwhile, the data processing and data analysis have also developed quite rapidly in many areas, such as media spreading [6,7], university management [8], information security [9–11], tourism brand management [12,13], city information technology [14], web portal evaluation [15], and other public domains. Whether this kind of informational data collecting and text data processing can be used to analyze the geopolitical events, and how to use the modern computer technology to service the international politics are what we want to explore in this study. After an unexpected political event occurs and spreads on the internet, the news reports and comments of net media can objectively reflect the concerns and opinions of the net media and citizens. Thus, the information collection and text analysis following political events are very important for the stability of network society and the security of national information. Though politics have been integrally involved in the hyperlinked world issues since the 1990s [16,17], and researchers have been explored automatic assembly of different geospatial web services to build distributed geospatial information systems utilizing service chaining for many years [18,19], it seems that little work has been performed in the area of political geography analysis.

When political events or controversial issues arise, new descriptions and opinions can be published without limits of location, time, age, political persuasion, subject, and form [20]. The abundance of information on the Internet presents a challenge concerning the quantitative analysis of political events. However, it also presents an opportunity with regard to the analysis of social movements, public sentiments, and social dynamics at the national level [21]. When political events spread on the network, they are always concerned by a majority of internet users. Once these events being to propagate in the social networking and network worlds, information dissemination develops quickly, spreading from its origin point to the entire Internet. As the time goes by and event develops, the spatial scale and content of the dissemination may change. What is more, the awareness to a political event is diverse because of its

complicated relationships with regard to the regional economy, regional security, social stability, production, and other socio-cultural factors. Public opinions on events always develop rapidly, change continuously, and transcend the limitations of time and space. As a result, the event analysis of geo-events is very different from the others. For instance, information collecting and data mining must be sufficiently comprehensive and fully focused on the event topic, and the text analysis should visually represent the characteristics of space and time relating to the event. The event analysis of geopolitical event must be traced back to the origin of the event (the location and time of the first network news report) and the related on-going reports in various network communities. It requires knowledge of the spatial location and spatial relationship of a geopolitical event and its reporting network. It requires to make clear the associated event (e.g., the related content and topics), as well as to locate and illustrate the temporal and spatial relationships between the spreading network and the developing event. A focused web crawler can automatically burrow into the link structure of web documents and index the hidden information in the document structure based on keyword definitions. With text analysis of the grabbing information and data, we can decode and monitor the public reactions to social controversial issues and track the disseminated information of breaking stories on networks. It also provides a visual representation of network media concerns, emotional trends and semantic networks concerning public consciousness, which can be helpful for knowledge-based decision making. The combination of web crawler for data collecting and text analysis for data analyzing in geopolitical research will be much more efficient, objective, and comprehensive than the traditional methods, which are based on manual collection of related information from newspapers and related statistical materials. In this study, the domestic internet news regarding Huangyan Island was tracked to provide a data source for text analysis. Meanwhile, the word frequency, net citizen emotional trends and network propagation of this political event were given vivid visual expressions by using network text analysis. The combination of web crawler technology and the text analysis method on Huangyan Island incident in 2012 will provide a preliminary exploration of the geopolitical event analysis based on computer technology.

2. Data Sources and Methodology

2.1. Background of Huangyan Island Incident

The Huangyan Island Incident (HII, for short) is a representative geopolitical event in 2012. On April 10, Chinese fishermen in a lagoon of China's Huangyan Island (also referred to as Scarborough Shoal before 1983, internationally) were harassed by a Philippine naval gunboat. The Philippine warship and navy attempted to arrest the Chinese fishermen and were stopped by a Chinese Maritime Surveillance ship in the South China Sea. Both sides claimed the Huangyan Island as their sovereign territory. A confrontation between the Philippines' largest warship and the Chinese fishery administration ship occurred. The Chinese Foreign Ministry suggested dealing with this issue through diplomacy, whereas the Philippines insisted on bringing the issue to the international court. They staged a protest by demolishing Chinese buildings, burning Chinese flags, invoking neighboring countries against China, and even intending to rename Huangyan Island as Panatag Shoal. China established Sansha City to consolidate management of the South China Sea. This conflict lasted for more than two months and was referred to as the HII in 2012. From April 11, 2012, at 10:36 a.m. to June 4, 2012, at 11:54 p.m., a total of

273 websites referring to the HII and 2855 news reports or hotspot comments on the Internet were found. Removing the redundant information, only 2589 new reports and commentaries are used for further text analysis. The websites that reported the above news more than 10 times are shown in Table 1.

Table 1. The list of websites that reported news on the Huangyan Island Incident (HII) for more than 10 times.

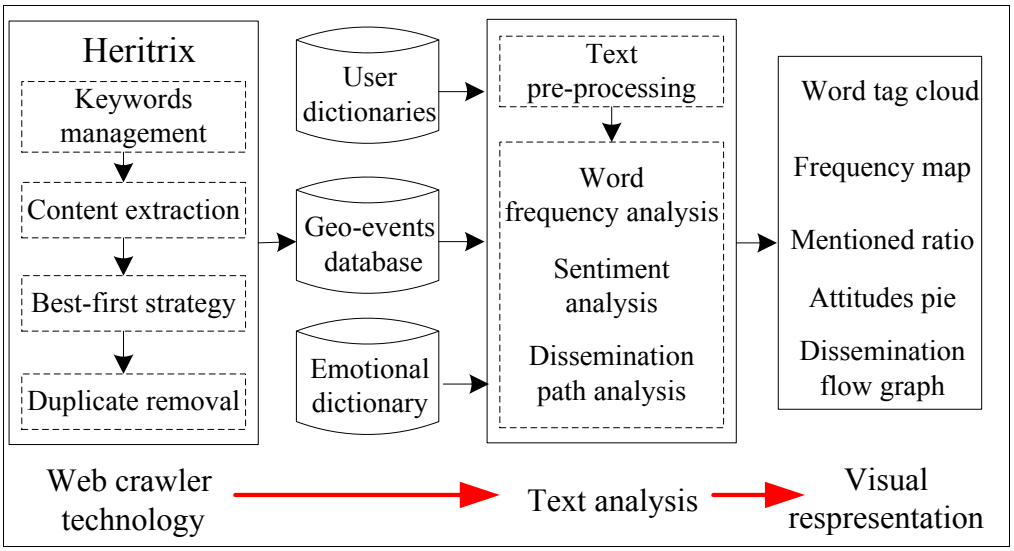
Website	New reports	Website	New reports	Website	New reports
Ifeng.com	484	china.com	43	stcn.com	17
sohu.com	156	cnfol.com	33	cutv.com	15
sina.com.cn	150	ce.cn	30	htexam.com	14
people.com.cn	138	cankaoxiaoxi.com	29	qianlong.com	14
kankanews.com	124	ws.hbtv.com.cn	29	eastday.com	12
caixun.com	114	s1979.com	29	huagu.com	12
hexun.com	99	china.com.cn	26	joy.cn	12
591hx.com	86	jrj.com.cn	25	163.com	11
21cn.com	74	huanqiu.com	23	cnr.cn	11
stockstar.com	74	chinanews.com	19	cntv.cn	11
xinhuanet.com	73	guancha.cn	18	cnstock.com	11
qq.com	60	cfi.net.cn	18	bjyouth.ynet.com	10

2.2. Web Crawler Technology

A web crawler, known as a network robot or spider, is a program, software, or programmed script that automatically fetches large collections of web pages according to http protocol and some clearly defined crawling strategies (e.g., depth-first strategy, breadth-first strategy, and best-first strategy) [22,23]. It can be categorized into general purpose web crawlers and focused web crawlers according to the purposes of the tools. The general purpose web crawlers, such as Heritrix [24], Nutch [25], and Labin [26], aim to download any web pages obtained through the hyperlinks, while focused web crawlers or topic-specific web crawlers intend to download as many web pages relevant to the given topic automatically and keep the amount of irrelevant web pages downloaded to the minimum [27,28]. Currently, web crawlers have been widely used in many fields, such as search engines, web data mining and business intelligence, *etc.*, [28,29]. However, web crawlers are seldom used in geopolitical events analysis. Therefore, in the case study of geo-event on Huangyan Island incident, we kept fetching and extracting relevant web pages in the duration of the geopolitical events to reasonably analyze the Huangyan Island incident by using the improved web crawler of Heritrix. We improved the Heritrix by adding keywords management module, content extraction module, best-first strategy, and duplicate removal module, as shown in Figure 1. Through keywords management module, we input defined keywords of Huangyan Island incident as filter conditions in best-first strategy and we extracted the title, abstract, sources, release time, and hyperlinks through content extraction module. We preferentially obtained the relevant text information of the geo-events from mainstream media with the help of Chinese segmentation on the Internet through the best-first strategy. Furthermore, we removed the relevant duplicate web pages and got their transfer number and propagation paths through duplicate removal module by using information fingerprint algorithm [30]. Finally, we indexed and ranked the

above information and stored them into a geo-events database, which was the up-to-date corpus of the text analysis.

Figure 1. The framework of the methodology.



2.3. Text Analyzing Method for Geo-Events

Text analysis of geo-events mainly includes text pre-processing, word frequency analysis, sentiment analysis, and dissemination path analysis (Figure 1). Text pre-processing contains sentence extraction and Chinese word segmentation based on the defined user dictionaries and emotional dictionary. Word frequency analysis includes the analysis of related topics, the production of a frequency map and the evaluation of the individual mention ratio. Sentiment analysis focuses on the comparison of different attitudes with regard to the HII, and dissemination path analysis represents the temporal and spatial relationship of a geopolitical event and its reporting network. A tag cloud is the most basic component for word frequency analysis, and the Likert Scale is the most important tool for sentiment analysis. Domain name confirmation and the IP address location are preconditions for information dissemination analysis.

A tag cloud can provide a vivid visual expression of word frequency. The greater the word frequency, the closer the word will be located to the central position and the higher the word position will be. Thus the most frequent word is the largest, the most eye-catching and located most centrally in the tag cloud. Due to the inherent limitations of Chinese segmentation and word frequency statistics in data processing, segmentation methods are improved by training the automatic results of Chinese network segmentation. After grabbing and crawling new reports of the HII, user dictionaries (e.g., a personal name dictionary, a geographic name dictionary, a proper nouns dictionary, a new internet words dictionary, the filter word group table, the reserved word group table, the emotional word group table, and the merge word group table) are established from the extracted information of crawling words and the statistic calculation of word frequency. Next, the processes of word segmentation and text filtering are adjusted by using the user dictionaries, text manipulations of titles, abstracts, and comments. Then word frequency and vocabulary analyses are repeatedly performed based on the ROST software (A content mining system

designed by the ROST virtual learning team in Wuhan University) [31]. Lastly, in order to provide text visualization [32] of the HII, the tag cloud is generated by using the ROST software.

The Likert Scale is one of the most widely used approaches to scale the responses in the surveys. Rensis A. Likert (1932), an American social psychologist, developed the principle of measuring attitudes by asking people to respond to a series of statements on a topic. Currently, the Likert Scale is often used to measure the attitudes or opinions of respondents and study the differences in personal subjective feelings. The scale has become an important measurement tool in social science research. According to the research needs, a graded assignment can be divided into models with three, five, and seven rating scales to express attitudes and measure the degree of a positive or negative attitude. Lubke asserted that the Likert Scale multistage rating model had a better imitative effect. In addition, the higher the level, the greater the measurement accuracy is [33]. The process of sentiment analysis for the HII mainly includes two steps. First, the various terms are classified into verbs, nouns, and adjectives by using an expert evaluation method and a constructed emotional dictionary based on the word frequency and the importance of words (in this study, we only invited Chinese diplomats to be the experts, so the result may only represent the opinions of the Chinese officials). Second, the results are divided into seven categories by Likert Scale, that is −3, −2, −1, 0, 1, 2, and 3 [34–36]. Positive or negative statements are analyzed by virtue of the Likert score and the average emotional value of the sentence. If a word is a derogatory term, the Likert score of the word is below zero; if the word is an appreciative term, its Likert score is above zero (e.g., war, fight −3; affect, propose 0; cooperation, respect 3). All the Likert scores and emotional values are evaluated by international relations experts. The process is as follows: (*i*) each new report title is turned into "one sentence in a line" to start the text manipulation process of grabbed information; (*ii*) the words with high-frequency are extracted from the results of the text manipulation process to update the emotional dictionary; (*iii*) the processes of word segmentation and segmentation correction are started according to the reserved word group table and the merged word group table in the user dictionaries; (*iv*) the final segmentation results of the news report are obtained after the insignificant auxiliary words and interjections being filtered; (*v*) the final segmentation result are matched with the words in the emotional dictionary, and the Likert score and emotional value of each word and report are obtained; (*vi*) the average emotional value of each sentence is calculated using the formula of the emotion analysis module to test and estimate the emotional analysis effect of the HII.

$$\mathrm{E}_{si} = \frac{1}{\mathrm{n}} \sum_{j=1}^{\mathrm{n}} \left[e_{ij} \times w_{ij} \right] \tag{1}$$

where E_{si} = average emotional value of sentence i

e_{ij} = emotional value of word j of sentence i

w_{ij} = weight of the word j

n = the number of words in the sentence.

Dissemination path analysis is based on the number of news reports and the domain names of news reports source which have a significant effect on the public opinions. It is well known that once a report be reprinted by other websites, the information is likely to spread rapidly to other network communities. The more the reprinted reports, the greater the bandwidth be occupied on the internet and the faster the speed of information dissemination is. Three steps are necessary to analyze how new reports and public opinion spread throughout a network. First, 36,323 news reports on the HII are crawled from the Internet

within a specific period of time. Second, 33,734 news reports that were reprinted from another website are collected, and all the sources (including the time and places) of the remaining 2589 new reports are obtained after information sort of the reported sources and reported time. Third, an information flow of reprints is drawn for the visualization of information dissemination after confirming the domain names of the remaining 2589 reports sites and locating the IP addresses of web servers with the help of a webmaster tools website [37].

3. Word Frequency Analysis of HII

3.1. Related Topics of the HII

The tag cloud of word frequency for the HII includes 210 words with high-frequency of more than 50 times. The cloud does not only highlight the subject vocabularies and the theme vocabularies, but it also vividly demonstrates a few issues and problems behind the HII. Overall, for the tag cloud, the closer a word is to the central position, the higher the frequency of that word will be. From the center to the border, the colors of the labels gradually change from red to green, with the red color representing greater word frequency and a higher grade (see Figure 2). Word frequency statistics and the tag cloud show that the frequencies of "China", "Philippines", "Huangyan Island" are the highest, between 3700 and 4100 times. This level is much higher than the following level between 1000 and 1200 times. Therefore, these words become the main body of the analysis of the HII because they are most frequently used in long-term representations of the Huangyan Island confrontation. The words "confrontation", "the South China Sea", "the United States", and "the Philippines" have word frequencies of more than 1000 times. The results reflect the fact that the interests of the net media and net citizens are not limited to the confrontation between the Philippines and China but also include the South China Sea issue and the international influence of economics and military power. Some words with high-frequency (e.g., "sea area", "problem", and "Huangyan Island incident", with frequencies between 500–800 times) and related reef sea area theme words (e.g., Diaoyu Islands, Nansha, the Pana Tug Reef, the Zhongsha Islands, Boracay, Subic Bay, Thitu Island, Scarborough Shoal, Xisha, middle ground, the Reed Bank, Luzon, Yongshu Reef, Scarborough Shoal, and Mischief Reef) demonstrate the logical relationship between the HII and the sovereignty dispute of the territorial sea. In addition, geopolitical-environment vocabularies, such as "countries", "sovereignty", "foreign affairs", "situation", "territory", "security", "peace", and "surrounding", also frequently appear in the cloud tag. These words reflect the different degrees of attention that net citizens pay to the affairs of the state, state sovereignty, national security, and boundary disputes. What is more, phrases, such as ocean surveillance ship, dialogue, protest, appeal to, solemn representations, and weather forecast, reflect the fact that the Chinese net media and net citizens pay close attention to Chinese diplomatic behaviors related to the HII. Other high-frequency words, such as Chinese, the Philippines, fishermen, fishing boats, economy, military project, tourism, banana, travel agency, petroleum, and Filipino domestic helpers indicate that international politics and international sovereignty disputes affect fishery productions, stock markets, arms sales, international travel and the import and export of products significantly. They also demonstrate that international friction has a significant influence on the regional economy, regional security, regional tourism development, and international trade. Words related to the South China Sea, such as "intra-area

countries", "the intervening countries outside the area" [38], "neighboring countries", and "international association countries", also frequently appear in the comments and show the complex international relations and their linkage to the development effect of hot issues in the context of globalization. Therefore, geo-events are not only associated with the geographical locations but also closely related to geo-political, geo-economic, geo-science, and technological factors. There is a big difference between the text analysis of geo-events and that of other events.

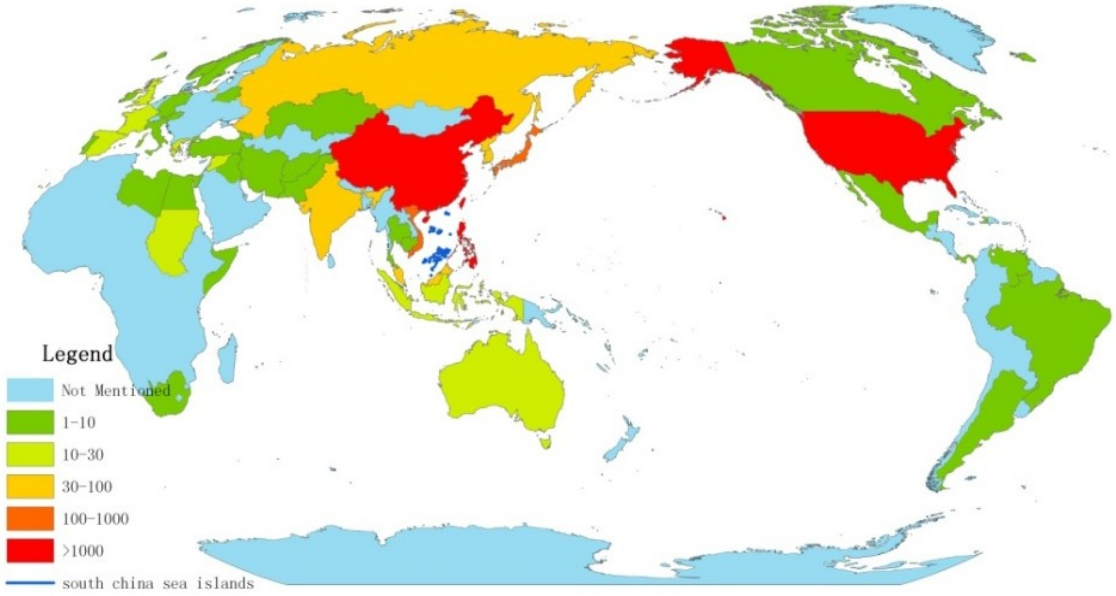
Figure 2. Tag cloud of HII.

3.2. Frequency Map of the HII

In addition to China and Philippines, the names of some other countries are also frequently mentioned. A total of 53 countries among the 193 existing countries appear in comments on the HII. As shown in Figure 3, China, Philippines, and the United States are repeated over 1000 times and are the hot countries related to HII. Although it is far from China and Philippines, the United States was directly involved in this event. To a certain extent, this finding provides an evidence of the United States' motivations and its national security strategy of "returning to the Asia-pacific region". Compared to the frequencies of the United States, main countries, such as Japan and Vietnam, ranked in the second layer with frequency of 100–300 times. These two countries were involved in the review reports on the HII because of their close proximity to the South China Sea. North Korea, India, Russia, South Korea, Malaysia, Singapore, Indonesia, Australia, and other neighboring countries to the South China Sea are also mentioned many times in the reported comments of the HII. The confrontation of the HII has triggered the discussions on the South China Sea, national sovereignty, territorial dignity, and the security concerns of the neighboring countries. All of these can be seen as a concrete manifestation of the social networking that expedites the spread of information and creates internet momentum. Greece, the United Kingdom, France, Spain, Sudan, and Syria, which are not directly connected to the HII, also

become the discussion subjects during the event due to their economic crises and political activities. Another 35 countries, including, Brunei, Iran, Thailand, Maldives, Germany, Canada, and Iraq are mentioned less than 10 times in the HII. The HII did not only provoke the discussions on neighboring countries around the South China Sea, but also cause great discussions on powerful countries and related countries and regions.

Figure 3. Frequency map of the HII.



3.3. Individual Mention Ratio of the HII

Among news reports obtained from web grabbing and crawling for text information on the HII, there were 996 personal names mentioned and 67 people involved in the confrontation from 11 April 2012 to 4 June 2012. Their individual effects on the HII were analyzed using the data of the mentioned names and the rates of the mentioned individuals (individual mention number divided by the number of all). The individuals and their rates of being mentioned in the HII are shown in Table 2.

Table 2. Individual mention ratio in the HII (unit: %).

Individuals	Ratio	Individuals	Ratio	Individuals	Ratio	Individuals	Ratio
Aquino III	9.74	Nicano Fildo	1.51	Jie Zhang	0.60	Jin Xun	0.20
Albert Del rosario	8.63	Bing Chen	1.41	Guoqiang Li	0.60	Kejin Zhao	0.20
Guanglie Liang	7.43	Ming Yao	1.00	Yoshihiko Noda	0.60	An Jiang	0.20
Lei Hong	7.23	Ningsi Lv	1.00	Xiaotian Ma	0.50	Walter	0.20
Weimin Liu	6.93	Guotu Zhuang	1.00	Zhongping Song	0.50	Xijin Hu	0.10
Ying Fu	5.72	Hao Zheng	1.00	Zhirong Yu	0.40	Xinyu Leng	0.10

Table 2. *Cont.*

Hillary Clinton	5.62	Yizhou Wang	1.00	Baosheng Xue	0.40	Xiaolin Ma	0.10
Leon Panetta	4.22	Xiaoling Tong	0.90	Xiaoying Sun	0.40	Liwei Zhang	0.10
Yuan Luo	3.31	Xiaojun Song	0.90	Cameron	0.40	Wuchang Zhang	0.10
Cishan Ruan	3.31	Liangliang He	0.90	Haiquan Ren	0.30	Yun Zhang	0.10
Xiaofeng Jiang	2.51	Binet	0.90	Bingguo Dai	0.30	Guoqing Zhang	0.10
Voltaire Gazmin	2.31	Guangqian Peng	0.80	Dongxing Huan	0.30	Jun He	0.10
Shicun Wu	1.91	Shintaro Ishihara	0.80	Wenlong Du	0.30	Yiming Guo	0.10
Zhaozhong Zhang	1.61	Kudashev	0.80	Naguib	0.30	Yongchun Liang	0.10
Putin	1.61	Brady	0.80	Viktor.n arches ii	0.30	Jiechi Yang	0.10
Obama	1.51	Ma Keqing	0.70	Medvedev	0.30	Yijian Ye	0.10
Zhonghua Deng	1.51	raul	0.70	Hao Su	0.20	Total	100.00

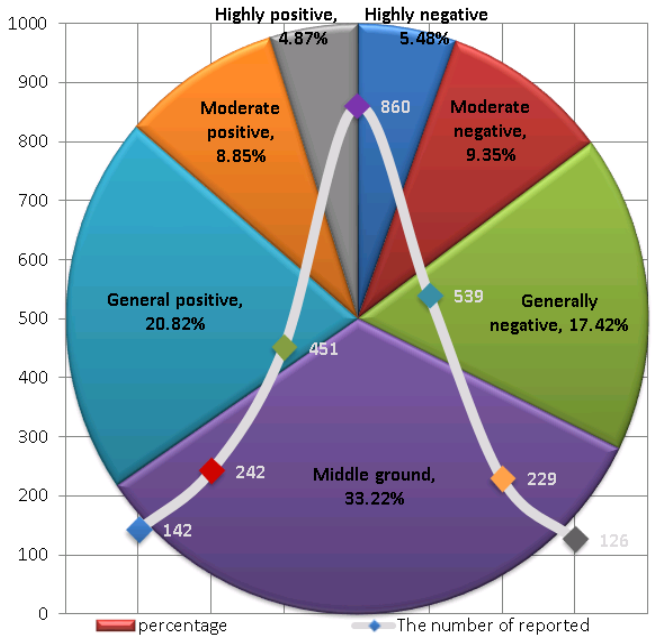
The President of Philippines, Benigno Aquino III, was the most frequently mentioned, followed by the Minister of Foreign Affairs of Philippines Del Rosario, the Chinese Defencs Minister Guanglie Liang, Foreign Ministry spokesman Lei Hong, Foreign Ministry spokesman Weimin Liu, Deputy Foreign Minister Ying Fu, and U.S. Secretary of State Hillary Clinton (ranked in the top 10 and a mention ratio of above 5%). The statement and behavioral responses of the President of Philippines and the Departments of Foreign Affairs of China and Philippines became hot topics in the network media and rolled into public concerns during the HII. In addition, the U.S. Secretary of State Hillary Clinton, the U.S. Secretary of Defense Panetta and the U.S. President, Obama, had unexpectedly high frequencies (ranked in the top 20 with mention ratios above 1.5%), which revealed the strategy of the US foreign policy that to strengthen the effect in the Asia-Pacific region. Russian President Putin was also reported with a high frequency (ranked top 15), indicating that Russia also had its international influence in the region disputes. The Governor of Tokyo Shintaro Ishihara and the Japanese Prime Minister Yoshihiko Noda attempted to increase the severity of the HII to provide an excuse for their dispute on the Diaoyu Islands. The former Philippine Marine Corps captain Nicano Fildo attempted to land on the Huangyan Island and raise the Phillipines' flag but blocked by Philippine President Aquino III. It indicated that Philippines were inconsistent with its attitude toward Huangyan Island and the attitude changes with the international behavioral development of great powers. It is an interesting finding that the sports star Ming Yao had a high frequency during the Huangyan Island confrontation (ranked in top 20), because Philippines once attempted to use basketball diplomacy with China to divert Chinese public attention from a political and military confrontation. Commentators on the HII, including Yuan Luo, Cishan Ruan, Xiaofeng Jiang, and Shicun Wu and other experts, as well as the leaders' suggestions from the authoritative network community, affected the viewpoints and opinions of the net media and net citizen. Surprisingly, the Minister of Foreign Affairs of China Jiechi Yang had little effect on the HII, with a low mention ratio and minor placing. All these factors provide the evidence of the complex

situation of the South China Sea and the critical issues of China's peripheral security. The findings show that the statement and behavioral responses of the diplomats became hot topics in the network media and filled with public concerns during the build-up of the HII.

3.4. Sentiment Analysis of the HII

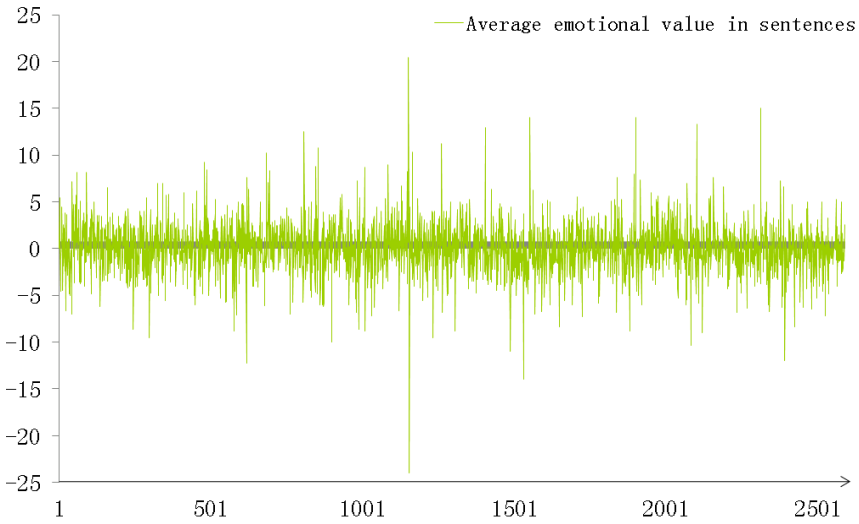
In this study, sentiment analysis of the HII is conducted by clustering analysis of the network text and network opinions on basis of weight assignment according to the Likert Scale. Of all the statements in the new reports, the neutral, positive and negative statements account for 33.22%, 34.53%, and 32.25% of the statements, respectively. The distribution of various attitudes in different network communities exhibited an upside-down "V" shape in the HII (See Figure 4). Most people are holding neutral opinions. The ratio decreased rapidly from neutral attitudes to unclear attitudes and from neutral attitudes to extreme attitudes. The generally positive statements, with an emotional value between 5 and 15, account for 20.82%, which is 2.4 times more than the moderate positive statements (emotional value between 15 and 25) and 4.3 times more than the highly positive statements (emotional value between 25 and 85). There were more negative attitudes in network media news reports than positive attitudes for the HII. Approximately, there were one third of the people commenting on the event had negative attitude. When the neutral opinions of new reports (*i.e.*, general positive emotions, general negative emotions, and neutral emotions) were ignored, the proportion of negative positions (*i.e.*, highly negative emotions and moderate negative emotions) reached as high as 14.83%. Whereas, the proportion of positive positions (*i.e.*, highly positive emotions and moderate positive emotions) accounted for 13.72% of all new reports. Additionally, the ratio of highly negative statements was 0.61% higher than that of highly positive statements. The percentage of moderate negative statements was 0.5% higher than that of moderate positive statements. These ratio differences suggest that the ideological content orientation and commentary guidance of news reports and net media could provide important contexts to the public opinions on the geo-events.

Figure 4. Comparison of attitudes pie toward the HII.



As the value of Likert score in the emotional dictionary and the quantity of matched words in the new reports may have some interference to the precision of emotional value in the grabbing information, error analysis must be done to measure all the error are acceptable in the sentiment analysis. The effect of emotional analysis of the HII was estimated by assessing the discrete average emotional value of each sentence. The average emotional value could be used to reflect the error size of the entire reported emotional analysis. The greater the average emotional value, the greater the error of the emotional analysis would be. As shown in Figure 5, the emotional value of the 2589 new reports fluctuated above and below the 0-axis. There were only 2, 3, 18, and 147 news reports with errors exceeding the Likert Scale values of 20, 15, 10, and 5, respectively. In total, there were 2442 news reports with errors of the emotional analysis ranging from −5 to 5, indicating that 94.32% of error could be controlled within a reasonable range.

Figure 5. Average emotional values of news reports for the HII.

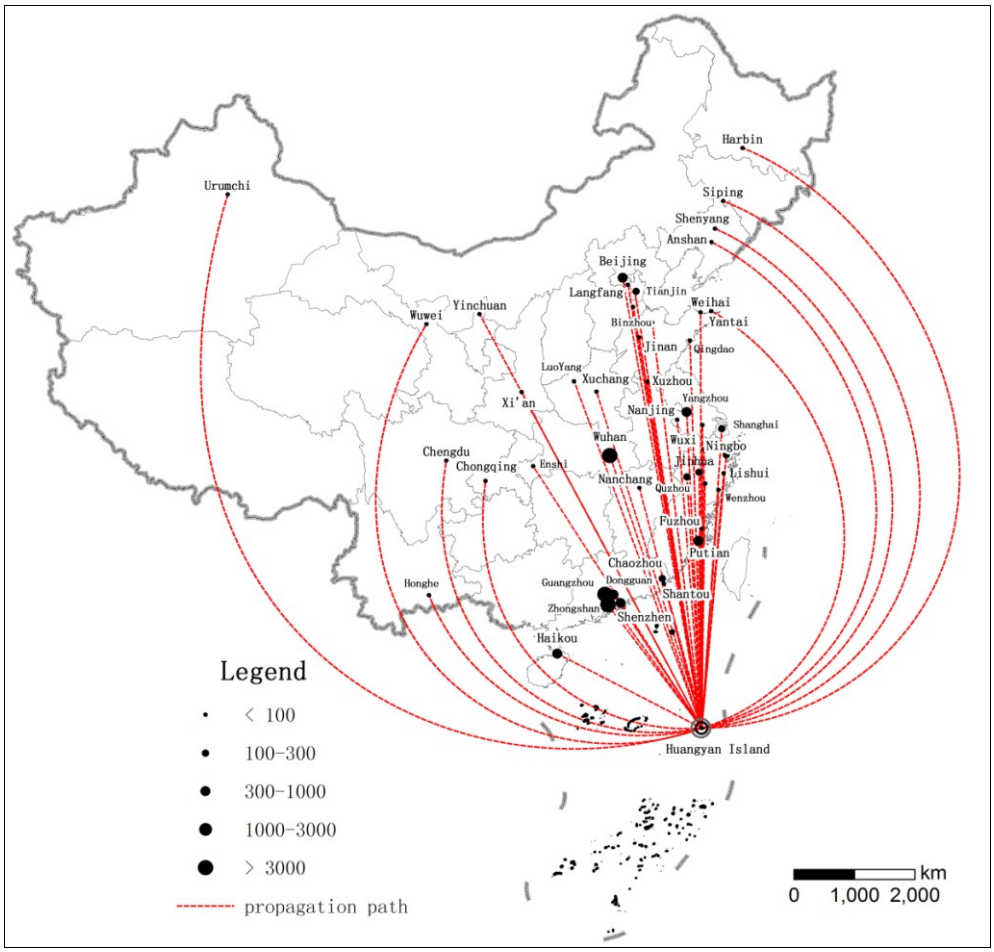


4. Dissemination Path Analysis of the HII

The HII was first reported by the website of Phoenix Net [39].Three hours later, other authoritative websites and local portal websites reproduced the news reports more than 100 times. Within the two months, there were a total of 36,323 news reports on the HII. However, 33,734 news reports were reprinted from other websites, which meant that 92.87% of the news coverage that affected the majority of internet users was reproduced from other sources. Among the remaining 2589 new reports, 1173 reports were not reproduced and 1416 news reports were reproduced for more than twice. The rate of reprint for new reports was 54.69%. In 33,734 news coverage reports, 33,428 were reprinted more than 10 times, accounting for 99.09%. Based on the distribution of the domain names and IP addresses of the web servers, the issue spread to 46 cities in 26 provinces following the outbreak of the HII. News reports on the HII first occurred in the coastal areas, and then there were reported and reproduced to a higher degree in the southern coastal areas and the eastern coastal areas than in other areas. At the national level, the highest density of reports and reprints occurred in the region of the Pearl River Delta, followed by the south region of the Yangtze River Delta. At the provincial level, coastal provinces, such as Guangdong, Zhejiang, and Shandong had much greater spread densities. Guangdong, Hubei, Jiangsu, and Beijing,

because of its location, technology, economy, and politics, had higher reprint rates for the precipitating events and political events. At the urban scale, the city of Guangzhou had the greatest frequency of reproduced news reports, up to 19,939, which was much more than the following cites, such as Zhongshan (4502 times) and Wuhan (3531 times). As shown in Figure 6, Guangzhou, Dongguan, Shenzhen, and Zhongshan, geographically close to Huangyan Island, had much more news coverage reports (more than 3000 times). The political, economic, and science and technology centers of China (e.g., Beijing, Shanghai, Wuhan, Yangzhou, Putian, *etc.*) also exhibited high rates of news reports (more than 1000 times).

Figure 6. Dissemination flow graph in the HII.



The dissemination characteristics of the HII also reflect the regional differences in the effect of geo-events due to the different consciousness in maritime rights and resource sovereignty. There are different response characteristics between land-locked and coastal provinces. In nine land-locked provinces, only half (approximately 5) responded to the maritime boundary dispute promptly, and not every province had the same response. For example, in the northeast and northern provinces, Heilongjiang, Jilin, and Liaoning, responded to the HII promptly. However, the response of Inner Mongolia was delayed. In the Northwest and northern part of China, Xinjiang responded, whereas Inner Mongolia did not. In the southwest, Yunnan responded, whereas Guangxi and Tibet did not. Among the 12 coastal provinces, only Guangxi province did not respond in the first instance (due to the restrictions of the gateway search and simplified Chinese segmentation, there was no crawling information for

Taiwan in this paper). The results show that there are regional differences in the response of the land-locked regions and coastal regions to the geo-events at the national, provincial and city levels. The levels of social and economic development and the marine consciousness of net citizens have effects on the dissemination of public opinion and the speed of response to maritime disputes significantly.

5. Discussions

In this paper, tag cloud of related topics, frequency map and individual mention ratios are used to provide a vivid expression of word frequency. The sentiment analysis and dissemination path analysis are conducted to estimate public opinion and the characteristics of communication socialization related to the HII. The results indicate that the differences of word frequency in news reports reveal the subject vocabularies (e.g., China, Philippines) and the theme vocabularies (e.g., reefs, sea area, geopolitical environment, and international relations) of related topics. The results also illustrate the related issues behind the HII (e.g., the South China Sea issue, the impact of international friction on the regional economy, regional security, regional tourism and regional international trade, government officials, and opinion leaders in the network community). The different attitudes of net citizens in various network communities are represented in the form of an upside-down "V" for the HII, and there are more neutral attitudes toward this political event than unilateral opinions. As this is a dispute regarding the sovereignty of marine land and resources, there are different response characteristics between the land-locked regions and coastal regions. The geopolitical event of Huangyan Island and its subsequent public opinion transmission on the network are shaped by geographical location and are closely related to the geopolitics, geo-economics, and geopolitics of science and technology, and the national defense consciousness of different regions.

The methods we used in this paper are useful tools to "map" the content of news reports on political event and track its diffusion of internet reactions. These methods are also able to "map" what is being said and what is being thought in different network community, which may be a valuable tool for understanding the public opinion and the way it is shaped in the social networking. The proposed crawler can achieve good performance in both crawling efficiency and results' coverage [40]. The text analysis of network information that collected by web crawler can be a tool for understanding the formation and diffusion of web-based public opinion to geopolitical events. The combination of web crawler technology and the text analysis can be developed to a useful policy tool for understanding the text information of the geopolitical events on the internet. The limitation of this tool should also be mentioned. One of the limitations is that this tool is not able to provide a clear indication of the motivations of the various actors in the political drama. It merely captures the public reactions to the event. In addition, due to the limitations of the current information processing and analysis techniques images, videos, folders, executive files, package files, and other non-text content of transmission information and online data are not included in this study. Furthermore, as the differences of morpheme characteristic and grammatical rules between Chinese and the other language, the method of word segmentation and text manipulation are very different from the other. Though we only grab the Chinese text on the internet, the completeness of internet coverage and the methodology performance of collecting information are well considered in this paper. The text analysis of multiple languages based on the data collection of web crawler will be further conducted in the future studies.

6. Conclusions

With the rapid development of network technologies and the blow-out growth of gate-kept news media sites, the influence of the net media becomes significant, especially in economic, social, and political fields. When an unexpected geopolitical event occurs and spreads on the internet, the news reports and comments of net media are the important pathways to reflect the concerns and opinions of the net media and net citizens. However, as it is difficult to gather and analyze all the ancillary data and theme events information on the internet, the demonstration of the spatial diffusion of internet attention always be blocked. This study found an efficient event analysis method which is composed of the web crawler for data collection and text analysis for data analysis of geopolitical events. After grabbing new reports and relevant information of the HII on the internet from 11 April 2012 to 4 June 2012, we explored data mining from the net media with the help of web crawler technology under the framework of Heritrix. The conclusion was drawn that web crawler technology, based on a web 2.0 framework structure, could grab and collect related topics of public opinion and information rapidly and effectively. After the event analysis and the text analysis of the HII including word frequency, emotional tendency and network propagation, we have explored data analysis based on computer language processing technology with the aid of text manipulation and user dictionary construction. It concluded that text analysis of network information could provide a vivid visual expression of the concerns of the net media, the emotional trends of the net citizens, the structure and content of the propagation model for political events. Therefore, the combination of web crawler technology and the text analysis method could be a good way to obtain the visual representation and provide an event analysis of geo-events.

Acknowledgements

This work was supported by National Key Technology R&D Program (2012BAK12B03), the Natural Science Foundation of China (41171097). The authors of this paper are grateful to A Xing Zhu, Robert Ostergren, Jim Burt and Jing Liu, Adam Mandelman in Department of Geography University of Wisconsin and Dong Chen, Yang Cheng, and Shufang Wang in school of geography Beijing Normal University for their constructive suggestions for this research.

Author Contributions

Hao Hu played an important role in the conception of the study, performing the data analyses, drafting and revising the manuscript. Yuejing Ge contributed to the conception of the study and played an important role in interpreting of the results and approved the final version. Dongyang Hou contributed a lot to framework of the methodology and the text pre-processing of acquired data, especially in the method statement of web crawler technology. He also approved the final version.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Wu, J.Y.; Jia, J.H.; Feng, X.F. Web Crawler Based on Agricultural Sector. *Comput. Dev. Appl.* **2012**, *25*, 30–32.
2. Goodchild, M.F.; Hill, L.L. Introduction to digital gazetteer research. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1039–1044.
3. Yang, D.Z.; Zhao, G.; Wang, T. Application of WebCrawler in information search and data mining. *Comp. Eng. Des.* **2009**, *30*, 5658–5662.
4. Zhang, C.J.; Zhang, X.Y.; Zhu, S.N.; Xu, X.T. Method of Toponym Database Updating Based on Web Crawler. *J. Geo-Inf. Sci.* **2011**, *13*, 492–499.
5. Liu, F.; Han, H.; Zhou, L.; Qi, J.Y.; Xu, B.L. Application of IT technology of global infection disease monitoring. *Chinese J. Front. Health Quar.* **2012**, *35*, 273–276.
6. Xu, X.; Zhang, C.Z.; Li, W.J. Research on the Public Opinion on the Internet in China: Review & Forecast. *Inf. Stud. Theory Appl.* **2009**, *32*, 115–120.
7. Zeng, R.X. A Review on Research and Development of China's Network Opinion. *Res. Libr. Sci.* **2009**, *29*, 2–6.
8. Tang, L.F.; Zhao, X.L. The response of Network public opinion and ideological work of university. *Res. High. Educ.* **2007**, *25*, 64–65.
9. Mei, Z.L. Research on the Technology of Internet Public Sentiment Analysis Based on Web Data Mining. *J. Chin. People's Public Secur. Univ. (Sci. Technol.)* **2007**, *13*, 85–88.
10. Dai, Y.; Yao, F. Research into Information Mining and Evaluation Index System Based on the Security of Public Opinion on the Internet. *Inf. Stud. Theory Appl.* **2008**, *31*, 873–876.
11. Huang, X.B.; Zhao, C. Application of Text Mining Technology in Analysis of Net-Mediated Public Sentiment. *Inform. Sci.* **2009**, *27*, 94–99.
12. Xiao, L.; Zhao, L.M. The Tourism Destination Image of Disseminated on Internet—Based on A Content Analysis of Travel–related Websites across Taiwan Straits. *Tour. Trib.* **2009**, *24*, 75–81.
13. Ma, Q.F. Regional tourism brand construction analysis based on symbolic communication. Ph.D. Thesis, Shaanxi Normal University, Xi'an, China, 2010.
14. Li, W.H.; Zhang, W.D.; Chen, Z.B. Study on Urban informatization development in china based on bibliometrics and social network analysis. *Libr. Inf. Serv. Online* **2011**, *2*, 12–22.
15. Yuan, H. Web Usability Evaluation of the University Portal Based on Web Content Analysis—Case Study of Jiangsu Province. *New Technol. Libr. Inf. Serv.* **2010**, *30*, 70–75.
16. Brunn, S.D.; Jones, J.A. Geopolitical Information and Communication in Shrinking and Expanding Worlds: 1900–2100. Available online: http://www.abebooks.com/geopolitical-information-communication-shrinking-expanding-worlds/5660014109/bd (accessed on 10 August 2012).
17. Brunn, S.D. The Internet as 'the new world' of and for geography: Speed, structures, volumes, humility and civility. *GeoJournal* **1998**, *45*, 5–15.
18. Alameh, N. Chaining geographic information web services. *IEEE Internet Comput.* **2003**, *7*, 22–29.
19. Rajasekaran, P.; Miller, J.; Verma, K.; Sheth, A. Enhancing web services description and discovery to facilitate composition. *Semant. Web Serv. Web Process Compos.* **2005**, *3387*, 55–68.
20. Brunn, S.D.; Dodge, M. Mapping the "worlds" of the World Wide Web: Restructuring global commerce through hyperlinks. *Am. Behav. Sci.* **2001**, *44*, 1717–1739

21. Hsinchun, C.; Larson, C.A.; Elhourani, T.; Zimbra, D.; Ware, D. The geopolitical web: Assessing societal risk in an uncertain world. Available online: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5984051&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5984051 (accessed on 1 March 2012).
22. Chau, M.; Zeng, D.; Chen, H.C.; Huang, M.; Hendriawan, D. Design and evaluation of a multi-agent collaborative web mining system. *Decis. Support Syst.* **2003**, *35*, 167–183.
23. Kausar, M.A.; Dhaka, V.S.; Singh, S.K. Web Crawler: A Review. *Int. J. Compt. Appl.* **2013**, *63*, 31–36.
24. Heritrix: Internet Archive Web Crawler. Available online: http://sourceforge.net/projects/archive-crawler/ (accessed on 10 July 2012).
25. Nutch. Available online: http://nutch.apache.org/ (accessed on 10 July 2012).
26. Labin. Available online: http://larbin.sourceforge.net/ (accessed on 10 July 2012).
27. Liu, J.H.; Lu, Y.L. Survey on topic-focused Web crawler. *Appl. Res. Comput.* **2007**, *24*, 26–29.
28. Batsakis, S.; Petrakis, E.G.; Milios, E. Improving the performance of focused web crawlers. *Data Knowl. Eng.* **2009**, *68*, 1001–1013.
29. Bedi, P.; Thukral, A.; Banati, H. Focused crawling of tagged web resources using ontology. *Comput. Electr. Eng.* **2013**, *39*, 613–628.
30. Kent, C.K.; Salim, N. Features based text similarity detection. *J. Comp.* **2010**, *2*, 53–57.
31. Shen, Y.; Fu, H.J.; Liu, P.P.; Wu, J. An Empirical Study on Virtual Learning Team. *Doc. Inf. Knowl.* **2009**, *25*, 103–107.
32. Wang, L. The typical social software tools and analysis method of network analysis. *China Educ. Technol.* **2009**, *29*, 95–100.
33. Lubke, G.H.; Muthen, B.O. Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Struct. Equ. Model.* **2004**, *11*, 514–534.
34. Flamer, S. Assessment of the multitrait-multimethod matrix validity of likert scales via confirmatory factor-analysis. *Multivar. Behav. Res.* **1983**, *18*, 275–308.
35. Liu, Q.Y.; Liu, B.H. *Strategic Management: Analysis, Formulation and Implementation*; Dongbei University of Finance and Economics Press: Shenyang, China, 2001; pp. 45–55.
36. Guo, Q.K.; Zhou, J. Effectiveness of Different IRT models in Likert-type Scale analysis. *Psychol. Explor.* **2004**, *24*, 67–70.
37. Webmaster tools. Available online: http://tool.chinaz.com/Ip (accessed on 10 August 2012).
38. Hu, H.; Ge, Y.J.; Hu, Z.D. The Geopolitical Environment of Greater Neighborhood of the South China Sea. *World Reg. Stud.* **2012**, *21*, 36–44.
39. Phoenix Net. Available online: www.Ifeng.com (accessed on 11 April 2012).
40. Li, W.; Yang, C.W.; Yang, C. An active crawler for discovering geospatial web services and their distribution pattern—A case study of OGC Web Map Service. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1127–1147.