The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-4/W3, 2013
ISPRS/IGU/ICA Joint Workshop on Borderlands Modelling and Understanding for Global Sustainability 2013,
5 – 6 December 2013, Beijing, China

USING WEB CRAWLER TECHNOLOGY FOR TEXT ANALYSIS OFGEO-EVENTS: A CASE STUDY OF THE HUANGYAN ISLAND INCIDENT

HU. Hao, GE. Yuejing

School of Geography Beijing Normal University 100875
No. 19 Xin Jie Kou Wai Street, Haidian District, Beijing, P.R. China -
bsdhao@163.com geyj@bnu.edu.cn

Commission

KEY WORDS: web crawler technology; text information; sentiment analysis; Huangyan Island incident

ABSTRACT:

With the social networking and network socialisation have brought more text information and social relationships into our daily lives, the question of whether big data can be fully used to study the phenomenon and discipline of natural sciences has prompted many specialists and scholars to innovate their research. Though politics were integrally involved in the hyperlinked word issues since 1990s, automatic assembly of different geospatial web and distributed geospatial information systems utilizing service chaining have explored and built recently, the information collection and data visualisation of geo-events have always faced the bottleneck of traditional manual analysis 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 is studied here by combining web crawler technology and the text analysis method. The results indicate that tag cloud, frequency map, attitudes pie, individual mention ratios and dissemination flow graph based on the data collection and processing not only highlight the subject and theme vocabularies of related topics but also certain issues and problems behind it. Being able to express the time-space relationship of text information and to disseminate the information regarding geo-events, the text analysis of network information based on focused web crawler technology can be a tool for understanding the formation and diffusion of web-based public opinions in political events.

1. INTRODUCTION

With the rapid development of network and communication technologies, websites, forums, blogs and twitter have become essential parts of people's lives. These network communication tools and the data information behind them have played increasingly important roles in the processes of social networking and network socialisation. As an important method of collecting information and data mining, web crawlers have been widely used in agricultural exploration (Wu et al. 2012), information searching and data mining (Goodchild et al. 2008; Yang et al. 2009), toponym database updating (Zhang et al. 2011), infection disease monitoring(Liu et al. 2012), network monitoring and management of cultural content (Yang 2009). Especially with regard to computers, there are many more research topics on the development of technology, and web crawler can collect information and data efficiently. At the same time, the data processing and data research aspects of text analysis have also developed quite rapidly in the areas of media spreading(Xu et al. 2009; Zeng 2009), university management (Tang et al. 2007), information security (Mei 2007; Dai et al. 2008; Huang et al. 2009), tourism brand management (Xiao et al. 2009; Ma 2010), city information technology(Li et al. 2011), web portal evaluation(Yuan 2010) and other public domains. Whether this kind of informational data collection and text data processing can be used to the events analysis of geopolitical research and how to use the modern computer technology to service the international politics is the aim of our study. After an unexpected political event occur and spread on the internet, the news reports and comments of net media always can reflect the concerns and opinions of the net media and citizens objectively. So the information collection and text analysis following political events can be very important for the stability of the network society and the security of national information. Though politics were integrally involved in the hyperlinked word issues since 1990s (Brunn et al., 1994; Brunn, 1998) , and researchers have explored automatic assembly of different geospatial web services to build distributed geospatial information systems utilizing service chaining (Alameh,2003; Rajasekaran et al.,2005), but 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 (Brunn et al.,2001). The abundance of information on the internet presents a challenge with regard to 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 a national level (Hsinchun et al. 2011). 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 point of origin to the entire internet. With the passing of time and the development of an event, 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 life. Public opinions regarding events rapidly and continuously change, transcending time limitations and spatial constraints. As a result, the event analysis of geo-events is very different than that of other events. Information collection 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 that relate to the event. The event analysis of geopolitical event must be traced back to the source of the event (the location and time of the first network news report) and the related on-going reports in different network communities. It requires knowledge of the spatial location and spatial relationship of a geopolitical event and its reporting network. It requires making clear the associated event (e.g., the related content and related topics), as well as locating and illustrating the temporal and spatial relationships between the spreading network and the developing event. A focused web crawler can automatically burrow deep 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 provide a visual representation of network media concerns, emotional trends and semantic networks based on public consciousness, which can be helpful for scientific decision making. The introduction of text analysis based on web crawler techniques in geopolitical research will be much more efficient, objective and comprehensive than the traditional analysis methods of geopolitical event, which are based on manual collection of related information from newspapers and statistical reference 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 using network text analysis. Our combining of web crawler technology and the text analysis method on Huangyan Island incident in 2012 will provide a preliminary exploration of the event analysis of geopolitical events based on computer technology.

2. KEY TECHNOLOGY AND PROCESS DESCRIPTION

2.1 Web crawler technology

A web crawler, also known as a network robot or spider, is a kind of script or programme that can automatically traverse the internet and grab text information based on the web link structure of World Wide Web and HTTP protocols (Chau et al. 2003). Web crawler search engines originated in 1990s and have become a hot topic in today's search engine and web mining fields (Liu et al. 2007). Web crawler tracking can be divided into three categories: (*i*) General web crawler, also known as the general crawler or the general purpose web crawler, such as AltaVista, Yahoo and Google; (*ii*) The focused web crawler, or focused crawler, also known as thematic network reptiles of topical crawlers, such as Lucene, Heritrix, Smart Crawler and web crawlers from the reptiles page mode; (*iii*) Other web crawlers, such as the incremental web crawler, deep web crawler and so on (Sun et al. 2010). More and more challenges have focused on the search engine index scale, the rapid updating of information and enhanced individual demands with the explosive growth of network information (Lee et al. 1999). Strategy optimisation for general web crawler grabbing (Yang 2009) (e.g., depth-first strategies, breadth-first strategies and best-first strategies), the application of web crawler technology and the user-oriented design of focused web crawlers (Li et al. 2003) are becoming the most popular issues in the research of search engines (Liu 2007). To reasonably analyse geo-events, related network information was automatically grabbed from mainstream media by a web crawler based on Heritrix's framework and an interaction-improved traversal strategy incorporating the depth-first and breadth-first principles. A detailed technical model of geo-events analysis based on a web crawler and text analysis is shown in Fig. 1.

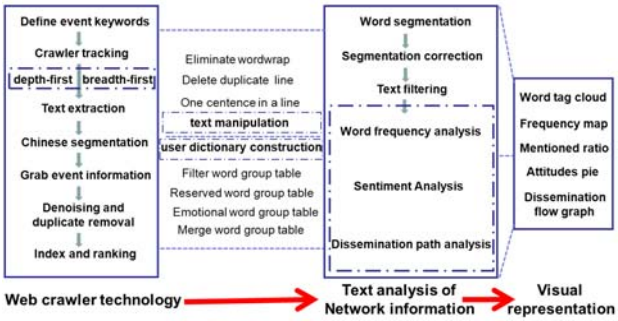

Figure 1. Text analysis based on Web Crawler Technology

2.2 The HII and its data processing based on a web crawler

The Huangyan Island incident (also called the HII for short) is a representative geopolitical event in 2012. On April 10, 2012, 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 the property and buildings of China, calling neighbouring countries against China and even intending to illegally rename Huangyan Island as Panatag Shoal. China reiterated that any of the Philippines' actions towards Huangyan Island were invalid and established the city of Sansha 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.

Through web crawler technology and the improved depth-first and breadth-first interactive traversal technique, internet information on new reports of Huangyan Island (the title, abstract, reported sources, reported time, transfer number, the propagation paths of information flow, the link of new reports and other relevant text information) are grabbed, denoised, indexed and ranked (Fig.1). From April 11, 2012, at 10:36:00 to June 4, 2012, at 23:54:00, a total of 273 websites referring to the HII and 2,855 news reports or hotspot comments on the internet were found. Because the above information is informational redundant, only 2,589 new reports and commentaries are employed for further text analysis. The websites that reported the above news more than 10 times are shown in Table 1. Data processing of crawled information were carried out as followed. First, wordwarp in the network text were eliminated, duplicate in the news reports were deleted, and after you make sure one sentence in a line, the foundation work of text manipulation were completed. Second, User dictionary was constructed based on the filter word group table, reserved word group table, emotional word table and merge word table to make the word segmentation more perfected. And after the segmentation correction and text filtering over and over again, the word segmentation and the crawled information will have more practicable semantic content and semantic structure for the text analysis in the next section.

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

Table 1. Websites that reported news on the HII more than 10 times

2.3 Text analysing method based on web crawler

Text analysis of geo-events using crawled information mainly includes word frequency analysis, sentiment analysis and dissemination path analysis, which are based on the results of mining and processing data from the internet. 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 a comparison of different attitudes with regard to the HII, whereas dissemination path analysis mainly 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 thing 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 to the centre position and the higher the position of the word will be. The most frequent word is thus 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 word frequency statistic calculation. Next, the processes of word segmentation and text filtering are adopted with the help of the user dictionaries, text manipulations of titles, abstracts and comments. Then word frequency and vocabulary analyses are repeatedly performed based on the ROST software. Lastly, to provide text visualisation (Wang 2009) of the HII, the tag cloud is generated using ROST software.

As a common psychometric scale in social science research, the Likert Scale is the most widely used approach to scaling responses in survey research. 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 differences in personal subjective feelings. The scale has become an important measurement tool of modern social science research. According to research needs, a graded assignment can be divided into models with three, five and seven different rating scales to express attitudes and measure the degree of a positive or negative attitude. Lubke asserted that the Likert Scale multistage rating model has a better imitative effect. The higher the level, the greater the measurement accuracy is (Lubke et al. 2004). The process of sentiment analysis for the HII mainly includes two steps. First, the various terms are classified into verbs, nouns, and adjectives with an expert evaluation method and a constructed emotional words 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 government). Second, the reports are divided into seven categories, that is -3, -2, -1, 0, 1, 2, and 3 (Flamer 1983; Liu et al. 2001; Guo et al. 2004), using the concept of the Likert Scale, and positive or negative statements are analysed based on the Likert score and the average emotional value of the sentence. If a word is a derogatory term, the Likert score of the word would be less than zero; if the word is an appreciative term, its Likert score would be greater than 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. More detailed steps on how this is performed are 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 high-frequency words are extracted from the results of the text manipulation process to update the emotional word dictionary; (*iv*) the processes of word segmentation and segmentation correction are started according to the reserved word group table and the merged word group table of the user dictionary; (*v*) the final segmentation results of the news report are obtained after the insignificant auxiliary words and interjections are filtered; (*vi*) the final segmentation result are matched with the words in the emotional word dictionary, and the Likert score and emotional value of each word and each report are obtained; (*vii*) 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.

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

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 number and the domain names of news reports source which have significance effect on the public opinions. It is well known that once a report is reproduced by a website, the information is likely to spread rapidly to other network communities. The higher the number of reproduced reports, the greater the bandwidth the disseminated information will occupy on the internet and the faster the speed of information dissemination will be. Three steps are necessary to analyse how new reports and public opinion spread throughout a network. First, 36,323 news reports on the HII is crawled from the internet within a specific period of time. Second, 33,734 news reports that were reprinted from another website are picked up, and all the sources (concluded time and place) of the remaining 2589 new reports are obtained after an information sort of the reported sources and reported time. Third, an information flow of reprints is drawn for information spread visualisation after confirming the domain names of the remaining 2589 reports sites and locating the IP addresses of web servers with the help of webmaster tools from http://tool.chinaz.com/Ip.

3. WORD FREQUENCY ANALYSIS OF HII

3.1 Related topics on the HII

The tag cloud of word frequency for the HII includes 210 high-frequency words, which had frequencies of more than 50. The cloud not only highlights the subject vocabularies and the theme vocabularies but also vividly demonstrates a few issues and problems behind the HII. Overall, for the tag cloud, the closer a word is to the centre position, the greater the frequency of that word will be. From the centre to the border, the font colours of the labels gradually change over a gradient from red to green, with the more red fonts representing greater word frequency and a higher grade (see Fig 2). Word frequency statistics and the tag cloud show that the frequencies of “China”, “Philippines”, “Huangyan Island” is the greatest and are between 3700 and 4100. This level is much higher than the next level between 1000 and 1200. Therefore, these words become the main body of the analysis of the HII because they are always 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 greater than 1000, reflecting the fact that the interests of the net media and net citizen 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. Some high-frequency words (e.g., “sea area”, “problem” and “Huangyan Island incident”, with frequencies between 500-800) 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 vocabulary, such as countries, sovereignty, foreign affairs, situation, territory, security, peace and surrounding, also appears frequently 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 paid close attention to Chinese diplomatic behaviours 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 suggest 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” (Hu et al. 2012), “neighbouring countries” and “international association countries”, also frequently appeared in the comments and show the complex international relations and their linkage to the development effect of hot issues in the context of globalisation. Therefore, geo-events are always not only associated with the geographical location of the event but also closely related to geo-political, geo-economic, geo-science and technology factors. There is a large difference between the text analysis of geo-events and that of other events.

Figure 2. Tag cloud of HII

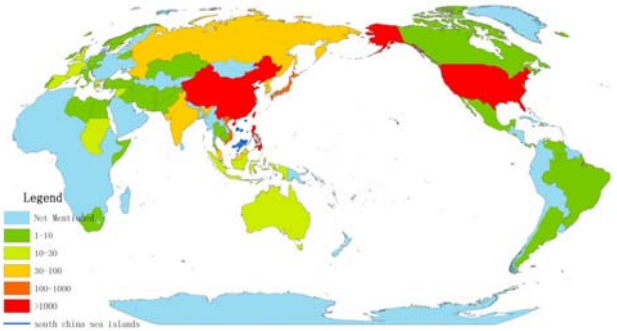

Figure 3. Frequency map of the HII

3.2 Frequency map of the HII

In addition to China and the Philippines, the names of other countries appeared as high-frequency words. A total of 53 of the 193 existing countries appear in comments on the HII. As shown in Fig. 3, China, the 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 the Philippines, the United States was directly involved in this event and this term appeared at a high frequency. This finding provides a certain degree of 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 and the main countries, Japan and Vietnam, which ranked in the second layer in the mentioned frequency structure, have reduced frequencies, in the 100-300 range,. 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 neighbouring 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 discussions on the South China Sea, thoughts on national sovereignty, territorial dignity and the security concerns of neighbouring 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 discussion subjects during the event due to their economic crises and political activities. Another 35 countries, namely, Brunei, Iran, Thailand, Maldives, Germany, Canada and Iraq are mentioned fewer than 10 times in the HII.

3.3 Individual mention ratio on the HII

Among new reports obtained from web grabbing and crawling for text information on the HII, there were996 personal names mentioned and 67 people involved in the confrontation from Apr.11.2012 to Jun.4.2012. Their individual effects on the HII are analysed using the data of the mentioned names and the rates of the mentioned individuals. The individuals and their rates of being mentioned in the HII are shown in Table 2.

The President of Philippines, Benigno Aquino III, was the most frequently mentioned, followed by the Minister of Foreign Affairs of the Philippines Del Rosario, the Chinese Defence Minister Liang Guanglie, Foreign Ministry spokesman Hong Lei, Foreign Ministry spokesman Liu Weimin, Deputy Foreign Minister Fu Ying, and U.S. Secretary of State Hillary Clinton (ranked in the top 10 and a mention ratio of more than 5%). The statement and behavioural responses of the President of the Philippines and the Departments of Foreign Affairs of China and the Philippines became hot topics in the network media and full filled public concerns during the build-up of the HII. In addition, U.S. Secretary of State Hillary Clinton, U.S. Secretary of Defence Panetta and U.S. President Obama had unexpectedly high frequencies (ranked in the top 20 with mention ratios greater than 1.5%), which may reveal a tilt of the foreign policy strategy of the United States to strengthen the effect of the Asia-pacific region. Russian President Putin also appeared with high frequency (ranked top 15), indicating that Russia also have its international influence in the region disputes though there is nothing to do with it. 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. Former Philippine Marine Corps captain Nicano Fildo attempted to land on the island and raise the Phillipines' flag but blocked by Philippine President Aquino III indicating that the Philippines was inconsistent in its attitude toward Huangyan Island and the attitude always change as the international behaviours development of great powers. Because the Philippines once attempted to use basketball diplomacy with China to divert Chinese public attention from a political and military confrontation, sports star Yao Ming had a high frequency during the Huangyan Island confrontation (ranked in top 20), which was a particularly interesting finding. Commenta tors on the HII, including Luo Yuan, Ruan Cishan, Jiang Xiaofeng, and Wu Shicun and other experts,as well as leaders' suggestions from the authoritative network community, affected the viewpoints and opinions of the net media and net citizen. Yang Jiechi had little effect on the HII, with a lower mention ratio and minor placing, even though he was the Minister of Foreign Affairs of China. All these factors provide evidence of the complexity of the situation of the South China Sea and the critical issues of China's peripheral security.

individuals	ratio	individuals	ratio	individuals	ratio	individuals	ratio
Aquino III	9.74%	Nicano Fildo	1.51%	Zhang Jie	0.60%	Xun Jin	0.20%
Albert Del rosario	8.63%	Chen Bing	1.41%	Li Guoqiang	0.60%	Zhao Kejin	0.20%
Liang Guanglie	7.43%	Yao Ming	1.00%	Yoshihiko Noda	0.60%	Jiang An	0.20%
Hong Lei	7.23%	Lv Ningsi	1.00%	Ma Xiaotian	0.50%	Walter	0.20%
Liu Weimin	6.93%	Zhuang Guotu	1.00%	Song Zhongping	0.50%	Hu Xijin	0.10%
Fu Ying	5.72%	Zheng Hao	1.00%	Yu Zhirong	0.40%	Leng Xinyu	0.10%
Hillary Clinton	5.62%	Wang Yizhou	1.00%	Xue Baosheng	0.40%	Ma Xiaolin	0.10%
Leon Panetta	4.22%	Tong Xiaoling	0.90%	Sun Xiaoying	0.40%	Zhang Liwei	0.10%
Luo Yuan	3.31%	Song Xiaojun	0.90%	Cameron	0.40%	Zhang Wuchang	0.10%
Ruan Cishan	3.31%	He Liangliang	0.90%	Ren Haiquan	0.30%	Zhang Yun	0.10%
Jiang Xiaofeng	2.51%	Binet	0.90%	Dai Bingguo	0.30%	Zhang Guoqing	0.10%
Voltaire Gazmin	2.31%	Peng Guangqian	0.80%	Huan Dongxing	0.30%	He Jun	0.10%
Wu Shicun	1.91%	Shintaro Ishihara	0.80%	Du Wenlong	0.30%	Guo Yiming	0.10%
Zhang Zhaozhong	1.61%	Kudashev	0.80%	Naguib	0.30%	Liang Yongchun	0.10%
Putin	1.61%	Brady	0.80%	Viktor.n arches ii	0.30%	Yang Jiechi	0.10%
Obama	1.51%	Ma Keqing	0.70%	Medvedev	0.30%	Ye Yijian	0.10%
Deng Zhonghua	1.51%	raul	0.70%	Su Hao	0.20%	Total	100.00%

Table 2. Individual mention ratio in the HII

4. SENTIMENT ANALYSIS OF HII

In this study, sentiment analysis of the HII is employed by clustering analysis of the network text and network opinions based on 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 different attitudes in different network communities exhibited an upside-down "V" shape for the HII (See Fig. 4). With respect to the distribution of extreme views, the more extreme view for this event, the fewer the number of people who hold it. There were many more people have neutral opinions, and from middle attitudes to general attitudes and from moderate attitudes to highly opinionated attitudes, the ratio decreased quickly. The generally positive statements, with an emotional value between 5 and 15, account for 20.82%, 2.4 times more than the moderate positive statements (emotional value [15, 25]) and 4.3 times more than the highly positive statements (emotional value [25,

85]). There were more negative attitudes in network media news reports than positive attitudes for the HII, approximately one third of the people commenting on the event had a negative attitude. When the neutral opinions of new reports were ignored, which included general positive emotions, general negative emotions and neutral emotions, the proportion of negative positions (i.e., highly negative emotions and moderate negative emotions) reached as high as 14.83%, and 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 higher than that of highly positive statements by 0.61%. The percentage of moderate negative statements was greater than that of moderate positive statements by 0.5%, suggesting that the ideological content orientation and commentary guidance of news reports and net media could provide important context to the public opinion in the geo-events.

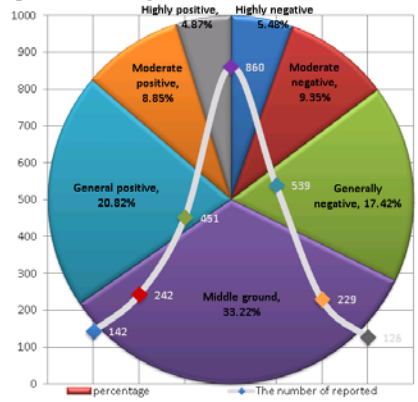

Figure 4. Comparison of attitudes pie toward the HII

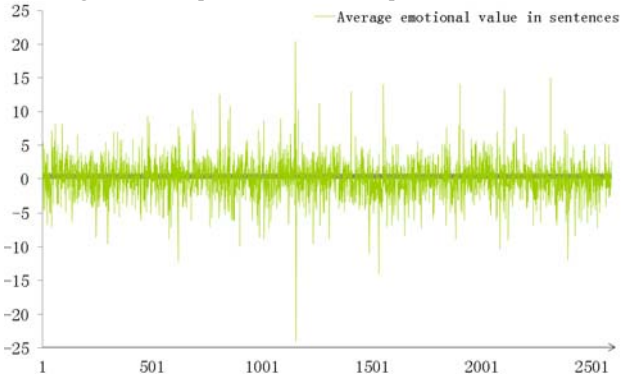

Figure 5. Average emotional values of news reports for the HII

To measure emotional affect and analyse the degree of emotional difference due to the words segmentation quantity found in the Likert score, the emotional word dictionary and the matched words quantity in the new reports, the emotional analysis effect of the HII is tested and 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 will be. As shown in Figure.5, the emotional value of the 2,589 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.

5. DISSEMINATION PATH ANALYSIS OF THE HII

The HII was first reported by the website of "www.Ifeng.com". Three hours later, other authoritative websites and local portal websites had reproduced the news reports more than 100 times. Within that specific time, there were a total of 36,323 news reports on the HII; however, 33,734 news reports were reprints from other websites, which means that 92.87% of the news coverage that affected the majority of internet users was reproduced from other sources. Among the remaining 2,589 new reports, 1,173 reports were not reproduced and 1,416 new reports were reproduced more than twice. That is, 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 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. On the national scale, 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. On the provincial level, coastal provinces, such as Guangdong, Zhejiang and Shandong had much greater spread densities. Guangdong, Hubei, Jiangsu and Beijing, which are famous for their locations, technology, economy, and politics, had higher reprint rates for the precipitating events and political events. On the urban scale, the city of Guangzhou had the greatest frequency of reproduced news reports, up to 19,939, which was far greater than those of following city Zhongshan (4502) and Wuhan (3531). As shown in Figure 6, Guangzhou, Dongguan, Shenzhen and Zhongshan, which are geographically close to Huangyan Island, have much more news coverage reports (more than 3000). The political, economic, and science and technology centres of China (e.g., Beijing, Shanghai, Wuhan, Yangzhou, Putian, etc.) also exhibited high rates of news reports (more than 1000).

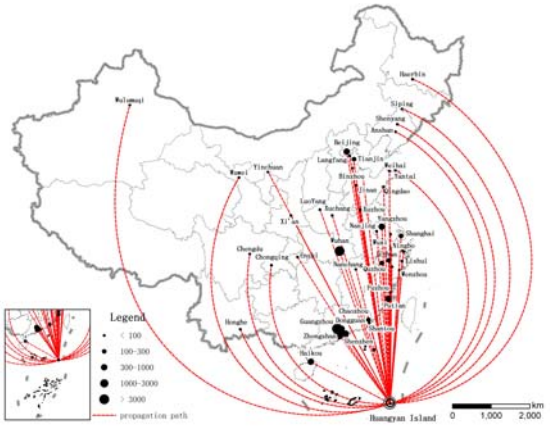

Figure 6. Dissemination flow graph in the HII

The dissemination characteristics of the Huangyan Island also incident reflect the regional difference on the effect of geo-events as the different consciousness over 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 regions, Heilongjiang province, Jilin province, and Liaoning province responded to the HII promptly; however, the response of Inner Mongolia might have been delayed. In the northwest and northern regions, Xinjiang

responded, whereas Inner Mongolia did not. In southwest China, 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 writing). This fact demonstrates not only that the level of social and economic development can influence the dissemination of public opinion but also that the marine consciousness of net citizens can contribute to the speed of response to maritime disputes significantly.

6. 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. Sentiment analysis and dissemination path analysis are also used to estimate public opinion and the characteristics of communication socialisation related to the HII. The results indicate that the differences of word frequency in news reports not only 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 but also reveal related issues and problems 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 different 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 was a dispute regarding the sovereignty of marine land and resources, there were different response characteristics between the land-locked provinces and coastal provinces. The geopolitical event of Huangyan Island and its subsequent public opinion transmission on the network were not only shaped by geographical location but also closely related to the geopolitics, geo-economics, and geopolitics of science and technology, and the national defence consciousness of different regions.

We have the technology to "map" the content and track the diffusion of internet reactions to and the reporting of a political event. We also have the technology to "map" what is being said and the relative frequency and associations of ideas, which may be a valuable tool for understanding public opinion and the way it is shaped through the processes of web networking. And the proposed crawler can achieves good performance in both crawling efficiency and results' coverage (Li et al., 2010). A text analysis of network information based on web crawler technology can be a tool for understanding the formation and diffusion of web-based public opinion to geopolitical events. The combining of web crawler technology and the text analysis may also quite possibly be of value as a policy tool. Because of we are interest in the text information of the geopolitical events on the internet, we inquired into the number of semantic topics and the word frequency among all the related websites and geographic entries in china, this capability does not necessarily provide a clear indication of the motivations of the various actors in the political drama by itself – it merely captures the public reaction to the event. In addition, due to the limitations of the current information processing and analysis techniques (i.e., search engines cannot crawl and process pictures, video, flash, executive scripts, executive programme and other non-text content files), images, videos, folders, executive files, package files and other online transmission information and data are not included in our analysis. And 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 grabbed the Chinese text on the internet, the completeness of internet coverage and the methodology performance of collecting information are well considered in this paper. Text analysis of multiple languages based on the web crawler will be studied in the future to erase the limitation of technology improvements.

7. CONCLUSIONS

With the rapid development of social media supported by network technology and network information transmission, the influence of network socialisation and social networking has gradually increased, 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 can objectively reflect the concerns and opinions of the net media and net citizens. But as it is difficult to extensive gathering and analysis of ancillary data and theme events information, the reveal and explain the spatial diffusion of internet attention always be blocked and insipid. In this paper, we used the web crawler technology and text analysis of geopolitical event on the internet to service the informational data collection and text data processing of international politics, and it presented relatively good results of the analysis and conclusion. After grabbing new reports and relevant information on the HII on the internet from April 11, 2012 to June 4, 2012, we have explored data mining based on web crawler technology under the framework of Heritrix. It concluded that web crawler technology based on a web 2.0 framework structures can grab and collect related topics of public opinion and information rapidly and effectively. After the event analysis and the text analysis of the HII that included word frequency, emotional tendency and network propagation, we have explored data analysis based on computer language processing technology with the help of text manipulation and user dictionary construction. It concluded that text analysis of network information can provide a vivid visual expression of the concerns of the net media, the emotional trends of the net citizens and the structure and content of the propagation model of political events. So the combination of web crawler technology and the text analysis method can be a good way to obtain the visual representation and provide an event analysis of geo-events.

REFERENCES

Alameh N., 2003. Chaining geographic information Web services. *Ieee Internet Comput*, 7(5): 22-29.

Brunn S D., Jones J A., 1994. Geopolitical information and communication in shrinking and expanding worlds: 1900-2100. *Reordering the World: Geopolitical Perspectives on the 21st Century*: 301-321.

Brunn S D., 1998. The Internet as' the new world' of and for geography: Speed, structures, volumes, humility and civility. *GeoJournal*, 45(1): 5-15.

Brunn S D., Dodge M, 2001. Mapping the "worlds" of the world wide web (Re) structuring global commerce through hyperlinks. *American Behavioral Scientist*, 44(10): 1717-1739

Chau, M., Zeng, D., Chen, H. et al., 2003. Design and evaluation of a multi-agent collaborative web mining system. *Decis Support Syst*, 35(1): 167-183.

Dai, Y., F., 2008. Research into Information Mining and Evaluation Index System Based on the Security of Public Opinion on the Internet. *Information Studies:Theory & Application*, (6): 873-876.

Flamer, S., 1983. Assessment of the multitrait-multimethod matrix validity of likert scales via confirmatory factor-analysis. *Multivar Behav Res*, 18(3): 275-308.

Goodchild, M.F., Hill, L.L., 2008. Introduction to digital gazetteer research.*International Journal of Geographical Information Science*, 22(10): 1039-1044.

Guo, Q.K., Zhou, J., 2004. Effectiveness of Different IRT models in Likert-type Scale analysis. *Psychology Exploration*, (3): 67-70.

Hsinchun, C., Elhourani, T. et al., 2011. *The geopolitical web: assessing societal risk in an uncertain world*: 60-64.

Hu, H., Ge, Y.J., Hu, Z.D., 2012. The Geopolitical Environment of Greater Neighborhood of the South China Sea. *World Regional Studies* (3): 36-44.

Huang, X.B., Zhao, C., 2009. Application of Text Mining Technology in Analysis of Net-Mediated Public Sentiment. *Information Science*, (1): 94-99.

Lee, G., Steve, L., 1999. Accessibility and distribution of information on the web. *Nature*, 400(8): 107-109.

Li, S.T., Yu, Z.H., Cheng, X.Q. et al., 2003. A Survey on Web Crawling. *Computer Science*, (2): 151-157.

Li, W.H., Zhang, W.D., Chen, Z.B., 2011. Study on Urban informatization development in china based on bibliometrics and social network analysis. *Library and Information Service Online*, (3): 12-22.

Li W, Yang C.W, Yang C, 2010. An active crawler for discovering geospatial web services and their distribution pattern--A case study of OGC Web Map Service. *International Journal of Geographical Information Science*, 24(8): 1127-1147.

Liu, F., Han, H., Zhou, L. et al., 2012. Application of IT technology of global infection disease monitoring. *Chinese Journal of Frontier Health and Quarantine*, 35(4): 273-276.

Liu, J.H., Lu, Y.L., 2007. Survey on topic-focused Web crawler. *Application Research of Computers*, (10): 26-29.

Liu, Q.Y., Liu, B.H., 2001. *Strategic management: analysis, formulation and implementation*: Dongbei university of finance and economics press, 45-55.

Lubke, G.H., Omuthen, B., 2004. Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, 11(4): 514-534.

Ma, Q.F., 2010. *Regional tourism brand construction analysis based on symbolic communication*. : ..

Rajasekaran P, Miller J, Verma K et al., 2005. Enhancing web services description and discovery to facilitate composition. In: Semantic Web Services and Web Process Composition: Springer, 55-68.

Sun, L.W., He, G.H., Wu, L.F., 2010. Research on the Web Crawler. *Computer Knowledge and Technology*, (15): 4112-4115.

Tang, L.F., Zhao, X.L., 2007. The response of Network public opinion and ideological work of university. *Researches on Higher Education*, (4): 64-65.

Wang, L., 2009. The typical social software tools and analysis method of network analysis. *Educational Technology*, (4): 95-100.

Wu, J.Y., Jia, J.H., Feng, X.F., 2012. Web Crawler Based on Agricultural Sector. *Computer Development & Applications*, 25(8): 30-32.

Xiao, L., Zhao, L.M., 2009. The Tourism Destination Image of Disseminated on Internet——Based on A Content Analysis of Travel-related Websites across Taiwan Straits. *Tourism Tribune*, (3): 75-81.

Xu, X., Zhang, C.Z., Li, W.J., 2009. The review and outlook of network public opinion research in china. *Information Studies:Theory & Application*, (3): 115-120.

Yang, D.Z., Zhao, G., Wang, T., 2009. Application of WebCrawler in information search and data mining. *Computer Engineering and Design*, 30(24): 5658-5662.

Yuan, H., 2010. Web Usability Evaluation of the University Portal Based on Web Content Analysis——Case Study of . *New Technology of Library and Information Service*, (10): 70-75.

Zeng, R.X., 2009. A Review on Research and Development of China's Network Opinion. *Researches in Library Science*, (8): 2-6.

Zhang, C.J., Zhang, X.Y., Zhu, S.N. et al., 2011. Method of Toponym Database Updating Based on Web Crawler. *Journal of Geo-Information Science*, 13(4): 492-499.

ACKNOWLEDGEMENTS:

This work was supported by National Key Technology R&D Program (2012BAK12B03), the Natural Science Foundation of China (41171097). Prof. A Xing Zhu, Prof. Robert Ostergren, Prof. Jim Burt and assistant JING Liu, assistant Adam Mandelman in Department of Geography University of Wisconsin and Chen Dong in school of geography Beijing normal university had been contributed lots of ideas and help. We are grateful to all of them.