A Study About Searching Behavior of Scientific Data User Based on Educational Background and Retrieval Capability

Guilan Zhang, Jian Wang, Guomin Zhou, Jianping Liu, Fei Gao, Caoyuan Wei

A study about searching behavior of scientific data user based on educational background and retrieval capability

Guilan Zhang (0000-0002-9153-3579), Jian Wang(✉)(0000-0003-4958-7669), Guomin Zhou, Jianping Liu (0000-0002-1817-5373), Fei Gao (0000-0002-7844-3464), Caoyuan Wei (0000-0001-5259-4187)

Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing, China
{1559943406@qq.com; wangjian01@caas.cn; zhouguomin@caas.cn; 1592126095@qq.com; 504668730@qq.com; 330823981@qq.com}

Abstract: As a new information carrier, the scientific data carries a lot of information. People who work in scientific research are more or less likely to search scientific data for work. We investigate the retrieval behavior of researchers in the process of searching data through questionnaires. The survey included the selection of searching data channels, using the clues and criterion in the process of retrieving data and some difficulties in the retrieval process. The results show that the retrieval behavior of scientific data users is influenced by educational background and retrieval ability. The education determines the breadth and depth of the data requirements, and the retrieval ability determines the access to data and the amount of data to use. So, in the process of designing data retrieval system, it is necessary to meet the needs of users in different cognitive contexts.

Keyword: scientific data, retrieval behavior, educational background, retrieval capability

1. Introduction

Scientific data plays an extremely important role in scientific research, more and more researchers rely on scientific data for scientific research. The sharing and service of data is not only the basic requirement of economic and social innovation, but also the inevitable product of information technology application. In the survey, Peng xiuyuan and others found that most researchers shared their scientific data, and they have the willingness and need to share data. What's more, they also want to share data through professional scientific data sharing platforms. Therefore how to better improve the scientific data sharing platform, to provide a high quality service for scientific research staff has become an urgent problem to be solved.

Information query is the process of discovering problems, solving problems, seeking information actively, meeting needs and reducing uncertainty. In the long-term retrieval activities, users will form a relatively stable information query route that suits their needs. Similar working environment and learning environment are similar to the group of the same type, whose information needs and information query behavior have similar characteristics. Information query behavior of different groups can be significantly different due to some factors such as personal cognition and environmental impact. Therefore it will appear the relative stable query behavior which is different from the cognitive background and external environment when the scientific data user makes the data query.

This study will focus on the students' behavior in the process of search and use of scientific data , mainly including the selection of searching data channels, using the clues and criterion in the process of retrieving data and some difficulties in the retrieval process. It is expected to find the different types of scientific data users' behavior characteristics and their differences when they complete specific tasks. So it provides theoretical basis for improving the efficiency and quality of data retrieval and contribution in order to meet the needs of different types of people.

In the rest of the article is organized as follow: in the section 2 we present previous related work on users' search behavior, details on methodology are explained in section 3. Results and analysis of the study are discussed in section 4 and 5. Finally ,we conclude in section 6.

2. Literature review

The goal of information query is seeking information, achieving happiness and

reducing internal unhappiness. Wilson believes that personality, relationships and environment are the main factors that influence information query behavior. Hu changping believes that in the process of acquiring information, the factors that influence user behavior mainly include subjective information consciousness and objective work needs. Many scholars also study the information retrieval behaviors of different types of users. Tang made a contrast research about information selection criteria used in empirical study and naturalism. The results show that criteria model used by two groups is obvious difference, and with the deepening of the task, the use of criteria also changed obviously. With the booming of the Internet, Li Fayun researched on information behavior of Internet users, and the result showed retrieval experience had some influences on retrieval behavior. Wu Jing and Li Shanshan believed the information behavior of university users is a hot topic, so the research on information retrieval behavior about students and teachers in digital environment was carried out. Results showed that it was the user's personal and external environment factors which affected the user's retrieval behavior, and the personal influence factors mainly included user background knowledge, computer application ability, knowledge structure and cognitive ability. Zhang Min and other scholars found that there was a distinct different retrieval behavior of different users, and the gender also had a significant influence on their behavior patterns. Thatcher believed that system experience and network experience had a positive impact on information retrieval, and later research also confirmed this conclusion. Dong Yuchao found that the information searching behaviors of scientific researchers with different ages, different education levels and different professional titles were significantly different. Wang Jiandong and others have shown that a user's degree has a significant impact on academic search in the journal database.

It can be seen that user information searching behavior is closely related to cognitive level, and the cognition level always reflected by educational background and retrieval ability. Scholars made a lot of researches on network information behavior and library information behavior, but there is no research on the behavior of scientific data. Therefore, we will study the impact of educational background and retrieval ability on data searching behavior in the scientific data environment.

3. Methods

3.1 Method of data collection

We designed the questionnaire in the web site, and the amount of information included the user's degree, the number of publishing papers, and the frequency of using data, searching way and its difficulties when users met a large amount of information. What’s more, we also survey the clues and criteria which users pay attention to in the retrieval process. The questionnaire was published on the official platform, and the questionnaires were completed on a computer or mobile phone. Each questionnaire is paid to ensure the quality. The system takes back automatically.

We received 671 questionnaires totally. We rejected invalid questionnaires depending on the time that filled in the questionnaire. It will take 8-10 minutes to finish a questionnaire for the designer, so we rejected those questionnaires that the using time less than 500 seconds or were repeated. The remaining valid questionnaires were 544, and the effective rate was 81.07%.

3.2 Participants

This experiment is mainly aimed at the graduate students (master and doctor) who participated in the "sharing cup" competition. There are 22 undergraduates, 421 master's degree students, 99 doctoral students and 2 postdoctoral fellows. As the proportion of master's and doctoral students reached 95.59 %, this study mainly analyzed graduate students. The percentage of subjects who regularly used scientific data in their work and study (50% of the work and study) was 68%; the percentage of subjects who sometimes used scientific data in their work and study (20%-50% of the work and study) was 23.9%. Therefore, the representative is strong and has certain reference value.

Statistical analysis was carried out in SPSS. We mainly used the independent t-test, with $P < 0.05$ as the standard to test the significant difference.

3.3 Data pre-processing

First of all, the data was preliminarily sorted out. According to the processing method when the data was too large, three groups of people were classified, including high retrieval ability, middle retrieval ability and low retrieval ability. When the retrieval of

information is too large, the people who mainly adjust the optimized retrieval words, transform the retrieval words and replace the retrieval tool were considered to be high retrieval ability; the people who just look at the first few items of the search results, pick a few rough views and browse through them were considered to be low retrieval ability; the people of middle retrieval ability fall in between. The detailed distribution is shown in table 1. The percentage of people with high retrieval ability and middle retrieval ability was 96.5%, so in the following analysis, the two groups were mainly analyzed. While the number of people with low retrieval ability was too small, and the data analysis was not done.

Table 1: Distribution table of different retrieval ability groups

	frequency	percentage	Cumulative percentage
high retrieval ability	333	61.2%	61.2%
middle retrieval ability	192	35.3%	96.5%
low retrieval ability	19	3.5%	100%
total	544	100%	

4. Results

4.1 Ways to get scientific data

The main way for users to get data is through a professional database, which accounts for 31.04% of the total information channel, followed by Baidu, Google and other search engines, published papers, experimental research, interpersonal network, and the last way is library. The proportion is 22.35%, 16.06%, 14.32%, 7.9% and 4.85% respectively. Thus, it is shown that professional databases, large search engines and other people's literature are the three main ways for scientific researchers to obtain scientific data.

We analyzed the different degree and different retrieval ability groups respectively, and the results show that no matter what kind of people, the proportion of using professional database is relatively highest, followed by Baidu, Google and other search engines. But there are some differences in the specific values. Therefore, the data about the information channel is compared and analyzed between two groups with different educational backgrounds and retrieval abilities to find out the differences of different groups.

For users of different educational backgrounds, independent t-test is conducted for master and doctoral students. The test results are as follows: Professional database $P=0.041<0.05$; Library $P=0.001<0.05$; Experimental research $P=0.028<0.05$. Therefore, there are obvious differences among the three kinds of data obtain channels. Masters are more likely to use professional databases and libraries, while PHDS are more likely to take data from their own experiments.

For users of different retrieval ability, independent t-test is conducted for high retrieval ability and middle retrieval ability. The test results are as follows: Professional database $P=0.001<0.05$; Library $P=0.008<0.05$; Published papers $P=0.018<0.05$; Other $P = 0.000 < 0.05$. Therefore, there are obvious differences among the four kinds of data obtain channel. The users of high retrieval ability are more likely to use professional databases, while the users of middle retrieval ability are more likely to take data from library and published paper.

Thus, the professional database is the most frequently used and reliable channel for users. Users can obtain the most professional and authoritative data through the professional database. When retrieval ability is limited, users are more dependent on libraries and published papers to obtain data. With the improvement of educational background and in-depth research, users will design their own experiment to obtain data more than professional search.

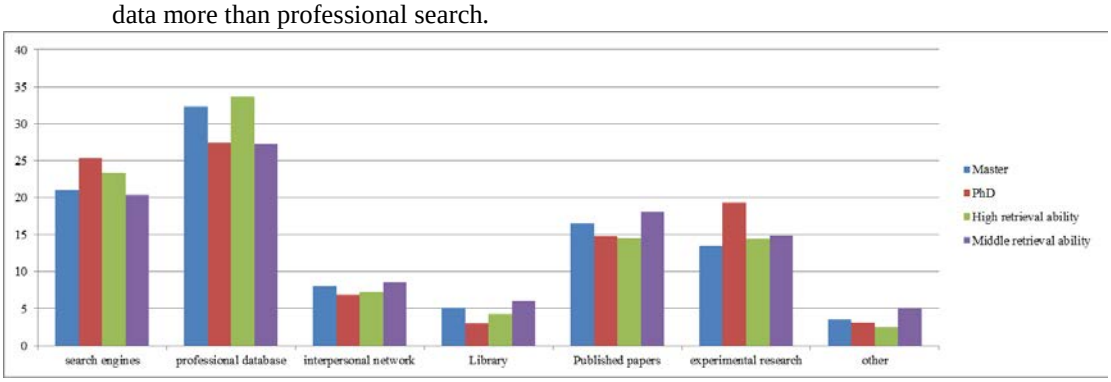

Figure 1：Bar charts of data sources in different groups

Table 2: T-test of data sources

T-test P-Value	search engines	professional database	interpersonal network	Library	Published papers	experimental research	other
Education	0.089	0.041	0.266	0.001	0.319	0.028	0.49
Retrieval ability	0.086	0.001	0.128	0.008	0.018	0.785	0

4.2 Clues to attention

The clue is that the user is visually stimulated by the external information during the retrieval of data. Through investigating the clues, we can better understand the focus on users' attention and make data relevant judgement during the retrieval, so as to improve data sharing platform. The user scored 19 clues in the questionnaire based on the importance. 0 is the lowest score, and 5 are the highest score. Through statistical analysis, it can be seen that the most important clues are whether to support the download, with an average score of 3.88. In the top five, the rest four clues are keywords, data quality, data abstract and data coverage time. The average score is 3.76, 3.63, 3.58 and 3.52. The least important clues are data size, whose m-value is 2.4. Thus, the most important thing for the user is that the data can be downloaded and used. If the data is not available, no matter how relevant it is useless. Keywords and abstracts can give users a preliminary understanding of data, and relevant quality instructions can guarantee the quality of data. Those all can affect users' retrieval and judgment. In the end, we found that users were least focused on the size of the data.

Then we repectively extracted the first five clues between master and PhD. It can be seen whether it can be downloaded and the keywords are the first and second, and obviously the doctor's score is higher than that of the master. The following three clues show a certain difference, which can be seen that the importance of clues in different educational background is different. Therefore, we conducted independent t-test for the importance of the clues to different educational background and different retrieval abilities. The significant difference was listed in the table. For all the clues, PhD scored higher than those of the master, and those with high retrieval ability scored higher than those in the middle. So the more cognitive level the user has, the higher the retrieval ability, the more important the clues are.

Table 2: Comparison table of top five clues scored by different education

	First	Second	Third	Fourth	Fifth
Master	Download（3.9）	Keyword （3.72）	Data quality （3.61）	Abstract （3.53）	Coverage time （3.52）
PhD	Download（4.08）	Keyword （4）	Abstract （3.88）	Data quality（3.8）	Title （3.75）

Table 3：Independent t-test about the clues

clue	Different education P-Value	Different retrieval ability P-Value
Title	0.007	0.040
Data author	0.008	
Keyword	0.033	0.000
Abstract	0.015	0.006
Format	0.021	
Full text	0.002	
Organization	0.005	
Data orders	0.014	
Download		0.002
Data sharing level		0.048

4.3 Selection of relevance criteria

The relevant criterion is the reason or factor that affects the user's relevant judgment[1].We can see from the scores made by uses depending on the importance, the most important criteria is quality(4.18),next is topical(4.12) and authority(4.06). The three lowest ratings are novelty (3.49), understandability (3.48) and convenience (3.48).First of all, the difference value between the top and bottoms was only 0.7. What’s more, the mean value of criterion all beyond 3. So we can see that all relevant

criteria play an important role in the relevant judgement process. Secondly, more than 50% of users scored five points on topic and quality, who think topic and quality are very important. Thus, it is generally accepted that both of them play an important role in the process of relevant judgment.

At the same time, we compared and analyzed the scores made by users of different education and retrieval abilities, and got table 5 through independent t-test. We can see from the M-value that the PhD scored higher than those of the master, and those with high retrieval ability scored higher than those in the middle. The results showed that there were significant differences in in the importance of topicality, accessibility, novelty, currency and quality between the doctor and master. There are significant differences in the importance of all criterion between different retrieval abilities, except accessibility, availability, and comprehensive.

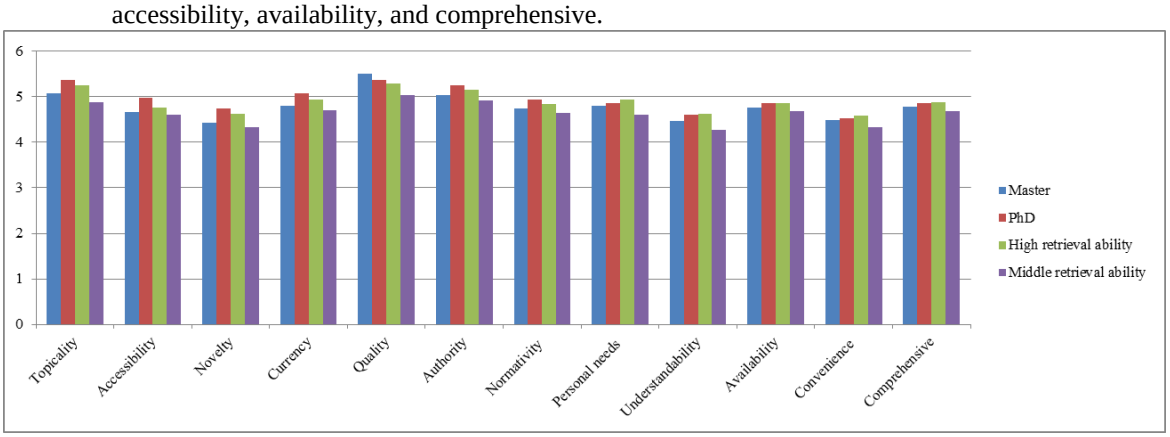

Figure 2：Bar charts of relevant criteria in different groups

Table 5: T-test of relevant criteria

T-test P-Value	Education	Retrieval ability
Topicality	*0.016*	*0*
Accessibility	*0.015*	0.136
Novelty	*0.024*	*0.005*
Currency	*0.048*	*0.026*
Quality	*0.04*	*0.004*
Authority	0.053	*0.022*
Normativity	0.092	*0.045*
Personal needs	0.706	*0.004*
Understandability	0.266	*0.001*
Availability	0.46	0.073
Convenience	0.764	*0.032*
Comprehensive	0.609	0.201

4.4 Difficulties encountered during retrieval

First of all, the biggest difficulty is that the results are too large to choose properly during retrieving data (16.57%). There are two reasons for this: On the one hand, the user has cognitive limitations, for example, users can't use the search term properly, can't understand the topic clearly, or can't use the relevant criterion comprehensively. On the other hand, the data sharing platform is not perfect. Users also meet the following difficulties, that there is no uniform data management platform（14.03%）, the data quality is not high(14.03%), the authors are unwilling to provide raw data(9.93%),no permission(9.01%).These four difficulties are all from the outside world. Since China's data sharing is still in its infancy, many experts and scholars are reluctant to share their data, resulting in a series of data usage difficulties.

Different people have different difficulties in retrieving data. For people with different degrees of education, the significant difference is that the data quality is not high (P=0.013) and the searching tool is not strong (P=0.007). The master will have more data quality problems, and the doctors are more likely to face the problem of searching tools. For people with different retrieval abilities, the significant difference is

that the lack of explicit copyright (P=0.015), the lack of access (P=0.021) and the lack of retrieval skills (P=0.008).What's more, people with middle retrieval ability are more likely to face these three kinds of difficulties than those with high retrieval ability. It can be seen that people with lower retrieval ability or lower cognitive level are more vulnerable to internal cognitive limitations, thus influencing their own retrieval. However people with higher education and high retrieval ability are more likely to face some external factors, such as lack of unified data management platform, weak retrieval tools, and lack of access.

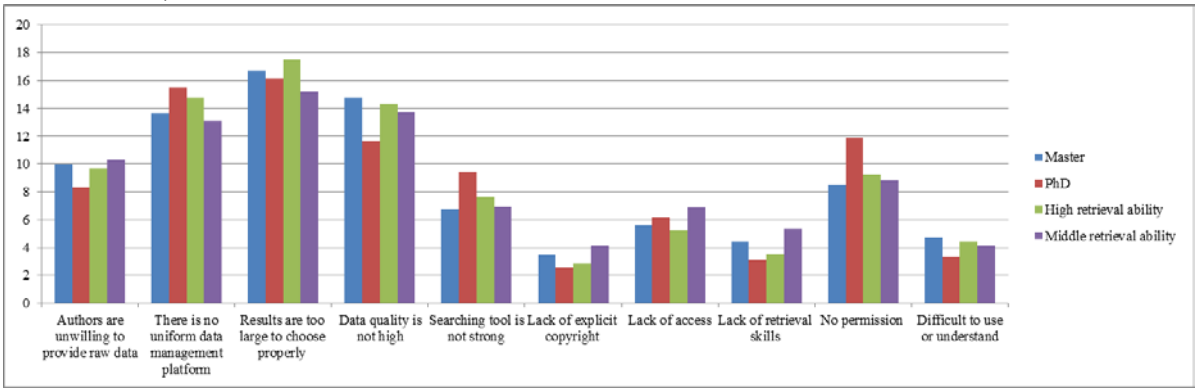

Figure 3：Bar charts of different groups about difficulties meeting in the retrieval process

Table 4: T-test of difficulties meeting in the retrieval process

T-test P-Value	Authors are unwilling to provide raw data	There is no uniform data management platform	Results are too large to choose properly	Data quality is not high	Searching tool is not strong	Lack of explicit copyright	Lack of access	Lack of retrieval skills	No permission	Difficult to use or understand
Education	0.233	0.344	0.759	*0.013*	*0.007*	0.135	0.542	0.098	0.072	0.1
Retrieval ability	0.576	0.179	0.06	0.637	0.349	*0.015*	*0.021*	*0.008*	0.699	0.665

5. Discussion

In the study of information behavior, scholars generally adopt methods of experiment, investigation (interviews and questionnaires) and log mining. With the development of technology and the increasing amount of data, the most popular method is log mining. Scholars analyze user's information behavior through a large number of user browsing records. It is because of the development of the Internet that the research method of information behavior has made a substantial breakthrough. Log mining makes data more objective and accurate, but the biggest drawback of this approach is that searching behavior cannot be associated with a particular users. It makes no sense to study the information behavior if the behavior is not connected to the users. In addition, considering the research cost and time, the study also did not adopt the method of situational experiment and interview. Finally, questionnaires were used to survey graduate students participating in the "sharing cup" competition. This method is not limited by time and region so as to cover research objects as thoroughly as possible, and can obtain data for quantitative analysis. But the flaw of this approach is obvious. Even if we work with officials to send links in official names and give back certain rewards, the quality of the data is guaranteed to some extent. However, the quality of the subjects and the accuracy of the data cannot be fully guaranteed in the network.

Through the analysis of the above four aspects, we will discuss the following four points.

(1)In the age of big data of Internet +, users have changed from negative information receivers to active participants with selective ability to meet their own data needs through all kinds of choices. The user's cognitive level influences the choice of information source. PhDs are more likely to rely on data from their own experiments, and master degrees are more likely to rely on others' data. In 1986, Isenberg found that experienced users were less likely to use available information than novices, and made more targeted conclusions based on a small amount of supporting information. Thatcher also confirmed that systematic experience and network experience had a positive impact on the retrieval success. It can be seen that educational background and

retrieval ability affect the user's choice of information. The education determines the breadth and depth of the data requirements, and the retrieval ability determines the access to data and the amount of data to use.

(2)The lack of information is the main cause of "information poverty" for people. So the ultimate goal of user retrieval is to get relevant data and make up for "information poverty". In the retrieval of scientific process, the criteria of judging data is mainly topicality, quality and authority, but the most important clue is whether to download. It can be seen that whether can be downloaded and acquired is more important than whether it is relevant. Data is useless if it is not available. As Savolainen mentioned in the study, unavailable let the user reject the information directly.

(3)The core criteria for selecting information in long memory are very different from the core criteria for judging information in a real environment. This study is aimed at the user relevant criteria in long-term memory only to find no difference in importance. The user thinks that each independent criteria is important to the relevant judgment. But in the context of the study of relevance criteria we will find that some of these criteria are used frequently, for example topicality, quality, and personal need（Taylor,2012;Savolainen,2006）.There are some criteria that are barely used, like authority（Abe Crystal）,normativity（Anastasias Tombros）and so on. It seems that some criteria are not used frequently , but it also very important. Frequency and importance are not positively correlated.

(4)There are many problems in the sharing of scientific data in China, which is an important factor influencing users' retrieval of scientific data. In the process of retrieving data, scientific data users often meet all kinds of problems, like authors are not willing to provide the data, or users have no access to the data. The reason is that data sharing work is not perfect and data is more in the hand of the experimenter. These problems have not only resulted in the duplication of work among the staff of different interest groups, but also resulted in the waste of resources and research funds.

6. Conclusion

With the arrival of the data era, scientific data is deeply embedded in scientific research in various fields. People who work in scientific research are more or less likely to search scientific data for work. So far, China's work on scientific data sharing has a lot of problems both from the technical level and from the government control level. In this study, we investigate the different behavioral characteristics of different types of users from the user's perspective, and analyze their commonality and specificity, so as to find the users' behavior rule of retrieving data. So this study provides a powerful theoretical basis for improving our scientific data search engine. Through this study, we better understand the development goals of scientific data retrieval. We need to provide different ways of retrieval for users who have different cognitive backgrounds.

Acknowledgments: This work was supported by a grant from the Social science fund-Scientific Data User Relevance Criteria and Use Model Empirical Study (14BTQ056), National High-tech R&D Program of China (863 Program No.2013AA102405) and Agricultural Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences(Project No.CAAS-ASTIP-2016-AII).

References

1. Wang Yanfei,Chen Meihua,Zhao Keran, el. Information Analysis on Scientific Data Sharing[J]. Journal of Intelligence, 01, 29-34 (2017).
2. Peng Xiuyuan, Wang Feng, Zhou Guomin. Investigation and analysis of agricultural scientific data sharing in Liao Ning province. Agricultural Economy, 01,59-61 (2017)
3. Qiao Huan. Information Behavior. Beijing Normal University Press, (2010).
4. Mcquail D, Windahl S. Communication models : for the study of mass communications. Communication models for the study of mass communication. Shanghai Translation Publishing House, 345-367 (1993).
5. Wilson T D. Models in information behaviour research. Journal of Documentation, 55(3), 249-270 (1999).
6. Hu Changping. Information Service and User Study. Wuhan University Press, (2008).
7. Tang R, Solomon P. Use of relevance criteria across stages of document evaluation: On the

complementarity of experimental and naturalistic studies. Journal of the Association for Information Science and Technology, 52(8):676–685 (2001).
8. Li Yunfa. Research on Internet user information retrieval behavior. Journal of Library Science in China, 2003, 29(2):64-67.
9. Wu Jing,Li Shanshan. Information retrieval behavior of university library users under digital environment. Journal of Architectural Education in Institutions of Higher Learning, 22(3), 139-141 (2013).
10. Zhang Min,Niu Rui and Luo Meifen. Analysis on the Characteristics of Network Health Information Retrieval Based on Demand Type and Gender Difference. Information and Documentation Services, 2, (2017).
11. Thatcher A. Web search strategies: The influence of Web experience and task type. Information Processing & Management, 44(3), 1308-1329 (2008).
12. Chen Yin, Sun Jianjun and Zheng Yanning. On Information Behavior of Academic Novice. Documentation,Information & Knowledge, (4), 63-70 (2011).
13. Dong Yuchao,Hu Dehua. Characteristics and differences of information searching behaviors in scientific researchers engaged in different scientific projects. Chinese Journal of Medical Library and Information Science, 23(7), 37-41 (2014).
14. Wang Jiandong,Wang Jimin. Study on Journal Database Retrieve Behavior of University Users Based on Log Mining. Acta Scientiarum Naturalium Universitatis Pekinensis, 48(1), 29-36 (2012).
15. Schamber L. Users' Criteria for Evaluation in a Multimedia Environment. Proceedings of the ASIS Annual Meeting, 28, 126-33 (1991)
16. Song Aibo, Hu Kongfa and Dong Yisheng. Research on Weblog mining. Journal of Southeast University (Natural Science Edition), 32(1), 15-18 (2002).
17. Wang Ruojia,Li Pei. A Study on Health Information Search Behavior Based on Log Mining. Library and Information Service, (11), 111-118 (2015)
18. Isenberg D J. Thinking and Managing: A Verbal Protocol Analysis of Managerial Problem Solving. Academy of Management Journal, ,29(4), 775-788 (1986).
19. Savolainen R, Kari J. User-defined relevance criteria in web searching. Journal of Documentation, 62(6), 685-707 (2006).
20. Taylor A. User relevance criteria choices and the information search process. Information Processing & Management, 48(1), 136-153 (2012).
21. Crystal A, Greenberg J. Relevance criteria identified by health information users during Web searches. Journal of the American Society for Information Science and Technology, 57(10), 1368-1382 (2006).
22. Crystal A, Greenberg J. Relevance criteria identified by health information users during Web searches. Journal of the American Society for Information Science and Technology, 57(10), 1368-1382 (2006).