2009 International Conference on Management of e-Commerce and e-Government

Empirical Research on E-Government Based on Content Mining

Yang SHEN[1], Zitao LIU[2], Shaoji LUO[3] ,Huijuan FU[1], Ye Li[4]

[1]School of Information Management
Wuhan University
Wuhan, China
yshen@whu.edu.cn

[2]International School of Software
[3]School of Economics and Management
[4]School of Electronic Information
Wuhan University

***Abstract*—According to acquiring data from the meta-search engine and getting information in specific websites, the author proposes an extraction model based on Web information which is used to construct network relationships of the subject based on its semantic link. Then based on the proposed model above, the author does content mining and semantic analysis on the Web data of five big cities (Beijing, Shanghai, Wuhan, Guangzhou, and Chengdu) with the help of self-made ROST Content Mining System, to get first 30 high-frequency e-government words respectively, and takes Shanghai for specific analysis; Meanwhile, the author, using ROST WebSpider to collect the web page from level 1 to 3 of governments' websites in Beijing, Shanghai, Wuhan, Guangzhou and Chengdu, constructs the evaluation model SCISS to do comparative analysis on the development of the five metropolis' e-government. Finally, the author comes up with some countermeasures, aiming to provide advice for the development of e-government in china, according to the empirical analysis.**

Keywords-social network; E-governmen; meta-search engine; content mining

I. INTRODUCTION

Researches of progressive analysis and content mining on the data generated in e-government network services have become the hot spot in information science. Jarl K Kampen and his partners did analysis on the citizens' preference data of Flanders's government website, coming to the conclusion that every government will provide more on-line services, but citizens need better ones[1]. Li, Honglai and Le, Zhongjian proposed the second relative evaluation method, in accordance with the importance and fairness of e-government evaluation [2]. Liu Honglu and Tian Zhihong put forward the personalized information services framework based on the Web content mining model after analyzing on these e-government systems that fulfill users' individualized needs and then designed and developed the experimental system for Web log mining, which established technical foundation for individualized e-government services[3]. Ou Jing-ying and his partners took advantage of the decision tree, association rules, clustering algorithm, to analyze on the hotline services provided by e-government, in order to provide reference for decision-making, managing, and serving[4] .

One informative government has become one of the key factors to enhance global competitiveness of a country or a region[5-6]. But in terms of Chinese e-government, there are many problems in urgent need of improvement and there are no researches focusing on overall data acquisition and deep analysis based on e-government in China[7]. This paper analyzes China's e-government situation based upon Chinese situation, aiming to provide reference for its development in both theory and practice.

II. INFORMATION EXTRACTION

A. Informaction Extraction Model

At present, the methods to acquire information related to some keyword belong to two ways: one is to search the information related to the keyword in search engines; the other is to acquire information in websites related to the keyword. The former method can help to get information in wide coverage, but has bad pertinency and depth, the latter helps to get accurate and deep information, but covers narrowly of the Internet. So, we decide to propose a model which combines information acquisition in search engines and information mining in specific websites with the help of our self-made software, ROST Content Mining System[8].

Firstly, we monitor web pages returned from the four huge search engines(Google, Baidu, Sougou, and Youdao), and then extract data from the returned pages; secondly, we get data of some specific websites; finally, we conduct the process of content analysis and social network mining using the acquired data. The information extraction model described above shows in Fig. 1:

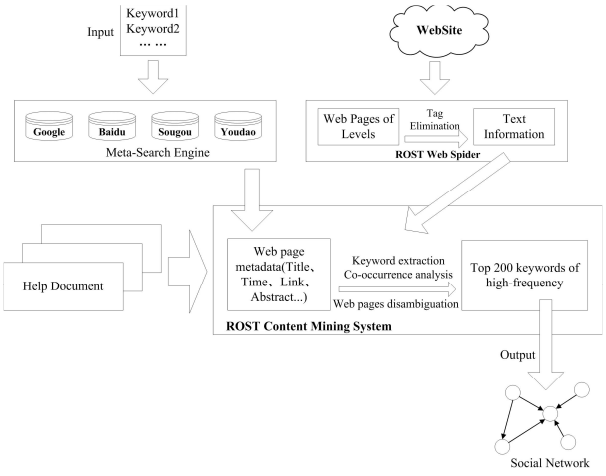

Figure 1. Information Extraction Model based on Web.

978-0-7695-3778-8/09 $26.00 © 2009 IEEE
DOI 10.1109/ICMeCG.2009.48

B. Basic Algorithm

Considering the fact that the data acquired from the meta-search engine is complex and pluralistic, we propose "adaptive text segmentation and co-occurrence algorithm" to achieve effective analysis on the data.

First, we extract the key content in corresponding sections according to the format of the acquired data, then we use the ROST Content Mining System to get the original Chinese words segmentation set, taking advantage of the software to keep the proper nouns in related fields with the help of its self-defined word list. After that we screen the data in the original Chinese words segmentation set, calling the filtering word list to get the rightsizing text data set. Second, we conduct "extraction non-repeated words in each line" which means exactly comparing the Chinese words in each line of the rightsizing text data set to get the non-repeated data set of each line. Third, we add one to the frequency of their relationship between two words if they appear in the same line, which causes to get the co-occurrence data set. Finally, we use ROST CM to get the first 200 (adjustable parameters) available words according to their frequencies in a descend order in the co-occurrence data set to form the visible semantic network graph.

The specific algorithm shows below:

```
1) Pre_ExtractContent (Data Set)
2) Div_Word (DataSet);
3) Get_CustomList (DataSet);
4) Get_FilterList (DataSet);
5) For i = 1 to DataSet.LineNum Step 1
6)     For j = 1 to DataSet.LineSet[i].WordNum Step 1
7)         If (Co_Occur())
8)             Cooccurrence_Count++;
9) Co_Occur_Sort (DataSet);
10) DrawFig (DataSet);
```

III. CASE STUDY

A. Analysis on Five Metropolis' E-governments based on Meta-Search Engine

Based on adaptive text segmentation and co-occurrence algorithm, we acquire e-government data of the five key metropolis in China respectively, and then we do content mining on the acquired data. During the process, we add one to the frequency of their relationship if one word appears together with e-government in each corresponding web page. So after filtering the meaningless words, we get the first 30 high-frequency words in semantic relationship respectively for these five metropolis to construct the table of e-government in Beijing, Shanghai, Wuhan, Guangzhou, and Chengdu. The table is shown in Table I:

According to the statistics of e-government in Beijing, Shanghai, Wuhan, Guangzhou, and Chengdu based on word frequencies, we come to the following conclusions:

- Words like "technology", "research", "university" and specialization field are hot in the five metropolis, indicating that technology researches and training are the core issues in the process of e-government construction.
- "enterprise" and "bid" are among the high-frequency words of Guangzhou and Chengdu, implying that other than government departments' direct constructing, enterprise outsourcing is a excellent way to solve the problems.
- "safety" appears in high frequency in four of the five metropolis, indicating the great importance of security in the process of e-government development. But "safety" does not exist in Wuhan's semantic high-frequency words, which shows that security should be taken into consideration more seriously by Wuhan government when developing e-government.
- How to make procurement more reasonable and efficient has been focused by the governments of Beijing, Guangzhou, and Chengdu, as procurement is governments' unavoidable behavior.
- "share" is only related to Guangzhou, reflecting that the idea of information resources sharing occurred to Guangzhou earlier than the other metropolis. Sixth, although one of the core goals to construct e-government is to enhance the communication between government departments and citizens, none of the five cities has words related to communication. So improving communication services should be one of the key issues to be done as soon as possible.

B. Analysis on Shanghai E-government based on Meta-Search Engine

Using the model proposed in section 2.1, we take Shanghai to do our case study. First, we input "Shanghai E-government" into the searching box of ROST Content Mining System, which causes the returning of 2665 pages of website abstracts from the meta-search engine (749 pages from Baidu, 662 pages from Google, 290 pages from Youdao, and 964 pages from Sougou). Then we do semantic

TABLE I. TOP 30 HIGH-FREQUENCY WORDS OF EACH METROPOLIS

Beijing	Shanghai	Wuhan	Guangzhou	Chengdu
information	construction	construction	information	construction
construction	information	information	construction	information
informationization	government	system	government	government
system	information ization	administration	system	information ization
Work	system	government	informationization	system
service	development	affairs	electron	network
government	administration	project	centre	service
administration	work	network	network	centre
project	network	platform	administration	application
online	website	centre	project	affairs

Beijing	Shanghai	Wuhan	Guangzhou	Chengdu
technology	China	informati onization	application	electron
engineering	application	software	website	safety
safety	centre	universit y	service	project
network	project	service	platform	developme nt
China	service	work	work	work
platform	safety	electron	affairs	administrati on
appliance	electron	technolog y	development	website
examine	affairs	administr ative	bid	China
development	technology	developm ent	enterprise	technology
electron	plan	software	technology	company
affairs	specializati on	city	security	platform
research	office	engineeri ng	purchase	engineering
Trade	online	appliance	city	enterprise
purchase	platform	design	share	impel
convene	software	country	office	software
standard	country	departme nt	data	purchase
office	company	office	engineering	bid
brace	administrati ve	plan	software	provide
centre	country	online	unit	unit
train	provide	research	country	department

analysis on each of the returned pages to get the first 200 high-frequency semantic relationships, on the assumption that if two words appear together once in any of the pages, the frequency of their relationship is added by one. So the visible semantic network graph based on Shanghai e-government is formed as below in Fig 2.

After analyzing on Fig 2, we come to the following

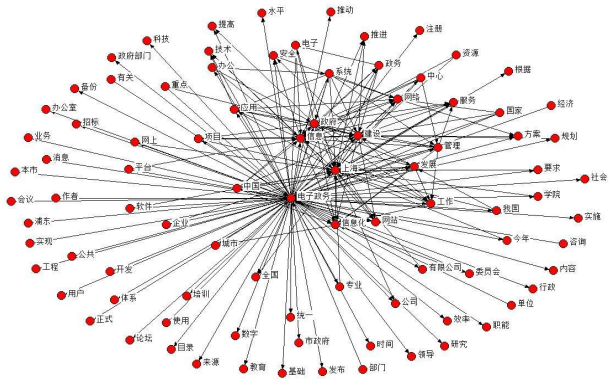

Figure 2. Semantic Network Graph based on Shanghai E-government

conclusions: first, since words like "construction", "development", "impel" appear in high frequency, it is obviously that e-government is in the developing stage. It needs not only to improve the system technique to make it more automatic, but also to increase the scope and depth of its offering services. Second, the appearing of words like "informationization" and "security" indicates that the most important and hardest point of developing e-government is how to informatize the mass data and how to solve the security authentication problem. Third, e-government has become an effective way for governments to communicate with their citizens. The discussion and consultation in network communicating platforms improve governments' work efficiency, and they reflect one government's image to some extent. Fourth, in order to improve efficiency, the standardization of e-government in different government departments is in great need for e-government development, which is beneficial to the sharing of information resources.

C. Analysis on Five Metropolis' E-governments based on Websites

The empirical process is as follows: firstly, we use ROST WebSpider to get the website data (altogether 2330 html web pages from level 1 to 3) of Beijing (http://www.beijing.gov.cn),Shanghai(http://www.shanghai.gov.cn), Wuhan (http://www.wuhan.gov.cn), Guangzhou (http://www.guangzhou.gov.cn), and Chengdu (http://www.chengdu.gov.cn); secondly, with the help of ROST CM , we do semantic analysis on the acquired webpage abstracts to extract the first 200 high-frequency words in semantic relationship, and then conclude the statistical table of word frequencies after removing the repeated words and efficient analyzing. At last, according to the semantic analysis on e-government, we construct the e-government evaluation model named SCISS which regards standardization (including specification, administration…), communication (including communicate, net fried…), informationization (including information, news, network…), service (including service, policy…) and security (including safe, guard, ensure…) degree as the main criterions[9-10],The table shows in Table II.

TABLE II. STATISTICS OF KEY WORDS TABLE OF FIVE METROPOLIS' E-GOVERNMENT

	Beijing	Shang hai	Wuhan	Guang zhou	Cheng du
Standardiz ation	2680	3570	3485	2410	2125
Communi cation	2905	2790	2655	2820	1830
Informatiz ation	3030	3090	2910	2915	2480
Service	2755	3465	3320	2860	2740
Security	2170	2885	2480	1590	1215

From the table above, we conclude that compared with other metropolis, Shanghai e-government develops in a fast pace, especially in standardization and service. In the aspect of informationization, Shanghai e-government and Beijing e-government are outstanding, deserving to be used for reference by other metropolis. Enhancing e-government security is still an important task for five metropolis. Communication is one of the key issues that should be developed in priority, where Beijing sets a good example for the other metropolis.

IV. CONCLUSION

The main work of this paper includes: We propose the new effective information extraction model which combines using meta-search engine and acquiring data in specific websites together and develop the content mining tool ROST Content Mining System after analyzing the research based on e-government home and abroad and mining on the information that both returned from search engines in width, and acquired in specific e-government websites in depth. Then through the process of semantic mining on Beijing, Shanghai, Wuhan, Guangzhou, and Chengdu, we propose a simplified evaluation model SCISS to evaluate e-government websites, which will surely help to give suggestions and opinions to achieve the goal of making a more efficient and effective government. In the next step, we will analyze the existing data more deeply, and mine data in more e-government websites that have typical representativeness. Also, we will monitor the acquired data set regularly in a dynamic way and conduct mining further on the real-time semantic network, and the ROST Content Mining System will be improved in a further way.

ACKNOWLEDGMENT

This paper is financially supported by National Natural Science Foundation of China (No. 60803080) and Ministry of Education of the P.R.C. Humanities and Social Science Youth Project (08JC870010), and the National Basic Research 973 Program of China (2007CB310806).

REFERENCES

[1] Jarl K Kampen, Kris Snijkers, and Geert Bouckaert, "Public priorities concerning the development of e-government in Flanders," Social Science Computer Review, vol. 23, Spring. 2005, pp. 136-138.

[2] Li Honglai, and Le Zhongjian, "E-government evaluation based on binary relative performance," Proc. International Conference on Management of e-Commerce and e-Government. (ICMeCG 08), Oct. 2008, pp. 115-119.

[3] Liu Honglu, and Tian Zhihong, "Research on application of web usage mining in e-government personalized information system," Proc. International Conference on Management Science and Engineering. (ICMSE 06), Nov. 2006, pp. 1297-1304.

[4] Ou Jing-ying, Yu Shou-hua, and Li Yin-mei, "The empirical study on the application of data mining in E-government," Proc. International Conference on Public Administration. (ICPA 07), OCT. 2007, vol. 1 pp. 455-459.

[5] G. J. Ramon, B.Sara A., P. Theresa A, B. G. Brian, and G.Ahmet, "Conducting Web-based surveys of government practitioners in social sciences: Practical lessons for e-Government researchers," Porc. Annual Hawaii International Conference on System Sciences (HICSS 09), Jan. 2009.

[6] Lappas and Georgios, "An overview of web mining in societal benefit areas," Proc. International Conference on E-Commerce Technology / International Conference on Enterprise Computing, E-Commerce and E-Services (CEC/EEE 07), IEEE Press, Jul. 2007, pp. 683-690.

[7] Ying Li, "Evaluation of public service maturity of e-government: A case in Chinese cities," Porc. International Conference on Service Operations and Logistics, and Informatics (SOLI 06), IEEE Press, Jun. 2006, pp. 209-214.

[8] Shenyang, "ROST Software and Tools Download List," http://hi.baidu.com/whusoft/blog/item/6259de2f9e7a2c3c1f3089f9.html. 2009-1-11/2009-3-20

[9] Vassilakis, C., Laskaridis, G, Lepouras, G., Rouvas, S., and Georgiadis, P., "A framework for managing the lifecycle of transactional e-government services," Telematics and Informatics, vol. 20, Nov. 2003, pp. 315-329.

[10] De Saulles M, "E-government and patterns of innovation in the public sector," Proc. European Conference on E-Government (ECEG 07), Jun. 2007, pp. 111-116.