Mining Web User Behavior: A Systematic Mapping Study

Nail Taşgetiren(✉) and Mehmet S. Aktas

Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey
nail.tasgetiren@std.yildiz.edu.tr, aktas@yildiz.edu.tr

Abstract. Nowadays, the number of people using the internet online increases day by day. Therefore, there is a growing need to analyze user behavior trends using navigational clickstream data these days. User behavior can be defined as the collection of the user's click sequence and transition movements between pages within site. Application areas of user behavior play an important role in many areas, from e-commerce to the banking sector. User behavior is a good argument for analyzing users' interests. Although there are many research studies in the literature related to user behavior mining, we could not identify a comprehensive literature review on this topic. This paper gives a systematic map of the analysis and mining of web domain users' activity in this research. We give exact user behavior measures, and publications that we anticipate will help other researchers discover study gaps and possibilities in this paper.

Keywords: Web mining · User Behavior Analysis · Clickstream Analysis · Systematic Literature · Literature Review

1 Introduction

The main purpose of mining web user behavior is to analyze and determine which aspects of the user behavior's results are most concerning to the user. Web usage data provides insight into the design of any website in terms of user behavior [9,11]. More efficient content can be suggested by analyzing users' navigation clickstreams and classifying them according to their interests. Therefore, mining web user behavior is an essential task in understanding the behavior of users. Olmezogullari and Aktas describe user behavior as "The navigational webpage patterns exhibited by clickstream sequences can be clustered to identify user activity." [15]. That's why collecting and clustering user behaviors are essential for useful mining results [10].

Over the last few years, for transferring information, the WWW has become increasingly popular and user-friendly [4,21] and the number of people who use the Internet has increased. In December 2021, Internet users were about 5 billion and the In 2022 will about 8 billion. This increase in the use of the Internet makes examining the behavior of users more critical. As a result, clickstream data is frequently noisy, and finding an optimum representation might be difficult.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 667–683, 2022.
https://doi.org/10.1007/978-3-031-10536-4_44

The importance of user behavior defined as "Understanding user behavior can dramatically improve a users' experience, either through better performance, customized user interface features, or better targeted ads." [25]. Therefore, user behavior attracts attention in the academic community and the private sector. Figure 1 shows the distribution of articles related to "web "user behavior" mining for the last five years. As seen in Fig. 1, the number of "Web "User Behavior" Mining" searches increases over time.

We discovered a survey study for web usage mining [17] and another survey for predicting user behavior [2] when searching the literature for secondary studies. Despite the topic's importance and popularity (Fig. 1), we were unable to locate any literature review studies on "Web "User Behavior" Mining" in general.

In this article, by examining the studies on the subject of mining web user behavior, these studies have not been studied before or which subjects have been studied more statistically. In line with this statistical information, we present the results on subject mining web user behavior to facilitate future research directions. These results provide academics with statistical data, enabling them to explore unstudied or understudied areas of mining web user behavior.

Following is the rest of the article: Sect. 2: Systematic Mapping Process explains how we conducted this literature mapping analysis. Section 3 presents synthesis data from primary investigations to answer the research questions. Section 4 analyzes the dangers to the mapping study's validity, and Sect. 5 wraps up the research and looks ahead to future projects.

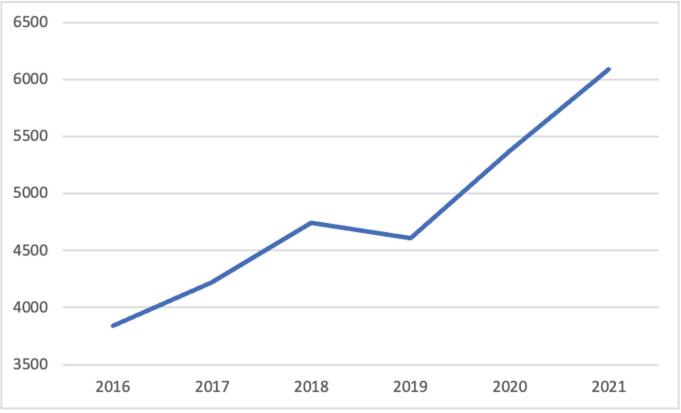

Fig. 1. Number of articles by year for "Web "User Behavior Mining".

2 Systematic Mapping Process

A preliminary study is primary empirical work in a particular field. On the other hand, a secondary study is defined as "a research that analyses all primarily related studies to a certain research issue to integrate/synthesize knowledge related to that research question." [7].

In our secondary investigation, we used a systematic mapping study method [7,19] on Mining Web User Behavior: A Systematic Literature Review.

2.1 Definition of Research Scope

To determine the body of knowledge regarding linked literature, we posed simple questions. The study questions and their preliminary research goals are listed in Table 1.

Table 1. Research questions

Research questions	Main motivation
RQ-1: What is the primary goal, necessity and basic process of mining user or customer behavior?	The main purpose is to analyze the clickstream and behavior of users or customers in the web environment
RQ-2: What research concerns in web user behavior mining are covered, and how many publications are there to cover them all?	Determine which research fields are frequently mentioned in articles about mining web user behavior mining
RQ-3: What are the most common used data types and methods for web user behavior mining?	Provide data and method for the most often studied web user behavior mining
RQ-4: Which embedding methodologies are used in the previous studies?	Determine which research embedding methodologies are frequently mentioned in articles
RQ-5: What is the paper-publication frequency?	Provide a graph with trend analysis
RQ-6: Which journals contain papers on mining web user behavior?	Determine which journals researchers should submit their research to

Figure 2 shows the basic map of the systematic literature review. The main purpose of this SLR is to identify Mining Web User Behavior methods, framework and data types used in Mining Web User Behavior.



Fig. 2. Basic mind map of the SLR

Table 2. Search strings

(Web)
AND
(User or Customer)
AND
(Behavior or Pattern)
AND
(Mining)

2.2 Search for Primary Studies

As part of a literature evaluation, we explored digital databases for preliminary investigations [19]. We chose the keywords web, user, behavior, and mining, and then created search strings using their synonyms (Table 2).

We applied the identical search strings to four digital libraries and gathered the outcomes without time constraints after settling on search strings (Table 3). We did not limit the search by year to track the topic's rise in popularity. As a result, we received the findings, released by February 2022. As a result, studies published after February 2022 were excluded. Downloading, maintaining, and working on a collection of documents are time-consuming operations that cannot be accomplished using spreadsheet apps alone. As a result, we deployed various reference management solutions, including Zotero. We used Zotero to handle all of the research we obtained, integrate them, apply the filter, apply exclusion criteria, and download references automatically from digital databases.

Table 3. Search results in digital libraries

No	Name	Result
1	IEEE	2.316
2	ACM	209
3	Science Direct	417
4	Springer Link	749
	Total	**3.691**

To avoid duplicate research acquired from different digital resources, we gathered search results and consolidated them. To find primary studies, we used three filters (Fig. 3). Table 4 lists the filters we used to narrow the search results (3691 papers). We got 86 primary studies at the end of the filtering procedure.

Table 4. Filters applied to search results

Filters	Definition
First filter	We omitted research from irrelevant journals and conferences
Second filter	To exclude papers that were not relevant, we examined the study title and abstract
Third filter	We read abstracts and titles, and in certain cases, whole papers, to apply our inclusion and exclusion criteria

2.3 Inclusion and Exclusion Criteria

We must choose some papers to answer the research questions in systematic mapping studies. In order to ensure that the research selection criteria is objective, we utilize inclusion and exclusion criteria. These criteria are established at the start of the selection process to eliminate selection bias [3,13].

We used these criteria to restrict two results to select our preliminary studies.

Inclusion Criteria: Studies that conclusively contributed to the mining of web user behavior were included. We considered short papers also traditional study papers. We accepted the new version of a study with multiple versions unless there was an expansion to main study. Therefore, both versions were included.

Exclusion Criteria: We did not include research that was not clearly linked to mining web user behavior or did not look into the matter thoroughly. We also did not include duplicate publications. Book chapters, secondary research, theses, and experimental investigations that did not suggest a solution or method were rejected.

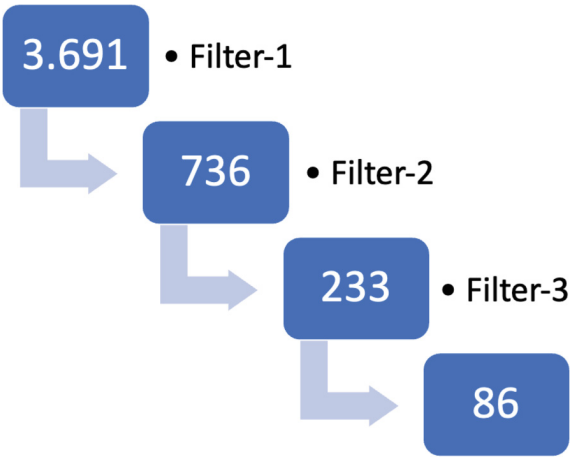

Fig. 3. The change in the number of main papers with the application of the filters in Table 4

2.4 Classification Schema

With a view of mining web user behavior as application level logs, application server logs and web server logs, we gathered 86 studies that we obtained after the filtering procedure. Figure 4 shows how we classified studies.

Web data mining is classified into three groups based on the different data mining objects: Weblog mining, web content mining, and web structure mining [26]. Our interest is Weblog mining. Web Usage Mining uses data mining techniques to find exciting usage patterns in Web content to understand better and satisfy Web-based applications' demands [12]. The practice of discovering browsing trends by studying a user's navigational behavior is known as web usage mining. This information takes as input usage data, which is data stored in web server logs that record user visits to a website [6]. The type of usage data being analyzed might further categorize web usage mining:

Web Server Logs: User information is logged by the Web server. These data can be the User's IP, page reference, and access time to the Web server [8].

Application Server Logs: A critical feature is the ability to track various business events and log them in application server logs.

Application Level Logs: It is the logging of the user's clicks and transitions between pages within the application. These logs can be categorized by dividing them into related categories. It may be necessary to use combinations of these logs in order to obtain more meaningful logs.

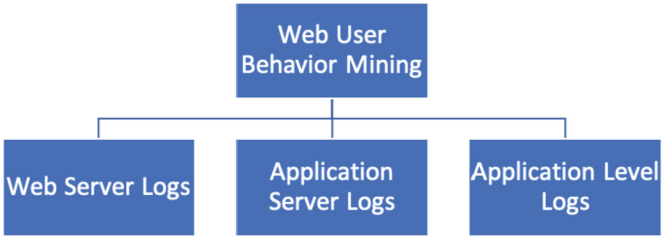

Fig. 4. Classification framework

We also classified in Table 5 the research issues as well as the contributions of the researchers.

Table 5. Research issues

User Behavior Analysis	The primary goal of the user's visit record analysis is to determine which aspects of the mining results are most concerning to the user [10]
Web (Usage or Data) Mining	Web usage mining uses data mining techniques to find patterns in Web data to understand better and service the demands of Web-based applications [23]
Clickstream Analysis	The clickstream data are made up of a jumble of misleading navigational URLs with no annotation information about their true identity, making them useless for any analytical activities [16]
Predict and Analysis User Behavior	One can remodel a website, portal, or e-commerce along with the behavior of online users by evaluating and predicting people's navigation patterns and their relationship with web content [22]

2.5 Data Extraction and Mapping of the Literature

Subsequently picking the preliminary study articles, we obtain the essential information from the studies to examine them. Following the creation of the data extraction form in Table 6, we referenced this form for the analysis procedures. If the content of an article is not fully understood or does not contain relevant information for our topic, we have omitted this section in order not to make any assumptions.

The following is our data extraction methodology: We chose a small dataset at random before beginning the data extraction process for 86 studies. Then, using Table 6 as a guide, we extracted data. We double-checked for the same data set. We produced this research to inspect a conflict and determine what created the conflict when we uncovered it. We used the data extraction technique to 86 papers to get the findings after satisfying all of these matters on the dataset.

Table 6. Form for data extraction

No	Data extraction columns
No-1	Study ID (example: S-1)
No-2	Title: Paper's title (example: Mining Web User Behavior: A Systematic Literature)
No-3	Author: Author's name
No-4	Year: Publication year (example: 2021)
No-5	Conference or journal name: Any conference or journal name (ex: IEEE Data Mining 2021)
No-6	Research Issues: User Behavior Analysis, ClickStream Analysis
No-7	Usage of Data Type (Application Level Logs)
No-8	Approach (Semi-supervised approach)
No-9	Research Method: Theory/Survey/Experiment/Own Experience/Review

3 Results

In this chapter, the study's findings are given and discussed the goal. The study's research questions were answered, and the results were given.

3.1 RQ-1: What is the Primary Goal, Necessity and Basic Process of Mining User or Customer Behavior?

User profiles can be used to summarize a user's enormous amount of data and achieve goals like product suggestions and personalized knowledge distribution. Users can contribute explicit information through registration and surveys, which can be included in a user profile [1,20]. Explicit information can also include simple facts about consumers' behaviors or transactions. For instance, such data could comprise;

- the number of times a visitor views a website.
- the average amount paid on each purchase, and
- the product category with the most sales.

The information listed above can process the user's data and prepare a better website usage experience for the user. A user browsing an e-commerce site may be presented with different products that he may need based on the products user's clicks.

One of the most important necessity for mining user behavior is a well-prepared data set. As the variety and range of the data set expands, the experience offered to the user increases.

The Fig. 5 shows that a basic level graph of the web mining flow to analyze user behavior.

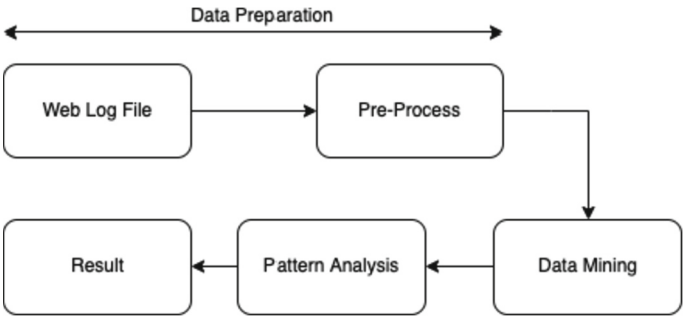

Fig. 5. Basic process of web mining user or customer behavior

3.2 RQ-2: Which Research Issues in Web User Behaviour Mining are Addressed and How Many Papers Cover the Different Research Issues?

We've covered a number of study topics connected to Mining Web User Behavior. The findings of our literature review are shown in Fig. 6. A total of 79 papers are grouped into four (4) different research issues after thorough evaluation and application of inclusion-exclusion criteria.

RQ2 aims to find new ways to address unsolved research problems. According to the findings, the top two topics addressed by the research community are user behavior analysis and web (usage or data) mining. Clickstream analysis and predicting and analyzing user behavior, on the other hand, are the most challenging difficulties. As a result, we consider clickstream analysis and predicting and analyzing user behavior the most promising research areas in the coming years.

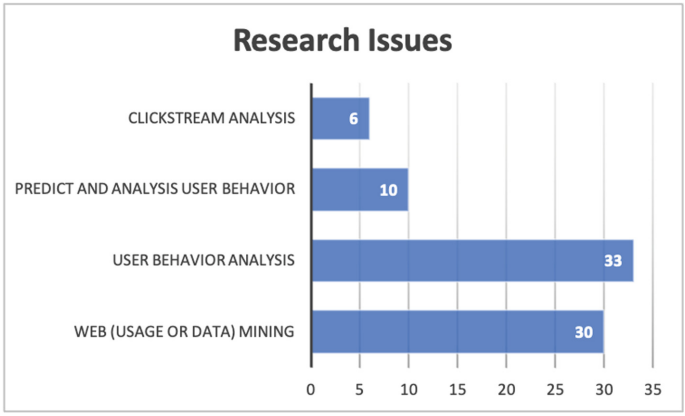

Fig. 6. Publications are distributed according to research topics.

3.3 RQ-3: What are the Most Common Used Data Types and Methods for Web User Behavior Mining?

When a Web user sends an HTTP request, the movement information is recorded in a Weblog file. A log file can usually be located in one of three locations: Application Level Logs, Application Server Logs and Web Server Logs.

The data types from the publications that remained after the most recent extractions were evaluated and classified using the three criteria listed above. The Fig. 7 shows the distribution of papers on the studied data. According to the result obtained from this graph, it has been observed that the Application Level Logs can be used more and more valuable and usable results are obtained through this data type.

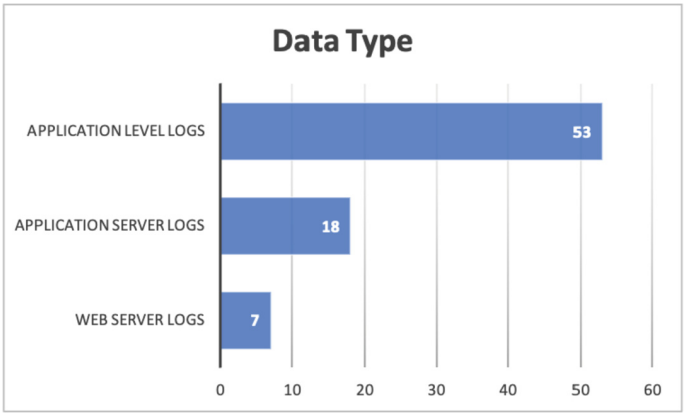

Fig. 7. Data types

The types of methods used in the research literature differ on paper. The methods used to measure user behavior are listed in the Table 7. The diversity of methods indicates that the research perspective is inclusive. This diversity shows that more papers will be added to the literature with different methods.

Hidden Markov Models (HMM) are used by Erdem et al. [5] to map topic-based unseen data sequences to URL sequences.

Uygun et al. [24] used both partitioning-based and hierarchical clustering approaches in the clustering phase of their studies. So they use K-means or Bisecting K-means, respectively.

Table 7. Method types

Support vector machines	DBSCAN&Chameleon
Kohanen Neural Network	Random Forest
LOGSOM (Log Self-Organizing Map)	Visual Attention Map(CNN)
Pattern-tree	SSPRA (Single Scan Pattern Recognition Algorithm)
Ant-Based Clustering	FP-Growth
K-Means	Apriori Algorithm
K Nearest Neighbor	Markov chain model
Matrix Decomposition	Frequent Pattern Analysis
M5 Model Tree	Web Map Service Platforms (WMSPs)
Word2Vec	Unsupervised Approach
Node2Vec	Semisupervised Approach
Fluxicon Disco	Cosine Similarity Measure
Pattern2Vec	Frequent Link and Access Tree (FLaAT)
LDA Model	ID3 algorithm
DBSCAN	Hidden Markov Model (HMM)
UpDown Tree for Finding Contiguous Sequential Patterns (CSP)	SMAP-Mine
ProCrawl (PROcess CRAWLer)	Markov Model with the Pre-Order Linked WAP-Tree Mining (PLWAP) Algorithm
Assoc. Rule Appr. (ARA) vs Cogn. Map Appr. (CMA)	Probabilistic Latent Semantic Analysis (PLSA)

3.4 RQ-4: Which Embedding Methodologies are Used in the Previous Studies?

In this chapter; we discovered that which embedding methodologies are used literature has exploded in recent years. We divided the embedding methodologies into two (2). These are; URL Embedding Methodologies and URL Sequence

Embedding Methodologies. The data in Fig. 8 show the embedding methodologies. Table 8 shows that list of the URL and URL Sequence Embedding Methodologies.

Oguz et al. [14] worked on predicting users' behavior on the e-commerce site based on clickstream sequences data in this study. This study discovered new datasets using a GAN (Generative Adversarial Networks)-based deep learning algorithm. They have developed a new prototype to demonstrate performance tests. They chose URL Sequence Embedding Methodology as their reference.

Oz et al. [18] worked on automating test scripting using LSTM (Long Short Term Memory). As the data set, they chose clickstream sequences data. They chose URL Sequence Embedding Methodology as their reference.

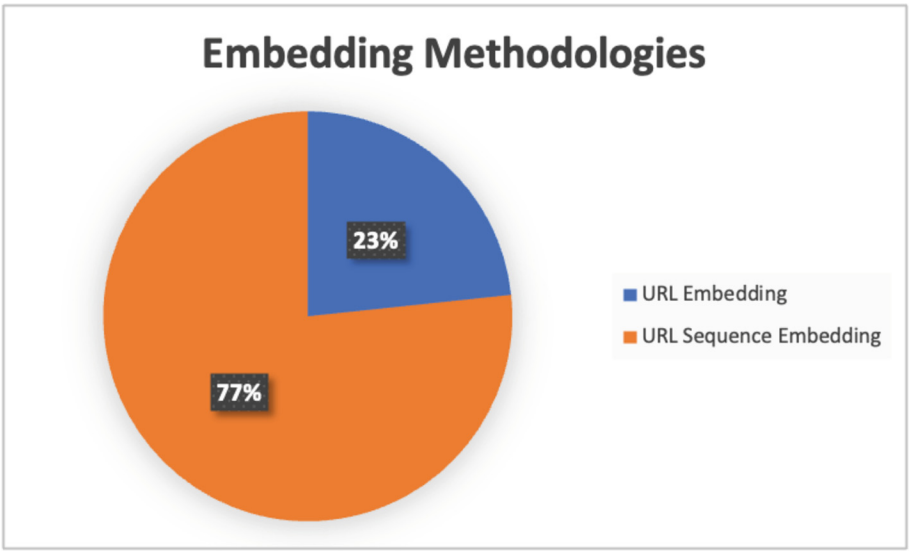

Fig. 8. Embedding methodologies

Table 8. List of embedding methodologies

No	Data extraction columns
URL	word2vec
URL	url2vec
URL Sequence	graph2vec
URL Sequence	DeepWalk
URL Sequence	subgraph2vec
URL Sequence	node2vec
URL Sequence	pattern2vec
URL Sequence	struc2vec

3.5 RQ-5: What is the Paper-Publication Frequency?

In this chapter; we discovered that mining web user behavior literature has exploded in recent years, owing to the widespread use of computers and other mobile phones in everyday (Fig. 9). The data in Fig. 9 show the overall number of publications per year. We believe that there is a relation between the maturity of the topic domain and this.

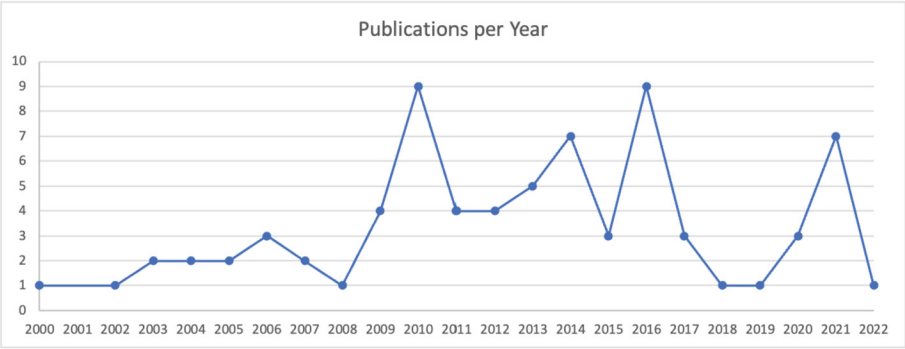

Fig. 9. Yearly publication frequency

3.6 RQ-6: Which Journals Contain Papers on Mining Web User Behavior?

Names of Journals are listed in Table 9.

Table 9. Journal names

Journal no	Names of journals
1	Concurrency and Computation: Practice and Experience
2	Journ. of Retailing and Consumer Services
3	Engineering Applications of Artificial Intelligence
4	Decision Support Systems
5	Materials Today: Proceedings
6	Inf. and Software Tech.
7	International Journ. of Circuits, Systems and Signal Processing
8	ACM Transactions on the Web
9	Procedia Engineering
10	SIGKDD Explor. Newsl.
11	SIGSOFT Softw. Eng. Notes
12	Physical Review E
13	Data in Brief

(*continued*)

Table 9. (*continued*)

Journal no	Names of journals
14	MethodsX
15	Procedia Computer Science
16	Information Processing and Management
17	Int. Journ. of Engineering Research
18	Expert Systems with App
19	Journ. of Systems and Software
20	Int. Journ. of Computer Applications
21	Int. Journ. of Human-Computer Studies

Table 10 shows the Scimago Journal Rank (SJR) values. Journal publications are ordered according to their SJR value.

Table 10. SJR of selected journals

Journal Publications	SJR
Journal of Retailing and Consumer Services	1,57
Decision Support Systems	1,56
Expert Systems with Applications	1,37
Engineering Applications of Artificial Intelligence	1,11
Information Processing and Management	1,06
Physical Review E	0,9
Int. Journ. of Human-Computer Studies	0,73
Journ. of Systems and Software	0,64
Inf. and Software Technology	0,61
ACM Transactions on the Web	0,44
MethodsX	0,36
Materials Today: Proceedings	0,34
Procedia Computer Science	0,33
Procedia Engineering	0,32
Concurrency and Computation: Practice and Experience	0,31
Int. Journ. of Computer Applications	0,29
Int. Journ. of Engineering Research	0,24
Int. Journ. of Circuits, Systems and Signal Processing	0,16
Data in Brief	0,12
SIGKDD Explor. Newsl.	0,00
SIGSOFT Softw. Eng. Notes	0,00

4 Threats to Validity of Research

Generally, some situations threaten the general validity of mapping studies. These; research questions, publication selection, and data extraction in selected publications. We have our suggestions for removing these three risks.

Research Questions: Research questions may not cover all mining web user behavior issues in detail. We searched only a portion of the literature on the mining web user behavior issue at a particular time. As a result, this literature study did not focus on a particular subject.

Publication Selection: Despite searching four big digital databases, we could not apply our criterion to all web user behavior mining topic. While we cannot eliminate the threat, we have attempted to decrease it by including as many publications.

Data Extraction: Because personal perspectives can be significant in the data extraction process, the data were double-checked to mitigate the risk, and the data were analyzed from a subjective standpoint. When an inconsistency was found, the data set was re-evaluated. It was included in the research was related to our topic.

5 Conclusion and Future Work

The literature review aimed to identify and analyze trends, datasets and methods in web mining user behavior. We conducted a literature review covering the years 2000-2021. We found 3691 published papers between these years. After applying the criterias for these papers. Finally, we analyzed and worked 86 papers.

By examining the outcomes in Fig. 6 and Fig. 7, we conclude that web mining user behavior is generally open to new contributions. Especially the studies on clickstream analysis show the lack of research studies in this area. This shortcoming may provide further research opportunities. The results also demonstrate the intensive research needs of clickstream analysis and predict and analyze user behavior research topics using application-level logs datasets. While analyzing the data, we see that prior studies have placed a high value on user behavior analysis.

This study is continued by expanding it to make a systematic literature review on the subject of web mining user behavior. We will also explore how to design a framework for clickstream analysis using application-level logs data in the future.

Acknowledgements. This study was supported by Getir Perakende Lojistik A.ş. company. We would like to thank Getir Perakende Lojistik A.Ş. for providing us with the necessary resources and funding.

References

1. Ahn, J., et al.: Open user profiles for adaptive news systems: help or harm? In: Proceedings of the 16th international conference on World Wide Web, pp. 11–20 (2007)
2. Anitha, V., Isakki, P.: A survey on predicting user behavior based on web server log files in a web usage mining. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE 2016), pp. 1–4. IEEE (2016)
3. Budgen, D., Brereton, P.: Performing systematic literature reviews in software engineering. In: Proceedings of the 28th International Conference on Software Engineering, pp. 1051–1052 (2006)
4. Dixit, V.S., Bhatia, S.K.: Refinement and evaluation of web session cluster quality. Int. J. Syst. Assurance Eng. Manage. **6**(4), 373–389 (2014). https://doi.org/10.1007/s13198-014-0266-x
5. Erdem, I., et al.: Test script generation based on hidden markov models learning from user browsing behaviors. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 2998–3005. IEEE (2021)
6. Kalaichelvi, S.: Web mining classification: a survey. Int. J. Eng. Res. Technol. (IJERT) **53**(3), 1–15 (2014)
7. Keele, S., et al.: Guidelines for performing systematic literature reviews in software engineering. Technical report, Ver. 2.3 EBSE Technical Report. EBSE (2007)
8. Lei, X.: Modeling and intelligent analysis of web user behavior of web user behavior. In: 2018 International Conference on Engineering Simulation and Intelligent Control (ESAIC), pp. 192–195. IEEE (2018)
9. Liao, Z., et al.: Task trail: an effective segmentation of user search behavior. IEEE Trans. Knowl. Data Eng. **26**(12), 3090–3102 (2014)
10. Luo, X., et al.: User behavior analysis based on user interest by web log mining. In: 2017 27th International Telecommunication Networks and Applications Conference (ITNAC), pp. 1–5. IEEE (2017)
11. Mahajan, R., et al.: Usage patterns discovery from a web log in an Indian e-learning site: a case study. Educ. Inf. Technol. **21**(1), 123–148 (2016)
12. Nguyen, S.T.: Efficient web usage mining process for sequential patterns. In: Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, pp. 465–469 (2009)
13. Novais, R.L., et al.: Software evolution visualization: a systematic mapping study. Inf. Softw. Technol. **55**(11), 1860–1883 (2013)
14. Oguz, R., et al.: Extracting information from large scale graph data: Case study on automated ui testing (2021)
15. Olmezogullari, E., Aktas, M.S.: Representation of click-stream datasequences for learning user navigational behavior by using embeddings. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 3173–3179. IEEE (2020)
16. Olmezogullari, E., Aktas, M.S.: Pattern2vec: representation of clickstream data sequences for learning user navigational behavior. Concurrency Comput. Practice Exp. **34**(9), e6546 (2022)
17. Omar, R., et al.: Web usage mining: a review of recent works. In: The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M), pp. 1–5. IEEE (2014)
18. Oz, M., et al.: On the use of generative deep learning approaches for generating hidden test scripts. Int. J. Software Eng. Knowl. Eng. **31**(10), 1447–1468 (2021)

19. Petersen, K., et al.: Systematic mapping studies in software engineering. In: 12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12, pp. 1–10 (2008)
20. Raghu, T., et al.: Dynamic profiling of consumers for customized offerings over the internet: a model and analysis. Decis. Support Syst. **32**(2), 117–134 (2001)
21. Romero, C., et al.: Applying web usage mining for personalizing hyperlinks in web-based adaptive educational systems. Comput. Educ. **53**(3), 828–840 (2009)
22. Sılahtaroğlu, G., Dönertaşli, H.: Analysis and prediction of e-customers' behavior by mining clickstream data. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 1466–1472. Ieee (2015)
23. Srivastava, J., et al.: Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsl. **1**(2), 12–23 (2000)
24. Uygun, Y., et al.: On the large-scale graph data processing for user interface testing in big data science projects. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 2049–2056. IEEE (2020)
25. Wang, G., et al.: Unsupervised clickstream clustering for user behavior analysis. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 225–236 (2016)
26. Yin, B., et al.: Research and application of data mining technology used in the analysis of smart home user behavior. In: 2014 Sixth International Conference on Measuring Technology and Mechatronics Automation, pp. 476–479. IEEE (2014)