Analyze user behavior patterns on academic search engines

Somayeh Fatahi [1], Amir Hossein Seddighi [2], Mohammad Rabiei [3]

[1] Assistant professor of Information Systems Research Group, Information Technology Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran,
E-mail: fatahi@irandoc.ac.ir

[2] Assistant professor of Information Systems Research Group, Information Technology Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran,
E-mail: Seddighi@irandoc.ac.ir

[3] Assistant professor of Electronic Business Research Group, Information Technology Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran,
E-mail: m.rabiei@irandoc.ac.ir

ABSTRACT

Nowadays, academic search engines have grown rapidly. Thus, understanding the users' information-seeking patterns has become one of the most important research topics. That is why by examining user interaction logs, developers can discover user behavior patterns to determine who they are and what they tend to do. Consequently, they can get guidance to design better academic search engines and improve their performance. In this paper, we analyze search engine users' logs gathered from the search engine of the Iran scientific information database (Ganj). The users are clustered into three distinct groups: fast surfing, broad scanning, and deep-diving, using the K-means clustering algorithm. After that, we investigate the frequent sequences of behavior patterns and the networks of search keywords for each cluster separately. The results show that users with similar information-seeking patterns have similar sequences of behavior patterns. The findings can help the developers of academic search engines and policymakers to identify users' needs and priorities and make better decisions.

Keywords
academic search engines, log mining, clustering algorithm, behavior analysis, information-seeking patterns

1. INTRODUCTION

Today, information is everywhere and plays an essential role in the quality of human life. Also, it is a critical asset for daily activities, and all decisions are influenced by Information (Pareek & Rana, 2013). Wiberley and Jones (1989) believe that information seeking is the primary activity of individuals. In recent years, technological advances of the Internet and search engines have helped people seek and find their information needs quickly and efficiently. A *search engine* is a service that allows users to search their keywords or key phrases and retrieve information. 1.7 billion people use Google as one of the famous search engines, and 7 billion searches are handled by it every day (ardorseo, 2020). Furthermore, academic search engines are no exception, and they are growing.
Due to the rapid growth of search engine including academic types or other, the users of these services are increasing. The first and foremost step in improving search services is to realize the users' needs based on analyzing their navigation behavior (Xie et al., 2012). Many studies indicate that individuals have different interests, needs, preferences, and navigation behavior in seeking information (Herder & Juvina, 2004; Kaspar & Müller-Jensen, 2019; Kim et al., 2014; Triantafillou et al., 2004). Therefore, analyzing user's navigation behavior would be helpful to cluster users based on their preferences in search, and it can be helpful to a personalized search engine.
To analyzing user's navigation behavior, search engine logs have to examine. Each search engine records a considerable amount of search logs containing the interactions between user and search engine. Search engine logs have valuable information such as submitted queries, viewing search results, and clicked URLs. Clicked URLs or clickstream data provide information about a user's goals, knowledge, and interests (Montgomery et al., 2004). It gives the website admins a better understanding of what users are doing on their sites (Hanamanthrao & Thejaswini, 2017). Generally, clickstream is a sequence of activities taken by users while visiting a website, represented by the sequence of links they click on (Filvà et al., 2019). The information hidden in clickstream can be used for different purposes, such as marketing (Koehn et al., 2020), e-commerce (Kumar et al., 2019), research management (Rabiei et al., 2017), and e-learning (Filvà et al., 2019;

[1] fatahi@irandoc.ac.ir

Hanamanthrao & Thejaswini, 2017). Eventually, search engine logs and clickstream data can help to discover user behavior patterns to determine who they are and what they tend to do.

As mentioned before, examining users seeking information behavior on an academic search engine is an important issue because the researchers, students, and faculty members widely use academic search engines as a vital source to obtain academic information. Therefore, in this paper, three critical contributions are made:

1) Extracting information-seeking behavior features on an academic search engine
2) Clustering the users based on information-seeking features
3) Investigate sequential patterns, and network of keywords search in each cluster and match it with information-seeking patterns

The remainder of this paper is structured as follows. Section 2 explains the related works and background. Section 3 describes a personalized search engine and information-seeking behaviors. The research design is introduced in section 4, and section 5 presents the experiments and results. Finally, discussion and conclusions are drawn in Section 6 and 7, respectively.

2. Related Works

Several studies have been conducted using pattern-mining methods for analyzing users' behavior patterns. In this section, we review the most important ones. Ting et al. (2007) propose a new Automatic Pattern Discovery method (APD), which automatically identifies a user's browsing patterns based on clickstream data. The goal of this research is sequence mining to pre-identify patterns. The results show that this approach is suitable for web usage mining. Chou et al. (2010) discover the user's navigation behavior in the navigation of websites based on the sequence mining technique. The authors collect a dataset from Dec. 1, 2004, to Dec. 31, 2005, on the EC website that sold soaps and skincare products in the UK. First, they visualize the user's clickstream data using the footstep graph. Then, they use a back-propagation network (BPN) model to predict and categorize the users' navigation behavior. Xie et al. (2012) analyze the log of a real commercial search engine, including 750 million search requests. They study users' behavior, including query length, the ratio of query refining, and recommendation access. Besides, they find that users with navigation search queries always click only one result, and users with sex-related queries act like surfing the results one by one. Also, the results indicate that more than 90% of users pay attention to the top five results the search engine returns; therefore, analyzing users' behavior can improve the performance of finding the top five relevant results. Shih et al. (2012) focus on the information needs radar model (INRM), which consists of users, content, and concepts to describe information needs and satisfies customer demands. The experimental results in this research prove that the information-seeking system can automatically extract the latest needed information and knowledge without manual manipulation. Furthermore, it increases the capability of the information-seeking system and causes users to be satisfied with new information needs. In addition, users' dependency and loyalty to the information-seeking system will be enhanced. Hu et al. (2016) propose a novel model to provide more personalized service for query suggestions. This model finds similar queries in similar query sessions from search engine logs and re-ranks them. Eventually, the results indicate that the model is useful and effective. Sisodia et al. (2018) extract frequently accessed patterns of users using a pattern-based parallel FP-growth (MIP-PFP) algorithm. The goal of this research is to predict user behavior by extracting frequent patterns from huge weblogs. The results suggest that the MIP-PFP algorithm running on Apache Spark reduces the execution time by a factor of more than ten times. Fan et al. (2018) propose an approach to learning search action sequence representation for the search task. The search task success rate is an essential metric to measure the performance of search systems. The modeling search action sequence help to capture the hidden search patterns of users in the successful and unsuccessful search tasks. The authors suggest a new data augmentation approach to increase the pattern variations on search action sequence data. The experimental results illustrate that the proposed approach achieves significant performance improvement compared to several excellent search task success evaluation approaches. Su et al. (2018) worked on the intents of users through users' search activities. They categorize users based on intents in product search into three categories: Target Finding (TF), Decision Making (DM), and Exploration (EP). Then, they analyze users' log behavior interactions. The results of this research confirm that users with different intents behave differently. TF users tend to issue few specific queries and browse only top-ranked results; DM users tend to issue short queries, browse deep, and click more results; and EP users issue many diverse queries, browse deep, but do not click often. Jin et al. (2019) propose a model of information-seeking behavior that uses data mining. The authors apply a socialized mobile network to test their model. Finally, the results prove that interaction has an impact on information-seeking behavior. Also, the findings show that the usefulness of information plays a mediating role in the path of information usability impact on information search behavior.

As mentioned above, several studies have been carried out to examine users' behavior patterns based on data mining techniques. None of these studies has tried to cluster academic search engine users' behavior based on their information-seeking patterns and search styles. In this research, we propose a novel approach to mine search engine users' logs. These users' logs gathered click log data from the Iran scientific information database (Ganj). In the first step, we extract information-seeking behavior features from the interaction logs of users. After that, we cluster users into three groups, including fast surfing, broad scanning, and deep diving. Then, we investigate sequences of behavior patterns in each cluster separately.

3. Materials

In this section, we explain the personalized search engine and information-seeking patterns.

3.1. Personalized Search Engine

In recent years, the personalized search engine has been the interest of researchers. They can learn a user’s preference based on their search history, including the keyword, time, clickstream, and download. Each action taken by the user presents his or her behavior pattern. The collection of behavior patterns indicates the user’s preference and guides the search engine's admin to retrieve customized results. Besides, in academic search engines, understanding users’ search styles and extracting information-seeking patterns are essential because we can cluster users based on their search behavior. Then, we can realize each user belongs to which group. Therefore, the academic search engine can be personalized and providing services to each user based on his/her search style.

3.2. Information-seeking behaviors

Heinström (2003) believes that there are three information-seeking patterns: fast surfing, broad scanning, and deep diving, which can categorize individuals based on them.
Fast surfers’ individuals prefer certain types of documents, for instance, articles rather than books. Also, they prefer recently written materials and overview materials. Besides, the way the document looks and the language of a document is essential for them. When they are searching in a database, if nothing is retrieved, they assume that nothing is written about their search topic; hence they give up searching quickly. The lack of time is a barrier for them to get information; thus, they do not search thoroughly and put effort into information seeking. Furthermore, it is tough for these groups of users to judge the relevancy between results retrieved and search keywords.
Broad scanners’ individuals are characterized by thorough information seeking. They use various sources, including internet sources (Journals on the Internet or other material on the Internet), media sources (TV, radio, newspapers), group sources (conferences, lectures, associations, companies), and written sources (encyclopedias, journals, books, brochures). They discover information accidentally. The most crucial point about these people is that they search enormous related documents rather than only a few precisely on target. Also, they do not like planning database searches in advance.
Deep divers’ individuals care about the author of the documents and acknowledged information sources. They emphasize the high scientific quality of the materials. Unlike fast surfers’ individuals, deep divers tend to thorough information seeking, and they like putting effort into searching.

4. Research Design

An approach to categorize and analyze the user behavior patterns in an academic search engine consists of three steps. These steps are as follows:

//The steps for analyzing user behavior pattern on an academic search engine
//Output: The clusters of users
Step 1: Collect data
For all users
a) Cleaning data
b) Extracting sequence of behaviors
Step 2: Running the K-means clustering algorithm
Step 3: Examine the frequent behavior patterns in each cluster

The proposed research architecture is shown in Fig 1. It consists of research steps, which are explained further as follows.

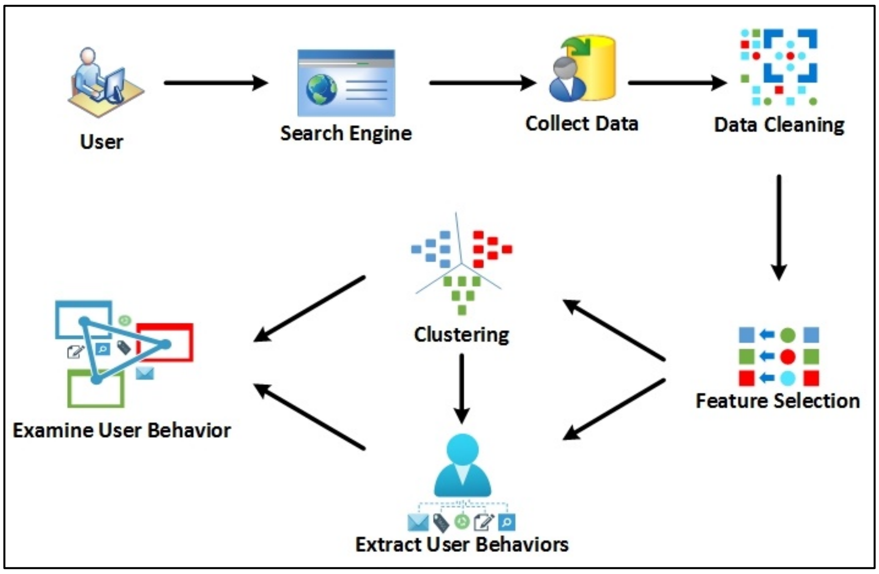

Figure 1. Research architecture

STEP 1: DATA COLLECTION

In the first step, the aim is to record users' navigation paths while visiting the website that provides access to a database of academic disciplines. In this paper, data is collected from the participants searching on the Iran scientific information database (Ganj). Ganj (Fig 2) results from the Iranian Research Institute for Information Science and Technology (IRANDOC) efforts in gathering, organizing, and disseminating scientific and technical documents in Iran from five decades ago. Although, nowadays, most updated data on Ganj belongs to dissertations, theses, and research proposals, the scope of its content is quite broad. According to the official announcement of IRANDOC, Ganj includes more than 1 million scientific records, in which more than 500,000 records of them are thesis. Searching on Ganj is free for all researchers, students, and faculty members and downloading the full text requires the users' free registration. In addition, more than 130,000 searches are done daily on this database (IranDoc, 2020). Many studies have carried out on the log file of users' search, which analyzed the users' behavior or assessed the user satisfaction of Ganj in recent years (Somayeh Fatahi & Ershadi, 2020; Somayeh Fatahi & Rabiei, 2020; Somayeh; Fatahi & Naimi-Sedigh, 2017; Rabiei et al., 2017). As Ganj is the official repository of scientific research in Iran and is widely used by Iranian researchers; thus, we used this database in our study.

Figure 2. The overview of Ganj

When users are searching on the Ganj, they have different kinds of interaction with this website. For example, they search some keywords to find their desire dissertations or click on the retrieved results. All of these search behaviors are logged in a raw format file. In this step, a dataset is generated from the raw format file based on behavioral features that include valuable information to categorize users. After that, the cleaning data is done. Details about the features will explain in the next section.

STEP 2: RUNNING THE K-MEANS CLUSTERING ALGORITHM

In the second step, the K-means clustering algorithm is run to the database created in the previous section. The goal of running a clustering algorithm is categorizing users based on their search behavior features. K-means is an iterative algorithm that tries to partition the dataset into K subgroups (clusters), where each record of data belongs to only one group. It attempts to make the inter-cluster data points similar while also keeping the clusters as different as possible. On the one hand, the popularity of this algorithm in different applications, simplicity, and efficiency are some reasons behind we chose this algorithm. Moreover, it works well with many variables, and it is fast (Han et al., 2011). On the other hand, one of the most important disadvantages of K-means is predicting the number of clusters (K-Value). Choosing k is often an ad hoc decision based on prior knowledge, assumptions, and practical experience (Hamerly & Elkan, 2004). In this research, we choose K=3 based on Heinström (2003) beliefs. He presents three information-seeking patterns: fast surfing, broad scanning, and deep diving.

STEP 3: EXAMINE THE FREQUENT BEHAVIOR PATTERNS IN EACH CLUSTER

In this section, we analyze each cluster based on the features, which separate the clusters. We choose one user as a data point from each cluster randomly. Then, we examine their activities, frequent sequential behavior, and search keywords network.

5. EXPERIMENT AND RESULTS

In this section, the above steps were performed on the Iran scientific information database (Ganj).

5.1. STEP 1: DATA COLLECTION

Fig 3 is an overview of users' logs called a raw format file in the previous section. A total number of 1,048,575 interaction records were collected from 28,526 unique users. Each record of users' interaction data includes "User_id", "Session_id", "IP_address", "Log_type", "URL", "Info", and "Created_at". The "Log-type" feature has 23 different values (Table 1).

User_id	Session_id	IP_address	Log_type	URL	Info	Created_at
183152	7077c01f51acc42515	192.168.12.250	2	8%B1%DA%AF%D9	2, search_title=	2019-05-26 00:00:05
10490	f2e699bf699d33f73c	192.168.12.250	5	c54e4457f0d290d8	rticle_id=>9567	2019-05-26 00:00:05
187102	ce2748d1afb26dcb48	192.168.12.250	1	v1/users/187102/p	{}	2019-05-26 00:00:05
25589	e00e0f7b598b774a1b	192.168.12.250	7	b872be5eddb08b8	ticle_id=>9386	2019-05-26 00:00:05
127339	cd1d30f4e0ca845ac4	192.168.12.250	12	0	2172, fulltext_a	2019-05-26 00:00:05
25589	e00e0f7b598b774a1b	192.168.12.250	7	72e63c9a4d723cd8	ticle_id=>6665	2019-05-26 00:00:05
42983	88ef33083a3b4568be	192.168.12.250	1	/v1/users/42983/pr	{}	2019-05-26 00:00:05
170903	06d6b6410c5b6aaabe	192.168.12.250	2	fis&limitation=&pa	4, search_title=	2019-05-26 00:00:05
151324	f971076dd91245176	192.168.12.250	2	08%A7%DA%A9%D8	15, search_title	2019-05-26 00:00:05
187609	9c3f26eef6815bc830	192.168.12.250	6	f21fd955a4f22cfd23	icle_id=>11485	2019-05-26 00:00:05
187609	9c3f26eef6815bc830	192.168.12.250	8	a4f22cfd23f1eed2e	icle_id=>11485	2019-05-26 00:00:05
187609	9c3f26eef6815bc830	192.168.12.250	9	22cfd23f1eed2e941	icle_id=>11485	2019-05-26 00:00:05
187609	9c3f26eef6815bc830	192.168.12.250	7	55a4f22cfd23f1eed	icle_id=>11485	2019-05-26 00:00:05

Fig 1- Overview of users' logs

Table 1- Description of different log types

Log_type	Description
1	Check the user profile (log in or check the user access)
2	Basic search
3	Show the user profile
4	Advanced search
5	Click on search results
6	View document info
7	Request to show keywords
8	Request to show abstract
9	Request to show additional fields
10	Request to show admin data
11	Show the document
12	View full text
13	Show the researcher data
14	RSS for a basic search
15	Export selected documents
16	Share the document on social networks
17	RSS for a keyword
18	RSS for a topic
19	RSS for a researcher
20	RSS for a subject
21	Report a bug
22	RSS for an advance search
23	RSS for an organization

A dataset generated from the raw format file based on the "Log_type" features called the behavioral features dataset. The dataset for each unique User_id includes the frequency of features in "Log_type", for example, the frequency of advanced search, the frequency of download a document, the frequency of click on the search results, and so on. We decided to focus on important features in the dataset, then the data cleaning is done, and the useless features are removed. Finally, we reached a dataset that includes nine features.
The average frequency of information-seeking behavior in these features is presented in Fig 4.

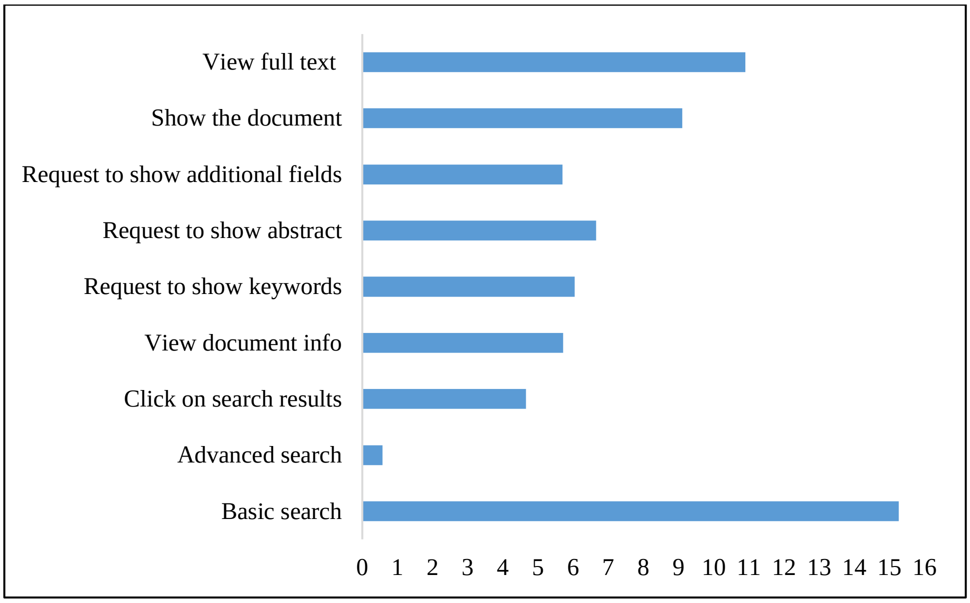

Fig 2- Average of information-seeking behavior features

In addition, we generated another dataset of users' navigation behavior from the raw format file based on a sequence of users' behavior and called it the sequential behavior dataset. For example, a user first login into a system, then type a query with the title "data mining"; some results show to him or her, and then the user clicks on one of them, and comes back and clicks the other one and finally views its full text. The sequence of his or her behavior is shown as <1, 2, 5, 11, 5, 11, 12> based on the "Log_type" in Table 1. This scenario is depicted in Fig 5.

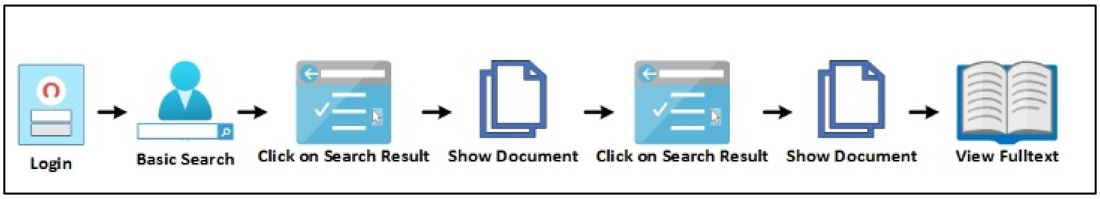

Fig 3- A sample scenario of information seeking behavior

5.2. Step 2: Running the K-means clustering algorithm

After preparing the behavioral features dataset, each of the feature columns is normalized based on the following formula:

$$R_{ij} = \frac{R_i - Min\ (R)}{Max\ (R) - Min\ (R)} \quad (1)$$

There are *j* features in the dataset, which have n records. R_{ij} is the value of i_{th} record for j_{th} feature. The K-means algorithm (k=3) is applied on behavioral features dataset to cluster the users. The results of the clustering are presented in Fig 6.

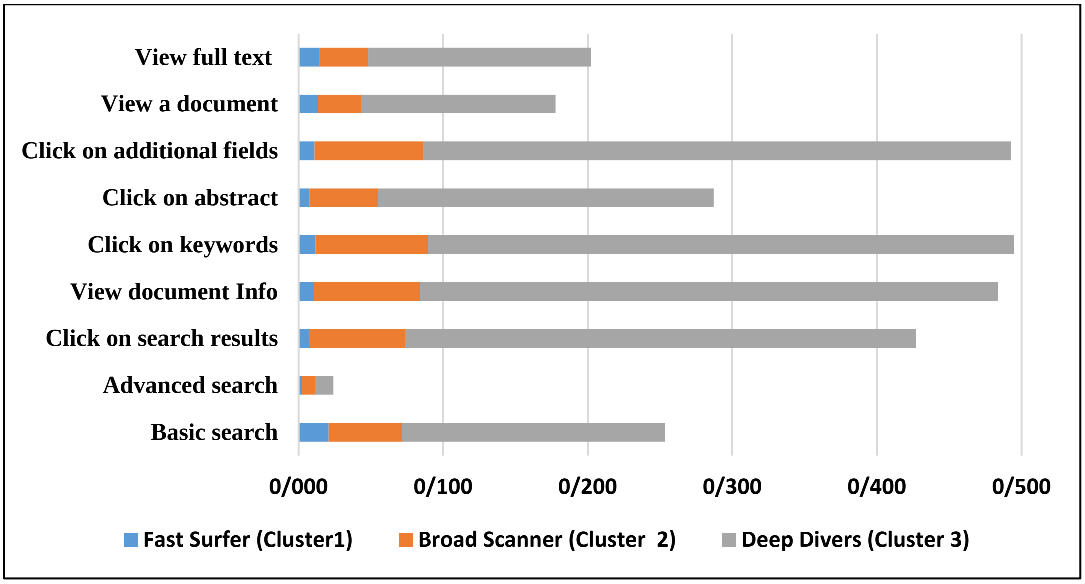

Fig 4- Clustering based on information-seeking behavior features

Based on Heinström (2003) definition and the value of features in each cluster, the appropriate label is assigned for each cluster. The details are explained in the discussion section.

5.3. Step 3: Examine the frequent behavior patterns in each cluster

In this section, we examine each cluster and investigate users' activities, frequent sequential behavior, and search keywords network as mentioned in the previous section. We choose a user as a sample for each cluster randomly.

First, the number of activities per day for each user is examined (Fig 7). As Fig 7 presents, the deep divers' user has a noteworthy difference with other users. The fast surfer's user and broad scanner's user have a similar pattern in using the search engine. It should be mentioned that the fast surfer's user is more active than the broad scanner's user.

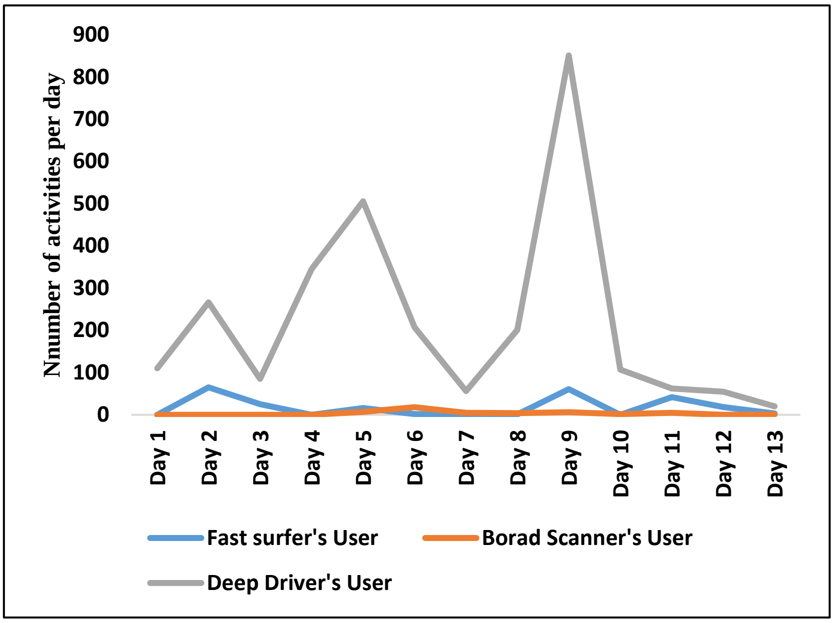

Fig 5- Comparing users in different clusters based on number of activity per day

Furthermore, the sequence of information-seeking behavior that users do on the academic search engine of Ganj is extracted from the sequential behavior dataset. Some of the frequent sequential behaviors are extracted from the log data (Table 2).

Table 2-Frequent sequential behavior in different clusters

Cluster	Frequent sequential behavior	
Deep Diver's user	*< View full text, View full text, View full text, Basic search, View full text, View full text, View full text, View a document, View full text>*	*< Basic search, View full text, View full text, View a document, Basic search, View a document, View full text>*
Broad Scanner's user	*< Basic search, Basic search, Basic search, Basic search, Basic search>*	
Fast surfer's user	*<View document Info, click on keywords, Click on additional fields, Click on abstract, Click on search results>*	*<View the information about a document, Click on search results, Click on search results, Click on additional fields>*

It is unquestionable, users in different clusters have various frequent sequential behavior when they are seeking information. It is seemingly, the shape of search keywords networks that are extracted for users in each cluster are different. Fig 8, Fig 9, and Fig 10 present the network of search keywords for each user separately.

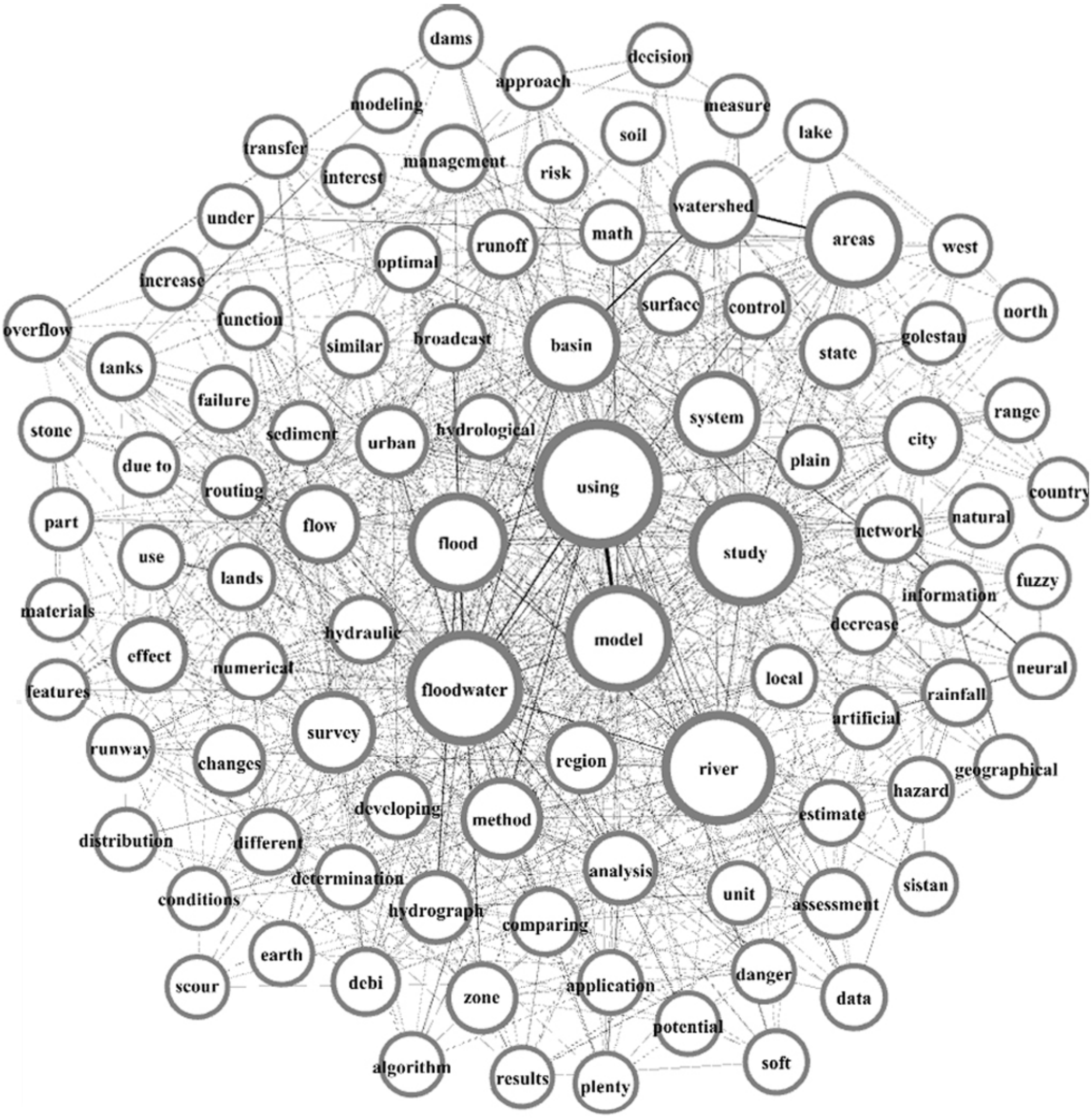

Fig 6-The network of search keywords of deep diver's user

Fig 7-The network of search keywords of fast surfer's user

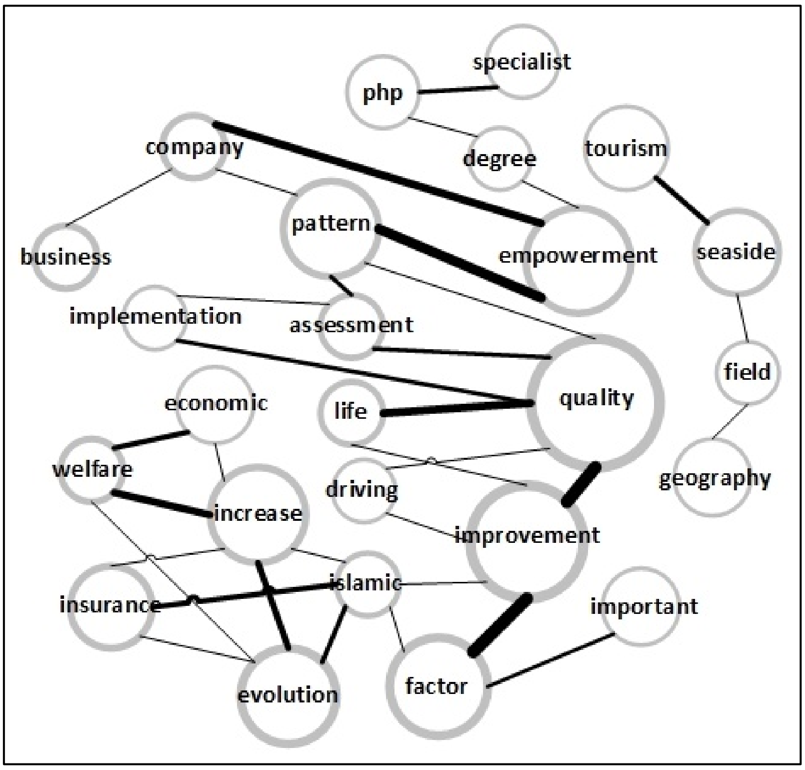

Fig 8-The network of search keywords of Broad Scanner's user

Fig 8, Fig 9, and Fig 10 present the network of search keywords for deep diver's user, fast surfer's user, and broad scanner's user. It is obvious that the size of these three networks, the number of nodes, and the edges are different because each is related to a specific cluster with specific features of seeking-information behavior.

6. Discussion

This section discusses our results based on three parts: users' activities, frequent sequential behavior, and search keywords network.

Based on Heinström (2003) beliefs, deep divers' individuals tend to comprehensive and complete information seeking; thus, they put their effort into searching. Besides, it is evident that they are more active than others (Fig 8). In contrast, the fast surfers' individuals will give up searching if they find nothing. In comparison with broad scanners' users, the fast surfers' users have a wave pattern, while the broad scanners individuals have a flat pattern because they search a lot of related documents rather than only a few precisely on target (Fig 9).

As mentioned above, in view of Heinström (2003), deep divers' people tend to seek information thoroughly, and they do their best to search. They search for more relevant online information and are more in-depth at reading. Therefore, it seems there is a significant relationship between this fact and the frequent patterns are extracted from our dataset. Table 2 shows that deep divers' individuals tend to view the full text of a document repeatedly. They prefer to download enormous documents as full text and read them later. On the other hand, the fast surfers' individuals scan much un-relevant online information without in-depth at reading them. Our results demonstrate that they look at different parts of a search engine web page and click on them. For example, they view the information about a document, click on keywords, click on additional fields, and click on an abstract. Besides, the fast surfers' users do not search thoroughly and prefer overview materials. The frequent sequence of their search behaviors in Table 2 shows that they overview materials and scan the pages. Broad scanner's individuals are characterized by thorough information seeking. They use Internet sources (Journals on the Internet or other material on the Internet), media sources (TV, radio, newspapers), group sources (conferences, lectures, associations, companies), and written sources (encyclopedias, journals, books, brochures) (Heinström, 2003). They also discover information accidentally. The most crucial point about these people is that they search enormous related documents rather than only a few precisely on target. The frequent search behaviors found in Ganj show that the broad scanner's user tends to basic search (Table 2).

Finally, looking at the network of search keywords for these clusters with different search styles reveals the other difference in users' behavior among these categories. Fig 8 belongs to the deep diver's user who is deep in searching and reading. This user tries to search many relevant keywords about one subject and knows what he/she wants; thus, the search keywords network for him is so extensive and associated with many related keywords. Fig 9 shows that the fast surfer's user searches a few keywords, scans the material and is not deep in reading. Eventually, this user is not putting effort into information seeking and gives up searching quickly. Fig 10 is not similar to Fig 9 and Fig 8 because the broad scanner's user is not like the fast surfer's user and the deep diver's user. The broad scanner's user is neither as deep in searching and reading as the deep diver's user nor as leaving search as the fast surfer's user. The network of search keywords for the broad scanner's user shows this user likes to search widely.

7. Conclusion

In this study, three critical contributions are made: 1) Extracting information-seeking behavior features, 2) clustering users based on information-seeking features 3) investigating frequent sequential patterns and search keywords network in each cluster. We choose a list of useful information-seeking behavior features and cluster users based on them. Three clusters of users are determined as deep divers, fast surfers, and broad scanners based on Heinström theory. Then, we examine users' activity, frequent sequential patterns, and the network of search keywords in each cluster. The results demonstrate that there is a significant relationship between users' activity, frequent sequential patterns, the network of search keywords, and users' search styles. This approach can apply to any academic search engine; however, we select the Iran scientific information database, which is called Ganj, as a testbed.
According to findings, Ganj managers could use some searching assistant tools such as autocomplete function in the searching box to improve the search performance of the fast surfers' users. Moreover, guiding these users in the different parts of the system via tooltip or online help and increase users' information literacy via different plans are practical. The broad scanners using selective dissemination of information (SDI) services have an important role in informing them about new scientific resources added to the database. Designing the professional search tool, along with the basic and advanced search, helps deep divers manipulate their searching queries to extract the best results from the database.
Managers and policymakers need to consider these characteristics in designing academic search engines. It definitely can improve the search quality for users, and their satisfaction will be increased.

8. Acknowledgment

We acknowledge the Iranian Research Institute for Information Science and Technology (IRANDOC) for providing log data of Ganj.

9. References

[1] ardorseo. (2020). How many Google searches per day on average in 2020? Retrieved from https://ardorseo.com/blog/how-many-google-searches-per-day/

[2] Chou, P.-H., Li, P.-H., Chen, K.-K., & Wu, M.-J. (2010). Integrating web mining and neural network for personalized e-commerce automatic service. *Expert Systems with Applications, 37*(4), 2898-2910.

[3] Fan, A., Chen, L., & Chen, G. (2018). *Search Action Sequence Modeling With Long Short-Term Memory for Search Task Success Evaluation.* Paper presented at the Thirty-Second AAAI Conference on Artificial Intelligence.

[4] Fatahi, S., & Ershadi, M. J. (2020). Assessment of User Satisfaction of Research Theses and Theses in Iranian Scientific Database (Treasure): Based on E-Qual Model. *Iranian Journal of Information processing and Management, 35*(2), 399-424.

[5] Fatahi, S., & Naimi-Sedigh, A. (2017). Analysis of Researchers Information Seeking Behavior in National Search Engine Thesis Information System in Iran. *Iranian Journal of Information Management, 2*(2), 31-58.

[6] Fatahi, S., & Rabiei, M. (2020). Users clustering based on search behavior analysis using the LRFM model (case study: Iran scientific information database (Ganj)). *Iranian Journal of Information processing and Management, 36(2)*, 419-442.

[7] Hamerly, G., & Elkan, C. (2004). *Learning the k in k-means.* Paper presented at the Advances in neural information processing systems.

[8] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

[9] Hanamanthrao, R., & Thejaswini, S. (2017). *Real-time clickstream data analytics and visualization.* Paper presented at the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).

[10] Heinström, J. (2003). Fast surfers, broad scanners and deep divers as users of information technology–Relating information preferences to personality traits. *Proceedings of the American Society for information science and technology, 40*(1), 247-254.

[11] Herder, E., & Juvina, I. (2004). *Discovery of individual user navigation styles.* Paper presented at the Proc. of the Workshop on Individual Differences in Adaptive Hypermedia (Adaptive Hypermedia 2004).

[12] Hu, F., Pu, C., & Peng, T. (2016). Recommendation in web search using similar query mining with big data. *Revista Tecnica de la Facultad de Ingenieria Universidad del Zulia, 39*(5), 238-243.

[13] IranDoc. (2020). Ganj. Retrieved from https://en.irandoc.ac.ir/service-product/94

[14] Jin, H., Miao, Y., Jung, J.-R., & Li, D. (2019). Construction of information search behavior based on data mining. *Personal and Ubiquitous Computing*, 1-13.

[15] Kaspar, K., & Müller-Jensen, M. (2019). Information seeking behavior on Facebook: The role of censorship endorsement and personality. *Current Psychology*, 1-12.

[16] Kim, K.-S., Sin, S.-C. J., & Tsai, T.-I. (2014). Individual differences in social media use for information seeking. *The Journal of Academic Librarianship, 40*(2), 171-178.

[17] Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning. *Expert Systems with Applications*, 113342.

[18] Kumar, A., Salo, J., & Li, H. (2019). Stages of user engagement on social commerce platforms: Analysis with the navigational clickstream data. *International Journal of Electronic Commerce, 23*(2), 179-211.

[19] Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing science, 23*(4), 579-595.

[20] Pareek, A., & Rana, M. S. (2013). Study of Information Seeking Behavior and Library Use Pattern of Researchers in the Banasthali University. *Library Philosophy & Practice*.

[21] Rabiei, M., Hosseini-Motlagh, S.-M., & Haeri, A. (2017). Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran. *Scientometrics, 110*(2), 815- 842.

[22] Shih, C.-W., Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2012). Enhancement of information seeking using an information needs radar model. *Information processing & management, 48*(3), 524-536.

[23] Sisodia, D. S., Khandal, V., & Singhal, R. (2018). Fast prediction of web user browsing behaviours using most interesting patterns. *Journal of Information Science, 44*(1), 74-90.

[24] Su, N., He, J., Liu, Y., Zhang, M., & Ma, S. (2018). *User intent, behaviour, and perceived satisfaction in product search.* Paper presented at the Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.

[25] Ting, I.-H., Clark, L., Kimble, C., Kudenko, D., & Wright, P. (2007). *APD-A tool for identifying behavioural patterns automatically from clickstream data.* Paper presented at the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems.

[26] Triantafillou, E., Pomportsis, A., Demetriadis, S., & Georgiadou, E. (2004). The value of adaptivity based on cognitive style: an empirical study. *British Journal of Educational Technology, 35*(1), 95-106.

[27] Wiberley, S. E., & Jones, W. G. (1989). Patterns of information seeking in the humanities. *College & Research Libraries, 50*(6), 638-645.

[28] Xie, K., Yu, H., & Cen, R. (2012). Using log mining to analyze user behavior on search engine. *Frontiers of Electrical and Electronic Engineering, 7*(2), 254-260.