Study on Information Collection Efficiency of Network Public Opinion in Unexpected Social Event

Qingguo Wang *
School of Management
Wuhan University of Technology, WUT
Wuhan, China
wqg1997@163.com
*Corresponding author

Zhixia Liu
School of Management
Wuhan University of Technology, WUT
Wuhan, China
604380095@qq.com

Qing Yan
School of Management
Wuhan University of Technology, WUT
Wuhan, China
263620881@qq.com

***Abstract*—Under the background of increasing unexpected social emergency events, this paper studied the research status of network public opinion information collection, analyzed its characteristics, container present form in unexpected social emergency, and gathering way, existing problems and causes of network public opinion information collection in social emergency. Then this paper verified information collection efficiency of network public opinion in unexpected social events, from the dimension of network public opinion fixed-point harvest engine and crawler search engines line programming tasks allocation. Static information acquisition model and crawler search engines model of network public opinion in unexpected social emergency are built.**

Keywords-network public opinions collection; unexpected social events; crawler search engines; fixed-point harvest engine

I. INTRODUCTION

With the rapid development of Internet, the interactivity of network platform makes people express personal views and opinions anonymously and uncontrollably. However, the increasing irresponsible comments lead to the formation of herding effect. Many netizens also follow the comments blindly. This accumulation of negative opinions will cause the outbreak of network public opinion inevitably. With the increase of unexpected social emergency and network public opinion, the governments need to focus on the development of network public opinion all the time besides master the mass-following psychology of the public, control the spread of negative information and panic mood and anxiety. To deal effectively with network public opinion crisis, the governments' primary tasks are collecting real-time data of public opinion effectively, without missing and accurately. By using the theory of the life cycle theory and herding effect theory of social psychology, this paper verifies information collection efficiency of network public opinion in unexpected social event, from the dimension of network public opinion fixed-point harvest engine and crawler search engines line programming tasks allocation. Governments can gather network public opinion information scientifically; provide the core foundation for subsequent policy transmission and decision-making.

II. NETWORK PUBLIC OPINION INFORMATION COLLECTION EFFICIENCY MODEL IN UNEXPECTED SOCIAL EVENT

Network public opinions collection containers in social accident include news message board, social network sites, instant messaging software, blog, E-mail, questionnaire.

This paper studies the collection efficiency base on the technology of search engine SEO, data warehouse and data mining, and information screening and filtering.

Network consensus information collection efficiency study's technology implementation in unexpected social event contains static information collection, network consensus crawler search engines model, and XML semi-structured documents collection.

A. Static Information Collection Research Model

Based on analysis of acquisition channels, information space, Acquisition principle and the characteristics of consensus, this paper built static information acquisition model through combining the correlation factors such as sampling methods, source and gathering tools, which is easy for government to acquire public opinion information's accurate distribution and build reliable and scientific decision support center and carrying decision support system. The result is shown in Figure 1.

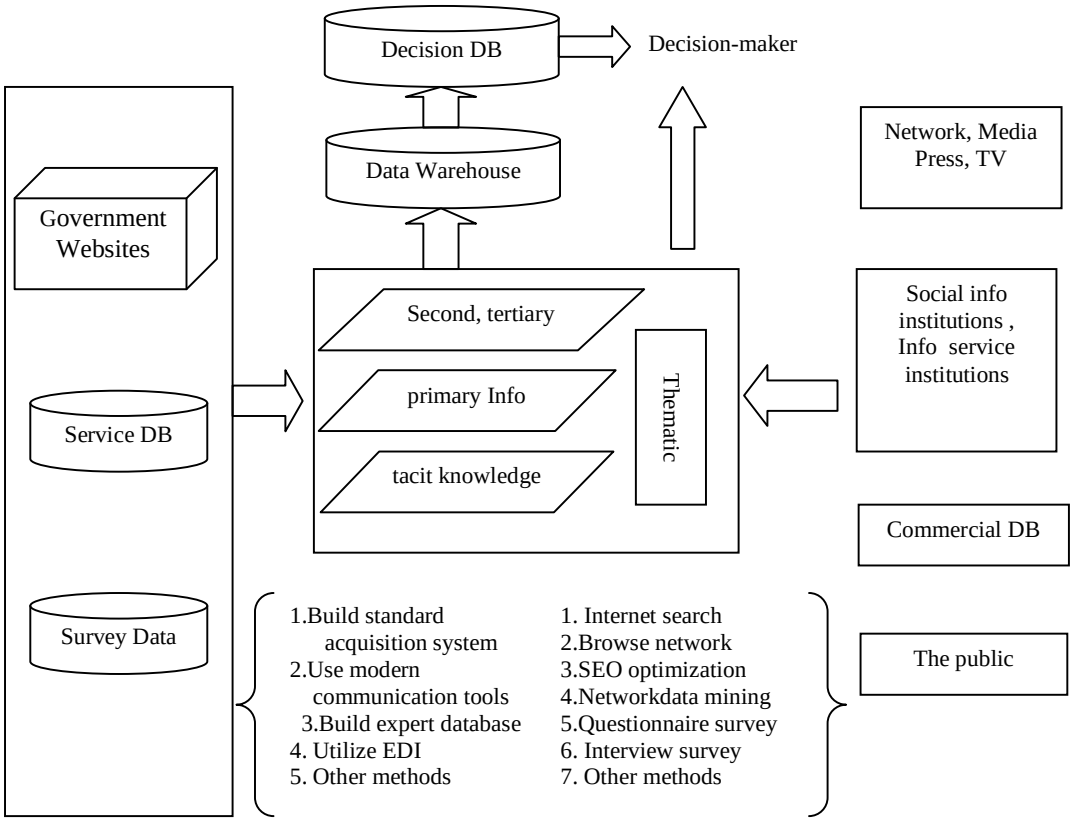

Figure 1. Network public opinion collecting mechanism.

Static information acquisition model consists of two parts: one part is to obtain information from external figureing to the society. The obtaining methods include direct labor, intelligent search engine's crawlers read in traditional media website, or purchase public opinion information and survey comments directly etc from the outside information collection agencies, which need to process re-integration analyze the public opinion information, and then store in their own data warehouse. Another part is to obtain net citizens' consensus comments from channels built by governments' branches and websites. And the data of public opinion are stored in e-government data warehouse.

1) Governments' search engine acquisition research

The network consensus spaces are almost the classification plate of several typical large portals. Using the know ability and unpredictability of intelligence gathering channel, collection method can roughly divide into network consensus fixed-point harvest engines and network consensus crawler search engines.

2) Network consensus fixed-point harvest engines research

For network public opinion caused by emergencies, we can make obtaining information list on a fixed portal, harvest the search engine, and collect public opinion information according to the list. Due to the fixed point harvest website format and contents subjects are known, the government can use fixed-point harvest information search engine to collect public opinion according to the established procedure.

a) Thread search engine time problems

Multi-thread approach can be implemented by fixed-point harvest search engine in the technology link. Web harvester sent multiple threads to reap established website listing content. The specific task allocation can be divided into: first is to create the harvest thread class, distribute a list for each harvesting thread, and a list correspond to a harvest thread. The second is to create multiple harvest thread instance, activate multiple harvest threads, share the URL list, then analyze collaborative harvest public opinion content based on the address class.

Set each web page consensus information amount as random scalar Random(x1), the rest search engine's working time caused by its randomness as random scalar Random(x2), the total search scalar as random scalar Random(x3). Uncertainty's existences cause the difference between scalar's quantity values. In the face of such phenomenon, to determine the task allocation is the entry point of bridging multithreaded harvest time difference.

b) Task allocation of threading search engine

In the Queuing Theory of operations research, D/M/c issue's D stands for that website lists' arrival process is a determined process, and its websites number is N. M stands for that harvest engine works obey negative exponential distribution, and harvest time of each site has no relevance. c stands for that search engine number is c. Assume that harvest engine to the site's harvest time obey negative exponential distribution which parameter is μ.

First. In the condition of D/M/1, the average collection time of website:

When c=1, the first site's collect time is zero. The first site's collection time is harvest engine service rate μ reciprocal , the number N website's collection time is

$$\frac{N-1}{\mu}$$

So the average collection time of total N websites is

$$W_{q,1}=\frac{N-1}{2\mu} \quad (1)$$

Second. Under the situation of D/M/c, the average collection time of website:

When $c \geqslant 1$, Assume distribute site name of c crawler engine averagely, so each crawler engine get $\frac{N}{c}$ site name. In the queue of number i crawler engine, i=1，…,c, the first site's search time is zero, the site's search time of number $\frac{N}{c}$ is

$$\frac{\frac{N}{c}-1}{\mu}$$

So the average search time of number $\frac{N}{c}$ websites is

$$\frac{\frac{N}{c}-1}{2\mu}$$

The current number of crawler engine is c, so the average search time of websites is

$$W_{q,1}=\frac{1}{c}c\frac{\frac{N}{c}-1}{2\mu}=\frac{\frac{N}{c}-1}{2\mu} \quad (2)$$

Notice that, when c=1, $W_{q,1}=\frac{\frac{N}{c}-1}{2\mu}=\frac{N-1}{2\mu}=W_{q,1}$, the result is consistent with the previous derivations formula (1).

If $c \geqslant 1$, so $< N$, thus to know that :

$$W_{q,c}=\frac{\frac{N}{c}-1}{2\mu} < \frac{N-1}{2\mu}=W_{q,1} \quad (3)$$

So it is known from formula (3) that the websites' average search time of adopting the method of multithreading public opinion harvest, a single listing is less than the average time of method of multithreading public opinion harvest, each harvest engine have a harvest listing respectively.

Therefore, task allocation method of multithreaded fixed-point engine harvest is adopting the method of single harvest listing, management in order by a harvest engine administrator. That is

$$\frac{Wq,c}{Wq,1}=\frac{\frac{\frac{N}{c}-1}{2\mu}}{\frac{N-1}{2\mu}}=\frac{\frac{N}{c}-1}{N-1}=\frac{N-c}{c(N-1)} \quad (4)$$

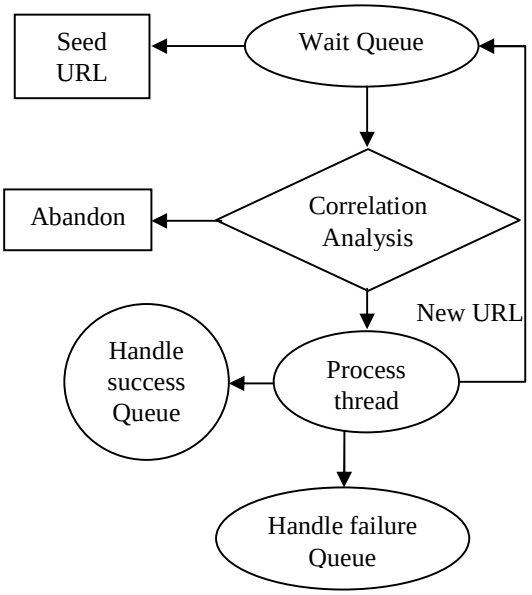

Figure 2. URL three states corresponding processing flow

B. Network Consensus Crawler Search Engines Model

The crawler is search procedure traversing the hypertext information on the Internet, visit sequentially inclusive links according to the traversal web URL, create index for gathered keywords, provide retrieval service to customers according to the index. Incremental crawlers build index directory for increased web pages On the Internet, and realize by URL and dispatching. The URL has three states, which Are waiting queue state, success and failure processing queue status. The three states conserve as way of queue, The crawler scheduling schemes achieve by using the three queue. The Figure 2 shows the corresponding process flow of URL three states.

URL processing procedures are as follows:

1) Put an initial seeds URL in the waiting queue;

2) Processing threads take out a URL to be processed from the waiting queue, call relevance judgment module to judge whether a URL is relate to the network public opinion corresponding to the specific unexpected incident., otherwise, abandon it and does not handle;

3) Processing threads obtain the web content directed by URL through the Internet, call the page processing module to handle;

4) Extract the page URL into the waiting queue;

5) Put URL handled successfully into success queue, and URL handled unsuccessfully into failure queue;

6) Return to operation（2）。

Multithreaded crawler engine task assignment

In M/M/c issue of queuing theory, the first M stands for that website list arrival process is a poisson process; the second M stands for that the crawler engine working times obey negative exponential distribution, and harvest time on websites are independent of each other; The c stands for that the number of crawler engine is c.

Assume that site arrived rules obey the poisson distribution process which parameters is λ.

First. In the condition of M/M/1, the average processing time for the site is:

$$W_{q,1}=\frac{\lambda}{\mu(\mu-\lambda)}$$

Second. In the condition of M/M/c, the average processing time for the site is:

$$Wq,c=\frac{(c\frac{\lambda}{\mu})^c}{\mu c!(1-\frac{\lambda}{\mu})^2[\sum_{k=0}^{c-1}\frac{1}{k!}(\frac{\lambda}{\mu})^k+\frac{1}{c!}\frac{\lambda}{(\mu-\lambda)}(\frac{\lambda}{\mu})^c]}$$

After proving, we can get that when $c>1$, $Wq,c>Wq,1$. Generally, $\frac{Wq,c}{Wq,1}$ is slightly larger than $\frac{1}{c}$. Therefore, the websites' average handling time of adopting the method of multithreading harvest, a single harvest listing is largely less than the average time of multithreading crawler, each crawler engine have a harvest listing respectively.

We can get multithreaded crawler engine task allocation method is adopting the method of single harvest listing, management in order by a crawler engine administrator.

Exactly:

$$\frac{Wq,c}{Wq,1}=\frac{(c\frac{\lambda}{\mu})^c}{\mu c!(1-\frac{\lambda}{\mu})^2[\sum_{k=0}^{c-1}\frac{1}{k!}(\frac{\lambda}{\mu})^k+\frac{1}{c!}\frac{\lambda}{(\mu-\lambda)}(\frac{\lambda}{\mu})^c]}\frac{\lambda}{\mu(\mu-\lambda)}$$

$$=\frac{(c\frac{\lambda}{\mu})^c(\mu-\lambda)}{\lambda c!(1-\frac{\lambda}{\mu})^2[\sum_{k=0}^{c-1}\frac{1}{k!}(\frac{\lambda}{\mu})^k+\frac{1}{c!}\frac{\lambda}{(\mu-\lambda)}(\frac{\lambda}{\mu})^c]}$$

So adopting the method of this search engine is reasonable.

Under the condition of a certain set of hardware and software, in view of the social emergency events which the theme is known--- H7N9 avian flu, large portal and online social community message board, BBS posts, and project module review information experiments, fixed-point harvest two threads compared with single thread running condition, easy to draw a conclusion that the efficiency of multi-threaded crawler read the webpage is significantly higher, mainly reflects in the crawling speed. In addition, in a multithreaded processing, difference in numbers of each thread crawl the page is obvious. For H7N9 avian flu, social emergency of which theme is known, provide a comprehensive range of Internet news portals, social networking sites, panel discussion platform specific sites.

Fixed-point harvest 10 threads compare with and single thread and two threads, through the empirical study can be concluded that using 10 threads crawlers read can give abuses in single thread and double threads to correct, can fully exert more number of public opinion, this is a single thread that cannot be achieved, and also can avoid that double thread only read the community of public information on interactive platform to, and underestimate the portal website of information access. The ten threads selections gather the advantages of the first two, and achieve the information mining quantity, accurate and avoid the opinion read incomplete.

III. CONCLUSION

This paper analyzes the present form of network public opinion container in social emergency events, it is concluded that the network public opinion through different channels to collect the characteristics of explanation. (1) Judgment module is used to limit the crawler crawl content, reduce the course of retrieval error, so as to reduce the deviation of data between accurately; (2) In the process of the crawler crawl using multithreading technology can improve the efficiency of the crawler.

Research on time efficiency in data acquisition based on operational theory ---multithreaded crawler search engines task allocation, for the pure digging in terms of information itself, it has certain significance. Turn it into a kind of production factor to analyze how to make the social emergency network public opinion data collection resource configuration optimization, so as to realize the comprehensiveness of data mining research.

ACKNOWLEDGMENT

This research is financially supported by the Fundamental Research Funds for the Central Universities (Grant No. 2012-IB-003).

REFERENCES

[1] Stanley DS.Stock price reactions to the Wall Street Journal's securities recommendations. Journal of Financial &Quantitative Analysis, 25(2):399-410, 2009.

[2] Shim J P, Warkentin M, Countney J, Power D, Tsoi A C.Incremental training of support vector machines, IEEE Trans on Neural Networks, 16(1):114-131, 2005.

[3] Stock D G, Allen J D. How to solve the N-bit encoder problem with just one hidden unit. Neurocomputing, 5:14-143, 2010.

[4] Peng Hwa Ang. How countries are regulating interact content[EB/OL]. 2010.8.22. htt://www. panAsia. org. sg.

[5] Duggan F, BanweUL. Constructing a model of effective information dissemination in crisis[J]. Inofonnation Research,2004,5(3)z:178-184.

[6] Simon French, Pmurray Turoff. Decision Support Systems[J]. Communications of the ACM, 2007, v01.50(No.3):39-40.

[7] Robrt Heath, Crisis Management for Managers and Executives. London: Financial Times/ Pitman Pub, 1998

[8] W . Timothy Coombs . Ongoing Crisis Communication Planning, Managing and Responding.NewYork:Sage Publications,Inve.1999