# Research on Application of Web Log Analysis Method in Agriculture Website Improvement

Jian Wang[1]
([1]Agricultural information institute of CAAS, Beijing 100081, China)
wangjian@caas.net.cn

***Abstract*:** With the advance of agricultural modernization, agriculture website was increasingly becoming a major tool for farmers getting information about life and production. How to make the analysis of the needs of farmers effectively to help them to find the information from the ocean of information and resources of the Internet they were interested in had become an urgent and important issue. In this paper, we used the website of Agridata as an example and focused on the solution for the problem. A way was proposed for analyzing and mining the web log of Agridata, which integrated statistical analysis and cluster analysis. By this method, could information behaviors of users be grasped when browsing the website. It was important significance in improving the structure and content of agriculture website, which could provide better services for farmers and improve the level of modernization of agricultural production.

**Key words:** Web log analysis; information behavior; agricultural users; agricultural website

## 1 Introduction

As the emphasis on the rural information continues to deepen in our country, the development environment for rural information was constantly optimized and farmers' awareness of information technology was growing. More and more famers knew the technology for getting information of production and living from Internet, which could fundamentally change the situation of information isolated from the outside world to remote villages. Therefore, agriculture web site provided a great convenience to farmers in eliminating poverty, agricultural sales, employment, medical treatment and education. Thus, agriculture web site played a positive role for improving living conditions in rural areas and safeguarding national unity and stability. However, during the process of building agriculture web site, as the special nature of agriculture, farmers' behavior of access information had some

special needs. How to grasp the information needs of agricultural users was important for improving agriculture web site. This was an urgent problem.

Web log was a kind of file which record some information of user access to web pages which included records of Browse, search, information downloads, and other message of site visits. We could obtain varies information about visits and utilization of website resource and properties of users by calculating and analyzing web server log. This information could help us optimize the information resources and improve quality and efficiency of web services. At present, many commercial Web sites had analyzed user behavior through the log analysis techniques to improve web services and this area had become a research focus. For example, Zhang xuehong(2005) discussed the actual value of user behaviors analysis, home page designing and customer service with the method of log analysis and introduced one process of homepage log analysis. Fengchunhui(2010) studied the web log applied in distance education using data mining technology and obtained a good result. Furthermore, Xuping (2010) and Wengchangping(2010) also studied web log playing the role of mining users information behaviors to improve website with their background.

However, the study background of user information behavior analysis based on web log which had referenced from existing literatures were for academic professionals such as library users and distance education students. This method was very little used for analyzing the information behaviors of web users in rural areas. To resolve this problem, in this passage, the Web access log of the agricultural scientific data center was used as an object and we discussed the method for studying the information behavior of users of agricultural methods with statistical analysis and cluster analysis. Then we conducted experiments with the Web access log of the national agricultural scientific data center and analyzed the pattern of information behavior of its user. With these results, several advices on this website of the agricultural scientific data center were made for providing a better service to agricultural users.

## 2 Materials and methods

### 2.1 web log of the agricultural scientific data center collection

The agricultural scientific data center (Agridata) is one of the experimental data

centers supported by National Facilities and Information (NFII) of Ministry of Science and Technology. Based on agricultural scientific data sharing Standard, Agridata Integrated 12 types of agricultural scientific data which included crop science, animal science & veterinary medicine, agricultural resource and environment, grassland science, food science and standards, etc. these data resources could support the agricultural technology innovation and management decision greatly. Up to 2009, Agridata Formed a stable user group including more than 150 group users and 8,000 individual registered users. The total visit count of Agridata was 1.8 million. It showed that Agridata was an important information source of network users in rural areas.

The web log of Agridata was some log files about website visits existing in the web server which included IP address of website visitors, URL of explored pages, date and time of visits, path of exploring, etc. Usually, as different setting of web server, the format of log files had three types: National Center for Supercomputing Applications Combined (NCSA), Microsoft IIS Format and W3C Extended Log File Format. The log file format of Agridata was W3C Extended Log File Format. The format and meaning of the fields in the log file was shown in table 1.

**Table 1**. The format and meaning of the fields

| D | Index | W3C Extended Log File Format | Meaning of fields |
|---|---|---|---|
| | Host IP | c-ip | IP address or domain name of the client computer which send HTTP request to sever. |
| | User name | cs-username | user ID and host of users who send |
| | | cs-host | HTTP request to sever |
| | Date and time | Date、Time | Date and time of HTTP request to sever send by user |
| | HTTP Request | cs-method、cs-uri-stem cs-uri-query | HTTP request to sever send by user |
| | HTTP service status | sc-status sc-substatus sc-win32-status | Service status codes of HTTP request which indicated success or failure |
| | The number of bytes | sc-bytes、cs-bytes Time-taken | The number of bytes and time taken of processing HTTP request and web server receiving and sending |

| | | | |
|---|---|---|---|
| | Backlinks | cs(cookie)<br>cs(Referer) | URL and cookie visited by user before the last HTTP request send |
| | User Agent | cs(User-Agent) | browser and operating system software information user used to visit web server |
| | Protocol Version | cs-version | format version of log file |
| 0 | Server circumstances | s-sitename<br>s-computername<br>s-ip<br>s-port | information of website sever included name of computer, name of website, IP address and port number |

## 2.2 Web log data processing

Web log file gotten from website sever directly contained very rich information. Only a part of this information was needed to analyze information behavior of website users. How to extract this useful information from the log file correctly and completely was the key for analysis of user information behavior. This process was called web log data processing.

As usual, there were four steps in web log data processing: data cleaning, user identification, session identification and access fragment identification. The flow chart of web log data processing was shown as figure 1.

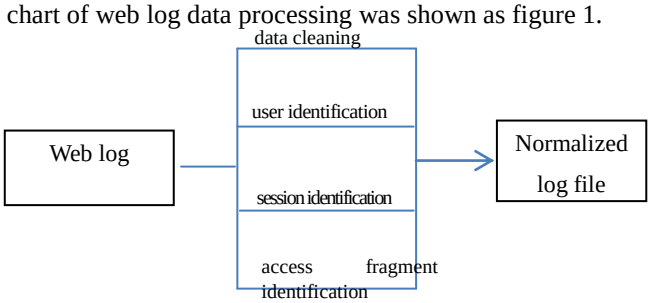


**Fig1.** The flow chart of web log data processing

### *2.2.1 Data cleaning*

The purpose of data cleaning was deleting the information unrelated to analysis the user information behaviors from web log files, combining similar records, summarizing and dealing with records of error accessing. As usual, HTML files in web log were associated with the user session. On the other hand, during the process of visiting page, images, video and audio files in the page were also downloaded and visit recorders of these files were written in the log. These information were useless for analysis the user information behaviors and should be

deleted with the method of URL suffix filter. But it was worth noting that the method of URL suffix filter was used with understanding website content to avoid the loss of some important user sessions. For example, for a page which main content was image, we should not simply delete records of images access from log file when using the method of URL suffix filter for data cleaning. Meanwhile, we should define a set of rules to retain this information as a basis for analysis. On the other hand, the operating records of network administrators, web spider and web crawler could be filter out during the process of data cleaning.

***2.2.2 User identification***

The purpose of user identification was analyzing information of users. As the local cache, proxy servers and firewalls were used in web, such as different users using the same proxy servers, the computer with the same IP address at the same time having different systems and browsers, etc. it was difficult for identifying users with IP address. Usually, during the process of research, we could set some rules to simplify the classification process: when the same IP address used different systems and browsers at the same time, users with the IP address could be considered as different users; when the same IP address had visited a group of pages which were not topological link, users with the IP address could be considered as different users and so on. On the other hand, as people in research institutions were relatively concentrated geographically while farmers were more scattered, we could set a rule to infer these two kinds of people: the grope with high concentration IP address were researchers, otherwise the group with low concentration IP address were farmers.

***2.2.3Session identification***

A session was a valid user to access the service. Details on session identification were processes of identifying users' requests for consecutive pages to get users' message of information behaviors and interesting objection. The purpose of session identification was creating page clustering for each user's information behaviors, which could provide "material" for analysis information behaviors of user. Usually, for a log file of long series access time, the same user may repeatedly visit the same page. Thus, we should set a rule: when time intervals of request for any two adjacent pages were over a set threshold, we could regard this process as restarting a session.

***2.2.4 Access fragment identification***

The nature of access fragment identification was supplementing the above process

of session identification. It could mean that this process determined whether any important request in file log could be identified. In this process, the method we applied was similar to the way of user identification which set a rule: when pages between current request and last request were not hypertext links, this page the user current request was called from the local cache. In this case, we should check the reference log to determine the source of the current user request. When several pages in visit history contained links to current requested page, we used the page that visit time was closest to the current one as a current source. If some web log was missing, we could add the missing page to the user's session file by the method of analysis of site topology.

## 2.3 Log data mining and user information behavior analysis

After data of web log processing, a large number of redundant information was filtered out and the useful information could be mined to get the information behavior of network users. Usually, there were two steps in this process: user behavior information mining and user information behavior analysis.

### *2.3.1 User behavior information mining*

User information behavior is a concept of diversity, on one hand, only when information was useful to users, had the information existence value; on the other hand, users could not do anything without information. It was meaning that information and users relied on each other. Thus, we can determine that during the process of users' information behaviors, users could absorb certain information and create information related to it. Information was given new life during the process. In general, the user's information behavior was determined by a variety of characteristics of the user, such as the user's knowledge structure, behavior, etc. In this paper, we focused on the need for Improve the Agridata website efficiency with the method of analysis agricultural network user's information behavior. So characteristics we chosen from the web log data mining in this study were as follows: click-through of a page, users' traffic, the view number of pages, traveling time of users, access duration, network traffic, the visit number of a single page, etc.

### *2.3.2 User information behavior analysis*

According to the characteristics of users' information behavior, we could analyze the information behavior of a user to grasp the user's psychology, which was a great significant for improving services effect of website. As usual, this analysis included two methods: statistical analysis and classification clustering.

(1) Statistical analysis

This was a most commonly method for analyzing user access patterns, which studied how to measure, observe, express and summary the number characteristics of objection. In this passage, we could compute parameters such as mean, maximum, minimum, etc, with a variety of information of users who visited Agridata website to estimate the behavior information pattern of users. To some characteristics that were not easy to see from data, we used graphics-assisted method for analysis, such as Scatter plots, histograms, line charts, box diagrams and so on, to show trend of changes intuitively. In the actual analysis process of this passage, we computed the distribution of site visits and page views changed with the length of visit. These results of statistical analysis could help to determine rationality of Agridata topology map and convenience of users' access to resources, which could useful for improving structure of the website.

(2)Cluster analysis

This was an important method for data partitioning or packet processing. Currently, the cluster analysis algorithm could be divided into the method, the level of methods, density-based methods, grid-based methods and model-based methods and so on. As there were large amounts of data resources in Agridata and the purpose of users was not clear (the meaning of users could not be accurately expressed with mathematical expressions) when they searched data in Agridata, in this passage, we applied fuzzy clustering analysis method for analyzing users' information behaviors. This method combined with fuzzy set theory and the traditional clustering method. The principle of this approach was as follows: first, the given objection was divided into a number of equivalence classes with the way of fuzzy equivalence relation, which was meaning that establishing fuzzy similar matrix to express similarity between each sample. Then, we clustered with these equivalence classes directly. During the process of specific implementations, we clustered pages of the same database users visited and regarded these pages as a whole for analysis. To these pages, we could compute some parameters such as click-through, access duration and navigation path, for analysis and displaying the users' information behaviors totally. Furthermore, this method for analysis could one-sidedness caused by reference to a single characteristic value

Based on the above analysis, in the cluster analysis, we let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of characteristic parameters which included

values of click-through, access duration and navigation path, etc. The set of pages of one database could be expressed as $W = \{W_1, W_2, \ldots, W_k\}$. The set of users could be expressed as $U = \{U_1, U_2, \ldots, U_m\}$. To any user in the set of U, that was $\forall U_i \in U(i = 1,2, \ldots, m)$, the process of the user visited pages could be expressed as $U_{iWj} = \{y_{i1}, y_{i2}, \ldots, y_{in}\}$, where $y_{ib}(b = 1,2, \ldots, n)$ represented the value of characteristic in the set of $X$. After normalizing the elements of $U_{iWj}$, we could get a multiple fuzzy set of view features for any user.

$$K = \{\{y_{i11}, y_{i12}, \ldots, y_{i1n}\}, \{y_{i21}, y_{i22}, \ldots, y_{i2n}\}, \ldots, \{y_{ik1}, y_{ik2}, \ldots, y_{ikn}\}\}$$

As a result, the interest value of a user $U_i$ to a set of pages could be expressed as (1).

$$\theta_{ij} = \frac{\sum_{t=1}^{k} y_{itj}}{\sum_{p=1}^{k} \sqrt{\sum_{t=1}^{n} y^2_{ipt}}} \quad (1)$$

Where $j = 1,2, \ldots, n$, $\theta_{ij}$ represented the interest value of a user $U_i$ to a set of pages $W$ for the characteristic $X_j$

## 3 Experiments and Analysis

In the experiment of this passage, we chose the website of Agridata as the representative of agricultural sites and analyzed information behaviors of its users. As thinking about the life and production situation of agriculture network users, in this passage, we chose web log files in July 2010 as the objection of experiments. Such sample we chosen could not only reflect needs of agriculture users during the busy farming season, but also avoid deviations of information behaviors in the general sense caused by various holidays.

In this passage, we used the analysis method combining with statistical analysis and cluster analysis. The flow chart of analysis was shown in figure 2.

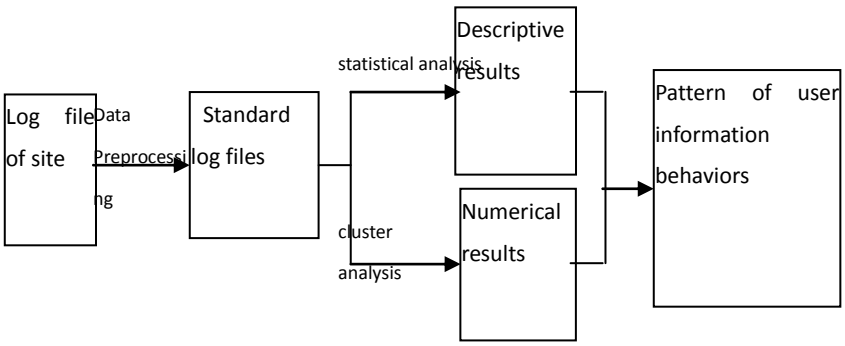


**Figure 2** . The flow chart of analysis

During the process of analysis, after processing 40873 records in the log file chosen as the objection of the experiment, we computed some parameters such as time distribution of access site, access duration, the amount of download data and the concentration of IP address, etc, to provide the basis for analysis information behaviors of agriculture users. These parameters also helped us to determine the future direction of development and related factors of agriculture users' information behaviors. In this passage, we analyzed the chosen log files with WebLog Expert, a common log analysis tools. Some results gotten from the experiment were shown in figure 3.

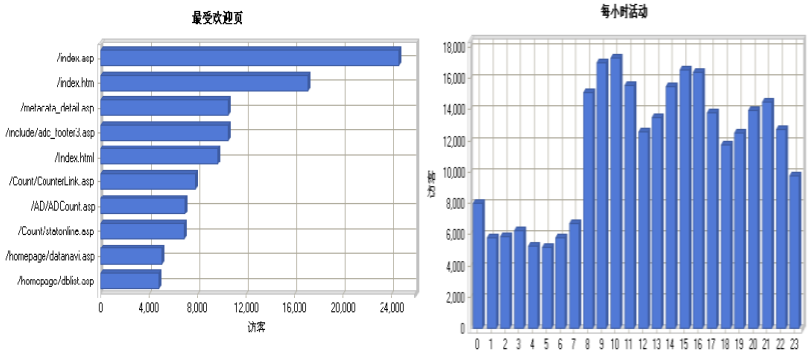


**Figure3.** Results from analysis with WebLog Expert

In this passage, to simplify the analysis, we chose 7 databases from 62 databases of Agridata as objections of cluster analysis during the process of cluster analysis. These 7 databases represented text-based database, numerical databases and image databases. Details of these databases were shown in table 2.

**Table 2**. Details of chosen databases

| Number | Name of the database | Number of recorders | Data type |
|---|---|---|---|
| 1 | agricultural insect in china database | 788 | Image |
| 2 | natural enemies spider database | 338 | image |
| 3 | Parameter of organic fertilizer database | 109 | data |
| 4 | Fertilizer parameters database | 145 | data |
| 5 | quality standard of rice database | 53 | text |
| 6 | quality standard of tea database | 55 | text |
| 7 | quality standard of green food database | 72 | text |

Records of these 7 databases which users visited from the website in log files should be extracted and clustered based on databases for the analysis. To simplify computing, we chose page residence time of users as a characteristic for analyzing. After clustered this feature, we computed the interest value for the database we

chosen as an objection for the experiment with these parameters. In the actual calculation, the interest value could be computed with formula (2) which derived from formula (1).

$$\theta_j = \frac{F_j}{T}, \quad (2)$$

Where $j = 1,2,3...,7$, $F_j$ represented the time a user stayed in the page set of database j; $T$ represented the total time a users used when he visited the website of Agridata.

The overall interest value of these seven databases could be gotten from the mean of all users' interest values for these seven databases, which were shown in table 3.

**Table 3**. the overall interest value of databases

| number | Name of database | the overall interest value |
|---|---|---|
| 1 | agricultural insect in china database | 0.51 |
| 2 | natural enemies spider database | 0.49 |
| 3 | Parameter of organic fertilizer database | 0.29 |
| 4 | Fertilizer parameters database | 0.35 |
| 5 | quality standard of rice database | 0.48 |
| 6 | quality standard of tea database | 0.35 |
| 7 | quality standard of green food database | 0.33 |

From analysis results above, we could infer information behavior characteristics of users who visited the website of Agridata.

(1) From data of table 3, we could infer that the interest value of images was the highest, which were 28% higher than the interest value for text database and 56% higher than the interest value for data database. It fully confirmed that acceptance of users for image was higher than that of text and data for data. After further studying data in table 3 and results of the statistical analysis, we could find that IP addresses of users who access image databases were many and scattered, the one for text databases was second and IP addresses for data databases were the most concentrated. It could mean that agricultural workers and farmers preferred image data while numerical data were often used by agricultural researchers.

(2) From results of the statistical analysis, we could determine that the access time period of users was very scattered, which could mean that users could visit the

website at any time of a day. In general, in the day time (about 8:00 to 18:00), click-through of the website was maximum, around 74% of all click-through and IP addresses of visited were more concentrated. So we could infer that agricultural researchers in institutions were more willing to use Agridata than that of farmers.

(3) From results of statistical analysis for main page and search pages residence time of users, we determined that IP addresses of users whose pages residence time was shorter had displayed more concentrated while longer pages residence time of users whose IP addresses were more scattered. Thus, we could infer that farmers were much weaker for adapting to main page and search pages than agricultural researchers did.

From results of the analysis in this passage, the website of Agridata should be improved in three aspects. Firstly, we should strengthen publicity efforts to farmers for Agridata; secondly, the main page and search pages of Agridata should be simplified to make them easier to use; finally, the number of image databases should be increased to meet the need for farmers. If more farmers visited the website of Agridata for life and production, we could sure that Agridata should provide better services for farmers and promote agricultural production and research.

## 4 Conclusions

In this passage, the Web access log was used as an object and we discussed the method for studying the information behavior of users of agricultural methods with statistical analysis and cluster analysis. Then we conducted experiments with the Web access log of the national agricultural scientific data center and analyzed the pattern of information behavior of its user. With these results, several advices on this website were made for providing a better service to agricultural users and promoting agricultural production and research.

**Acknowledgements:** Funding for this research was provided by “basic scientific research special fund of nonprofit research institutions at the central level” (2011j-1-06)

## References


1. DING Jing-da. Analysis and Research on the Web-journal of University Library [J]. Sci-Tech Information Development & Economy,, 2006,16(6):5-6.
2. FENG Chun-hui. Research on Application of Web Log Mining in Network Teaching [J]. Computer Technology and Development, 2010,20(6):183-187.
3. Li Junjun and Sun Jianjun. Empirical Study on Website Quality,User Perception and Technology Adoption Behavior [J]. Journal of the China Society for Scientific and Technical Information，2011，28(3):227-236.
4. LIANG Xiao-xue and WANG Feng. A survey and prospect of log analysis based on clustering [J]. Journal of Yunnan University(Natural Sciences Edition),2009,31 (S1): 52-55.
5. Qian Peng. Research of Library Web- based Resources Log Analysis [J]. The Journal of the Library Science in Jiangxi,2004,31(1):30-31.
6. Ping tu. The application of Server Log Analysis Method in Website Improvement [J]. Journal of Jiujiang University(Natural Science Edition),2010,(4):31-33.
7. Chang-pin Weng. A Web Log Based University Library Users Information Behavior _ Taking the AnHui University Library as the Example [D].Anhui: AnHui University ,2010.
8. Zhang Xuehong. Log File Analysis of Peking University Library Homepage [J]. New Technology of Library and Information Service,2005,(5):81-83.
9. RONG TANG, PAUL SOLOMON. TOWARD AN UNDERSTANDING OF THE DYNAMICS OF RELEVANCE JUDGMENT: AN ANALYSIS OF ONE PERSON'S SEARCH BEHAVIOR[J]. Information Processing & Management, Vol. 34, No. 2/3, pp. 237 256, 1998.