Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Database-Specific Keyword Frequency Analysis in Merged Web Log Data: A Preprocessing Method

Wan Hussain Wan Ishak[1 ,*], *Nurul Farhana Ismail*[2], *Fadhilah Mat Yamin*[3]*Abdullah Husin*[4]

[1] *School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
[2] *School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
[3] *School of Technology Management & Logistic, Universiti Utara Malaysia, Sintok, Kedah, Malaysia*
[4] *Department of Information System, Faculty of Engineering and Computer Science, Universitas Islam Indragiri Indonesia*
email: [1,*]*hussain@uum.edu.my,* [2]*nurulfana.ismail@gmail.com*, [3]*fmy@uum.edu.my*
** Correspondence*

ARTICLE INFO

Article history:
Received January 20, 2024
Revised February 22, 2024
Accepted February 25, 2024
Available online February 28, 2024

Keywords:
Data Preprocessing, Web Log Analysis, Electronic Resources Module, Keyword Extraction, Database Usage Patterns

Please cite this article in IEEE style as:
W.H.W. Ishak, et al., , " Database-Specific Keyword Frequency Analysis in Merged Web Log Data: A Preprocessing Method" *Data Science Insights*, vol. 13, no. 2, pp. 45-52, 2024.

ABSTRACT

This study delves into the complex intricacies of web log data within the Electronic Resources module of the Perpustakaan Sultanah Bahiyah (PSB) website at Universiti Utara Malaysia (UUM). Serving as a cornerstone of academic infrastructure, the Electronic Resources module acts as a vital gateway, seamlessly connecting the UUM academic community to a vast repository of scholarly information. However, the search process is often tedious, as users must browse each repository separately without prior knowledge of which repository contains the desired information. Studying the web log is one approach to revealing historical search patterns, which can guide new users towards the most relevant repository for their needs. However, the size and complexity of the web log present challenges in this study. Therefore, meticulous preprocessing methods are employed, involving the restructuring of raw data, outlier cleaning, and user session identification, laying the foundation for a comprehensive analysis. The study further explores the identification of search keywords embedded in the log file, employing a systematic process that transforms data into a structured format. Subsequent extraction of databases and keywords yields intriguing findings, prominently highlighting IEEE and Serial Solution databases. The analysis of 19,146 keywords associated with 11 databases offers valuable insights into user behaviour, preferences, and the overall effectiveness of the Electronic Resources module. Identifying frequent keywords not only provides analytical insights but also accelerates users' search processes, reducing cognitive load and fostering a more efficient research experience. This research contributes to optimising user experiences and refining digital library services, aligning them with the evolving needs of the academic community.

Correspondence:
Wan Hussain Wan Ishak
School of Computing
Universiti Utara Malaysia,
Sintok, Kedah, Malaysia
hussain@uum.edu.my

1. Introduction

In the digital landscape of academic institutions, the utilization of electronic resources (e-Resources) has become integral to scholarly pursuits. This study focuses on web log data from the Electronic Resources module in the website of Perpustakaan Sultanah Bahiyah (PSB), Universiti Utara Malaysia (UUM).

The Electronic Resources module serves as a crucial conduit, providing access to a wealth of scholarly information tailored to meet the diverse needs of the academic community at UUM. Specifically designed for registered PSB users, it facilitates seamless access to a broad range of subscribed databases and electronic resources. Operating as a virtual portal, it effectively bridges the gap between scholars and academic content, empowering research, facilitating journal access, and supporting database exploration. Together, these features enhance the overall academic experience for the UUM community.

Understanding user interactions within this module is pivotal given the dynamic nature of digital scholarship and the myriad electronic resources available. This study delves into web log data complexities, unraveling patterns and trends among registered users' navigation, engagement, and utilization of resources facilitated by the UUM Library. Focusing on this module, the research contributes subtle insights into user behavior, preferences, and the overall efficacy of the Electronic Resources module as a gateway to online databases and scholarly content.

This study also holds promise for practical outcomes that enhance the user experience and streamline scholarly pursuits. A paramount advantage lies in identifying frequent keywords derived from web log data. Scrutinizing patterns within user queries, this research aims to unearth keywords commonly employed by registered users during searches.

In data-intensive environments, the role of data analytics is paramount for informed decision-making [1]. The identification of frequent keywords goes beyond providing analytical insights, delivering tangible advantages to users exploring extensive databases and electronic resources. This importance is underscored by the fact that keywords serve as vital connectors between the content of webpages and user inquiries [2]. Moreover, discerning frequent keywords contributes to a reduction in users' cognitive load. Optimized search functionalities spare individuals from exhaustive database browsing, resulting in a more user-friendly and intuitive exploration of electronic resources. This approach results in swifter access to pertinent information [3], ultimately saving users time and enhancing the efficiency of their research experiences.

2. Literature Review

Data mining, a fundamental aspect of data science, is a multidisciplinary field that extracts valuable insights, patterns, and knowledge from extensive datasets. Operating as a pivotal step within knowledge discovery in databases (KDD), data mining employs algorithms to reveal meaningful patterns and relationships within large data repositories [4].

Web mining is an application of data mining that utilizes various data mining techniques to explore and extract valuable information from the vast expanse of the World Wide Web [5]. This concept plays a pivotal role in facilitating navigation, search, and visualization of web content. Web mining is categorized into three major types [6]: web content mining, web structure mining, and web usage mining. Web content mining involves extracting useful information from diverse types of web content, aiding in classification and pattern discovery. Web structure mining focuses on extracting knowledge from hyperlinks that represent the web's structure. Lastly, web usage mining observes user access patterns on the internet, providing valuable insights for organizations studying user behavior on a large scale.

Methods for extracting patterns from web access log data can be classified into three main categories, that are; pre-processing, pattern discovery tools, and pattern analysis tools [7]. The pre-processing phase of Web Usage Mining involves data cleaning, user identification, session identification, path completion, session reconstruction, transaction identification, and formatting [8]. Some researchers define additional tasks, such as conceptual hierarchies construction and URL classification [9;10]. The goal of the pre-processing phase is to provide a structural, valid, and integrated data source for the pattern discovery phase. Moreover, intelligent tasks, such as forecasting, necessitate clean and preprocessed data to ensure accurate predictions [11].

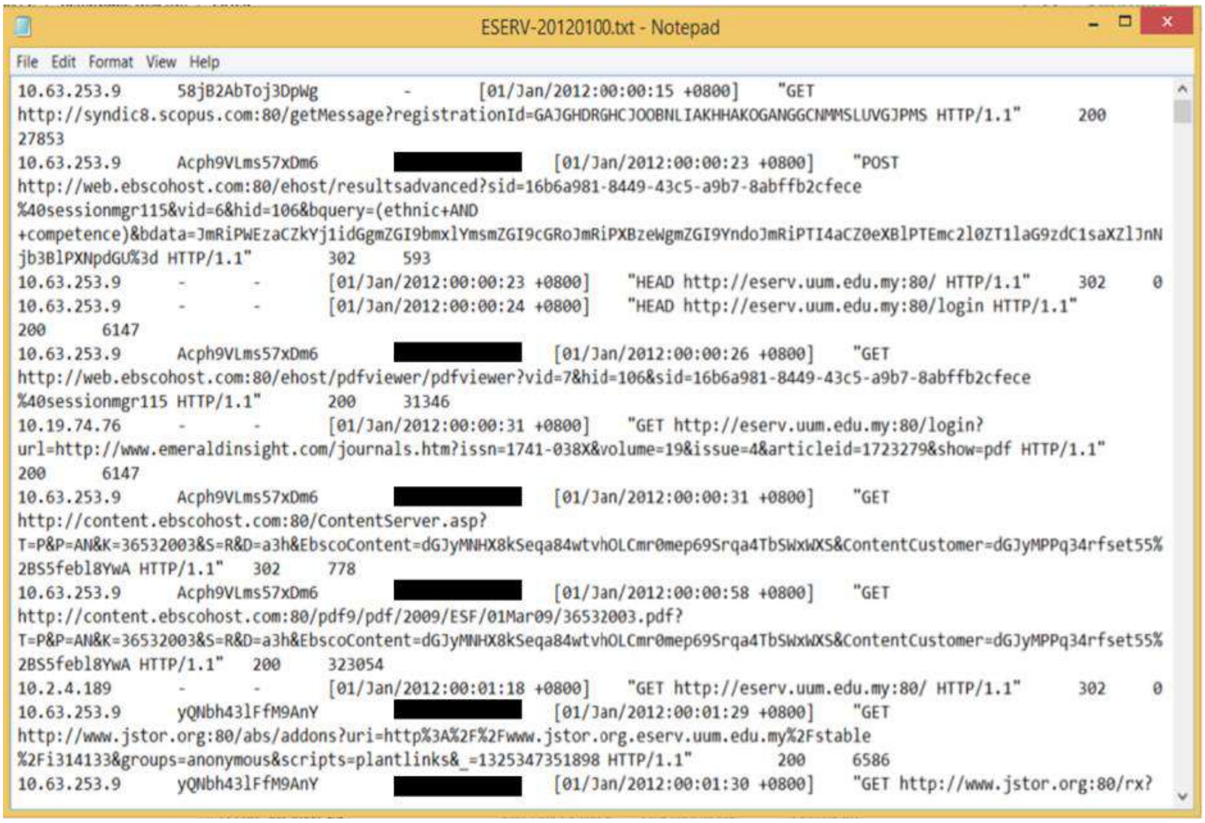
Thakare & Gawali [12] propose an effective pre-processing process, starting with data collection from primary sources like server log files. After data collection, field extraction is performed using Java code, followed by data merging and sorting. Data cleaning removes unnecessary information, and in the conversion phase, the TransLog algorithm is employed to convert log files into Access or Oracle tables. The session identification phase then reconstructs clickstream data into the actual sequence of actions performed by a user during a site visit.

Deepa & Raajan [13] implemented pre-processing techniques to convert log files into user sessions suitable for mining. This process also reduces the size of the session file by filtering the least requested pages. Pattern discovery, a crucial phase in web usage mining, involves extracting behavioral patterns from formatted data [14]. Elhiber & Abraham [15] highlight various methods for pattern discovery, with clustering, association rules, and statistical analysis being the most frequently used.

Advancements in clustering methods include Aghabozorgi & Wah's [16] hybrid approach, leveraging usage data and data domain for recommendation design, and Alphy & Prabakaran's [17] Ant ClusterTrack algorithm for cluster optimization using ants' nest mate identification techniques. Association rules have seen improvements with Islam & Chung's [18] improved FP tree and algorithm, providing efficient mining of all possible frequent item sets without generating the conditional FP tree. Statistical analysis continues to be a robust method for knowledge extraction, aiding in improving system performance and security, as well as facilitating site modification [19;20]. Overall, these advances underscore the ongoing refinement and application of web usage mining techniques.

3. Method

This study focuses on consolidating web log data encompassing users' search activities across a collection of databases owned or operated by various service providers. The challenges encountered in this research primarily revolve around the size and complexity of the data. The log file, rich with attributes like IP addresses and URLs, poses a valuable resource. However, the data is currently disorganized and irregular, necessitating restructuring before any subsequent processing can take place. Example of the raw data is presented in Figure 1.

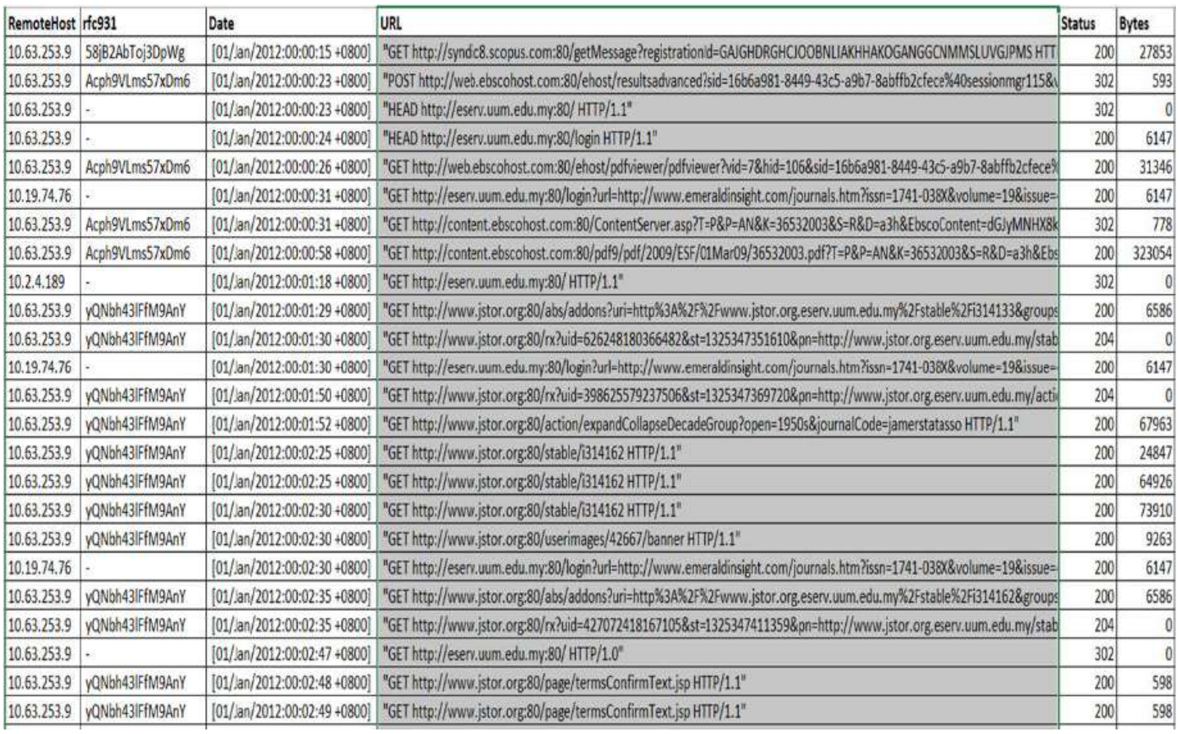


```
10.63.253.9      58jB2AbToj3DpWg          -        [01/Jan/2012:00:00:15 +0800]     "GET
http://syndic8.scopus.com:80/getMessage?registrationId=GAJGHDRGHCJOOBNLIAKHHAKOGANGGCNMMSLUVGJPMS HTTP/1.1"      200
27853
10.63.253.9      Acph9VLms57xDm6     [redacted]     [01/Jan/2012:00:00:23 +0800]     "POST
http://web.ebscohost.com:80/ehost/resultsadvanced?sid=16b6a981-8449-43c5-a9b7-8abffb2cfece
%40sessionmgr115&vid=6&hid=106&bquery=(ethnic+AND
+competence)&bdata=JmRiPWEzaCZkYj1idGgmZGI9bmxlYmsmZGI9cGRoJmRiPXBzeWgmZGI9YndoJmRiPTI4aCZ0eXBlPTEmc2l0ZT1laG9zdC1saXZlJnN
jb3BlPXNpdGU%3d HTTP/1.1"        302     593
10.63.253.9      -        -        [01/Jan/2012:00:00:23 +0800]     "HEAD http://eserv.uum.edu.my:80/ HTTP/1.1"      302     0
10.63.253.9      -        -        [01/Jan/2012:00:00:24 +0800]     "HEAD http://eserv.uum.edu.my:80/login HTTP/1.1"
200      6147
10.63.253.9      Acph9VLms57xDm6     [redacted]     [01/Jan/2012:00:00:26 +0800]     "GET
http://web.ebscohost.com:80/ehost/pdfviewer/pdfviewer?vid=7&hid=106&sid=16b6a981-8449-43c5-a9b7-8abffb2cfece
%40sessionmgr115 HTTP/1.1"       200     31346
10.19.74.76      -        -        [01/Jan/2012:00:00:31 +0800]     "GET http://eserv.uum.edu.my:80/login?
url=http://www.emeraldinsight.com/journals.htm?issn=1741-038X&volume=19&issue=4&articleid=1723279&show=pdf HTTP/1.1"
200      6147
10.63.253.9      Acph9VLms57xDm6     [redacted]     [01/Jan/2012:00:00:31 +0800]     "GET
http://content.ebscohost.com:80/ContentServer.asp?
T=P&P=AN&K=36532003&S=R&D=a3h&EbscoContent=dGJyMNHX8kSeqa84wtvhOLCmr0mep69Srqa4TbSWxWXS&ContentCustomer=dGJyMPPq34rfset55%
2BS5febl8YwA HTTP/1.1"   302     778
10.63.253.9      Acph9VLms57xDm6     [redacted]     [01/Jan/2012:00:00:58 +0800]     "GET
http://content.ebscohost.com:80/pdf9/pdf/2009/ESF/01Mar09/36532003.pdf?
T=P&P=AN&K=36532003&S=R&D=a3h&EbscoContent=dGJyMNHX8kSeqa84wtvhOLCmr0mep69Srqa4TbSWxWXS&ContentCustomer=dGJyMPPq34rfset55%
2BS5febl8YwA HTTP/1.1"   200     323054
10.2.4.189       -        -        [01/Jan/2012:00:01:18 +0800]     "GET http://eserv.uum.edu.my:80/ HTTP/1.1"      302     0
10.63.253.9      yQNbh43lFfM9AnY     [redacted]     [01/Jan/2012:00:01:29 +0800]     "GET
http://www.jstor.org:80/abs/addons?uri=http%3A%2F%2Fwww.jstor.org.eserv.uum.edu.my%2Fstable
%2Fi314133&groups=anonymous&scripts=plantlinks&_=1325347351898 HTTP/1.1"        200     6586
10.63.253.9      yQNbh43lFfM9AnY     [redacted]     [01/Jan/2012:00:01:30 +0800]     "GET http://www.jstor.org:80/rx?
```

Figure 1. Sample of raw data

The initial step of data preprocessing involves utilizing spaces to segregate each attribute in the log file into individual columns within an Excel sheet. Subsequent data cleaning procedures are then applied to filter out outlier data, such as entries containing file formats like JPEG, PNG, and PDF, along with data marked as failed. Records falling into these categories are discarded. Following this, in the user differentiation phase, a unique user ID is assigned to each IP address, facilitating user identification for further analysis, particularly in the session identification process. Each identified session is grouped by its unique user ID in the session clustering phase. The final refinement in this phase involves correcting data spelling during the data formatting process.

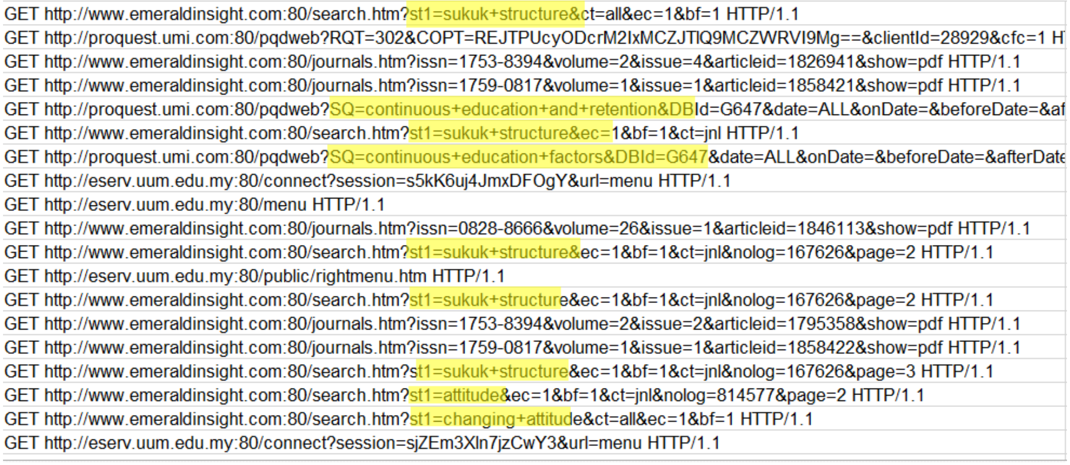
For the identification of search keywords within the log file, a transformation is applied to structure the log into a formatted layout, exemplified in Figure 2. Subsequently, only the URL column is retained for analysis. In this phase, irrelevant and outlier data are eliminated, specifically by discarding records with failed codes less than 200 or greater than 299. Following the completion of data cleaning, the extraction of database and keyword information is performed.



RemoteHost	rfc931	Date	URL	Status	Bytes
10.63.253.9	58jB2AbToj3DpWg	[01/Jan/2012:00:00:15 +0800]	"GET http://syndc8.scopus.com:80/getMessage?registrationId=GAJGHDRGHCJOOBNLIAKHHAKOGANGGCNMMSLUVGJPMS HTT	200	27853
10.63.253.9	Acph9VLms57xDm6	[01/Jan/2012:00:00:23 +0800]	"POST http://web.ebscohost.com:80/ehost/resultsadvanced?sid=16b6a981-8449-43c5-a9b7-8abffb2cfece%40sessionmgr115&v	302	593
10.63.253.9	-	[01/Jan/2012:00:00:23 +0800]	"HEAD http://eserv.uum.edu.my:80/ HTTP/1.1"	302	0
10.63.253.9	-	[01/Jan/2012:00:00:24 +0800]	"HEAD http://eserv.uum.edu.my:80/login HTTP/1.1"	200	6147
10.63.253.9	Acph9VLms57xDm6	[01/Jan/2012:00:00:26 +0800]	"GET http://web.ebscohost.com:80/ehost/pdfviewer/pdfviewer?vid=7&hid=106&sid=16b6a981-8449-43c5-a9b7-8abffb2cfece%	200	31346
10.19.74.76	-	[01/Jan/2012:00:00:31 +0800]	"GET http://eserv.uum.edu.my:80/login?url=http://www.emeraldinsight.com/journals.htm?issn=1741-038X&volume=19&issue=	200	6147
10.63.253.9	Acph9VLms57xDm6	[01/Jan/2012:00:00:31 +0800]	"GET http://content.ebscohost.com:80/ContentServer.asp?T=P&P=AN&K=36532003&S=R&D=a3h&EbscoContent=dGJyMNHX8k	302	778
10.63.253.9	Acph9VLms57xDm6	[01/Jan/2012:00:00:58 +0800]	"GET http://content.ebscohost.com:80/pdf9/pdf/2009/ESF/01Mar09/36532003.pdf?T=P&P=AN&K=36532003&S=R&D=a3h&Ebs	200	323054
10.2.4.189	-	[01/Jan/2012:00:01:18 +0800]	"GET http://eserv.uum.edu.my:80/ HTTP/1.1"	302	0
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:01:29 +0800]	"GET http://www.jstor.org:80/abs/addons?uri=http%3A%2F%2Fwww.jstor.org.eserv.uum.edu.my%2Fstable%2Fi314133&groups	200	6586
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:01:30 +0800]	"GET http://www.jstor.org:80/rx?uid=626248180366482&st=1325347351610&pn=http://www.jstor.org.eserv.uum.edu.my/stab	204	0
10.19.74.76	-	[01/Jan/2012:00:01:30 +0800]	"GET http://eserv.uum.edu.my:80/login?url=http://www.emeraldinsight.com/journals.htm?issn=1741-038X&volume=19&issue=	200	6147
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:01:50 +0800]	"GET http://www.jstor.org:80/rx?uid=398625579237506&st=1325347369720&pn=http://www.jstor.org.eserv.uum.edu.my/acti	204	0
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:01:52 +0800]	"GET http://www.jstor.org:80/action/expandCollapseDecadeGroup?open=1950s&journalCode=jamerstatasso HTTP/1.1"	200	67963
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:25 +0800]	"GET http://www.jstor.org:80/stable/i314162 HTTP/1.1"	200	24847
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:25 +0800]	"GET http://www.jstor.org:80/stable/i314162 HTTP/1.1"	200	64926
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:30 +0800]	"GET http://www.jstor.org:80/stable/i314162 HTTP/1.1"	200	73910
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:30 +0800]	"GET http://www.jstor.org:80/userimages/42667/banner HTTP/1.1"	200	9263
10.19.74.76	-	[01/Jan/2012:00:02:30 +0800]	"GET http://eserv.uum.edu.my:80/login?url=http://www.emeraldinsight.com/journals.htm?issn=1741-038X&volume=19&issue=	200	6147
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:35 +0800]	"GET http://www.jstor.org:80/abs/addons?uri=http%3A%2F%2Fwww.jstor.org.eserv.uum.edu.my%2Fstable%2Fi314162&groups	200	6586
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:35 +0800]	"GET http://www.jstor.org:80/rx?uid=427072418167105&st=1325347411359&pn=http://www.jstor.org.eserv.uum.edu.my/stab	204	0
10.63.253.9	-	[01/Jan/2012:00:02:47 +0800]	"GET http://eserv.uum.edu.my:80/ HTTP/1.0"	302	0
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:48 +0800]	"GET http://www.jstor.org:80/page/termsConfirmText.jsp HTTP/1.1"	200	598
10.63.253.9	yQNbh43IFfM9AnY	[01/Jan/2012:00:02:49 +0800]	"GET http://www.jstor.org:80/page/termsConfirmText.jsp HTTP/1.1"	200	598

Figure 2. Data in Structured Format

Following the completion of pre-processing, the detection of search keywords employed by users is initiated. Typically, these keywords are embedded within the URLs, concluding with terms like "Query," "Keyword," and so forth. An illustrative example of a keyword in the log file is presented in Figure 3. The extraction process involves isolating these keywords, and subsequently, associating them with the respective databases. As demonstrated in Figure 3, keywords such as "sukuk+structure" and "continuous+education+and+retention" stand out among those successfully identified from the log.



```
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ct=all&ec=1&bf=1 HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?RQT=302&COPT=REJTPUcyODcrM2IxMCZJTlQ9MCZWRVI9Mg==&clientId=28929&cfc=1 H
GET http://www.emeraldinsight.com:80/journals.htm?issn=1753-8394&volume=2&issue=4&articleid=1826941&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1759-0817&volume=1&issue=1&articleid=1858421&show=pdf HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?SQ=continuous+education+and+retention&DBId=G647&date=ALL&onDate=&beforeDate=&af
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl HTTP/1.1
GET http://proquest.umi.com:80/pqdweb?SQ=continuous+education+factors&DBId=G647&date=ALL&onDate=&beforeDate=&afterDate
GET http://eserv.uum.edu.my:80/connect?session=s5kK6uj4JmxDFOgY&url=menu HTTP/1.1
GET http://eserv.uum.edu.my:80/menu HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=0828-8666&volume=26&issue=1&articleid=1846113&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=2 HTTP/1.1
GET http://eserv.uum.edu.my:80/public/rightmenu.htm HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=2 HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1753-8394&volume=2&issue=2&articleid=1795358&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/journals.htm?issn=1759-0817&volume=1&issue=1&articleid=1858422&show=pdf HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=sukuk+structure&ec=1&bf=1&ct=jnl&nolog=167626&page=3 HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=attitude&ec=1&bf=1&ct=jnl&nolog=814577&page=2 HTTP/1.1
GET http://www.emeraldinsight.com:80/search.htm?st1=changing+attitude&ct=all&ec=1&bf=1 HTTP/1.1
GET http://eserv.uum.edu.my:80/connect?session=sjZEm3Xln7jzCwY3&url=menu HTTP/1.1
```

Figure 3. Sample keyword in log file

Figure 4 illustrates the comprehensive steps involved in the database and keyword extraction process. The initial phase involves converting the Microsoft Excel file to CSV format. Subsequently, the "GetDatabases" function is implemented to systematically organize the URL column into an array based on the type of databases. For clarity, variables are declared to distinctly identify each database. Another crucial function, "CheckDataType," is developed to categorize the type of data, distinguishing between string, numeric, or alphanumeric formats. Example of extracted alphanumeric data is "challenges+to+sukuk&ct=all&ec=1&bf=1." Subsequently, any extraneous characters from the sentence are meticulously removed, leaving behind only the string of keywords.

Figure 5 shows examples of the successfully extracted keywords. To optimize their quality, all identified keywords undergo meticulous data formatting and spelling correction. Following this refinement, the subsequent phase involves the systematic removal of stop words from the keywords, setting the stage for comprehensive data analysis activities. The complete steps in databases and keywords extraction is outline in Figure 6.

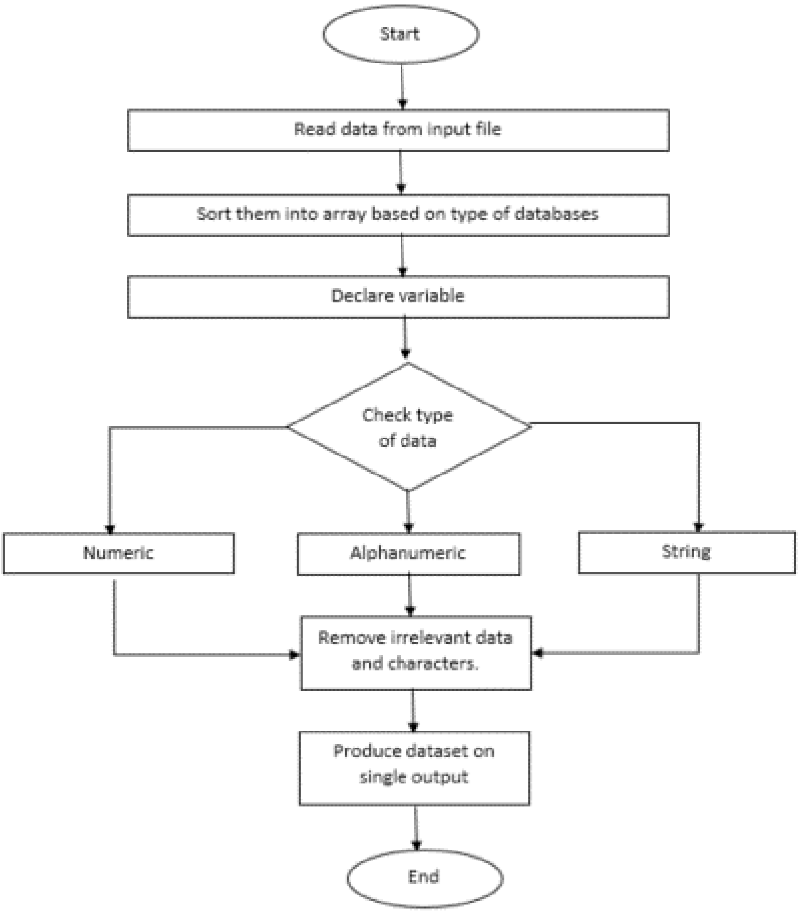
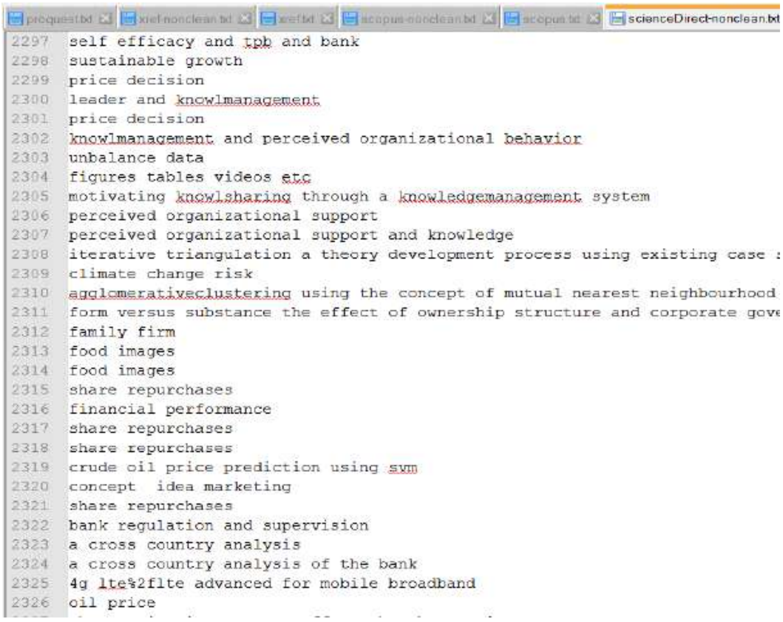

Start
Read data from input file
Sort them into array based on type of databases
Declare variable
Check type of data
Numeric
Alphanumeric
String
Remove irrelevant data and characters.
Produce dataset on single output
End

Figure 4. Keyword Identification Steps



```
self efficacy and tpb and bank
sustainable growth
price decision
leader and knowlmanagement
price decision
knowlmanagement and perceived organizational behavior
unbalance data
figures tables videos etc
motivating knowlsharing through a knowledgemanagement system
perceived organizational support
perceived organizational support and knowledge
iterative triangulation a theory development process using existing case :
climate change risk
agglomerativeclustering using the concept of mutual nearest neighbourhood
form versus substance the effect of ownership structure and corporate gove
family firm
food images
food images
share repurchases
financial performance
share repurchases
share repurchases
crude oil price prediction using svm
concept  idea marketing
share repurchases
bank regulation and supervision
a cross country analysis
a cross country analysis of the bank
4g lte%2flte advanced for mobile broadband
oil price
```

Figure 5. Extracted Search Keyword

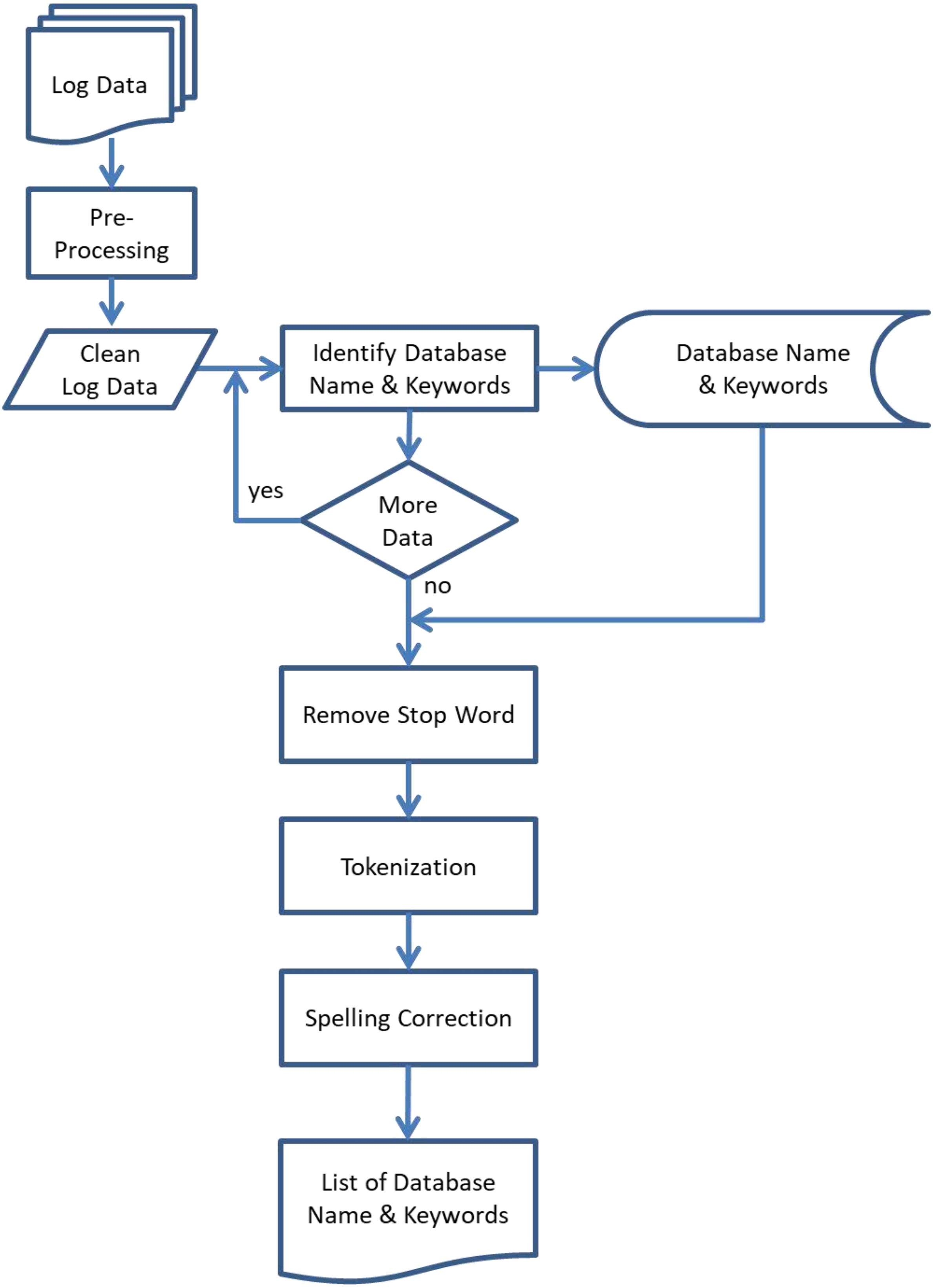


Figure 6: Overall Database and Keyword Extraction Process

4. Results and Discussion

This study successfully extracted a total of 19,146 keywords from the dataset, each intricately linked with its respective databases. The analysis of the log file revealed the presence of 11 databases, and Table 1 succinctly summarizes the obtained keywords. Notably, the largest number of keywords is associated with the IEEE database, totaling 4,862. Within this count, 3,418 keywords stand out as unique, while 808 appear with frequent recurrence. Serial Solution emerges as the second-largest contributor, with a total of 3,043 keywords. Other databases yielded less than 2,000 total keywords each. These findings underscore the prevalence of IEEE and Serial Solution databases within the UUM community. Figure 6 illustrates the trends in keyword distribution, encompassing counts of total keywords, frequent keywords, and the overall keyword landscape.

Table 1. Databases and Keywords

No.	DATABASE	NUMBER OF KEYWORDS	FREQUENT KEYWORDS	SUM OF KEYWORDS
1.	ebscohost	1552	217	1835
2.	Emerald	912	74	1011
3.	Eric	462	341	1081
4.	ieee	3418	808	4862
5.	JournalCambridge	1080	67	1148
6.	JSTOR	1321	150	1503
7.	OnlineLibrary	956	80	1044
8.	ScienceDirect	1561	255	1906
9.	SCOPUS	1226	124	1379
10.	SerialSolution	2408	363	3043
11.	XREF	307	18	334
TOTAL		6143	2497	19146

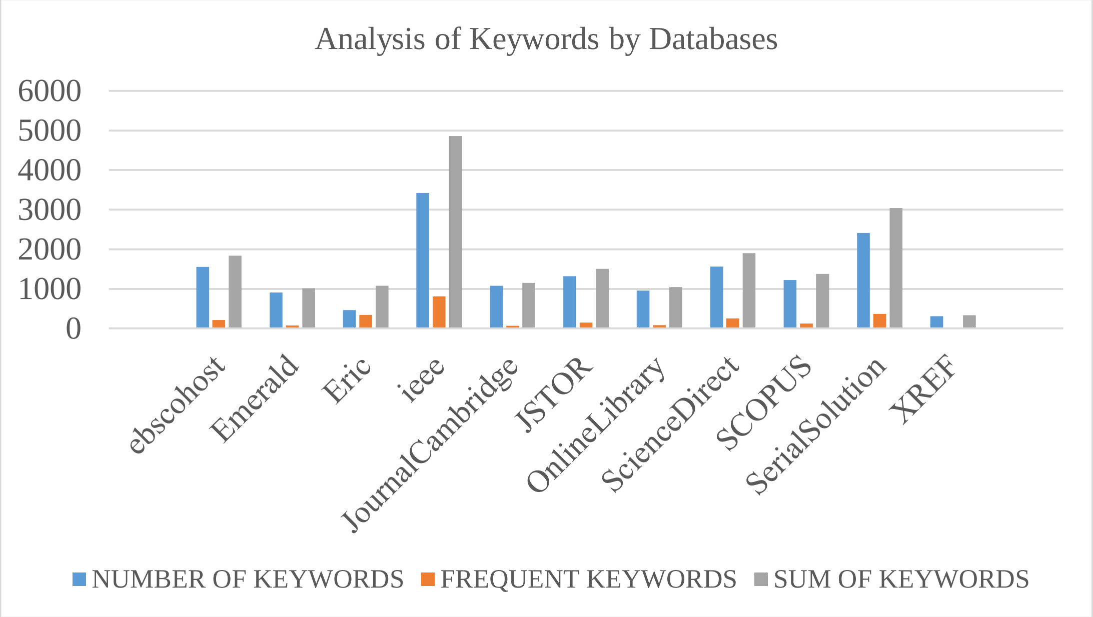


Figure 6. Analysis of Keywords by Databases

5. Conclusion

The Electronic Resources module serves as a digital hub that not only consolidates and organizes scholarly content but also democratizes access to this wealth of information. Its pivotal role in providing a seamless and efficient avenue for scholarly engagement underscores its importance as an integral component of the academic ecosystem at UUM.

The preprocessing method employed in this study has successfully addressed the challenges posed by the size and complexity of web log data. The transformation of raw data into a structured format, coupled with meticulous data cleaning and user differentiation processes, has laid the foundation for meaningful analysis. Additionally, the identification and extraction of search keywords, including the removal of irrelevant data and the systematic correction of spelling, have enhanced the quality of the dataset.

The practical implications of this research extend beyond academic analysis, offering tangible benefits for users navigating extensive databases and electronic resources. The identification of frequent keywords not only provides analytical insights but also accelerates users' search processes, reducing cognitive load and fostering a more efficient research experience. This study contributes valuable perspectives for optimizing user experiences and refining the management of electronic resources within academic libraries.

In essence, this research lays the groundwork for continued refinement of digital library services, aligning them more closely with the evolving needs and expectations of the academic community. As digital resources continue to play a crucial role in contemporary academia, the insights gleaned from this study contribute significantly to the ongoing enhancement of digital library services and the overall scholarly journey at UUM.

Acknowledgment

This work was financially supported by Universiti Utara Malaysia (UUM) under the university grant funding. Additional thanks to the Research Management Centre (RMC) for the technical support and Perpustakaan Sultanah Bahiyah, the library of UUM, for sharing the web log data.

References

[1] R. Basu, W. M. Lim, A. Kumar, and S. Kumar, "Marketing analytics: The bridge between customer psychology and marketing decision-making," *Psychology & Marketing*, vol. 40, pp. 2588–2611, 2023, https://doi.org/10.1002/mar.21908.

[2] C. Wu, K. Jenab, S. Khoury, and S. Moslehpourd, "A quality analysis of keyword searching in different search engines projects," *Journal of Project Management*, vol. 3, pp. 89–104, 2018.

[3] F. M. Yamin, T. Ramayah, and W. H. W. Ishak, "Information Searching: The Impact of User Knowledge on User Search Behavior," *Journal of Information and Knowledge Management*, vol. 12, no. 3, p. 1350023, 2013.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, pp. 37–54, 1996.

[5] K. K. Ibrahim and A. J. Obaid, "Web Mining Techniques and Technologies: A Landscape View," *Journal of Physics: Conference Series*, vol. 1879, no. 3, p. 032125, May 2021, https://dx.doi.org/10.1088/1742-6596/1879/3/032125.

[6] P. Shah and H. B. Pandit, "A Review: Web Content Mining Techniques," in *Data Engineering for Smart Systems*, P. Nanda, V. K. Verma, S. Srivastava, R. K. Gupta, and A. P. Mazumdar, Eds. Springer, Singapore, 2022, vol. 238, https://doi.org/10.1007/978-981-16-2641-8_15.

[7] M. Dhandi and R. K. Chakrawarti, "A comprehensive study of web usage mining," in *Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016, pp. 1-5, doi: 10.1109/CDAN.2016.7570889.

[8] H. Gu, "Data mining in the application of e-commerce website," *Adv. Intell. Syst. Comput.*, vol. 180 AISC, no. 8, pp. 493–497, 2013.

[9] G. R. Bharamagoudar, S. G. Totad, and P. Reddy, "Literature Survey on Web Mining," *IOSR Journal of Computer Engineering (IOSRJCE)*, vol. 5, no. 4, pp. 31-36, 2012.

[10] K. Dharmarajan and M. A. Dorairangaswamy, "Current Literature Review - Web Mining," *Elysium Journal of Engineering Research & Management*, vol. 1, no. 1, pp. 38-42, 2014.

[11] M. A. I. Aquil and W. H. W. Ishak, "Predicting Software Defects using Machine Learning Techniques," *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol. 9, no. 4, pp. 6609-6616, EISSN: 2278-3091, 2020.

[12] S. B. Thakare and S. Z. Gawali, "A effective and complete preprocessing for Web Usage Mining," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 3, pp. 848–851, 2014.

[13] A. Deepa and P. Raajan, "An Efficient Preprocessing Methodology of Log File for Web Usage Mining," in *Computer Science*, 2015, pp. 13–16, 2015.

[14] M. Jafari, F. S. Sabzchi, and A. J. Irani, "Applying Web Usage Mining Techniques to Design Effective Web Recommendation Systems: A Case Study," *Advances in Computer Science:an International Journal*, vol. 3, no. 2, pp. 78-90, 2014.

[15] M. H. A. Elhiber and A. Abraham, "Access Patterns in Web Log Data: A Review," *Journal of Network and Innovative Computing*, vol. 1, pp. 348-355, 2013.

[16] S. R. Aghabozorgi and T. Y. Wah, "Recommender Systems: Incremental Clustering on Web Log Data," in *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, 2009, pp. 812–818, https://doi.org/10.1145/1655925.1656073.

[17] A. Alphy and S. Prabakaran, "Cluster optimization for improved web usage mining using ant nestmate approach," in *Int. Conf. Recent Trends Inf. Technol. ICRTIT 2011*, 2011, pp. 1271–1276.

[18] A. B. M. R. Islam and T. Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique," in *International Conference on Information Science and Applications*, 2011, pp. 1-8, doi: 10.1109/ICISA.2011.5772412.

[19] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proceedings of the 20th VLDB Conference*, 1994, pp. 487-499.

[20] R. Mishra and A. Choubey, "Discovery of frequent patterns from web log data by using Fp-Growth algorithm for web usage mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 9, pp. 311-318, 2012.