Research of Web Information Quality Control based on Informatics Theory

Applied in Water Resource Search Engine

Gao yimin
Information research center
Hohai university, Xikang road 1
Nanjing, China
Emails:gaoyimin@hhu.edu.cn

Ye feng
Computer and Information Engineering College
Hohai university, Xikang road 1
Nanjing, China
e-mail: yefeng1022@ hhu.edu.cn

Abstract—**With the fast growth of network information, it is urgent to control network information quality and to grasp the information efficiently. Webpage control research can be classified in to two stages: before-search quality control and in-search quality control. Before-search stage includes website quality control and webpage quality control. This paper applies intelligence theory to web site quality control. We try to filter kernel web site through analysis by Budde Raffo's law, to set up water resource search words by Zipf's law ,and build the model of webpage quality control. and Price's law was applied to in-search quality control by user's search behavior. This research method has now applied in water resource search engine.**

Keywords-Information quality control;Intelligence theory;Search engine

I. INTRODUCTION

With fast increasing quantity of web page and no auditing process which usually used in ordinarily publication, it becomes difficult to find valuable information on the web.

Currently there are many researches on web page quality control which are focus on webpage quality control before-search stage. Researches rarely apply intelligence theory, but statistics theory. Intelligence theory is summarized from literature work and the web page is literature work in web. So this paper presents an automatic web information control methods which go through the whole search process and apply intelligence theory.

Researches of web information quality control could be divided to manual control and automatic control. Let's introduce them respectively. Automatic control methods researches are as bellow:

Automatic control methods focus on hyperlink, URL, the length of text, etc. For example, the most popular search engine Google adopted the web page quality control method—PageRank. PageRank believes that the more the page was linked by other page, the more important t he page is; and the more important the page was linked by, the more important the page is. The famous Chinese search engine Baidu also has its web page quality control algorithm --Hits. The Hits believes that the importance of one page can be divided into two indices, one is authority, and the other is hub. Both Pagerank and Hits evaluate the quality of web page based on traditional literature citation analysis. However, hyperlink is different from traditional literature citation, because some hyperlinks are used to organize web site, some are directed to the web page that is related to this page, and even some are just advertisement. Also, there are some search engines which use cheating to promote the lower quality page to higher ranking. Chen xiaofei in Fudan University proposed improving Pagerank algorithm. He thinks the quality of webpage A, which was linked by high PR rank webpage, is higher than the quality of webpage B, which was linked by lower PR rank webpage through webpage A and B gets the same PR rank.

Manual control methods includes: University of Michigan set up guide to subject resource. The web site evaluation includes formulation of web resource, content, navigation, organization structure. American researcher Smith·Alasatir G brought forward the index of web resource, which includes scope of information, content of information, graph and multimedia design, comment, charges and so on. And other researchers evaluate website with accuracy, coherence, safety, timeliness, wholeness, accuracy, reality, Availability, relativity, impartiality, usability, intelligibility, and so on. Shirlee-Ann Knight divided this index into three class: subjective, objective, and process. James A proposed that website intelligibility evaluation, such as webpage views, and through web blog data mining, information of users who provide feedbacks, promote website quality.

Stuart Barnes suggest that website quality evaluation must listen to user's opinion. He studied E-business website and discovered that service quality of website is key factor. Service includes navigation, communication, amalgamate, information quality, individuality, payment platform, and so on. Tsinghua University Ma shaoping tried to find high quality web page through URL, length of text, the number and length of link. Shanghai website construction site believes that reputation could be used to evaluate web page quality. The indices include trust, authority, correlativity, revisit, recommend, and ranking. Li Shuqing in Nanjing University of Finances and Economics put forward index "availability" in information search. It includes popularity, webpage links, website traffic, user visiting model, the webpage author's reputation. WuHan University Ma Feicheng discovered that the research of user's behaviour could promote web service quality effectively. He put forward that through analysis of web blog, set up user's behavior pattern and improve website quality. Ma Feicheng also applied information half-value period to web information quality control.

It shows that early research of manual web quality control focused on web information construction. Recent researches focus on user’s behavior. User’s behavior research is to find user’s information through web blog data mining, and then improve web site quality. Also there is research which applied half-value period to web information quality control. Manual web quality control is more reliable, but it is more suitable for small scale webpage but not for large scale as World Wide Web.

Most of automatic web quality control research relay on analysis of hyperlinks. As hyperlinks could be used as website organization for commercial benefit, it could not be used as web quality control precisely. Many search engines apply words frequency to web information quality control. Most of them apply statistics theory as data analysis, but not intelligence theory.

II. RESEARCH ON INFORMATION QUALITY CONTROL BASED ON INTELLIGENCE SUBJECT THEORY

Informatics is the research of the rule of information’s production, transmission, and the method to keep the intelligence system in best state by using the modern information technology. This research try to explore the rule of information based on informatics. As a new application, intelligence theory hasn’t been applied in subject other than intelligence. Most of intelligence theory was emerged in conditional document, and would be revised in web environment. Search engine research is lack of theory support because most research methods applied in search engine are statistics method. This paper combined intelligence theory with web information character, and explored information quality control theory based on practice.

A. Web quality control before search process

1) web site level quality control

Different types of websites use different quality control model. Different types of sites have different contents. If apply the same quality control standard to all web information, it cannot give consideration to both recall ratio and precision ratio. Webpage of scientific website quality control can be appropriately looser and focus on removing irrelevant information. As to the portal site and blogs, quality control would be more strictly than scientific website. For example, news about storm from portal site can only reported Beijing City waterlogged conditions and the technology level is not high, while the storm information in hydrology website must be detailed precipitation information which had high technology level. Therefore, it is needed to establish suitable standard for different types of websites control model.

It is important to screening core web site which applied “Budde Raffo's law ". Web site quality control could improve the search results priority.

2)Webpage level quality control

In the webpage level of quality control, it is based on search thesaurus and analysis the relationship between keywords in thesaurus.

Water conservancy search thesaurus, which embodies the concept hierarchy, was applied “Zipf’s law " and composed of Water Conservancy Key words and Terms.

The index which extracted from relationships between key words and terms includes content coverage, content depth, content density, correlation degree, content innovation degree and etc.

Both the index of content quality and hyperlink analysis together build webpage level quality control model.

B. Web quality control in search process

Through the research of science user’s behavior, effectuate the quality control of both before search process (webpage collection processing) and in search process. In the process of webpage quality control, user's search behaviors (such as the hit ratio) are main affect factors. Combined the User search behavior and “Price law”, constructed the search process quality control model, which promote high quality webpage in search results ranking.

III. WEB SEARCH QUAILTY CONTROL APPLIED IN HYDRAULIC SEARCH ENGINE

Water conservancy is priority in our development strategy. It is badly in need of the support of network information. Hydraulic search engine have been developed and implemented by the work group. This search engines using webpage quality control, will greatly enhance the searching efficiency of the users in water conservancy field.

This work group is on the tentative practice. Because of small scale, the results are not mature. The analysis provides experience for further study. The work group conducts quality control on the web site level in the first step, and then collects water conservancy webpage as seed site, crawling webpage, a total of 14815, mostly from government web sites and large research institutes web sites.

The 14815 web pages are from about 5 websites which everyone has over 1000 web pages be crawled; and from about 16 websites having 100 web pages; and the remain web pages from other 10 sites . l of interdisciplinary study arranged in descending order: core area: Area: non related area : 1:n:n^2.

Contrast to “BOD La law”, the ratio of water conservancy web pages was arranged in descending order: 1:3:2.

There is large difference between application of ”Bo La law” in tradition paper magazine and web environment.

Through author’s analysis, in water conservancy field, the cross among the different classes of websites, such as government, scientific research and commercial websites, blogs and other websites, is not so much. Therefore, in the crawling process, it is difficult to obtain web pages other than water conservancy web sites. This led to the test result. The core websites are below:

TABLE I Core websites

	Web site	Affiliations	The number of web pages
1	China water potential	Research Center of development of irrigation works ministry	3803
2	China South-to-North Water Transfer Network	The State Council South-to-North Water Diversion Project Construction Committee Office	2077
3	The Yellow River Water Conservancy Commission	The Yellow River Water Conservancy Commission	1771
4	The ecological construction of soil and water conservation network	Dept. of water and soil conservation	1228
5	China Water Conservancy and international cooperation and technology network	International cooperation and the Department of science and technology, Ministry of water resources	1081

In the layer of webpage control, this work team applied “Zipf's law” to build water conservancy web search thesaurus. The webpage collected was divided to three layers: water resource, water conservancy and integration. The analysis of the web page among layers is bellow in table 2.

TABLE I. WEB PAGE KEYWORDS IN DIFFERENT LAYERS

	The number of webpage	The total of key words	lower limit of high frequency key words	high limit of lower frequency key words
Water resource	230	9711	5276	1
water conservancy	500	8978	180	3
Integration	600	34973	588	8

The frequency of key words of water resources layer can better reflect the content of the article and classification, even the frequency of keywords is higher. Keywords volume in integration layer increased significantly, in which there was 16813 key word’s frequency being one. Research group built thesaurus which eliminate high frequency key words and lower frequency key words. To compare the search efficiency using thesaurus which use stop list combined three levels.

This research results applied in water conservancy search engines. Compare with Baidu, Google and other well-known search engines, the analysis of effect of different search engine as follows:

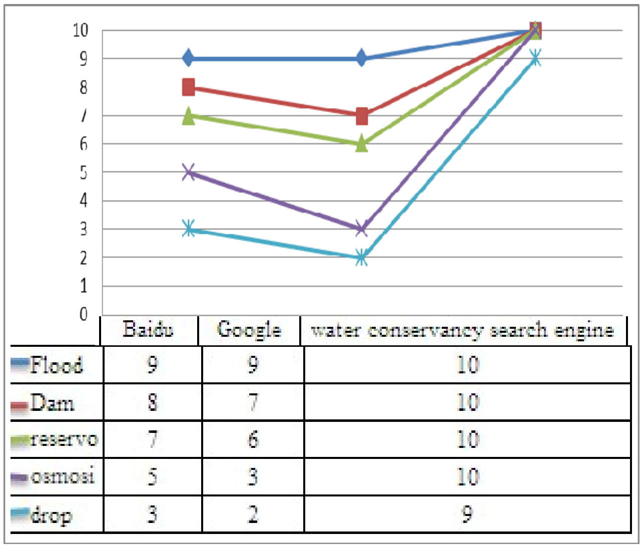

Figure 1. The first page of search results contrast among search engines

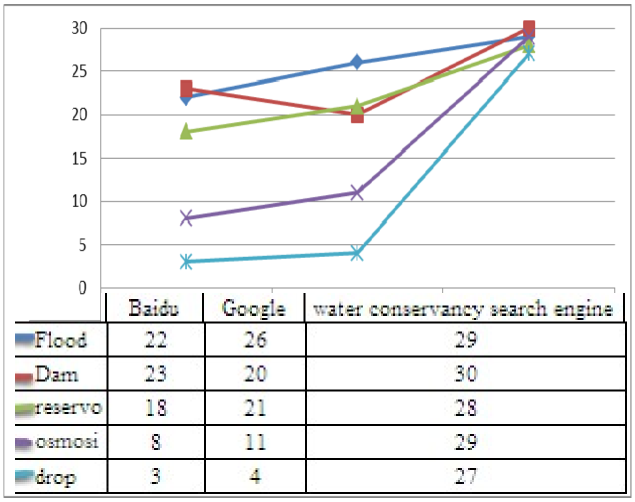

Figure 2. The first three pages of search results contrast among search engines

IV. Prepare Your Paper Before Styling

Search engine, such as Baidu, Google, all adjust algorithm to improve the quality of webpage. They all use limiting keywords frequency method to eliminate the lower quality webpages. This topic applied "Boyd La law ", " Zipf's law " and other intelligence theory to web information quality control . Through the application of "Boyd La law " , it shows that there is a gap between water conservancy field sites and other field sites. To solve the problem, spide needs more different kinds of webpage as seeds. The application of "Zipf's law” can improve the search performance of water conservancy.

From the experiment, it shows that, after revision, the theory of information science could improve the network search efficiency, and to form the network search theory.

As the research on web quality control in search process must be after a long time of search engine test run, this topic have no experiment data.

Based on the research, this team established water conservancy search engine. In the practice, water conservancy search engine increased the precision ratio.

REFERENCES

[1] Ma Feicheng, Gao Jing. “Empirical Study of Impact Factors of Web2.0 Information Half Life”. Information Studies:Theory & Application,2010(11) ,pp. 1-6

[2] Discard the false and retain the true discard the dross and select the essential——Page quality assessment and its application in Web Information Retrieval”. http://www.sewm2006.sdu.edu.cn/ppt/Experts/Evaluation.ppt

[3] Chen Xiaofei,Wang Zhitong, Feng Xiaojun. “An Improvement of Page Rank Algorithm Based on Page Quality”. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT. 2009(z2) , pp. 756-762

[4] Li Shuqing, Chui Huizhi. “The Evaluation of Web Pages Quality Analysis Methods in Web Information Retrieval System”. INFORMATION SCIENCE, 2008,(05), pp. 729-734

[5] Yu Xiaosheng; Ma Feicheng. “Research on Construction Methods of Network User Behavioral Mo del”.Information Science,2011（04）, pp. 605-608

[6] Suto, Hidetsugu and Kawakami, Hiroshi and Handa, Hisashi, “A Study of Information Flow Between Designers and Users Via Website Focused on Property of Hyper Links”, booktitle. “Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design”, Heidelberg, Springer Berlin, 2007,pp.189-198

[7] James A. Ogle MSc, MBA ; ”Improving Web Site Performance Using Commercially Available Analytical Tools” ; Clinical Orthopaedics and Related Research, 2010-04-02, 10, 2604～2611

[8] The web site KPI's quality control, http://webdataanalysis.net/en/web-quantitative- analysis/kpi-quality-control/