Clustering Users by Information-Seeking Style: An empirical study on an Academic Search Engine

Somayeh Fatahi
Department of Computer Science
University of Saskatchewan
Saskatchewan, Canada
somayeh.fatahi@usask.ca

Amineh Setayesh
Department of Computer Engineering
Shahid Bahonar University of Kerman
Kerman, Iran
aminehsetayesh@eng.uk.ac.ir

Julita Vassileva
Department of Computer Science
University of Saskatchewan
Saskatchewan, Canada
jiv@cs.usask.ca

***Abstract*—Nowadays, academic search engines have become indispensable tools for getting important online scholarly information. User differences are important factors that influence the use of information systems. The way people use academic search engines to find information varies depending on their information-seeking style. Therefore, finding and understanding different information-seeking behaviors has become an important line of research. User behavior patterns can be discovered by examining user interaction logs to determine who the users are and what they intend to do. These insights can be useful in designing more optimized academic search engines. In this paper, we analyze the user interaction logs collected from the Iranian scientific information database. The Ganj database is the official repository for collecting and organizing theses and dissertations in Iran. Many researchers search scientific and research resources from the Ganj database daily. We use a sequential pattern mining approach to extract frequent sequential behavior patterns on user interaction logs and to cluster users into three groups based on their frequent behavior patterns, using the K-means clustering algorithm. Cluster analysis shows that users with similar frequent behavior patterns have similar information-seeking styles. Finally, we found three clusters and named them: fast surfers, deep divers, and broad scanners. Our findings can help developers of academic search engines and policymakers to identify users' needs and priorities to make better decisions to design a reasonable page layout and well-organized website for all users based on their search styles.**

Keywords—Information-seeking style, Frequent sequential behavior, Pattern mining, Academic search engines, Clustering

I. Introduction

We live in an age when access to information plays an important role in our daily activities, affecting our decisions. With the increase in information and communication technology, utilizing the Internet to search for information has become prevalent behavior. Recently, seeking information online has become a preferred way to satisfy our information needs, due to the availability and coverage of information, the convenience of searching, affordability of access, and interactivity. Thus, studying information-seeking behaviors would be beneficial. The importance of information-seeking becomes clear when it comes to academicians, researchers, and students due to their demand for updated information for their research needs. Studying researchers' information-seeking behavior on search engines is important for several reasons such as meeting the demands of research, personalizing search results, developing new features or algorithms to improve search experience, improving search engine quality, and identifying trends and gaps in research. Today, there are many academic search engines, including Microsoft Academic Search, Google Scholar, Web of Science Core Collection (WoS), Scopus, and so on. They allow free and easy access to scientific literature and publisher independent information on the web [1]. Studying information-seeking behavior in these search engines gives us better insight into the expectations of the users and the existing information services [2].

Information-seeking behavior is described as the way individuals get information and put it to use [3]. Many factors influence the information-seeking behavior, including demographic, psychological, role-related, and environmental factors [4] [5]. People with different characteristics access information in particular ways, which we call "information-seeking style" [6]. Marchionini [7] defined four levels of description of information-seeking behaviors: moves, tactics, strategies, and patterns. Moves are defined as single behavioral actions such as clicking on a hypertext link, typing a query into a search engine, and using the ''return'' button. Tactics are a group of behaviors, for example, typing a search query using general search terms and then using successively more specific query terms to narrow the search domain. Strategies are defined as approaches to finding specific information through sources, for example, using only a search engine to search for information. Finally, a pattern is defined as automatic and unconscious interactions used for all information-seeking tasks, not just those on the WWW (i.e., probably something like a searching style). In this paper, we focus on users' behavior patterns to extract their information-seeking style. It is important to be aware of the users' search behavior and their needs in order to improve search engines [8]. A good way of analyzing users' search behavior and getting an understanding of their information-seeking style is to examine the log files of search engines because they provide rich information about systems and the activities occurring on them [9]. Web Mining is a method of finding information from the Web; it can be classified into three large types depending on which part with Web is exploited: Web content mining, Web structure mining, and Web usage mining [10]. Web usage mining focuses on users' interaction with websites [11]. The main aim of web usage mining is to investigate user navigation and browsing through web pages, extract users' behavior patterns, and understand various other requests made by users.

We use web usage mining to determine users' information-seeking style based on frequent sequential behavior. In the first step, we extract frequent sequential behavior for each user, then run a clustering algorithm and then analyze each cluster. The results show that users with similar frequent patterns during their research have the same search style. This finding can help developers in personalizing search engines based on users' search styles.

II. Related Works

The problem of identifying user search behaviors relates to several distinct research areas, including information-seeking

979-8-3503-2759-5/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCE60160.2023.00295

behavior, extracting frequent sequential behavior, and segmentation of users. The novelty of our research is in addressing the lack of previous work that analyses frequent sequential user behavior in academic search engines and clustering users based on their behavior.

In 2006, Thatcher defined some search tasks for 80 participants and investigated their strategies during their search [24]. These tasks are categorized into two groups, including two researcher-defined tasks and two participant-defined information-seeking tasks using the WWW. The participants were given a maximum of 15 min each for each task and could terminate the search at any point before the 15 min. Log-file data and observations were examined and analyzed. There were 28 types of search moves identified and based on Marchionini's [7] search model; the search moves were classified into five areas: task initiation behaviors, search terms, sustaining search behaviors, task termination behaviors, and unusual behaviors. The results showed that the search strategy was dependent not only on the type of search task but also on whether the task was defined by the researcher or the participant. In addition, at least two of the cognitive search strategies (i.e., ''search engine narrowing down'' and ''search engine player'' strategies) were utilized by participants to exploit the different perceived qualities of search engines.

In 2013, researchers studied user behavior on Spotify [14]. They used a massive dataset collected between 2010 and 2011. They found that there is a daily pattern for users and behavior of switching between desktop and mobile devices for using Spotify. Their results showed that users have their favorite times of day to access the service. Moreover, the results demonstrated that there are correlations between the session length and downtime of successive user sessions on single devices. These results helped developers understand user behavior in Spotify and provided new insights into user behavior in other music streaming services.

In 2016, Han & Wolfram analyzed users' patterns for understanding user search and browsing behavior [15]. They applied three analytical techniques: network analysis, sequential pattern mining, and k-means clustering. They gathered log files from an image collection digital library for three months. The results showed that even though users had different search lengths, their behaviors were uniform. Moreover, users were not interested in using all search features. Finally, they concluded that developers of image-based digital libraries should consider design features that support rapid browsing.

In 2018, Nguyen et al. proposed a visual analytics approach to gain deep insight into the expected and unexpected user behaviors through an analysis of their action sequences [16]. Their contribution was designing a method to identify high-level, semantically relevant patterns, activities, from raw action sequences. They evaluated the applicability of their approach in a case study that involved tasks requiring effective decision-making by a group of domain experts. In 2019, Liu et al. investigated information-seeking behavior on Web searches. They gathered 693 query segment data from 40 participants and analyzed their intention and search behavior [13]. The results showed that task features significantly impacted the frequency of occurrence of most of the information-seeking intentions.

In 2018, Ndumbaro conducted a study on the information-seeking behavior of users in the Online Public Access Catalogue (OPAC) of the University of Dar es Salaam library [36]. The study aimed to identify the reasons for search failures and to shed light on the effects of default search options on users' search experiences. The research findings were based on empirical evidence from search logs and previous related studies. The study found that users preferred to use titles, author names, subject terms, and keywords in their searches, which accounted for over 83% of all search queries. On the other hand, phrase search accounted for only 13.3% of all searches, while ISSN, ISBN, abstract, publisher, and year were among the least used search options. One possible explanation for this finding is that ISSN and ISBN may not be widely known to users. The study also identified the reasons for search query failure, which were related to both human errors and system ineffectiveness. Human errors included spelling mistakes, typos, inappropriate use of Boolean operators, wrong syntax, and inappropriate use of search fields.

In 2020, researchers suggested using sequential search pattern analysis and clustering to analyze customers' search behavior throughout an entire shopping process[12]. They used maximal repeat patterns (MRPs) and lag sequential analysis (LSA) to analyze the sequence of search paths and identify significant repeated search patterns. Based on their research, there are four groups of customers who browse for information, adopt recommendations, consult reviews, and conduct searches with different levels of goal-oriented or exploratory-based need-states. The findings demonstrated that customers with strong goal-oriented need-states have the simplest search paths compared to other groups, whereas exploratory-based customers have the most complicated search paths. Moreover, customers who are goal-oriented prefer to search directly, consult reviews carefully, and have stored sequential search patterns, while customers with exploratory-based need-states have a tendency to explore the categories of products and adopt product classification hierarchy.

In 2022, Zerhoudi et al. suggested a Markov approach to simulate user sessions to protect users' privacy. They use their methods to simulate log data of search sequences performed by users in a digital library [29]. They used the Sowiport User Search Session data set (SUSS) which contains over 48000 individual search sessions and around 8 million log records. Their results show that their approach reliably simulates global and type-specific search behavior.

In 2019, a study was conducted by Barifah and Landoni which aimed to identify hidden usage patterns by analyzing a dataset of 28 million records from the RECO Doc DL log files [35]. They used four main features: session starting points, content discovering actions, types of functions used (if any), termination actions, and session duration to recognize usage patterns. Based on these features they created three datasets from the population with different sizes (10%, 5%, and 2%) through a process of random selection without replacement. After that, the clustering algorithm was run, and the optimum number of clusters was determined. The findings revealed that there are three distinct usage patterns: item seekers, navigators, and searchers. Item seekers are distinguished by performing a single action, such as viewing or downloading items, without engaging in any additional interactions with the system. Navigators, on the other hand, spent more time on the system and performed activities such as clicking on a collection, viewing a result page, evaluating the snippet, and browsing the result pages. Finally, searchers were users who submitted queries, evaluated the results, browsed the result pages, and clicked on items.

The novelty of our research lies in using frequent sequential behavior as a means of identifying similar search styles. Considering the information-seeking style of the users while designing the academic search engine will help managers and policymakers of academic search engines, such as the Ganj, create a better and more efficient search experience for the users.

III. Research Design

We took a three-step approach to analyze and categorize the users' frequent sequential behavior patterns in an academic search engine. The steps are:

A. Data collection

The purpose of this step is to record users' navigation paths when they visit the Iranian scientific information database (Ganj). The idea behind the creation of Ganj originated five decades ago when collecting, organizing, and disseminating scientific and technological documents began in Iran. Currently, most of the up-to-date data of Ganj comes from theses, dissertations, and research proposals, but its scope is quite broad and includes research plans, government reports, and scientific articles. According to IranDoc's official report, Ganj contains over 1 million scientific records, over 500,000 of which are theses. All researchers, students, and faculty members can use the search engine of Ganj; free user registration is required to download the full text of the document. In addition, over 130,000 searches are performed daily on this database (IranDoc, 2022). In recent years, a lot of research has been done on users' search logs, analyzing users' behavior and evaluating users' satisfaction with Ganj [17][18][19][20]. Since Ganj is the official archive of Iranian scientific research and is widely used by Iranian researchers, we used this database in our study. When users search on Ganj, they have different types of interactions with this website. For example, researchers can search literature by entering keywords related to the search topics. The search results will display a list of theses, dissertations, research proposals, and scientific articles that match with search criteria. Once researchers are logged in, they can enter their search criteria into the search bar. This could include keywords, author names, or article titles. Users can also use advanced search options to refine their search results based on factors such as publication date, subject area, and type of publication. After displaying the search results by Ganj, Users can review the search results and filter them further based on relevance, and other factors. They have two options for accessing the documents: Downloading the first 15 pages of the documents or downloading the full text of the materials. Also, they can just click on Abstract, keywords, RSS, and so on. All search behaviors are recorded in a raw format file. In this step, for data collection, a dataset containing valuable information is generated from the raw format files based on users' behavioral activities. Next, data cleaning is performed on the dataset. The details of the activities are described in the next section.

B. Extracting frequent sequential patterns for each user

We defined a sequence of users' behavior patterns based on their activities. We consider a sequence of behavior patterns as $X=<x_1, x_2, \ldots, x_m>$ where x_i is a user activity during their search (session). For example, the sequence <*Advanced search, Click on search results, Show the document, Basic search, Request to show abstract*> is a sequence of user behavior patterns. In this step, based on previous research [21][22], we considered a 30-minute time interval for the segmentation of sessions and extracted the sequences of behavior patterns for all users. After that, we extracted all subsequences of behavior patterns for all users. We define a subsequence of behavior patterns as follows. Let $X=<x_1, x_2, \ldots, x_{k-1}, x_k>$ and $Y=<y_1, y_2, \ldots, y_m>$, $m<=k$ be two sequences of behavior patterns, $Y\subseteq X$ indicates that X is a super behavior pattern or Y is a subsequence of X if there exists $x_1=y_1$, $x_2=y_2$, $\ldots$, $x_j=y_m$ where $j<=k-1$. A new dataset is generated based on all subsequences of behavior patterns along with their frequency. Then, the most frequent sequential behavior across all users is determined.

C. Running the K-means clustering algorithm on frequent sequential patterns and analyzing clusters

In this step, we apply the K-means clustering algorithm to the frequent sequential behaviors that were obtained in the previous step in order to categorize users based on their frequent sequential patterns. The k-means algorithm was first introduced by Lloyd and then by MacQueen [23]. It is an iterative algorithm that attempts to divide a dataset into K pre-defined distinct subgroups (clusters) where each data point belongs to only one cluster. It tries to make the data points in a cluster as similar as possible but keeps the clusters as distinct as possible. The popularity, simplicity, and efficiency of this algorithm in various applications were some of the reasons for choosing this algorithm. In addition, it works well with many variables and is fast. However, one of the main drawbacks of K-means is that the number of clusters (the value of K) needs to be chosen in advance. Choosing k is often an ad hoc decision based on prior knowledge, assumptions, and practical experience and it is more challenging when the data has many dimensions, even when clusters are well-separated [25].

IV. Experiment and Results

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

A. Data collection

A total number of 683,932 interaction records were collected from 1000 unique users in 183 days (from September 2021 to February 2022). Each record of user interaction includes "id", "log_type", "date", "url", and "info". The "id" feature indicates ids of users. The "log_type" feature shows the behavior of users and has 12 different values. In Table 1, the distribution of percentage for different log_type is listed. As shown in Table 1, Log type 2 which refers to "*Basic search*" was the most used log type comparing with other log types. Nearly half (49.2%) of users' behaviors are log type 2 (basic search), followed by log type 1 (*Log in to the system)* with 12%, *Show the document* with 10.1%, and *view full text* with 10% as the next most frequent behaviors.

Not all of the 1000 users were active every month. Some users were only active for some months from September 2021 to February 2022. February is the month with the least number of active users, whereas October was the month with the most active users. In Iran, the academic year starts on the 23rd of September, and because of that, there is a high number of users in October. In February most grad and undergrad students and faculty members are busy with the final exams, they only have a little time to do research, which explains why the number of users decreased in February.

TABLE I. DESCRIPTION OF DIFFERENT LOG TYPES

Log_Type	*Description*	*Distribution of Percentage*
1	Log in to the System	12%
2	Basic search	49.2%
3	Show the user profile	0.61%
4	Advanced search	0.84%
5	Click on search results	2%
6	View document info	3.1%
7	Request to show keywords	4%
8	Request to show abstract	4.6%
9	Request to show additional fields	3%
10	Request to show admin data	0.15%
11	Show the document	10.1%
12	View full text	10%

B. *Extracting frequent sequential patterns for each user*

As mentioned in the previous section, we extracted a dataset from the raw data, which includes users' id and a sequence of behavior patterns in a 30-minute time frame. The generated dataset has 1000 records representing 1000 users. For each user, there is a lot of sequential behavior which shows their activities in the system. As an instance, the sequence *<1, 2, 5>* is a sequence of user behavior patterns and is described based on log_type in Table 1 *< Log in to the System, Basic search, Click on search results>*. Because there are many unique sequences with different lengths, we extracted all subsequences for all users along with their frequency. For example, for a user who has a sequence *<2,3,4>*. There are sequential subsequences such as *<2,3>* and *<3,4>*. In this step, we extracted all subsequences for all users. Since the aim of this research is to investigate frequent sequential behavior, we just focus on the subsequences which are more frequent rather than others. Moreover, we are looking for meaningful subsequences therefore we limit our dataset to subsequences with lengths between 3 to 7. It is clear that the long and short subsequences are not interpretable. Finally, we chose 100 of the subsequences that are the most frequent. The total number of subsequences is 10, 202, 019. The frequency of the top 20 frequent subsequences for all users is shown in Fig 1.

C. *Running the K-means clustering algorithm on frequent sequential patterns and analyzing clusters*

In this step, we ran the k-means algorithm to cluster users. Our aim was for users with the same frequent sequential behaviors to be placed in the same cluster. The number of clusters (value of K) should be specified. As mentioned before, choosing the value of K is a bit challenging. We used different measures to determine the K value.

The first measure is Silhouette which uses the compactness of individual clusters (intra-cluster distance) and separation amongst clusters (inter-cluster distance) to measure an overall representative score of how well the clustering algorithm has performed [26]. A higher score indicates better clustering. Therefore, the silhouette score for K = 3 is the best for our study. The other measure is Davies-Bouldin which is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances [27]. Based on this measure, lower values indicate better clustering. The lowest score belongs to K=3 which suggests the best number of clusters for our data. Another measure is the Elbow method which is based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids [28]. The best K is the K at the spot where SSE starts to flatten out and form an elbow. We evaluated SSE for different values of K on our dataset. The result shows that the curve forms an elbow and flattens out at K = 5.

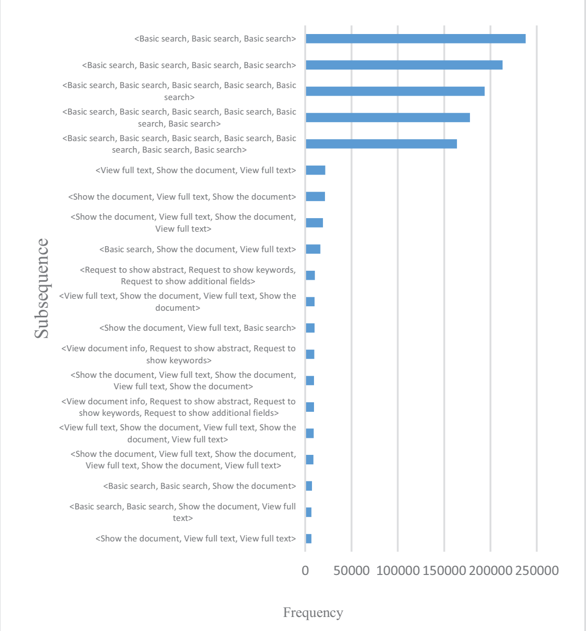

Fig. 1. Frequency of top 20 most frequent subsequences for all users

We ran the k-means algorithm for K=3 and K=5 on the dataset that was generated in the previous step. The result showed that the best K is 3 because for K=5, three clusters almost have the same frequent sequences, thus, we consider them the same. The results of clustering show there are three clusters with different frequent sequential behaviors. In the first cluster, there are 478 users whose frequent sequential behaviors are shown in Fig 2. In the second and third cluster, there are 358 and 152 users whose frequent sequential behaviors are shown in Fig 3 and 4.

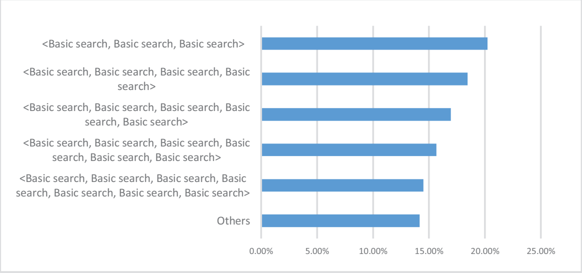

Fig. 2. Percentage of frequent sequential behaviors in Cluster 1 (Fast surfers)

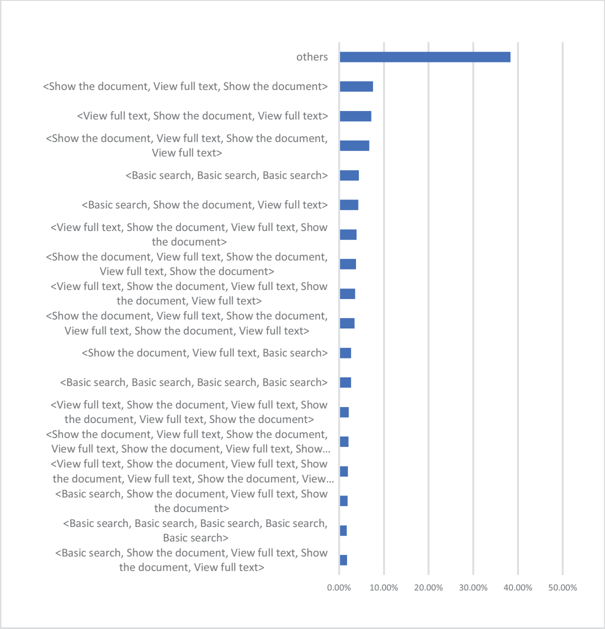

Fig. 3. Percentage of frequent sequential behavior in Cluster 2 (Deep divers)

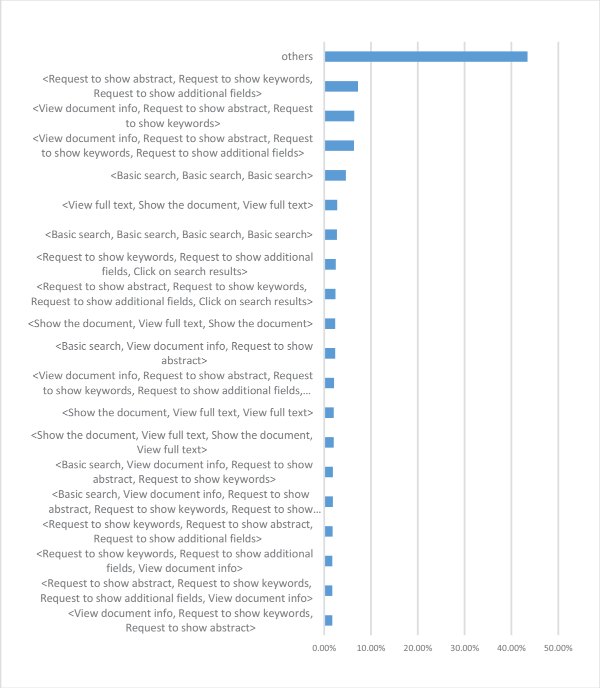

Fig. 4. Percentage of frequent sequential behavior in Cluster 3 (Broad scanners)

V. Discussion

Based on Heinström's view [34], there are different categories of people according to their information-seeking style: fast surfers, broad scanners, and deep divers. We used this view to label our clusters. In our opinion, users in cluster 1 are Fast surfers. Fast surfers tend to judge a document by its characteristics, such as looks, type, easy availability, etc., rather than the content of the document. Additionally, they prefer recently written and overview materials. They tend to search for quick answers; hence, they do not use web material unless it can be found with the least effort. Lack of time acts as a barrier to thorough information seeking. In Fig. 2, frequent subsequence behavior patterns for users show many basic searches. It means these users only use the basic and simple search on the search engine instead of advanced search. Moreover, they do not view the full texts and do not read them. They only perform a quick search with the least effort. For example, a frequent behavior of users in this cluster is *<Basic search, Basic search, Basic search, Basic search, Basic search>*.

Deep divers tend to consider the validity of the information source important. It's important for them that the material has high scientific quality. These individuals put much effort into seeking information; therefore, they value thorough material. As the results in Fig. 3 show, users in Cluster 2 behave as deep divers because most of the frequent subsequences of their behavior are viewing documents for downloading the full text. It is a sign that they are interested in fully reading the documents. As an illustration, a common behavior among users in this cluster is *<Basic search, Show the document, View full text, Show the document, View full text >*.

Broad scanners support the idea of using various sources. They search widely and access many sources. These individuals thoroughly seek information and discover information accidentally. As a significant characteristic, they tend to search for numerous slightly related documents rather than a small number of exactly on target documents. Additionally, they are opportunistic and do not plan their database searches in advance. As the frequent sequential pattern shows in Fig. 4 users in Cluster 3 are broad scanners which is why their frequent behaviors are based on broad search. They present different behaviors, such as clicking on the abstracts, clicking on the search results, viewing the information about the document, and viewing the full text. We conclude that people in Cluster 3 show broad scanner behaviors because they have different sequences of behaviors compared to people in Cluster 1 and Cluster 2. The most frequent behavior in Cluster 1 (fast surfers) is only basic search and in Cluster 2 (deep divers) is basic search, request for showing the document and view full text for downloading it. For instance, a prevalent behavior observed among users in this cluster is *< View document info, Request to show abstract, Request to show keywords, Request to show additional fields >*.

Considering the results described above, managers and designers of academic search engines can create different versions of the search engine interface and options to match the needs of each search behavior cluster and support rapid browsing.

Fast surfers tend to give up on their search quickly and prefer overview material; therefore, using autocomplete function in the search box and other search assistance tools along with retrieving shorter and recently written articles can lead to a better search experience for them. Deep Divers prefer to carefully read the full text, so showing the most cited documents would create a better searching experience for these users as they value the validity of the article and the author. Providing an advanced search tool can help deep divers design their search queries to retrieve their desired results from the database. Broad scanner users tend to click on many items in the search results and they prefer to search broadly. An option to save useful finds for later could facilitate their search by allowing them to keep track of their discoveries.

Mapping the user's search behavior to the appropriate cluster will allow using a tailored version of the system interface that to optimizes the functionality to the needs of the user.

VI. Conclusion

According to some research, users' loyalty to digital libraries is declining. Researchers have found that factors such as user satisfaction [31], quality of services [33], customer value [32], ease of use, perceived usefulness and digital libraries' affinity influence customer loyalty directly or indirectly [30]. Based on the results, users' differences also impact user satisfaction and loyalty. Unfortunately, building a fully personalized user model is challenging due to the lack of sufficient daily user interaction on an academic search engine in digital libraries compared to a web search engine. To address this problem, academic search engines need to be smart and identify users' needs and styles based on their behavior. In most of the research related to digital libraries and search engines, researchers focused on analyzing log data to report information such as query length, time of the search, and so on, but not on identifying user behavior patterns. Extracting sequential behaviors has been of interest to researchers in the online shopping research area and clustering customers.

The novelty of our approach is using sequential behavior clustering to find different search styles of researchers in research papers repositories. We believe that if academic search engines can detect users' search styles, and if developers present a personalized version of a digital library to each cluster, they can increase users' satisfaction without incurring the cost of personalization for an individual. Therefore, we focused on clustering users based on their search style or information-seeking behavior. We categorized users into three groups based on their information-seeking behaviors. First, frequent sequential behavior patterns of each user were extracted. Next, users were clustered into three clusters based on their frequent sequential behavior patterns. We named the three clusters "deep divers", "fast surfers", and "broad scanners", by mapping them to Heinström's information seeking personality types [34]. The method we propose was applied to the Iranian research repository Ganj, but it is applicable to any other academic search engine database.

Considering technology and internet improvements in the information era, demands for easier and more efficient information retrieval are increasing. Providing search engines based on users' needs and styles can help to increase user satisfaction.

When an academic search engine is evaluated, users evaluate its performance based on their needs and styles. If developers consider what features are attractive and useful for each group of users, they can design a reasonable page layout and well-organized website for all users based on their search styles. Finally, designing an academic search engine considering the users' information-seeking style helps managers and policymakers bring a better and more efficient search experience for the users.

Acknowledgment

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

References

[1] J.L. Ortega, "Academic search engines: A quantitative outlook", 2014, Elsevier.

[2] V. Chopra, "Information seeking behavior of library users in select PG Degree Colleges of Chhattisgarh State". *Journal of Library & Information Science*, *8*(1), 2018.

[3] A.K. Pareek and M.S. Rana, "Study of information seeking behavior and library use pattern of researchers in the Banasthali University". *Library Philosophy and Practice (e-journal)*, *887*(9), 14-35, 2013.

[4] X. Niu and B.M. Hemminger, "A study of factors that affect the information-seeking behavior of academic scientists". *Journal of the American Society for Information Science and Technology*, *63*(2), 336-353, 2012.

[5] Z. Zhang, X. Yao, S. Yuan, Y. Deng, and C. Guo, "Big five personality influences trajectories of information seeking behavior". *Personality and Individual Differences*, *173*, 110631, 2021.

[6] J. D. Rivera, "Factors influencing individual disaster preparedness information seeking behavior: analysis of US households". *Natural Hazards Review*, *22*(4), 04021042, 2021.

[7] G. Marchionini, "Information Seeking in Electronic Environments." Cambridge University Press, New York, 1995.

[8] D. Lewandowski, "Search engine user behavior: How can users be guided to quality content?" *Information Services & Use*, *28*(3-4), 261-268, 2008.

[9] P. Nimbalkar, V. Mulwad, N. Puranik, A. Joshi, and T. Finin, "Semantic interpretation of structured log files". In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)* (pp. 549-555). IEEE, 2016.

[10] K. K., Ibrahim and A.J. Obaid, "Web Mining Techniques and Technologies: A Landscape View". In *Journal of Physics: Conference Series* (Vol. 1879, No. 3, p. 032125). IOP Publishing, 2021.

[11] B. Bhavani, V. Sucharita, and K.V.V Satyanarana, "Review on techniques and applications involved in web usage mining". *International Journal of Applied Engineering Research*, *12*(24), 15994-15998, 2017.

[12] I. C. Wu and H. K. Yu, "Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states". *Information Processing & Management*, *57*(6), 102323, 2020.

[13] J. Liu, M. Mitsui, N.J., Belkin, and C. Shah, "Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search", In *Proceedings of the 2019 conference on human information interaction and retrieval* (pp. 123-132), 2019.

[14] B. Zhang, G. Kreitz, M. Isaksson, J. Ubillos, G. Urdaneta, J.A. Pouwelse and D. Epema, "Understanding user behavior in spotify". In *2013 Proceedings IEEE INFOCOM* (pp. 220-224). IEEE, 2013.

[15] H. Han and D. Wolfram, "An exploration of search session patterns in an image-based digital library". *Journal of Information Science*, *42*(4), 477-491, 2016.

[16] P. H. Nguyen, C. Turkay, G. Andrienko, N. Andrienko, O. Thonnard, J. Zouaoui, "Understanding user behavior through action sequences: from the usual to the unusual". *IEEE transactions on visualization and computer graphics*, *25*(9), 2838-2852, 2018.

[17] S. Fatahi, A.H. Seddighi and M. Rabiei, "Analyze user behavior patterns on academic search engines", 2021.

[18] S. Fatahi and M. Rabiei, "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, 2020.

[19] S. Fatahi and A. Naeimi Seddigh, "Analysis of Researchers""Information Seeking Behavior in National Search Engine Thesis Information System in Iran". *Iranian Journal of Information Management*, *2*(5-6), 31-58, 2017.

[20] S. Fatahi and M.J. Ershadi, "Assessment of User Satisfaction of Research Theses and Dissertations in Iranian Scientific Database

(Ganj): Based on E-Qual Model". *Iranian Journal of Information processing and Management*, *35*(2), 399-424, 2020.

[21] H. Xinhua, and W. Qiong, "Dynamic timeout-based a session identification algorithm". In *2011 International Conference on Electric Information and Control Engineering* (pp. 346-349). IEEE, 2011.

[22] S. Dhawan and M. Lathwal, "Study of preprocessing methods in web server logs." *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(5), 430-433, 2013.

[23] J. Han, J. Pei, and H. Tongm, "*Data mining: concepts and techniques",* Morgan kaufmann.

[24] A. Thatcher, "Information-seeking behaviors and cognitive search strategies in different search tasks on the WWW". *International journal of industrial ergonomics*, *36*(12), 1055-1068, 2006.

[25] G. Hamerly and C. Elkan, "Learning the k in kmeans". Advances in Neural Information Processing, 2004.

[26] S. Rousseeuw, "A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*, (20), 53.

[27] D.L. Davies and D.W. Bouldin, "A cluster separation measure". *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227, 1979.

[28] M. A. Syakur, B.K. Khotimah, E.M.S. Rochman, and B.D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster". In *IOP conference series: materials science and engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing, 2018.

[29] S. Zerhoudi, M. Granitzer, C. Seifert, and J. Schlötterer, "Simulating user interaction and search behaviour in digital libraries", In *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 2022.*

[30] F. Xu and J.T. Du, "Factors influencing users' satisfaction and loyalty to digital libraries in Chinese universities", *Computers in Human Behavior*, *83*, 64-72, 2018.

[31] C. T. Townley and K.B. Boberg, "The changing role of technical university libraries: 1983-1996". In *Proceedings of the IATUL Conferences*, 1997.

[32] S. McKnight, "Envisioning future academic library services: initiatives, ideas and challenges", Facet Publishing, 2010.

[33] S. Ullah, "Customer satisfaction, perceived service quality and mediating role of perceived value". *International journal of marketing studies*, *4*(1), 2012.

[34] J. Heinström, "Fast surfers, broad scanners and deep divers: Personality and information-seeking behaviour". Turku: Åbo Akademi University Press. Retrieved April 15, 2010, from http://www.abo.fi/~jheinstr/thesis.htm.

[35] M. Barifah, and M. Landoni, "Exploring usage patterns of a large-scale digital library". In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 67-76). IEEE, 2019.

[36] F. Ndumbaro, "Understanding user-system interactions: An analysis of OPAC users' digital footprints", Information Development, 34(3), 297-308, 2018.

[37] O. Raphaeli, A. Goldstein, and L. Fink, "Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach", *Electronic commerce research and applications*, *26*, 1-12, 2017.