CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

A Framework for Mining Trends in Web Clickstreams with Particle Swarm Optimization

by

Tasawar Hussain

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Computing

Department of Computer Science

2018

A Framework for Mining Trends in Web Clickstreams with Particle Swarm Optimization

By
Tasawar Hussain
PC103007

Foreign Evaluator 1
Dr. Muhammad Younas
Oxford Brooks University, Oxford, United Kingdom

Foreign Evaluator 2
Dr. Andrea Ko, Ph.D., CISA
Corvinus University of Budapest, Hungary

Supervisor Name
Dr. Nayyer Masood

Dr. Nayyer Masood
Head, Department of Computer Science

Dr. Muhammad Abdul Qadir
Dean, Faculty of Computing

DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD
2018

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD
Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled "**A Framework for Mining Emerging Trends in Web Clickstreams with Particle Swarm Optimization**" was conducted under the supervision of **Dr. Nayyer Masood**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science.** The open defence of the thesis was conducted on **28 May, 2018**.

Student Name : Mr. Tasawar Hussain(PC103011)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Ehsan Ullah Munir,
Associate Professor
CIIT, Wah Cantt
Ehsan Munir

(b) External Examiner 2: Dr. Ayyaz Hussain,
Associate Professor
IIU, Islamabad

(c) Internal Examiner : Dr. Muhammad Tanvir Azal,
Associate Professor,
CUST, Islamabad

Supervisor Name : Dr. Nayyer Masood,
Professor,
CUST, Islamabad

Name of HoD : Dr. Nayyer Masood,
Professor,
CUST, Islamabad

Name of Dean : Dr. Muhammad Abdul Qadir,
Professor,
CUST, Islamabad

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "**A Framework for Mining Emerging Trends in Web Clickstreams with Particle Swarm Optimization**" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.

(Mr. Tasawar Hussain)
Registration No. PC103007

Dated: 28 May, 2018

AUTHOR'S DECLARATION

I, **Mr. Tasawar Hussain (Registration No. PC103007)**, hereby state that my PhD thesis titled, '**A Framework for Mining Emerging Trends in Web Clickstreams with Particle Swarm Optimization**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

(Mr. Tasawar Hussain)
Registration No : PC103007

Dated: 28 May, 2018

Acknowledgements

In the name of Allah, the Most Gracious, and the Most Merciful Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people including my well-wishers, my friends, colleagues and various institutions. At the end of my thesis, I would like to thank all those people who made this thesis possible and an unforgettable experience for me. To express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me. First of all, I pay my gratitude to my supervisor,*Prof. Dr. Sohail Asghar* for his continued guidance, commitment, support, and encouragement on pursuing my thesis on a priority basis. In fact, I chose data mining as a research area because of his lectures during my MS program. Due to his guidance, I got the right direction in the ongoing research work in pursuit of a Ph.D. degree in computer science. Despite his busy schedule, he used to review my thesis progress, give his valuable suggestions and made my corrections. His unflinching courage and conviction always inspired me. I also want to pay my gratitude to my second supervisor Prof. Dr. Nayyer Masood Dar for providing necessary infrastructure and resources to accomplish my research work. Under his guidance, I successfully overcame many difficulties and learned a lot. I would also like to thanks to Higher Education Commission, Islamabad, Pakistan, for providing financial assistance in the form of Fellowship which buttressed me to perform my work comfortably. I am really grateful to all my group members and colleagues especially Chaudhry Muhammad Yaqub, Ejaz Hussain and Sajjad Ahmad for their support. Now, I think of my parents whose selfless sacrificial life and their great efforts with pain and tears and unceasing prayers have enabled me to reach the present position in life. I should not forget to acknowledge my brother Sabir Hussain for his continuous love, support, understanding, and good wishes whenever I needed. Finally yet importantly, I want to express my deepest gratitude to my family, My wife Khair un Nisa and children Muhammad Haider, Maryam Noor, Wania Noor and Muhammad Hashim who always welcomed me with smiling face even when I returned home very late and with the tired face.

Tasawar Hussain

PC103007

List of Publications

Journal Papers

1. Hussain, T. and Asghar, S. (2016) Chi-square based hierarchical agglomerative clustering for web sessionization. Journal of the National Science Foundation of Sri Lanka, 44(2) pp. 211-222.

2. Hussain, T. and Asghar, S. (2013) Evaluation of similarity measures for categorical data. The Nucleus, The journal Pakistan Atomic Energy Commission, Pakistan, 50(4) pp. 387394.

3. Hussain, T. and Asghar, S. 2013. 'Web mining: approaches, applications and business intelligence.' International Journal of Academic Research Part A, 5(1), pp. 211-217.

4. Hussain, T., Qadir, M. A. and Asghar, S. 2012. 'Fuzzification of web objects: A semantic web mining approach.' International Journal of Computer Science, 9(1), pp. 61-67.

Conference Papers

1. Hussain, T., Asghar, S., and Masood, N. (2010a) 'Hierarchical sessionization at preprocessing level of wum based on swarm intelligence.' In Emerging Technologies (ICET), 2010 6th International IEEE Conference on. IEEE, pp. 2126.

2. Hussain, T., Asghar, S., and Masood, N. (2010b) ' Web usage mining: A survey on preprocessing of web log file.' In Information and Emerging Technologies (ICIET), 2010, IEEE International Conference on. IEEE, pp. 1-3.

3. Hussain, T., Asghar, S. and Fong, S., 2010, November. 'A hierarchical cluster based preprocessing methodology for Web Usage Mining.' In Advanced Information Management and Service (IMS), 2010 6th International Conference on. IEEE, pp. 472-477.

Abstract

The expansion of World Wide Web (WWW) in its size and exponential growth of its users has made the web most powerful and dynamic medium for information dissemination, storage, and retrieval. Moreover, the improvements in data storage technology have also made it possible to capture the huge amount of the user interactions (clickstreams) with the websites. The availability of such a huge amount of web user clickstreams has opened the new challenges for researchers to explore the weblog for the identification of hidden knowledge. For the last decade, web usage mining is playing a crucial role in the identification of trends from user clickstreams such as web personalization; user profiling; and user behavior analysis. These trends are beneficial in many ways such as information retrieval, website administration and improvement; customer relationship management; e-marketing; and recommender systems. The plenty of techniques are available in the literature, however, the accuracy, correctness and validity of the generated trends is totally relying on the proper selection of web mining process such as web sessionization, which is the benchmark for the later web usage mining stages. For the promising and optimized results, weblog sessionization is the eventual choice. Moreover, the extraction of proper, accurate and noise free sessionization is a demanding and challenging job in the presence of huge web clickstreams. The sessionization problem may fail to identify the focused and visualized groups from clickstreams records with high coverage and precision. Even though the well-known web session similarity measures such as Euclidean, Cosine, and Jaccard are prevalent in literature for mining process at the early learning stages. The web sessionization must take account of the validity of generated trends, which entirely depends upon the correctness and credibility of web sessions. To overcome the limitations of existing web sessionization techniques, we propose a Framework for Mining Trends (F_MET) that empowers us to gauge the user activities on the website through evolutionary hierarchical Sessionization. Hierarchical Sessionization enhances the visualization of user click data to improve the business logic and mines the focused groups for scalable tracking of user activities. The foundation of the proposed framework is the swarm based optimized clustering technique along with a proposed web session similarity measure ST_Index to address the Hierarchical Sessionization problem. The proposed web session similarity measure ST_Index for hierarchical sessionization overcomes the limitations of Euclidean, Cosine and Jaccard measures, which may have failed to explicitly seek the proper and accurate trends. The Euclidean measures are of numerical in nature while the weblog data is of mixed nature. Moreover, existing measures are best for independent and isolated clustering groups. The proposed similarity measure ST_Index computes the similarity among the user sessions through the common features (pages) shared among the sessions while assigning weight to uncommon features among the given sessions along with the minimum time shared by the given sessions time ratio. We validated and verify the proposed framework on three different datasets. The proposed ST_Index measure produced the accurate and valid relationship among the sessions against common web session similarity measures. Furthermore, framework also produced the correct, accurate and valid trends. The performance of the proposed framework is validated against the well-known data analysis metrics such as VC (visitor coherence), accuracy, coverage and F1 Measures. The results show the significance improvements over the existing techniques of hierarchical Sessionization

Contents

List of Figures

List of Tables

Abbreviations

WWW	World Wide Web
WUM	Web Usage Mining
WCM	Web Content Mining
WSM	Web Structure Mining
PSO	Particle Swarm Optimization
HAC	Hierarchical Agglomerative Clustering
PSO-HAC	Particle Swarm Optimization based Hierarchical Agglomerative Clustering
F_MET	Framework for Mining Emerging Trends
ST_Index	Session and Time Index
WSS	Web Session Similarity
CRM	Customer Relationship Management
KDD	Knowledge Discovery in Databases
WebKDD	Web Mining Knowledge Discovery in Databases
URL	Uniform Resource Locator
URLID	Uniform Resource Locator Identification
IP	Internet Protocol

Symbols

symbol	**Description**
L	Weblog record of n transaction
$\tilde{L}$	Clean and Preprocessed Weblog
S_i	i^{th} session of n transaction such that $S_i \subset L$
CS_j	Collection of clickstreams in j^{th} session
F_{jp}	Features of j^{th} session in t_j transactions
$\tilde{\phi}$	User defined threshold
χ^2	Chi-square metric

Chapter 1

Introduction

The World Wide Web (WWW) is rapidly growing for the last two decades and is a powerful medium to disseminate informations. Millions of people are availing the services of the web. Due to advancements in data capturing technologies, user interactions and browsing clickstreams are captured in the form of weblog. The availability of such a huge amount of web user clickstreams has opened the new challenges for researchers to explore the weblog for the identification of hidden knowledge. The objective of this chapter is to present the overview and motivations behind our research. At the end of the chapter, we are presenting the problem statement and summary of the chapter.

1.1 Overview

The web is a powerful and cost effective medium to deliver services to its users [1]. Mostly web is explored for simple information and information is part of web services. Consequently, business community prefers the Internet for their services and users feel free to avail the web services [2, 3]. The web is providing its services in all most every walk of life and department irrespective of geographical boundaries [4]. Due to the stateless protocol http and https, the Internet is simple in nature to deliver the web services to its users and it motivates the organizations for online e-business for the more competitive environment, business opportunities and challenges [5]. For the last two decades, the IT paradigm has been changed and the huge amount of data acquisition is made available for research and knowledge visualization [6, 7]. This huge amount of web data can be mined in three different dimensions such as Web Content Mining (WCM); Web Structure Mining (WSM); and Web Usage Mining (WUM) [8–10].

In traditional data mining, WUM is a classical approach to explore the interesting and useful patterns (trends) from user clickstreams [11]. The crucial steps involved in WUM process are Preprocessing; Sessionization [12]; Pattern Discovery; and Knowledge Visualization [13]. The WUM techniques are helpful in various web applications such as website improvement; website administration; web server performance improvement; information retrieval; web personalization [14]; Customer Relationship Management (CRM) [11]; predictions; and recommended system [15, 16]. Clustering; association rule mining; sequential pattern mining; and classification are the most common data mining techniques which are in practice for knowledge extraction process from weblog data [17, 18]. The successful accomplishment of these web applications is totally relying on the proper selection of WUM process such as sessionization, which is the benchmark for the later WUM stages. According to Bayir and Toroslu [19] sessionization is a first major step to address the web usage mining and its applications. Consequently, weblog sessionization is the eventual choice to address the core issues of web sessionization for the promising and optimized results. However, the extraction of proper, accurate [20]and noise free sessionization is a demanding and challenging job [21].

The WUM process is divided into four steps such as preprocessing; web sessionization; pattern identification; and knowledge visualization. At preprocessing level, accuracy and session identification issues are significant. At web sessionization level, the identification of proper and accurate session relationship among the sessions is play vital role in WUM process. At pattern identification level, focused and visualized groups are important to generate the interesting patterns. All the issues at different level of WUM are composed in the form of web sessionization problem to take account of validity and correctness of trends generated for weblog.

The sessionization problem may fail to identify the focused and visualized groups from clickstreams records with high coverage and precision [22]. Even though the well-known web session similarity measures such as Euclidean [23], Cosine [22], and Jaccard are prevalent in literature for mining process at the early learning stages. The web sessionization must take account of the validity of generated trends, which entirely depends upon the correctness and credibility of web sessions. To overcome the limitations of existing web sessionization, we required a web usage mining framework for the identification of trends from the weblog. At the preprocessing stage, the framework must account the valid and noise free session construction. For web sessionization, the web session similarity measure must be capable of identifying the true close relationship among the sessions. The correct and properly identified relation is the base for the discovery of valid and credible trends. Hierarchical sessionization further enhances the visualization of user click data to improve the business logic and mines the focused groups for scalable tracking of user activities. Figure 1.1 briefly highlights the complete life cycle for web usage mining from the end user to service delivery. The figure also shows the various components and mechanism of a website.

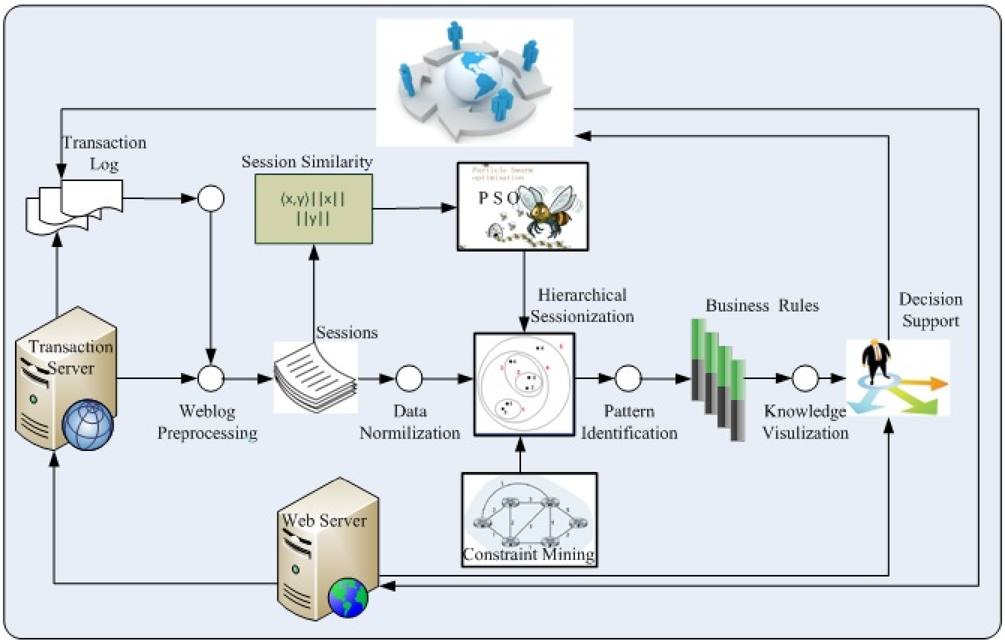


FIGURE 1.1: Lifecycle for Web Usage Mining.

1.2 Motivation

The expansion of World Wide Web in its size and exponential growth of its users has made the web most powerful and dynamic medium for information dissemination, storage, and retrieval [1, 24]. Around 40% of the world population has the Internet access today and a number of websites have reached to 1 billion [25]. Moreover, the improvements in data storage technologies have also made it possible to capture the huge amount of the user interactions (clickstreams) with the websites [26]. The availability of such a huge amount of web user clickstreams has opened the new challenges for researchers to explore the weblog for the identification of hidden knowledge.

Plenty of web mining techniques are available in the literature to mine the trends from weblog [27–29]. However, the accuracy, correctness and validity of the generated trends is totally relying on the proper selection of web mining process such as web sessionization, which is the benchmark for the later web usage mining stages. For the promising and optimized results, weblog sessionization is crucial choice. Moreover, the extraction of proper, accurate and noise free sessionization is a demanding and challenging job in the presence of huge web clickstreams.

1.3 Research Objectives

The main of aim this research is to address the web sessionization problem and contribute the knowledge in different steps of sessionization.The research objective of this dissertation is the identification of trends from weblog to address the sessionization problem in order to achieve the visualized and focused groups (sessions).

The web data is unstructured, heterogeneous, and dynamic in nature [28]. The exploration of such a diversified web data for information retrieval is a challenging and complex task. Besides this, we also lack the proper web data structure for the information management. The websites are also unable to capture the user feedback. Nowadays, the Internet resources are facing the challenges of knowledge visualization, relevant and required information serving mechanism. To overcome these challenges, the web data exploration plays a crucial role. The user interactions with website are recorded in server weblog. The weblog can be studied to track the user traversing aptitude and behavior. There is a close relationship between the weblog data and its users. Furthermore, the researchers are adapting different Web Mining and Knowledge Discoveries in Databases (WebKDD) procedures to address the web sessionization problem due to improper web structure. The traditional clustering approaches failed to identify the valid and accurate user patterns from weblog due to the selection of improper web session similarity measures such as Euclidean; Cosine; and Jaccard.

The above-mentioned web mining issues are composed in the form of web sessionization problem and traditional web mining techniques and methodologies are unable to deliver the reliable and accurate solution. Consequently, there is a growing demand to produce a mining framework to address the web sessionization problem based on reliable similarity measure. The proper and accurate weblog analysis is the key to many web usage mining applications such as user analysis; profiling; prediction; and recommender system. Moreover, the introduction of evolutionary approaches such as particle swarm in web usage mining are delivering efficient results. One of the major objective of this research is to propose a complete mining framework based on particle swarm to cater the web sessionization issues.

1.4 Scope of the Research

The web usage mining is an attractive and active research area for the researchers due to the rapid expansion of web. The number of its users and web data is expanding around the clock. There is a close affinity between the web users and web data. The web data is a major source to analyze user traversal mood and behavior. The identification of similarity relation between the website users based on their clickstreams is a key to many websites related applications. The web data has three major characteristics such as the bulk of web data availability; unstructured; and dynamic; in nature and according to Xu [1], there is no proper data structure available to organize the data for information management. Interestingly, researchers are applying their own data paradigms to organize the web data for information retrieval to fill the gap. This dissertation is also an effort to streamline the web usage mining process to mine the trends based on the accurate and valid session similarity from such type of web data. In the following section, we aim to discuss the scope of this dissertation which is as follows:

One of the major objectives of the web is the delivery of seamless services and information to its users. The web designer and owners are trying to gain the satisfaction and confidence of end users. As a result, they are trying to improve the web services to attain the goal. The researchers are bridging the gap between the web applications and web users through delivering the various research methodologies and approaches. Web Sessionization is one of the methodologies to tackle the web issues at different phases of web usage mining. Whereas the crux of sessionization is the identification of accurate and noise free similarity among the users (sessions). The identification of strong relationship among the sessions is a key solution to the web sessionization issue under the umbrella of web mining techniques at all the phases. Weblog preprocessing will be performed to get the noise free weblog data and accurate session identification. Different data filtering algorithms are introduced for preprocessing to cater the dynamic nature of weblog data. Session construction is helpful for us to identify the true user in the presence of firewall and cache. The session construction is a non-trivial issue in the presence to the proxy server. The session construction is composed of weblog attributes such as IP host; User Agent; User OS; and Referrer Page.

The web session similarity is a crucial step and computed from the sessions construction at preprocessing level. The web session similarity is based on two very important attributes of the web such as web pages (Uniform Resource Locator)(URLs) and time spent in a session by a user [11]. These two attributes are the most significance for sessionization and how the two sessions are similar. In this research, we will apply the notion of Session Index (SI) based on the common pages traversed between the sessions along with the uncommon web pages in respective two sessions [30]. The second notion is Time Index (TI) that will be computed from the minimum time of both the sessions. The common web pages are computed by assigning unique URLID (Uniform Resource Locator Identification) to all filtered weblog. The similar sessions are further used for tends identification to analyze the user behavior through proposed framework F_MET by applying hierarchical sessionization. In this dissertation, we are applying particle swarm optimization based agglomerative clustering for trend mining and knowledge visualization.

1.5 Research Questions

The following research questions have been raised during the research meetings with my supervisor throughout the Ph.D. period. The focus of this dissertation revolved around these research questions.

- Why weblog preprocessing is significant in web usage mining for the extraction of accurate, correct and interesting patterns?
- In the presence of large number of preprocessing tools in data mining such as Weka, XL Miner etc, why do we need to develop a preprocessing methodology for web usage mining?
- The web session construction is a significant step in weblog sessionization. How can we identify the true user in presence of firewall and cache?
- Do all the sessions generated are accurate and can lead the promising results for later stages of WUM?
- There are a number of web session similarity measures available such as Euclidean, Cosine, Jaccard and along with a number of proposed measures. Why do these measures fail to deliver the accurate, correct, and valid relationship among the web sessions?
- Can we achieve the credible weblog analysis that can ensure the validity of the results?

- How we can mine the trends from weblog. Do we need a complete mining framework to address the web sessionization problem?

1.6 Problem Statement

Given a weblog of n transactions $L = \{t_1, t_2, t_3, \ldots, t_n\}$ where the i^{th} transaction t_i is defined as a user transaction containing User IP; Client Identifier; User Name; Time Stamp; Request Method; URL Resource; HTTP Protocol; Status Code; Data Transferred; Referrer URL; and User Agent. Let $S = \{S_1, S_2, S_3, \ldots, \ldots, S_n\}$ be collection of n sessions and $S_j = \{t_{j1}, t_{j2}, t_{j3}, \ldots, t_{jk}\}$ be the k clickstreams in S_j session where each $S \subset L$ containing transactions of j^{th} session and CS_j be the collection of clickstreams in j^{th} session is defined as be the user clickstreams where C_{jk} be the k^{th} click of j^{th} session. In fact $C_{jk} = \{F_{j1}, F_{j2}, F_{j3}, \ldots, F_{jp}\}$ where F_{jp} is the user transaction features extracted from the transactions t_j related to the j^{th} session. The similarity function $f : S * S \rightarrow C$ for the given two sessions S_i and S_j based on predefined threshold $\widetilde{\phi}$ is defined as:

$$\phi(S_i, S_j) = \sum_{|t|>0} S_i(C_i) * S_j(C_i) \geq \widetilde{\phi}$$

The sessionization problem arises that how the sessions S_i andS_j w.r.t. click-streams features C_{ij} traversed by a users in a time spent in a session are *similar, accurate* and *noise free*? These problems are defined as:

- The given two sessions S_i and S_j are similar $S_i \cong S_j$ if both S_i and S_j share the common features while surfing the given website and having the common score threshold $\geq \widetilde{\phi}$.

- The session $S_i = \{C_{i1}, C_{i2}, C_{i3}, \ldots, \ldots, C_{im}\}$ comprises of user clicks. If $\sum_{|t|\geq 0} S_i(C_i) \notin S_i$ but placed inaccurately in S_i then the **accuracy** of the session is questionable and may lead to misleading WUM process and results.

- The web log data contains 80% irrelevant data that is also a big hurdle for the construction of quality sessions.

- The web sessionization must take account of the validity of generated sessions, which entirely depends upon the correctness and credibility of sessions. This problem is mostly hindering by the proxy server, cache, and firewalls in the client web browser.

- The sessionization problem may fail to explicitly seek user profiles with high coverage and precision even though well-known measures such as Euclidean, Cosine, and Jaccard are prevalent in literature for WUM process in the early learning stages [22].

- The evolving and dynamic nature of WWW leads to enormous challenges for web usage mining such as web clickstreams for extraction of patterns; user behavior [27]; targeted and focused visualization of coherent sessions.

1.7 Research Methodology

To address the research questions raised in this dissertation, we are adapting the following research methodology.

- **Research Area Selection:** Web mining is active research area and web usage mining always attracted me to study about user behavior and their traversing pattern. To study more about the web usage mining, we studied different approaches from literature and came up with an idea to study the trends from weblog.

- **Literature Review:** Relevant literature gathering on the topic is a hectic job when abundance of literature is available. We used different available sources such as Google Scholar, IEEE, ACM, Elsevier, Springer, DBLP, and Research Gate to gather the relevant literature. We gathered the literature in three directions web session similarity; web sessionization techniques; and hierarchical clustering techniques.

- **Development of Conceptual Model:** To address the web sessionization problem in its true sense, an efficient working model was necessary. In this regard, we studied different models and come up with solution that end to end model should be developed. We developed the conceptual model which provides the systemic solution to sessionization problem.

- **Formulation of Research Questions:** During the literature review and research meetings with supervisor, we studied the pros and cons of various web sessionization techniques. Different research questions on the web sessionization problem were discussed. We also formulated the seven research questions on the topic and literature was reviewed to answer these questions.

- **Research Solution and Implementation:** Without the proper implementation strategy, the web sessionization problem can not be addressed. We designed and developed the conceptual model to reach at the solution by completing all the tasks. The whole problem was divided into different coherent steps and tasks. The algorithms were also designed and developed. These algorithms were implemented in PL SQL through oracle 10g and were evaluated on datasets.

- **Data Collection:** The weblog contains the sensitive data of users and professional websites hesitate to share their weblogs. The datasets used for this experiment are two university weblogs and third dataset was shared by [31].

- **Result Evaluation:** We applied the accuracy, precision, recall and page visit as evaluation metric to validate the hierarchical clustering classifier.

- **Result Presentation and Conclusion:** The results of experiments were presented in graphical mode. Chapter summary at the end of each summarizes the chapter.

1.8 Research Contributions

In this research work, we propose a **F**ramework for **M**ining **E**merging **T**rends (F_MET) based on the weblog for the hierarchical sessionization to address the issue of sessionization that how the given two sessions under study are similar. The F_MET is a unique framework for the generation of hierarchical sessionization from weblogs and delivers a complete solution from end to end. It takes raw weblog data as an input, processes to produce the trends based on the proposed web session similarity Session Index and Time Index (ST_Index). As the flow of data is increasing day by day, the F_MET is designed to manage a large data to overcome the scalability issues of existing frameworks. Another prominent feature of F_MET is to manage the diversified weblog file formats such as Apache log Format; IIS Log Format; and Common Log Format. The contributions of the dissertation are as under:

We applied the four-step research mechanism in this research: in the first step, we applied the preprocessing algorithms on the weblog to prepare it for high coverage and noise-free as the rest of the web usage mining procedures are totally relying on it. Without proper preprocessing, the weblog directly cannot be used for web mining as weblog contains around 80% irrelevant entries. In the presence of such a huge amount of irrelevant entries, the WebKDD process cannot deliver the desired objectives of mining. The second crucial step of web Sessionization is session construction. The construction of accurate web sessions is another significant fact of this research as without having the proper and accurate web sessions, the WebKDD cannot be resourceful for mining purpose.

The third important step of this research is the introduction of novel web session similarity measure. Besides the well-known web session similarity measures, a number of proposed measures are also available in the literature. The existence of such a huge number of measures is an evidence of unsatisfactory results of the web usage mining process. The identification of true close relationship among the web sessions is the cornerstone of this research. The proposed web session similarity measure finds the relationship among the sessions based on the web pages visited by a user in sessions along with the uncommon web pages in two sessions respectively. The web proximity matrix computes the common and uncommon web pages among the session accordingly. The second feature of proposed web session similarity measure is time indexing that is based on the time utilized among the sessions. The heuristic used in time indexing is that if two sessions have similar traversing behavior than minimum time both have spent is shared. Based on session indexing and time indexing score, the final similarity is computed among the sessions.

The fourth core step of this research is proposed mining framework. The proposed framework delivers the complete mining solution to the web sessionization problem by adopting the WebKDD process. The framework follows all the steps of web usage mining from raw data intake to final delivery of knowledge visualization. The three main steps of web usage mining such as Preprocessing; Pattern Discovery; and Knowledge Visualization; are the main feature of this framework. The proposed framework is based on the particle swarm optimization for pattern identification and knowledge discovery in an optimized way to address the sessionization problem. The framework is equally beneficial in knowledge analysis through the hierarchical particle swarm based sessionization. The proposed framework is simple in implementation and delivers the high-end promising results. The results produced by the framework can apply in any domain of web usage mining such user behavior analysis; prediction; recommender system; online fraud detection system; e-applications; and website improvements. Specifically, the framework is tested against the three university datasets and delivered the optimized results. Following are the main contributions of this dissertation.

- **In literature review** work, we aim to investigate the merits and demerits of the existing web mining literature. The review was intensively performed in three directions covering following segments of web usage mining:

  - web session similarity measures
  - web sessionization techniques
  - hierarchical sessionization

  The plethora of web session similarity measures is available. However, the effectiveness of measures to seek the true relationship among the sessions is the ultimate goal. The coverage and precision are the two major artifacts for web session similarity measure. The existing measures were unable to deliver the required accurate results at an early stage of WebKDD learning process. To fill the research gap, we introduce the ST_Index, web session similarity measure to find the precise and accurate relationship among the weblog sessions. The results showed the better results in comparison with existing renowned web session similarity measures.

Almost all the data mining techniques are being tried to discover the hidden patterns from weblog, however, a clustering technique is a most common strategy to cluster the sessions with similar behavior [32, 33]. The research investigated that traditional clustering techniques are unable to address their legacy limitations such as number clusters, the center of the cluster, initialization [34, 35]. Furthermore, evolutionary approaches are also facing issues of feature selection, local maxima, efficiency, quality [36], visualization and reliability. To be more specific, particle swarm based clustering approaches are more appropriate for sessionization as the best matching pairs are grouped together after the number of iterations. In this research, we opted the particle swarm optimization technique along with hierarchical agglomerative clustering. Such kind of hybrid clustering techniques delivered the accurate, valid and correct web session patterns and these patterns are helpful in weblog data analysis.

- **The preprocessing** is a vital step in data mining for quality and noise free results. Mostly, this step is ignored and deliver misleading results at later stages [37].The preprocessing techniques include Data Cleansing; Data Filtering; Path Completion; User Identification; Session Identification; and Session Clustering [14, 38]. There is almost a consensus that all the researchers in literature review have performed Data Cleansing to remove the irrelevant entries from weblog [39]. Moreover, there are numbers of data mining tools available that can perform weblog preprocessing. However, the weblog data is dynamic in nature, the tools and existing techniques are unable to deliver the required noise free weblog for upcoming phases of web usage mining. In this research, we are proposing a complete weblog preprocessing methodology that delivers the noise-free data for the upcoming phases. The most sensational part of the proposed preprocessing methodology is session construction algorithm. There are various heuristics approaches are available for session construction, we applied the IP; User Agent; OS; and Referring Page weblog attributes based heuristic to cater the issues of firewall and cache. This approach constructs the accurate and precise sessions with high coverage. The generated sessions delivered the true website users.

- **The web session similarity measure** is very important to identify the relationship among the web sessions [22, 40]. The sessionization problem arises that how the sessions traversed by users are similar, accurate and noise free? The web sessionization must take account of validity of generated sessions, which entirely depends upon the correctness and credibility of sessions. The sessionization problem may fail to explicitly seek user profiles (behavior) with high coverage and precision even though well-known measures such as Euclidean [23, 28, 41], Cosine [22], Jaccard and Longest Common Sequence [42] are prevalent in literature for WUM process in the early learning stages of WebKDD. Due to the criticality of the session similarity issue in web sessionization, the appropriate session similarity measure is vital and keeping in view the limitations of existing measures, we introduce the novel session similarity measure ST_Index based on Session Index (SI) and Time Index (TI). The synopsis of proposed similarity measure is to cater not only the shared web pages and shared time between the two sessions; in fact, it also assigns the weights to the unshared pages with respective to the sessions each other as users have the same pool of web pages to traverse them with different objectives.

- **A Framework for Mining Emerging Trends (F_MET)** is a conceptual and structured web sessionization solution with specific functionalities to explore the user clickstreams with intended mining objectives. The proposed F_MET delivers the complete solution of web sessionization problems and the F_MET is a set of interrelated WebKDD processes that work iteratively for the knowledge visualization that was previously unseen in user clickstreams by covering all aspects of the WebKDD process. F_MET takes the raw weblog as input data and delivers the hierarchical sessionization of the weblog as output. The objectives of F_MET have been defined for each step that provides the dynamic solution of the challenges and issues at each step. The major components of F_MET are Preprocessing; Web Session Similarity; and Particle Swarm Optimization based Hierarchical Sessionization.

- **The Particle Swarm Optimization** based Hierarchical Sessionization Clustering Algorithm (PSO-HAC) is simply working like agglomerative hierarchical clustering in an optimized way. PSO-HAC takes all the sessions as particles, as single clusters and merge them into pairs based on ST_Index criteria. The merging of sessions (clusters) continues until the construction of a complete web session hierarchy is achieved in an iterative mode. The sessions adjust their best position during the iterations for an optimized solution. Both the hierarchical and partitioning clustering algorithms suffer initialization, local maxima of particles and efficiency deficiencies by default. The proposed PSO-HAC is the combination of swarm particles optimization and agglomerative algorithm to overcome the above-cited issues in an efficient and optimized way.

1.9 Significance in Industry and Academia

Data mining is the backbone of the current software industry and research academia. There are numerous data mining applications in the software industry. Web mining is one of the most influential data mining application areas where we apply the data mining techniques on the huge web data available across the world. The web data is available in structured; semi-structured; and unstructured formats. Web mining techniques are dynamic and robust that these techniques are not only effective on structured data but can also be applied on semi and unstructured data. Consequently, web mining is playing a critical role from converting man-understandable web contents to machine-understandable semantics. Arotaritei and Mitra [43] referred the web mining as the application, implementation, and usage of data mining techniques and algorithms to retrieve, extract and evaluate information for knowledge from various web resources and web data. Web Mining is further classified into three broader categories, commonly distinguished as Web Content Mining; Web Structure Mining; and Web Usage Mining [44].

Web mining is a tool that links the business applications to its users. It not only helps to manage business applications but also focuses on obtaining the in-depth tracking and analysis of the business improvements. Web mining is quite different from website visitor counter and tracking system. In visitor hit counter, the website only provides counter that how many visitors are visiting the website. The counter can be overall or on daily basis. This counter cannot provide the business analysis, business trends, and future business growth. It provides no knowledge about any user whether this user is useful to business or a just simple visitor. Moreover, web mining techniques provide sufficient knowledge to trace the visitor origin and can suggest the business organization about the business trend for different cities and locations throughout the world. Similarly, web mining provides a mechanism to business application owner to be dynamic at different visiting hours. Web mining provides business solutions by suggesting business intelligence in business applications cater the growing awareness among the visitors. By adding intelligent to business applications, business owners can easily study the market trends and can take the appropriate measure to save the stakes.

By introducing web mining techniques and approaches in the web based business application, systematic market analysis encourages the competitor environment to boost the business. These techniques help the organizations to assess the performance evolutions of their products. On any website, the intelligent approaches can create the mega difference. For example, there are two books selling website business A and B. A has adapted the web mining intelligent approaches and when someone searches a particular book such as Intelligent Mining, from A. A can offer the visitor with some other related books such as Data Mining, Intelligent Data. In this way, A can exploit the visitor interests intelligently. It is possible that visitor may cart Intelligent Mining along with Intelligent Data. While B did not implement any intelligent mechanism to study the interest and behavior of its users and is unable to cash the visitor interests and confidence.

Just like the terms, e-business, e-applications, the another common term is getting popularity is e-education. The web has become the most charming source for eduction. There are number of online courses and even online degree programs are available globally. The service providers are interested to get the prior knowledge that from where, they are getting more students and on which course or degree the students are more interesting. Besides, online educational services, the whole academia has also shifted on the web. From student intake to degree convocation, every thing is web based. In this regard, the web usage mining is providing the in-depth analysis about the students and their preference trends. The inference can be made through the web usage applications to capture the e-education business.

1.10 Structure of the Dissertation

This dissertation consists of eight chapters. The focus of this dissertation is to produce the viable solution to the web sessionization problem. The document presents a framework F_MET for the web sessionization and incorporates the results of research from the weblog. In order to organize a self-contained document on data mining web usage research, various web mining techniques are being reviewed to highlight the significance of sessionization problem in the presence of such a mega-repository. Finally, the framework is presented and experimental results are examined.

The descriptions of the individual chapters are:

- **Chapter 1, Introduction:** In this chapter, we presented an overview of web usage mining. How the web usage mining is playing the key role in making the information retrieval system more reliable. In motivation section, the aspects of motivation for this dissertation are discussed. The chapter also includes detailed research objectives and research contributions. The problem statement is composed from the literature review Chapter 2 and is presented in this chapter. The significance of web usage mining in industry and academia is also the part of this chapter.

- **Chapter 2, Background and Literature Review:** In this chapter, we briefly described background knowledge about the Web Usage Mining (WUM) and its taxonomy. The different web techniques available for WUM process have been discussed along with pros and cons. In literature review section, we reviewed the literature in three directions such as web session similarity measures; web sessionization techniques; and hierarchical sessionization. We summarized the literature with the findings of limitations and contributions of various web mining techniques. These research gaps helped us to formulate the problem statement.

- **Chapter 3, Proposed Framework F_MET:** One of the outcomes of the literature review was that the alone web session similarity measure is insufficient to address the web sessionization problem. In this chapter, we proposed a framework F_MET, a complete working solution to the sessionization problem.

- **Chapter 4, Weblog Preprocessing and Web Session Similarity:** In this chapter, we presented the importance of preprocessing of weblog in web usage mining. We also presented the various state of the art weblog cleansing and filtering algorithms. These algorithms are applied to the datasets and produced the results for next phases of web usage mining. In section web session similarity, we proposed the ST_Index web session similarity measure to address the web sessionization problem. The proposed ST_Index also overcome the limitations of well-known existing web session similarity measures.

- **Chapter 5, Results and Evaluation:** In this chapter, we performed the hierarchical sessionization through the proposed ST_Index and Particle Swarm Optimization for the efficient extraction of useful knowledge from the weblog. We also presented the results along with the comparison and evaluation of results.

- **Chapter 6, Conclusions:** This chapter summarizes the overall thesis and highlights the research contributions along with the recommendations and claims. We also highlighted the significance of the proposed research. We also described the future directions that could further strengthen the proposed research.

1.11 Summary

In this chapter, we synopsis and composed the overview of the web usage mining and its effectiveness in industry and academia. The trend identification in a weblog is a challenging and complex phenomenon. There is a close relationship between web users and web data. The identification of relationship among the users from web data is beneficial in web many ways. The sessionization problem statement has been composed to address the sessionization problem. This dissertation highlights the path and action in the form of proposed F_MET to address the sessionization problem. The research motivation paved the path for the research objectives and research scope in the field of web usage mining. The problem statement highlighted the web sessionization problems that are being faced in web mining and are essential to be tackled in the form of framework. The abstract level of F_MET was discussed as a solution of web sessionization. At the end of the chapter, we also elaborated the significance of web usage mining in industry and academia. In the next chapter, the foundation of web usage mining and web sessionization will be discussed.

Chapter 2

Background and Literature Review

The World Wide Web is the high-flying and the largest data repository, which is serving the millions of people around the clock. In recent years, the web is dominating the web-based e-business. The volume of web-based applications is increasing rapidly. The user reliability and confidentiality on web applications is the ultimate goal of web services providers. The web is a crucial clickstreams databank as millions of people exchange the bulk of information while communicating with web applications. To explore the trends from this databank through data mining techniques is a crucial and complex phenomenon and requires the proper web usage mining process to address the web sessionization issue. In this chapter, we are describing the theoretical foundation of web usage mining and literature review. The objective of literature review is to review the different web sessionization approaches to highlight the sessionization problem and to identify the gaps, weaknesses, issues and sessionization controversies that need to be essentially addressed.Furthermore, this literature review is a meta-analysis of web sessionization. It will enable us to integrate the findings to enhance the understanding about the sessionization problem and research gaps in the review to formulate the problem statement.

2.1 Introduction

In recent years, the data mining is playing a key role in the software industry and academia for the extraction of the novel, potentially useful, and knowledgeable patterns. The presence of such a huge amount of web data has opened the new research challenges for researchers to deliver an efficient information dissemination and retrieval mechanism to gain the confidence of web stakeholders. Data mining is providing a strong mechanism as a solution for web data analysis and delivering the path of action that converts the raw web data into useful information. Furthermore, it helps companies to understand customer behavior to introduce the competitive marketing strategies and decision support systems [45]. The data mining tools and techniques are used for the identification of hidden patterns and their analysis for knowledge visualization. In their research [46–48], explained the road map and data mining life cycle for the extraction of knowledge. The data-mining life cycle consists of iterative and interactive steps, when applied sequentially, produce the viable solution. The complete data-mining life cycle is as shown in Figure 2.1. The data mining has various applications in various disciplines of real life. The prominent data mining area such as bioinformatics, medical, mining, pattern identification, machine learning, data visualization, cyber crimes, and statistics. In recent years, the e-applications such as e-learning, e-medical,

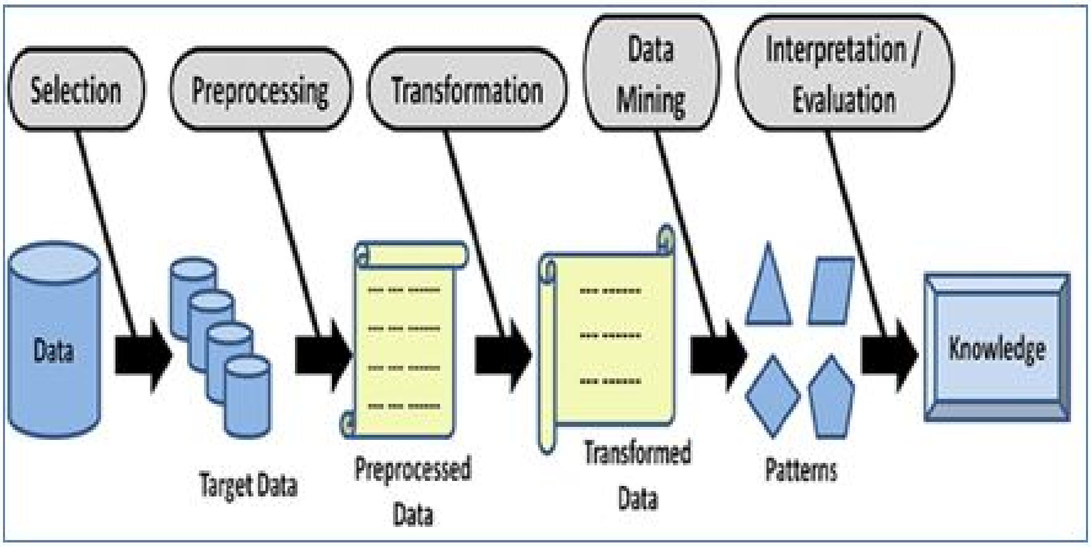


FIGURE 2.1: The Knowledge Discovery in Databases Gullo [48]

e-business and e-commerce are exponentially growing and web data has become the leading raw data repository for information retrieval. The focus of data mining techniques has also been enhanced and extended to cover the end user analysis for smooth execution of e-applications and better way to understand the requirements of customers. In this regard, web usage mining is playing a pivotal role to convert the classical data mining to emerging and evolving data mining to safeguard the interests of all web stakeholders.

2.2 Web Usage Mining (WUM)

The process of knowledge mining from huge data repository is known as Data Mining (DM) or knowledge-discovery in databases (KDD) [49]. The process of knowledge discovery from web data by applying data mining techniques is web mining (WM) and the process of knowledge discovery from web usage (user click-streams) is known as web usage mining (WUM) [50]. Web usage mining (WUM) is the application of data mining techniques to explore the weblog data for the pattern and knowledge discoveries. The role of web usage mining is expanding day by day due to the expansion of web size and its users. The web-based applications are also increasing and the user confidence and reliability is mainly relying on the web usage mining techniques and processes. In web usage mining, the focus of mining approaches is weblog, where the website visitor's clickstreams are recorded. The web usage mining is serving the web users and web administrators (web developers, web designers, web owners) in parallel by upholding the interests of all the stakeholder intact. There are numerous applications that are directly relying on the proper implementation of web usage mining process.

The web usage mining is serving the web users by delivering them the quality, accurate and focused information from the ocean of information (Internet). Information retrieval is the main area where web usage mining techniques actively and continuously working to improve it. Profiling; Prediction; Personalization; Recommender System; User behavior analysis are the trends for information retrieval.

On the other hand, the web administrators are also using these trends to improve the web applications and to gain the confidence of its customers.

2.2.1 Web Usage Mining Example

The effectiveness of web usage mining can be highlighted via visiting an online book selling website. The available books are categorized accordingly. Mining the weblog of the booking selling website may discover the different interesting patterns. The online customer, who is buying Web Mining book, may also be interested in buying the Data Mining book. This association between web mining book and data mining book can only be traced through web usage mining of the weblog of that particular website. There exists no other mechanism to exploit the user behavior and to find out such a correspondence and association between the users and users interests in various books can be used for web personalization by applying web mining techniques. Through web personalization, the website can offer data mining book to visitors who are buying the web-mining book [27].

There are numerous examples of web usage mining algorithms and its techniques in literature where we can find most appropriate applications such as online bookstores and other online shopping malls. Similarly, we can establish a suitable relationship between usage and user where we may anticipate the different usage relationships. For example, by applying sequential mining techniques on a given weblog, we can have result such as, that most of the users who visited page A to page B, due to the existing path between A and B. Such results can be misleading and indication of these relations among the different web pages clearly shows the usability problem of website structure. Since there is no link between two pages and the user may want to visit page B. In the absence of a proper link, the user may take the browsers search support to visit page B. Such minor errors are absolutely wastage of resources on both sides: user and website. Web usage mining techniques are very fruitful to indicate such type of flaws in website design and implementation.

Data mining overall has a strong potential to expose the relationship between business and its customers. For web based business and applications, this association is exposed through the web mining by applying data mining techniques in its true flavor. Web usage mining has a dual edge to expose the relationship between web-based business and its customers. Consequently, web usage mining is an ideal solution for web-oriented applications. Web usage mining is a multipurpose approach to data mining. It not only helps to manage the website but also provide a legitimate solution to business users as well.

There are a number of techniques that being applied to explore the web usage data from the web usage mining platform. These techniques are association rule mining; clustering; classification; and sequential pattern mining are the most frequently applied techniques for the pattern and knowledge discovery in the WebKDD. The selection of web mining techniques is a complex job as the proper selection of technique is mandatory and essential for the accurate and correct results with high coverage and precision.

2.2.2 Sessionization

Another key term is web sessionization coined by Nasraoui and Petenes [51], Roman et al. [52] and defined the sessionization as part of the WebKDD process for the extraction of precious knowledge from the weblog. The web sessionization and web usage mining are interchangeable words with similar mining strategies. The umbrella of web sessionization covers the weblog preprocessing; pattern (trends) extraction, and knowledge visualization. The term sessionization is also used to compose the web usage issue such as session construction, web user analysis, pattern extraction, and knowledge discovery. The main source of web sessionization is weblog and on the basis of weblog data capturing strategy, the sessionization is divided into proactive sessionization and reactive sessionization. In proactive strategy, weblog directly records and manipulate the user click record. The direct invasive into weblog data, compromises the user privacy. While, in reactive sessionization, post recorded weblog is used for mining purpose that is less sensitive to the user privacy policies. Due to privacy concerns, in research, mostly reactive sessionization is performed.

2.3 Web Usage Mining Taxonomy

The accelerated growth of the web and accessibility of large amount of unstructured user clickstreams data, data mining techniques are acting as a software industry backbone to deliver the dynamic and efficient results to gain the confidence of their clients (users). Researchers and academia are also putting high-ranking data mining techniques to fulfill the industry requirements. The web mining is one of the most dominating and effective data mining application areas where we apply the data mining techniques on the huge weblog data available across the world. The raw weblog data is available in structured; semi-structured; and unstructured formats. Web mining techniques are dynamic and robust that these techniques are not only effective on structured data but can also be applied on semi and unstructured data. Consequently, web mining is playing a critical role from converting man-understandable web contents to machine-understandable semantics. Arotaritei and Mitra [43] referred the web mining as the application, implementation, and usage of data mining techniques and algorithms to retrieve, extract and evaluate information for knowledge from various web resources and web data. Web Mining is further classified into three broader categories, commonly distinguished in following categories. 1. Web Content Mining 2. Web Structure Mining 3. Web Usage Mining.

2.3.1 Web Content Mining (WCM)

The WCM is well known branch of web mining. In literature, it is also known as text mining. In this phase, mining is performed on the contents such as text, images, graphics, audio and video of the website. The content mining establishes the relationship between the available contents and user queries [53, 54]. The content mining improves the information retrieval mechanism and helps the search engines to rank the lists as per user requirements and queries. By applying clustering, the top ranked contents can be grouped from the pool of websites for users. It also links the web pages level wise and reduces the irrelevant in response of user query.

The delivery of proper and accurate information is the key objective of the content mining. The content mining is active research area and very helpful to the industry for delivering to the point and accurate information to the web users. It also helps to create relevant databank for future use for the search engines. The content mining helps to build automatic categorization of contents, storage and disseminate the information in an organized manner [55]. The main use of this type of data mining is to gather, categorize, organize and provide the best possible information available on the WWW to the end user.

2.3.2 Web Structure Mining (WSM)

The WSM, is the second important category of web mining to implement the data mining techniques. In web structure mining, the structure of the website is linked (web pages) with the information linked (web contents) at the page level. Web pages are the basic information storage cell. By incorporating, the web mining techniques, the relationship between the links and contents is explored through the search engines. The web data structure holds the precious data linked web pages and web mining approaches build the connection to explore the website by the end user in the form user query. The simple mechanism used by the search engine is the use of robots, spider, and crawler to explore the website and the linking structure of the website to satisfy the end user request [56, 57].

The Internet is a free information mega-repository. The proper, accurate and relevant information retrieval is the one of the biggest challenges of the current global world. In this regard, web structure mining is minimizing the information retrieval gap between end user and the Internet. As the bulk of information is available on the web, searching the proper information in minimum time is also a big task. The structure mining, is helpful to index the huge information to optimize the retrieval time of end user query [58].

2.3.3 Web Usage Mining (WUM)

The WUM is the third and most important category of web mining. The mining is performed on the weblog data available in text format at web server. The weblog contains the user click history [59]. The weblog is the entity of web server OS and it captures the user clickstreams data automatically. The weblog is configured by the web administrator for analysis of OS errors. The weblog stores the user click paths which are the rich source of the web user browsing trends informations. The web usage is an attractive area for research and industry. It not only helps the individual users but also to the web designers and owners at the same time [60].

The companies get the traversing weblog of users and perform the mining to explore the data for multiple purposes such as prediction, profiling, personalization. The companies also get the first-hand knowledge for the future production and resources management. The web usage mining is beneficial not only to companies who have opted online business, but also to those who are managing the web services overall. The customers are the best critic and companies usually exploit the user click history to increase the business opportunities. The websites have no feedback mechanism or user mostly have not enough time to give the proper feedback to companies. However, the web usage mining provides the impartial and unbiased user feedback for future forecasting of web based business and e-applications [58, 61].

2.4 Preprocessing

The weblogs are basic and major raw source for WUM process [62] and are stored in plain text file (ASCII) [63]. The common weblog are Access Log; Agent Log; Error Log; and Referrer Log. Referrer log file contains the information about the referrer page or link. As someone jumps from any website to www.google.com by clicking the link, referrer log of Google server will record a referrer entry that a user came from that particular website. The referrer URL may be the linked web pages within the websites. Number of commercials, marketing, and advertising website use referrer log for their purpose.

Error weblog records the errors of the website especially when the user clicks on the particular link and the link does not locate the promised page or the website and the user receives *”Error 404 File Not Found”*. Error weblog is more helpful for the web page designer to optimize the website links. Agent weblog records the information about the website users browser, browser’s version and operating system [63]. This information is again utilized by the website designer and administrator for the analysis that users are using which specific browser to access the website. There are number of browsers available to users and each browser has its own properties and advantages to their users. Different version of same browser can have various added utilities and benefits to its users, so website can be modified accordingly. Information about the users operating system is also helpful for designer and website changes are made accordingly.

Access weblog or weblog is a major log of web server, which records all the click-streams, hits, and accesses made by any website user. There are number of attributes in which information is captured about users. Table 2.1 elaborates the different attributes of access log along with their description. The data is not

Table 2.1: Weblog Attributes, Format in which the attributes are stored and their Description

Attribute	**Format**	**Description**
Client IP	Customer IP	Client Machine IP Address
Client Name	CS User Name	Client Name and Password if Provider by Server otherwise Hyphen ”–”
Date	Date	Date on which client accessed the website
Time	Time	Time along with date on which client accessed the website
Server Site Name	S- Sitename	Internet Service Name on Client Machine
Server Computer Name	S-Computer Name	Web Server Name
Server IP	S-IP	Host Machine IP
Server Port	S-Port	Host Machine Port for Data Transmission (80/8080)
Client Server Method	CS- Method	Client Method of Request (GET/POST/HEAD)
Client Machine URL Stem	CS-URL-Stem	Targeted Default Web Page of website
Client Server URL Query	CS-Method	Client Query after ”?”
Server Client Status	SC-Status	Status Code returned by the Server (200,404, 300)
Server Client Win32 Status	SC-Win32 Status	Windows Status Code
Client Server Bytes	CS-Bytes	Number of bytes received by client
Server Client Bytes	SC-Bytes	Number of bytes sent by Host to Client
Time Taken	Time Taken	Time Spent by Client to perform any action
Client Server Version	CS-Version	Protocol Version as HTTP
Client Server Host	CS-Host	Host Header Name
User Agent	User Agent	client Browser
Cookies	Cookies	Cookies Contents
Referrer	Referrer	Link Page of Client Request

captured in all the available weblog attributes. The server administrator captures the weblog clickstreams only in mandatory attributes to save the server resources. Secondly, the nature of HTTP protocol is stateless and design of websites is responsible for storing all the objects (audio, video, images, css) available on each web page. The crawler, robots and administrator's actions (update, insert, delete) are also the part of weblog. Due to these discrepancies, weblogs contains around 80% raw data. For web usage mining, only the web pages (URLs) visited by users are helpful for weblog analysis. The presence of such a huge amount of irrelevant entries requires that the noise free preprocessing is a must. Following are the preprocessing techniques which are in practice.

Data Cleansing: In this technique, the irrelevant entries are removed. The entries such as audio, video, images, style sheets available on the each page are stored in weblog. These entries are removed. Similarly, the administrative tasks and crawler entries are removed. When successful web page is delivered to the user, in status code attribute the "200" is stored otherwise, the error code is recorded. All the other status codes are removed as we are interested in only those web pages that have been successfully delivered to end user.

Path Completion: The excessive use of technology has made the web access more users friendly. Some of the web pages visited by users are provided through the cache and the cookies of local machine. This technique reduces the web server load. On the other hand, the web server machine weblog does not has the record of these web pages. With the help of structure mining, the broken links are completed through path completion technique.

User Identification: The weblog records the client IP and this IP is unique. The IP is user identity to trace the user clickstreams. However, the use of proxy server and firewall, the single IP is issued to many clients (Users). The capturing of the true user is a cumbersome job. There are different mechanism adapted to identify the genuine user. These heuristics are:

- IP Based
- IP and User Agent Based

- IP, User Agent, and OS Based
- IP, User Agent, OS, and Referrer Page
- Back Button or Click

Web Session: The weblog is the primary and basic data input for the web sessionization. The weblog contains the user activity and clickstreams history. The weblog can be directly used in web usage mining process. The raw weblog is preprocessed to filter around 80% irrelevant entries. The processed weblog is then converted into sessions as necessary step. The IP is key attribute of weblog to identify the individual users from weblog. However, the IP is complex and do not represent the single user due to proxy server, and firewall. Consequently, sessions are constructed from the weblog to prepare the weblog for web sessionization process. Session is the group of user activities (clickstreams) within the login and logout time. Login time is the time when user arrived on the website first time on any link (page/URL). Onward the user traverses various pages on that website as per requirement and desire. Every click of user is recorded along with timestamps in weblog. The logout time is the time when user leaves the website and logout time is picked from the last activity of user on that particular website. All the user activates are grouped in the form of sessions. The average session time is 30 minutes. However, the user may spend more than 30 minutes and extra time is converted into multiple sessions of that user.

2.5 Web Session Similarity

A similarity measure symbolizes relations among the objects, which can be either, documents, queries, attributes, and features of any database. Similarity measure helps to rank the objects in accordance with their importance in specific data mining application. A similarity measure is defined as a function that computes the degree of similarity between a pair of objects [64]. The similarity or dissimilarity between two objects or entities plays core role in data mining applications for knowledge discovery where the objects have to be classified on the basis of distance computations [65]. Data mining applications such as clustering; classification and distance based outliers detection require the similarity or distance measure between their objects. If we are able to find out how much similar are the data objects, we can have better results of classifier. Similarity measure gives us the precision and accuracy of closeness of relationship between objects. It can be apprehended that proper selection of similarity measure is key process.

The Web Session Similarity, computation among the web sessions (data objects) is although a complex, however, a significant sessionization problem in the web usage mining process at the early learning stage of WebKDD [66]. Can we obtain the web sessions with high coverage and precision from preprocessed user clickstream? The valid and accurate session construction requires the proper and quality web session similarity metric for enhanced analysis of the web usage mining process to address the web sessionization problem.

2.5.1 Euclidean Distance

In web usage mining, the Euclidean distance measure is frequently applied. It is the best-suited distance measure for numerical data. The Euclidean distance is very effective and produces excellent results where the clusters are independent and isolated [29]. Even though, the Euclidean metric is defacto measure in web session clustering, however, it has drawbacks in applying on weblog data [67]. The nature of weblog data is unstructured and categorical. If the given two web sessions have no web pages in common, that session pair have short distance than the pair have more common web pages. Another problem with Euclidean distance is the convergence of categorical web data type to numerical data format. This convergence effects the nature of the web sessionization. The standard computation of Euclidean Measure is given in Eq 2.1 [29].

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (S_{ik} - S_{jk})^2} \tag{2.1}$$

Where $S_{ik} = (S_{i1}, S_{i2}, S_{i3}, ..., S_{in})$ and $S_{jk} = (S_{j1}, S_{j2}, S_{j3}, ..., S_{jn})$ be the two given sessions with k weblog attributes.

2.5.2 Cosine Similarity

Cosine similarity measure is frequently applied in web sessionization. It is simple in implementation and takes only the common web pages with relation to the total number of web pages present in both the sessions. The Cosine measure is also used in clustering for content mining. The similarity of two web sessions correlates between the cosine vector of two sessions. It has various applications in data mining and machine learning. The following Eq 2.2 is showing the computation details of Cosine similarity measure [68].

$$Cosine(S_a, S_b) = \frac{|\vec{S_a} \cap \vec{S_b}|}{\sqrt{|S_a|.|S_b|}} \tag{2.2}$$

Where $|S_a|$ and $|S_b|$ give the number of web pages traversed in each session.

2.5.3 Jaccard Coefficient

The Jaccard coefficient is also known as Tanimoto coefficient. It computes the web session similarity by taking the common web session and dividing them by the web pages available in both the sessions. It belongs to the Cosine similarity family. In Cosine similarity, we divide the common web pages in given two sessions by the total number of pages available in both the sessions [69]. The computation formula for the Jaccard similarity given below in Eq 2.3. It is commonly used in web session clustering for pattern discovery in weblog data [70].

$$Jaccard(S_a, S_b) = \frac{|\vec{S_a} \cap \vec{S_b}|}{|\vec{S_a} \cup \vec{S_b}|} \tag{2.3}$$

where S_a and S_b are two given sessions and denominator is union of both the sessions to keep the Jaccard coefficient within range between 0 and 1.

2.5.4 Canberra Distance

The Canberra distance (CD) is used to compute the distance between the two given web sessions. It computes the distance numerically and coverts the categorical data from weblog into quantitative like Euclidean distance metric. It is used number of data mining techniques such clustering and classification. It is effective for large datasets and scalable in attribute coverage [71, 72]. The computation of Canberra distance shown in Eq 2.4.

$$d(S_i, S_j) = \sum_{k=1}^{n} \frac{|S_{ik} - S_{jk}|}{|S_{ik}| + |S_{jk}|} \tag{2.4}$$

Where $S_{ik} = (S_{i1}, S_{i2}, S_{i3}, ..., S_{in})$ and $S_{jk} = (S_{j1}, S_{j2}, S_{j3}, ..., S_{jn})$ be the two given sessions with k weblog attributes.

2.5.5 Angular Separation

Angular Separation (AS) measures the cosine angle between the two given sessions and measures the similarity rather than distance like Euclidean [72]. It is computing the web session similarity like the cosine metric. Its value ranges $[-1, 1]$ and higher the angular value between the session , higher the similarity. The formula for the angular separation is given in Eq 2.5 [73].

$$d(S_i, S_j) = \sum_{k=1}^{n} \frac{S_{ik} * S_{jk}}{[\sum k = 1^n S_{ik}^2 \sum k = 1^n S_{jk}^2]^(1/2)} \tag{2.5}$$

Where $S_{ik} = (S_{i1}, S_{i2}, S_{i3}, ..., S_{in})$ and $S_{jk} = (S_{j1}, S_{j2}, S_{j3}, ..., S_{jn})$ be the two given sessions with k weblog attributes.

2.6 Web Session Clustering

Clustering is a well-known unsupervised technique [74] and cluster is the collection of similar items (objects) within the same cluster while dissimilar to the items of the other clusters. The focus of clustering algorithm is to group the most similar data objects in a cluster. Clustering helps us to organize the data objects into different groups based on certain similarity among the data objects. The key factor for quality clustering is depending on the proper selection of similarity measure [75]. Mainly clustering is used for data analysis in various applications such as marketing analysis, e-business, pattern recognition, data visualization. It is also used as preprocessing step for other algorithms.

The quality clustering classifier not only produces the high quality clusters, but also clusters with maximum intra-cluster similarity and minimum inter-cluster similarity. The quality clustering classifier can also discover the trends from the data. Another feature of quality clustering algorithm is selection of similarity measure and implementation of that similarity measure in relevant data context. In web mining domain, the clustering algorithm must be efficient, high coverage, scalable to handle the features and attributes of data with minimum domain expertise.

In web usage mining clustering is also practiced widely for the web Sessionization clustering [76]. In web session clustering, either it is item based clustering or user based clustering, the similarity measures is an important factor for grouping the users. According to Forsati et al. [77] the user clickstreams are rich source information and knowledge. We can extract the knowledge from the clickstreams by applying the web session clustering. The different data mining techniques are being applied to the pattern and knowledge identification in weblogs, however, web session clustering has advantages over the rest of techniques such as data analysis and data visualization. In following sections, we discussed briefly the major clustering techniques that are being practiced in web session clustering.

2.6.1 Partition Clustering

In partition clustering technique, the set of items/objects is divided into n number of predefined clusters. Each cluster is at least has one data item and every data item belongs to exactly one cluster. No data item can be the member of two or more cluster at a time. In web usage mining, sessions are treated as data objects or data items. The sessions are partition into predefined k clusters. Each cluster has a centroid. K-MEANS and K-Medoids are most commonly practices clustering approaches for web sessionization. There are number of variants available in web session clustering such as k-modes, frequency based. The variant are easily developed through the selection of different distance metrics. The computation of centoids and selection of initial seed may also produce the variants of partition clustering [78, 79].

The partition clustering approaches are simple in implementation and appropriate for large datasets. The partition clustering techniques suffer the global maxima issue. Mostly delivers best results when applied to quantitative datasets with numeric distance measures such as Euclidean; Minkowski, Mahalanobis, Canberra, and angular measures.

2.6.2 Density Based Clustering

Density based clustering approaches produce the dense clusters in the sparse datasets. This approach identifies the unique and distinct clusters with minimum noise in the dense region. The clusters produced as maximal set of connected point. Density based approaches are helpful to identify the noise and outliers. The points that do not belong to any clusters are marked as noise and outliers. DBSCAN, CLIQUE, and OPTICS are the commonly practiced density based clustering techniques. In web usage mining, its application are very rare due to nature of weblog data [80, 81].

2.6.3 Hierarchical Clustering

Hierarchical clustering approaches are prominent in data mining due to its simple nature and helpful in data analysis. Unlike the partition clustering techniques, hierarchical clustering is not required the prior knowledge of number of clusters and seeds (centroid). However, the similarity metrics are used for intra-cluster similarity. The hierarchies of data items can be computed through agglomerative and decisive techniques. The agglomerative clustering technique is known as bottom-up approach, in which every session is treated as single cluster and then these clusters are merged based on similarity measure until all the clusters are merged in the form of tree structure. In decisive approach, all the data items (sessions) are considered as one single cluster. Then in successive step, the single cluster is split into sub clusters based on similarity measure to reach at the root of tree. The merging or unmerging criteria for the inter cluster distance can be the single linkage (nearest cluster), complete linkage (furthest cluster) and average linkage (average of two clusters). Hierarchical clustering techniques are slow and time consuming. The outcome of dataset is presented in the form of tree (dendrogram) [82–84].

2.7 Overview of Web Sessionization

Web sessionization is an active research area to obtain the unbiased and focused groups from weblog for the identification of interesting patterns, which are previously unknown [42, 52]. Moreover, the WUM is a complete process for mining hidden knowledge from weblog and Sessionization is a very important step as the rest of WUM process steps are solemnly depending on this step [12]. The expansion of the web is an emerging challenge and researchers are interested in coping with the size of web, accuracy, quality; noise and scalability issues to help web customers to safeguard their web interest accordingly by applying various web mining techniques [85]. Since the inception of the web in the nineties, the role of web usage mining became the necessity to address the above-mentioned issues. The researchers have proposed various techniques to deal with the sessionization challenges and issues. However, researchers are still striving hard to deliver the viable framework based on web mining techniques to address the sessionization.

The different phases of WUM [35] have already been discussed in Chapter 2. The sessionization is composed of various sequentially inter-linked steps such as web session identification; web session similarity; web sessionization technique; and knowledge visualization from the weblog. The web user sessions are constructed from weblog by applying the heuristics such as IP Address, User Agent, User OS and Referrer Page, after preprocessing step [8]. According to Roman et al. [52]

sessions are the primary input for the complete web Sessionization and is a crucial step as well [86]. The identification of pair-wise relationship among the web sessions is an essential and decisive step for the analysis of web user sessions [87–89]. The rest of WUM processes exclusively rely on the proper implementation of Sessionization steps [20]. Weblog Sessionization is achieved by applying various Sessionization techniques such as Classification; Clustering; Association Rule Mining; and Sequential Pattern Mining; [2, 90–95].

In the following sections, we are reviewing the literature in multi fold directions to investigate the sessionization problem. In the first part, we are reviewing the various session similarity measures to come up with the best option to identify the pair-wise relationship between the web sessions and session matching. In the second part, the web mining techniques are being reviewed with pros and cons and how these techniques are helpful to overcome the issues of web sessionization. In the last part, we are analyzing the hierarchical sessionization techniques for the focused and visualized pattern discovery from weblog and the effectiveness of these patterns in knowledge visualization in WebKDD.

2.8 Review on Web Session Similarity

With the advancement of data storage technology, bulk of user transactions are captured in the form of weblog. Information and knowledge retrieval from the web has become a challenging research. The web session similarity computation among web sessions is although a complex, however, a significant problem in unsupervised learning. The identification of similar sessions from the weblog data is a non-trivial solution to the sessionization problem and to capture the users with similar traversing behavior is further applied to the various web applications such personalization, recommender systems, decision support systems, prediction and system improvements. Moreover, web session similarity is also a primary artifact used in clustering and classification for pattern discovery and pattern analysis. In following paragraphs, we are discussing the web session similarity measures for identification of research gaps and their limitations. Table 2.2 summarizes the literature on the web similarity measures.

According to Sisodia et al. [96], augmented web session similarity is more useful as it is based on the user interest. Furthermore, session similarity is the core of the web session clustering to understand and capture the user behavior from the weblog [40]. In this research, Sisodia et al. [86], computed the web page interest (relevancy) in a web session by incorporating the interest of user on any page and the frequency of page visited in Eq 2.6.

$$RoP_{pi} = \frac{2 \times DoP_{pi} \times FoP_{pi}}{DoP_{pi} + FoP_{pi}} \tag{2.6}$$

Where $(DoP)_{pi}$ page duration in ith session and, $(FoP)_{pi}$ is the frequency of the page in ith session, and is the relevance of page in ith session. The authors computed the relevance matrix (RM_{mxn}) based on Eq 2.6 through page stay time (duration)$(DoP)_{pi}$ in a session and frequency of page $(FoP)_{pi}$ in a session. After calculating relevance matrix, authors applied traditional Cosine Similarity measure to find out the different flavors of web session similarity and produced the different outcomes. In this research, the authors applied the use of page duration and page frequency. To compute the page interest from the time consumed by a visitor is a weak parameter as it cannot be justified that time spent by the user was due to page interest? Similarly, page frequency computation is useless as we have a number of tools that can provide the number of page hits. Based on these two parameters, the sanctity of the WUM process is not sure and even invalid. The authors also claimed to identify the realistic relationship between the sessions is also questionable due to frequency of page hit that might be wrong in case of index page or link page from referred page. The authors also applied the different combinations of classifiers and this is clear indication that proper selection of web session similarity is vital and important matter in web sessionization.

A recent research on the sessionization investigated the effectiveness of web session similarity to group the similar web pages visited in an order (sequence) in a session [97, 98]. The authors combined the techniques of Needleman-Wunsch (NW) and Smith-Waterman (SW) to propose the new similarity measure to overcome the limitations local and global maxima of Euclidean, Manhattan, Levenshtein, Hamming Distance and Longest Common Sequence (LCS). The proposed similarity measure considered the maximum size (l) of the LCS of the two given sessions and the respective similarity score is being calculated for matching and mismatching through Eq 2.7.

$$S(s_i, s_j) = [\frac{NW(s_i, s_j)}{l}] + [\frac{SW(s_i, s_j)}{(2 * l)}] \tag{2.7}$$

Where $NW(s_i, s_j)$ NW and $SW(s_i, s_j)$ SW are two scores between given two sequences s_i, s_j and l is the length of longest sequence. The proposed web session similarity measure strengthens the idea that traditional similarity measures are inappropriate to gauge the user traversing behavior in web usage mining. The authors Luu et al. [97], gave the due importance to the Dynamic Time Warping (DTW) and compared the proposed measure results with it. However, the authors were unable to address the time factor in the proposed measure. Furthermore, the authors weighted the matching and mismatching but are silent on the issue of uncommon pages traversed between two sessions. Here arises the question of accuracy and precision of sessions generated?

According to Yu et al. [75], the growth of the Internet has increased the demand to group the similar users exhibiting similar traversing behavior. The authors proposed a novel session similarity measure Minimum Support for Large Web Page (MSLWP) and calculated the threshold support of each page in different sessions Eq 2.8.

$$Sup_{p_{ij}} = \frac{N_{p_{ij}}}{N\_Session_i} \tag{2.8}$$

Where $N_{p_{ij}}$ is the number of time page visited by a user in different sessions and $N\_session_i$ be the total number of user sessions. The authors applied the threshold based technique to identify the users traversing the similar web page and fixed the threshold value to 0.25. The concept of dynamic threshold values is being applied in Aprori based pattern identification techniques and were failed to get the researcher and industry appreciation. Fixation of threshold might be unable to find proper link between the user sessions and that may be drastic to the upcoming steps of WUM.

Dixit and Bhatia [99] discussed the two major challenges of web sessionization, such as the quality outcome of the clustering algorithms and similarity/dissimilarity of the web sessions [66, 100]. The authors applied the Jaccard and Cosine similarity measure and the combination of both for the session similarity to manage the refinement and quality sessionization. The major artifacts for web session were accessed time and web pages viewed by the users in a session. The authors applied the evolutionary approaches to tackle the issue of scalability, quality, and refinement of web sessionization and produce the series of research work based on the cluster refinement issue. Both the Jaccard and Cosine measures computation cycle is similar except minor change in denominator. The authors worked out with the traditional similarity measures and these measures are the overload of the web sessionization process.

According to Pai et al. [2] organizations are using web based applications to improve their business by expanding their business area and reducing their cost. In order to achieve this, organizations are interested in studying and analyzing their web customers (visitors) behavior. Pai et al. [2] proposed a mechanism to find out the similar visitor sessions and applied the Large Margin Nearest Neighbor (LMNN) technique. To address the Sessionization, feature selection was made from click-streams based on URL visited by users in a 30-minute session. For each feature, the Mahalanobis learning metric was used that transforms Euclidean Metric to overcome the equal weight issue of Euclidean Measure. Pai et al. [2] applied the Mahalanobis metric to overcome the limitations of Euclidean Metric for session similarity. As the weblog data is not of numeric data type, by just converting the number of pages traversed in a session to a number is insufficient and by this way, every session has a common page with others. The single large website can offer different categories to their relevant users. Hence, the Euclidean family of measure identifies the wrong relationship among the various sessions and that leads to the weak and even poor quality results at a later stage of WUM process.

Roman et al. [52] categorized the WUM as an important process to address the web sessionization as a proven strategy in e-business for user pattern extraction and for the improvement of user navigation and asserted the notion that the accuracy of session identified process is a complex phenomenon. Moreover, sessionization is an important step in the WUM process for pattern extraction from large data repositories. In this research Roman et al. [52], addressed the problem of estimating accurate user sessions from a weblog. To address the sessionization problem, Roman et al. [52] presented Sessionization Integer Program (SIP) and Bipartite Cardinality Matching (BCM), optimized models. Integer programming and navigation-oriented heuristics were used for the session construction. The graph-based techniques required that weblog must be transformed into a graph. For a large weblog of any large dynamic websites which may contain millions of web pages, the transformation and then graph matching may not be able to seek the accurate [101] and noise free sessionization even though the integer programming has the linear complexity for SIP. Secondly, for the verification of results, no comparison was performed with any other existing techniques of sessionization.

Han and Xia [102] highlighted the importance of WUM for web server log analysis, single user analysis and proposed the user characteristics based session similarity between the weblog sessions. The concept of time was used to find out the long and the short user interest. The normal threshold value of $\theta > TTFIV$ for long term interest and $\theta < TTFIV$ for short-term interest was applied and the usual threshold value is 30 minutes. Han and Xia [102] applied the threshold base approach for session identification and frequency interest based similarity was calculated between users sessions. The threshold-based techniques are failed to depict the true relationship between the web sessions in either form. Whether threshold is dynamic of static; the web session similarity results are unable to deliver a viable solution to the sessionization issue. Furthermore, user interest was calculated based on time attribute only, while the pages visited are also key log attribute that was totally ignored.

Alam et al. [41] categorized the weblog data as heterogeneous because it is composed of both numeric and categorical data types. The session time, a number of pages traversed by a user in a given session and data downloaded during that particular session are numerical while the pages visited are categorical. To find out the similarity among the web sessions in such a sparse nature of data is a difficult task. Alam et al. [41] used two-step techniques to calculate the session similarity among the sessions. In the first step, the user session is marked as $XS_{ai} = XS(a_1, a_2, a_3, a_4, a_5)$. Where a_1 first attribute and dissimilarity among two sessions is calculated as under Eq 2.9.

$$d(XS, YS) = (\sum_{i}^{n}(XS_{ai} - YS_{ai})^2)^{\frac{1}{2}} \tag{2.9}$$

Where $d(XS, YS)$ is dissimilarity between two sessions XS and YS. In the second step, a hybrid of Boolean and Euclidean distance was used to compute the Boolean distance among the user sessions Eq 2.10.

$$b(XS, YS) = (\sum_{i}^{n}(XS_i \bigcap YS_i)^2) \tag{2.10}$$

The final distance among the various sessions is computed in following Eq 2.11.

$$Dist(XS, YX) = d(XS, YS) + b(XS, YS) \tag{2.11}$$

Alam et al. [41] used the Euclidean measure to compute to dissimilarity among the sessions, which is proven failed measure for web session similarity to address the accuracy of generated sessions. The authors claimed the heterogeneity of the weblog file and even then applied the numeric data type of metrics for session similarity. This will generate mock results of web session similarity as well as WUM process. Chen et al. [17] computed the similarity among the users (sessions) through the fractures and each web user is represented as a set of FRACTURES and the user similarity (US) is computed in the range [0,1] in following Eq 2.12.

$$US(u_1, u_2) = \frac{\sum_{k=1}^{n} \delta_k FS_k(u_1, u_2)}{\sum_{k=1}^{n} \delta_k} \tag{2.12}$$

Where δ_k are shared fractures of two users. The two major limitations were discussed for proposed similarity such as common fractures and the denominator as total shared fractures. For the large and dynamic websites, the fracture based similarity measure is not an optimized solution to tackle the quality and scalability issues of web sessionization. Nasraoui et al. [11] computed the web session similarity score between session and profile by cosine similarity and web session similarity was computed from URL to URL based on overlapping Profiles P_i and P_j in the following Eq 2.13.

$$S_u(i,j) = \begin{cases} 1 & \text{if } i = j \\ \frac{\min(1, |P_i \bigcap |P_j)}{\max(1, \min(|P_i|, |P_j|) - 1} & \text{otherwise} \end{cases} \tag{2.13}$$

Nasraoui et al. [11] applied the Cosine similarity measure while in their previous work [22] bluntly criticized the Cosine similarity for web sessionization. In their previous research, [51], applied the Euclidean Distance to find the similarity among the session. The Euclidean measure is widely criticized due to its nature and its application in web usage [22, 44, 51, 103], the Euclidean Distance, Cosine, and Jaccard measures are not suitable measures for web session clustering due to the nature of user clickstreams data. Li [103] proposed the time based and URL page similarity among the page visited by different users. For any two web pages visited, the page viewing time was [0, 1] and for matching similarity, the similarity score is 20 and for mismatch and in between the gap, the similarity score is -10. The session similarity was computed by utilizing the total web page viewed time through dynamic programming. The only issue of match and mismatch among the sessions were considered while the similarity must be relative to sessions. The mechanism to calculate the total page viewed time not explained. The cache issue and broken paths of web log file can deliver the misleading results and this will affect the quality of the results produced by such a similarity measure.

Duraiswamy and Mayil [104] carry out the session clustering through the agglomerative hierarchical clustering algorithm. Duraiswamy and Mayil [104] used alignment score S_a and local similarity S_b to compute the similarity between the given two sessions (Eq 2.14 and 2.15).

$$S_a(S_1, S_2) = \frac{v}{(S(m) * M)} \tag{2.14}$$

$$Sim(S_1, S_2) = Sa * Sb \tag{2.15}$$

Duraiswamy and Mayil [104] performed sessionization by utilizing dynamic programming and hierarchical clustering approach. The accuracy of session produced was not managed as weblog data is prone to noise and assigning equal weight to all the matching session sequence irrespective of a number of pages visited. Furthermore, similarity measure justification was not discussed. In Table 2.2, we are summarizing the web session similarity measures from literature discussed.

2.9 Review on Web Sessionization Techniques

The Internet is the mass transit route for delivering the various services such as e-commerce, e-learning, entertainment, social linkages and much more. The millions of web users interact with the web and take the advantages of this mega knowledge repository. The exploration of this huge data requires the scalable data mining tools and techniques [6]. The Web Usage Mining techniques are applied as an analytical tool from the WebKDD platform for investigating the usefulness of this huge data. The user interactions with the web are of mega-worth to understand the user likeness and dislikeness for the improvement of the web and its services [105]. These user interactions with the web are locked in a weblog on a web server that contains precious hidden knowledge about the user. To extract the useful patterns from weblog about the user behavior, choices, prediction, and recommendations, researchers have applied the various data mining techniques such as Clustering; Classification; Association Rule Mining; and Sequential Pattern Mining [106, 107]. In the following section, we are reviewing these web mining techniques that are frequently used in web sessionization. The literature is summarized in Table 2.3 along with the limitations of existing web sessionization techniques.

Table 2.2: Summary of Web Session Similarity Measures

Authors	Technique	Advantages	Limitations
Sisodia et al. [86]	K-Mediods	Applied the different flavors to compute the session similarity such as ASS, AUSS, IASS, AHSS	Time spent on Page (Duration)
Sisodia et al. [96]	Fuzzy C-Means	Critically analyzed the existing measures.	Frequency of Page, Cosine Measure
Luu et al. [97, 98]	Hierarchical Agglomerative Clustering, Dynamic Programming	Highlighted the issue of Accuracy; Precision; Noise, Proposed Hybrid measure	Matching and mismatching scoring criteria, Time factor not considered, Accuracy and precision unattended
Yu et al. [75]	K-Means	Labeling of Clusters with TF-IDF, Preprocess the weblog	Threshold base, IP-based user grouping, Page link graph
Dixit and Bhatia [99]	K-Mean, GA, PSO, MKRA	Refinement, Scalability, Quality	Jaccard Measure, Cosine Measure, Sessions IP and Timeout
Pai et al. [2]	Large Margin Nearest Neighbor (LMNN)	Comparison with HMM and SVM, Sessionization problem identification	Mahalanobis metric failed to overcome inadequacies of Euclidean Metric , The accuracy and precision of session generated were not concentrated
Roman et al. [52]	Session Integer Program (SIP), Bipartite Cardinality Matching (BCM) optimized models	The integer programming constructs sessions. Accuracy of sessionization discussed.	SIP and BCM for large data are inappropriate. Session accuracy was not compared with any other published accuracy measures
Alam et al. [41]	Euclidean distance, PSO for hierarchical clustering	Assigning weights to Euclidean Measure, Addressing the heterogeneity of web data	Taking an equal number of visited pages in each session. Accuracy of sessions ignored. Quality of sessions not discussed
Chen et al. [17]	COWES web user clustering	The clustering was performed agglomerative algorithm. Proposed new similarity measure	The complicated similarity computation by use of fractures. Quality of sessions produced was not discussed

According to Shivaprasad et al. [108], the traditional web clustering approaches were unable to cater the issue of overlapping user behavior. Only the Fuzzy approach is capable of managing the overlapping behavior as hardcore clustering techniques construct the crisp clusters and these clusters are unable to manage the user behavior in its true form. Furthermore, authors claimed that single clustering techniques are also unable to deliver the satisfactory results. These two issues are being managed by the authors by incorporating the neuro-fuzzy technique for web sessionization. A fuzzy algorithm is versatile clustering approach and fuzzy alone is sufficient to manage the quality clustering, however, authors were unable to point out the issues which fuzzy failed to maneuver and combined the neuro to construct the session clustering. The results were not compared to the existing Fuzzy based web sessionization techniques and the quality of the results might not be as fruitful as claimed. Forsati et al. [101] investigated in their research that user traversing patterns can be modeled in a session format to extract the user behavioral patterns and recommended the system. For the identification of these patterns from web sessions, the authors applied the binary session clustering to partition the sessions into a number of fixed clusters. Forsati et al. [101] applied the hybrid technique to tackle the issue of scalability through binary clustering technique and k-means to cater the issue of local maxima of the evolutionary approaches by the harmony search to produce the quality clusters. The fitness function used for binary session clustering is explained in Eq 2.16.

$$fitness(C) = \frac{1}{k}\sum_{i=1}^{k}\frac{(\sum_{SeS_i} D(S, C_i))}{n_i} \tag{2.16}$$

The authors applied the binary partition clustering along with the combination of k-means for pattern identification. The partition clustering techniques are not scalable as claimed by the authors, and for quality clustering, k-means was used to handle the issue of local maxima. This is a complex hybrid technique which is again an overhead to the mining process. The K-means is one of the prominent partition clustering technique, however, partition clustering approaches are unable to deliver the focused groups in web sessionization.

In their research, Mishra et al. [109], intended to deliver the right information to the right user through web recommended a system with the help of website contents and sequential information of user traverse in the form of weblog. For pattern identification, the authors applied clustering through the classifier, rough set upper approximation technique to manage the overlapping behavior of users. The session length was varying from 1-500 and authors applied the average length of 6 during the experiments and ignored the other sessions with any proper justification. Such sorts of data modifications are unable to deliver the precise and accurate results as claimed by the authors, even though how strong clustering algorithm may be incorporated. Moreover, the results were not compared with any well-known soft clustering algorithm such fuzzy approximation. Web usage mining is playing a pivotal role in e- commerce for identifying buying pattern and suggesting ways improve the user surfing [52]. The authors used integer programming for Sessionization and proposed bipartite cardinality matching algorithm to produce accurate sessions and presented two optimized Sessionization models Sessionization integer program (SIP) and bipartite cardinality matching (BCM). The authors used the SIP binary variable X_{ros} where r is log register and o is the o^{th} position in sessions. The objective function is defined as Eq 2.17.

$$Z(X) = \sum_{ros} C_{ro} X_{ros} \tag{2.17}$$

Where C_{ro} is coefficient for register r in o^{th} session. In BCM model, Sessionization was performed through matching cardinality in the bipartite graph by connecting nodes through edges in the undirected network. The bipartite cardinality is defined as $C_{r,0=1,s} = -1 \forall rsandC_{r,0>1,s} = 0 \forall rs$ and . Both the techniques SIP abed BCM are good effort to generate sessions, however, both the techniques suffer the performance and scalability issues. Both the techniques can work efficiently on small data sets. In today's world, the number of web users and web-based applications are growing exponentially whereas technology has also enabled us to capture the millions of user access. The high dimensional data cannot be accommodated through these techniques. Secondly, research was unable to answer the

Sessionization technique to cater the issue of the proxy server to depict the real users from web log file where users are hidden behind the single IP.

According to Kotiyal et al. [110], web log file contains useful user access data that can be mined for precious information. The authors applied Naive Bayesian (NB) technique to classify the user access behavior through Weka. The authors also discussed the importance of user classification for the efficiency and effectiveness of the system overall by reducing the browsing time. The classification was performed on the basis of the training data set and in next step, the trained data were used to classify the future browsing. The CSV file was loaded into Weka and Naive Bayesian algorithm performed the classification. The authors did not mention the parameters used in NB. No scoring function was discussed which play a key role in NB to classify the data. How the prior, likelihood and posterior probabilities were calculated, these points were not discussed in this research. The authors evaluated the accuracy and performance of the classifier through Precision, Recall, and F-Measure, which strengthen the research work.

Bayir et al. [20] termed the WUM as an important technique to discover interesting patterns from weblog and these patterns help the administrators to understand the web server needs and web domain design. The authors termed the session reconstruction as an important and challenging activity in WUM process and mandatory step before discovering the user patterns from the weblog. For the session reconstruction, the authors applied the Smart-SRA [111] technique to cater the limitations of time-based heuristics and used page-stay and time of the session by a Topology rule: $\forall i : 1 \leq i \leq n$, there is a hyperlink from $P_i\mathrm{P}_{i+1}$. After the session reconstruction step, the authors applied the Sequential Apriori Algorithm for pattern identification. It is no doubt that the session reconstruction is a tedious job in WUM process and authors well managed the backward browsing and page-stay time, however, the use of Apriori and support has their own limitations such as database scanning in each iteration. Moreover, the support base heuristics approach can miss the few important patterns. Furthermore, Apriori is also unable to handle the big data issue and scanning database is another overhead to deliver the accurate results in due time.

According to Vellingiri et al. [33], web log file provides useful knowledge about the user behavior and to study the website structure. Furthermore, web log file is mined to predict the next user click to help the user to be more focused on the topic in cyberspace navigation. In this research, the authors performed the complete WUN process. For pattern discovery, the Weighted Fuzzy-Possibilistic C-Means (WFPCM) algorithm was used to cluster the users having similar behavior. In pattern analysis phase user behavior was analyzed through Adaptive Neuro-Fuzzy inference System with Subtractive Algorithm (ANFIS-SA). The objective function JWFPCM for WFPCM was calculated through FCM and is defined in Eq 2.18.

$$J_{WFPCM}(U,T,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}(S_{ij}^{m} + t^{n})d^{2}(X_j, v_i) \tag{2.18}$$

For pattern analysis, the combination ANFIS and SA were used to get the optimal number of data clusters with the adaptive layers of ANFIS. The authors were able to carry out the complete process of WUM and evaluated the results with existing approaches and accuracy of the results was measured through parameters such as Prediction accuracy; convergence behavior; and Execution time. This is a good effort overall; WFPCM is normally used for large data sets that cover the huge web log data. It seems that the authors adopted the renowned techniques for experimentation with any novelty. Sessionization is a challenging problem in the present scenario of the web-based era, this aspect was ignored in this research. Ant Colony Optimization (ACO), a heuristic technique of swarm intelligence family, has been applied to predict the user behavior by combining the user provided heuristics [42]. The authors criticized the traditional approaches to depicting the user behavior in WUM and these techniques are unable to converge the time in different sessions. The ACO can be incorporated to find out the new ways to persuade the user behavior. The authors applied the longest common subsequence (LCS) with ”r” real sessions and ”c” artificial sessions given below in Eq 2.19.

$$sim(r,c) = \frac{LCS(r,c)}{max(||r||,||c||)} \tag{2.19}$$

The website was presented by a directed graph through web pages as nodes and transition probability of randomly linked edges were calculated as explained in following Eq 2.20.

$$p_{ij} = \frac{p * g_{ij}}{c_j} + \frac{1-p}{n} \tag{2.20}$$

ACO is one of the practical approaches, inherits parallelism to adapt the dynamic changes in sessions. The authors have done the marvelous job by incorporating ACO for web sessionization. The authors failed to deliver the complete working model based on ACO. Web usage patterns were not discovered nor was even analysis performed. Similarly, no comparison was done with the existing approached. The major limitation of ACO is a change in probability in each of the iteration; the authors did not mention how to address the change in probability as it affects the final results.

According to Ying‘ [112], the web is expanding in all its aspects and analyzing user behavior is becoming a central part of the current web mining research. Various techniques are being applied to study the user behavior and author has applied Web user interest-based Fuzzy Clustering Model (WFCM). The user and session identification is key to find out the user interest goal of web surfing. The session identification process is mainly relying on visitor IP and comparison with locally stored web data. To model the similar users, the direct Hamming distance method was used given below in Eq 2.21.

$$r_{ik} = 1 - c \sum_{k=1}^{m} |x_{ik} - x_{kk}| \tag{2.21}$$

Where $r_{ik} \in [0,1]$ For session similarity, the threshold , was adjusted for clustering as the different values of threshold defined the different sets of clusters and threshold reflects the strong correlation between user behavior and web pages visited [113]. Fuzzy logic is well renowned technique and various applications in different domains, particularly, to address the Sessionization issue. The authors used the Hamming distance to produce the fuzzy similarity matrix. The Hamming distance has its own limitations with respect to the nature of web log data. Hamming distance is insensitive to the page visited and works with only the number of visited pages. This approach might not work in this web usage domain to understand the true user behavior.

To improve the competency of web mining algorithms and to achieve the accuracy in WUM process and its results, evolutionary techniques are playing an important role [28]. Furthermore, Hierarchical Particle Swarm Optimization (HPSO) based clustering helps reduce noise in big data and delivers efficient working models. The authors applied the HPSO in web usage mining domains such as recommended system which is the backbone of web based applications. In the implementation of PSO in WUM process, initialization of particles (sessions) is an important step and authors applied the following Eq 2.22 to initialize the particles.

$$loc(X(i)) = i * (\frac{N}{k} - 1) \tag{2.22}$$

After initializing the particles, the velocity of each particle is calculated and pBest and gBest of the swarm are calculated. For the session similarity, the authors applied the hybrid of Euclidean and Hamming Distance. The technique presented by the authors is an excellent contribution in WUM research. It will be more appropriate if the authors could have assign weights to the session similarity. The Euclidean family measure is proven inappropriate for the session similarity. Furthermore, the authors utilized of stronger particles in the next step. There are two observations that weak the research; particles may produce the interesting patterns which are being dropped. Secondly, Genetic Algorithm (GA) could be the better option to apply the fitness function for particle selection.

Awad and Khalil [114] categorized the prediction of next page problem as classification problem of web sessionization. The authors applied the modified Markov Model for the next page prediction. The Markov model incorporates the recorded user clickstreams data for the prediction in off line mode and predicts the next page to users online in minimum time.

Table 2.3: Summary of the Web Session Techniques

Technique	Authors	Classifier	Domain	Advantages	Limitations
Clustering	Shivaprasad et al. [115]	Neuro-Fuzzy Clustering	User behaviour, Administrator, Personalization	Hybrid Model Neuro-Fuzzy, Session construction IP, User-agent, Referrer Field	No comparison with any renowned technique, The Accuracy and precision of clustered not considered, The Accuracy and precision of clustered not considered.
	Forsati et al. [101]	Binary Partition-K-means Clustering	User behaviour, Recommended System	Hybrid Model Binary and K-means, The issues of Accuracy, High Coverage, and Scalability highlighted	Use of binary clustering and K-Means, Hamming distance as the similarity metric, No Comparison with any other technique.
	Vellingiri et al. [33]	Weighted Fuzzy-Possibilistic C-Means	Navigation of user interest	Pattern analysis through Adaptive Neuro-Fuzzy inference System with Subtractive Algorithm (ANFIS-SA)	How the membership function was improved, Results were not compared with any other existing technique.
Classification	Mishra et al. [109]	Rough set Upper Approximation	User behaviour, Recommended System	Tacke the overlapping behavior, Applied the Rough Set Upper Approximation Technique	Use of S3M as similarity, Modification in data set, No Comparison with any soft cluster technique.
	Kotiyal et al. [110]	Classification Naive Bayesian (NB)	User information, System Administrator	Small Size Training data set, accuracy and performance of the classifier through, use of metric Precision, Recall, F-Measure	The authors did not mention the parameters used in NB, No scoring function was discussed which play a key role in NB to classify the data, How the prior, likelihood and posterior probabilities were calculated
Sequential Pattern Mining	Patil and Khandagale [116]	General Sequential Pattern Mining (GSP)	Navigation Usability	Sequential Pattern Mining, Targeted the usability and accuracy of sessions	Database Scanning like Apriori approach, Use of threshold mechanism for pattern discovery
Association Rules Mining	Malarvizhi and Sathiyabhama [117]	T+Weight Tree Algorithm	Frequent Page Mining	Used the dwelling time of page visited	Storage of database in memory and website scalability issue, Use of threshold and Confidence, Use of threshold and Confidence

In this sessionization technique, authors applied the classification, while in clustering techniques with were not tested as clustering algorithms are commonly used for predictions in various web applications. The scoring function for the Markov model was used support and confidence while these threshold base techniques are generally unable to discover all the interesting patterns from web sessionization. Furthermore, authors applied the N-grams of Markov model for accuracy tests while only 3-Gram and 4-Gram were sufficient. The parameters used in training set were not explained and in the case of dynamic websites, the predicted web pages may not be accurate and it will compromise the accuracy claim of the proposed methodology.

2.10 Review on Hierarchical Sessionization

The exponential growth of web-based applications is posing the challenges to most web usage mining session clustering techniques. Moreover, to analyze the user behavior for the improvement of decision support and recommended systems are the open challenges to secure the interests of web stakeholders. The performance, security, and reliability of web-based applications enhance the user confidentiality. The research community has tried the various clustering techniques to address the sessionization issues such as density based clustering; model-based clustering; partition based clustering; fuzzy based models; grid-based clustering; and agglomerative based clustering techniques. In the previous section, we have already reviewed the literature about the various Sessionization techniques in practice nowadays with pros and cons. In this section, we are interested in reviewing the literature on hierarchical sessionization. The review is summarized in Table 2.4.

Hierarchical sessionization is an extension of WUM techniques to enhance the visualization of weblog sessionization in an iterative manner [118]. The Web is compiling the huge amount of unstructured user transaction data [119] and hierarchical clustering technique is an important tool for the analysis of weblog for focused and visualized identification of unbiased previously unknown groups [84]. The literature review on hierarchical Sessionization is summarized in Table 2.4.

In 2015, Kundra et al. [67], have investigated that the accuracy and stability of the whole WebKDD process can be improved through evolutionary approaches such as Efficient Hierarchical Particle Swarm Optimization (EHPSO) which can further reduce the complexity of Markov Model for online navigation prediction. The proposed EHPSO is capable of catering the issue dense sessions and accuracy. In the proposed algorithm, the authors used two similarity measures. The Euclidean distance metric to cover the numerical portion of the log file and Boolean Metric to cover the non-numeric attributes of web log data. The working steps and functionality of the EHPSO were not discussed and even how it will help Markov Model for online prediction. The authors were failed to give the complete working methodology for hierarchical web session clustering.

According to [120], web session clustering is an important step in the web usage mining to identify the visitors choices during the web page traversing. The authors used the Fast Optimal Global Sequence Alignment Algorithm (FOGSAA) for web session hierarchical (single link) clustering. For the alignment of web sequences, FOGSAA technique overcomes the time complexity issue of existing sequence alignment algorithms. The similarity between pages was calculated on the basis of optimal sequence alignment defined in Eq 2.23.

$$lAl * (A, B) = argmax(SC(Al(A, B))) \tag{2.23}$$

On the basis of this similarity function, hierarchical clustering was performed on the criteria of single linkage. The authors used the FOGSAA only to achieve the time complexity while in web Sessionization, the quality of results is more important that time efficiency. The correct identification of sessions is the first fundamental step to implement the WUM process. The FOGSAA result with other traditional alignment techniques were not compared for the noise free and quality of sessions. For an in-depth study of clusters and make the whole web mining process more efficient, Hawwash and Nasraoui [5] proposed the Hierarchical Unsupervised Niche Clustering (HUNC) algorithm. One of the objectives of this research was to spot the changing behavior of users on a website by evolving user profiles. The density function applied in HUNC for the scalar measure of a cluster is given in following Eq 2.24 with win robust weight.

$$\sigma_i^2 = \frac{\sum_{j=1}^{N} w_{ij} d_{ij}^2}{\sum_{j=1}^{N} w_{ij}} \tag{2.24}$$

To compute the session similarity with existing profiles Cosine measure was used. The authors used the variation of the agglomerative algorithm to find out the evolving user profiles to analyze the user behavior. Overall, this research was a remarkable addition for the researchers, developers, and website owners. It would be much better to use the simple agglomerative algorithm instead of Niche, as the Niche is of a complex nature and others claimed to address the evolving nature of user profiles in a speedy way. According to Hussain et al. [72], WUM is an important data mining area for the research community and weblog preprocessing is an important step to guarantee the noise free and quality clusters at later stages of the WebKDD process. For the session similarity, Angular separation (AS) and Canberra distance (CD) were used. The results were comparatively better than Euclidean Distance measure. However, the quality of the hierarchical clusters remained in question while preprocessing step was supported with stepwise algorithms to obtain the noise free swarm particles for hierarchical Sessionization. There was no standard metric used to ensure the sanctity of hierarchical clusters. The similarity measures are not suitable for the weblog data due to the nature of data as weblog data is unstructured [119].

Alam et al. [28] proposed the recommender system based user click streams by applying hierarchical particle swarm optimization (HPSO) to cluster the web sessions and categorized it as a complex job due to noise and distortion in data. The authors explained the three major components of HPSO such as initialization of swarm particles, learning of swarm particles and the velocity of swarm particles in detailed along with local and social components. The authors also provided the pseudo code of proposed HPSO. The fitness function of swarm particles is calculated after specified number of iterations and only stronger particles move to the next generation while the weak particles are removed. To calculate the session similarity between two sessions, the authors proposed the new similarity measure (Eq 2.27) by merging Euclidean Distance (Eq 2.26) and Hamming Distance (Eq 2.25).

$$d(XS,YS) = (\sum_{i}^{X_n}(XS_{ai} - YS_{ai})^2)^{(\frac{1}{2})} \tag{2.25}$$

$$h(XS,YS) = \sum_{i}^{n}(XS_i - YS_i) \tag{2.26}$$

$$Dist(XS,YS) = d(XS,YS) + h(XS,YS) \tag{2.27}$$

The authors proposed research addresses the noise and quality of cluster issues in web usage mining by applying the HPSO. This is motivational research and confirmed our claim that the hierarchical Sessionization issue requires the proper attention to address it. However, the research work has few fatal limitations and these limitations must be addressed in its true spirit to make the research innovative and useful for the research community. There was no strong arguments or justification to remove the weak swarm particles to take part next generation. The weak swarm particles may be useful and can lead for discovering of interesting patterns. Furthermore, the authors took the 21 pages per session, while the average web pages in a session are 20 to 40. This trimming can be disastrous for whole research and can lead to poor results at the end and where the quality of clusters may be compromised. Moreover, the authors proposed new similarity measure by combining Euclidean Distance and Hamming Distance. Both the measures belong to the same family and it is proven the fact that Euclidean Distance measure is not suitable for click streams due to the nature of web usage data. The authors used the weblog attributes such as a number of pages in a session, time utilized in a session and data downloaded. The URL (pages) in a session and time factors are very important attributes in Sessionization. We require such a similarity measure that can find the similarity of web pages visited by two sessions rather than the number of pages visited.

Table 2.4: Summary of Hierarchical Sessionization Techniques

Author	**Technique**	**Domain**	**Advantages**	**Limitations**
Chakraborty and Bandyopadhyay [120]	Fast Optimal Global Sequence Alignment, Algorithm (FOGSAA)	The Web Session Clustering, User Behavior	FOGSAA technique overcomes the time complexity issue of existing sequence alignment algorithms of Needleman-Wunsch	The overlapping issue of clusters not discussed, How the noise and scalability issues will be addressed
Hawwash and Nasraoui [5]	Hierarchical Unsupervised Niche Clustering (HUNC)	Profiling	Proposed the framework to manage the scalability, and noise issues.	GA based HUNC, how to cater the new users and changes, Multiscan to cater the changes in profiling
Kundra et al. [67]	Efficient Hierarchical Particle Swarm Optimization (EHPSO)	Online Prediction	Applied the Markov Model for prediction	Applied the combination of two similarity measures, The accuracy and quality of clusters are not tackled.
Alam et al. [28]	HPSO	Recommended System	Address the local maxima issue, Detail algorithms of HPSO	Why remove the weak sessions, Use of inappropriate similarity measure
Kumar et al. [121]	Hierarchical Clustering	User Behavior, Prediction	Address the limitations flat clusters such as Number of clusters, Centroid, Sequence of access pages	Same score of similarity in more than two sessions, No suitable methodology for optimized hierarchical sessionization.

The prediction of accurate user behavior for web usage is a challenging clustering task to group the similar sessions in the current era [121]. K-means and other flat partitioning clustering techniques suffer the limitations such as a number of clusters, centriod and sequence of web page access. However, these limitations can be overcome through the hierarchical sessionization by incorporating the sequence of web pages accessed in different sessions. The Page Rank algorithms are biased to predict the next page for user and authors proposed modified Lavenshtein distance based hierarchical sessionization for session clustering and Markov Model for the next page prediction. The authors proposed the complete web sessionization architecture and model along with the supporting algorithms.

Hierarchical sessionization interrelates at most two sessions based on similarity criteria. If more than two sessions have same similarity scoring value, then authors were unable to select the most suitable pair of sessions, which may compromise the claimed accuracy. To overcome such type of inconsistencies in the results, the evolutionary approach may provide the optimized select of the best pair of sessions. Moreover, the proposed hierarchical sessionization technique was not compared to any other existing technique.

2.11 Critical Review: Web Sessionization

There are various web sessionization techniques available in the literature for WUM process. Mostly the research community focused the web session similarity measure as a core for session identification. Nasraoui et al. [22] used Cosine Measure and criticized the Euclidean and Jaccard Measures because of the suitability due to weblog data type and nature of both the measures. Euclidean distance computes the distance between sessions and Jaccard computes the common web page among the session based the number of pages in a session. Whereas, Alam et al. [23] applied the Euclidean Measure and clipped the use of Cosine measure for sessionization as cosine measures in the range of [0, 1]. For absolute matching, sessions are assigned 1 and for mismatching, Cosine assigns 0. However, Euclidean measure was unable to address the noise issue in the identification of similar sessions and Alam et al. [41] used various heuristic approaches to minimize the noise in the web log data and claimed the accuracy up to 95% and proposed the similarity measure by combining the Euclidean and Hamming distance measures to overcome the limitations of Cosine and Jaccard measures. Even though Cosine measure was not delivering accurate results but still used for session similarity and in 2016, [86] produced the series of research on web sessionization and categorized the similarity measure as a core for capturing the true user behavior [86, 96, 122]. However, authors monopolized the research and used the Cosine Measure in different flavors to produce the results with different clustering algorithms such as K-Medoids, Fuzzy C-Means and Fuzzy C Medoids. Pai et al. [2] also criticized the Euclidian Metric for assigning equal weight to the all the features and is incapable of identifying the accurate sessionization. Pai et al. [2] overcome the Euclidean limitation by transforming it to Mahanlobis metric. The actual problem remained as it is due the nature of web log data. As most of the researchers were converting the non-numeric data into numeric before computing session similarity and that was producing biased sessions without accuracy. To overcome this limitation, Luu et al. [97, 98] proposed the web session similarity technique by combining Needleman-Wunsch (NW) and Smith-Waterman (SW) approaches to overcome the limitations of Euclidean, Manhattan, Levenshtein, Hamming Distance and Longest Common Sequence (LCS). The authors applied the modified form of LCS by taking the size of longest sequence and composed the proposed measure. Duraiswamy and Mayil [104] proposed the same web sessionization technique by utilizing the LCS through dynamic programming without involving the complexity of Needleman-Wunsch (NW) and Smith-Waterman (SW). Duraiswamy and Mayil [104] categorized the user interactions with a website is a key problem and described the click stream data presents the unique patterns and clustering provides the unusual knowledge about the user behavior. In 2011, Azimpour-Kivi and Azmi [123] criticized the all the well-known measures and defined that measure plays a key role to identify the relationship among the web sessions. Azimpour-Kivi and Azmi [123] applied the sequence alignment for session similarity. Bianco et al. [124, 125] discussed the threshold mechanism for session clustering and claimed that the threshold value is significant to understand the user behavior statistically and it optimizes the performance of session identification techniques in terms of speed and precision. Bianco et al. [124] applied the Poisson and correlation measure to identify the session similarity. Bianco et al. [124] produces another research on web Sessionization and further criticized the threshold base techniques [81] and proposed the technique that works without requiring a prior threshold value. Huidrom and Bagoria [81] compared the threshold-based techniques and suggested threshold free session similarity measures are more effective for sessionization. Li [103] also criticized the traditional measures used for web session similarity and proposed URL visited and time-based similarity measure to reduce the shortcomings of prevalent measures. The quality and accuracy of session generated affects the WUM process and particularly the knowledge extraction step [19]. The authors reviewed the various session construction techniques such as time heuristics, navigation heuristics, and integer programming [52] and discussed their limitations. Bayir and Toroslu [19] proposed the Smart-SRA algorithm for session similarity. The drawback of Smart-SRA is how to cater the dynamic nature of the web and how the noise will be eliminated by linking pages. Pai et al. [2] proposed that selected features based sessionization for user behavior analysis and applied LMNN algorithm as a similarity metric to identify the pair-wise relationship among the sessions. Table 2.2 provides the detailed analysis of existing web session similarity measures.

Roman et al. [52] applied integer programming technique to identify the patterns through the Bipartite Cardinality Matching. The BCM technique is not beneficial in dynamic websites as it is unable to address the dynamic scenarios. Moreover, for huge websites, weblog, the production and managing the huge graph data is also overhead that will affect the accuracy and genuinely of sessionization. Dixit and Bhatia [99] discussed the two major challenges of web sessionization techniques, such as the quality outcome of the clustering algorithms [100] and similarity/dissimilarity [66] of the web sessions. The authors proposed the Modified Knockout Refinement Algorithm (MKRA) along with Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to overcome the local maxima issue of KMeans. Shivaprasad et al. [108] proposed the hybrid of a Neuro-Fuzzy clustering technique for session clustering and claimed that single clustering technique cannot produce the quality results. However, authors failed to convince empirically and practically that their technique is producing accurate and quality results. On the other hand, Ying‘ [112] applied the fuzzy without the support of any other algorithm to find out the system with high accuracy and performance wise scalable. Vellingiri et al. [33] used the Weighted Fuzzy Probabilistic C-Means to cluster the web data for the extraction of user interest. The authors also implemented the inference system for the analysis of pattern generated. However, Vellingiri et al. [33] were unable to define the fuzzy membership function, like the membership function defined by Ying‘ [112]. Classification is another popular web mining technique. Nasa and Suman [126] evaluated the different classification techniques for web data, however, the authors were unable to pinpoint any suitable classification technique for web sessionization. Kotiyal et al. [110] performed the weblog classification to get the user information for the system administrators. The authors were unable to manage the parameters and scoring function for Naive Bayesian. In 2013. Chand et al. [127] also tried the classification through CART and was unable to produce any promising results for both researchers and web administrators. The main advantage of applying fuzzy for web sessionization is to handle the overlapping behavior Udantha et al. [128] and proposed the combination of DBSCAN and Expectation Maximization (EM) to overcome the limitations of PSO and K-Means that the number of cluster count by experts in prior knowledge. The authors claimed the accuracy of results generated through DBSCAN+EM, while the number of database scans and iteration were the two main limitations of the proposed work. These limitations can only be resolved through the evolutionary approaches [129] and investigated that Artificial Bee Colony (ABC) produce the high similarity clustering groups and according to Forsati et al. [77] ABC can manage the limitations of partition clustering algorithms such as K-means and DBSCAN. Loyola et al. [42] predicted the web user behavior through Ant Colony Optimization and implemented the concept of soft computing through evolutionary approaches as compared to Fuzzy clustering approaches for soft computing. In their research [23, 119, 130], the researchers performed the web session clustering through PSO to overcome the limitations of K-Means and both the researchers applied the different similarity measures. The authors claimed the accuracy and point out that with the improvement of similarity measure, the PSO generated session clusters can be more accurate and scalable with high coverage. Dou and Lin [131] criticized the PSO for its poor search and efficiency and produced the hybrid IPSO with Genetic Algorithm (GA). The hybrid idea of PSO and GA was implemented in 2014 by [132] to generate the web session clusters with improved efficiency, quality, and accuracy. Table 2.3 provides an analytical review of web session techniques.

Hawwash and Nasraoui [5] generated the hierarchical sessionization by applying GA-based Hierarchical Unsupervised Niche Clustering (HUNC) Algorithm for user profiling. The proposed framework works efficiently by delivering accurate clusters with improved visualization. The GA suffers from the fitness function for sessions and in the case of updating the user profiles, the database scan can further affect the efficiency of the WebKDD process. On the other hand, Chakraborty and Bandyopadhyay [120] applied the FOGSAA technique to address the scalability and overlapping user behavior. Sequence alignment was used overcome the matching hurdle of Needleman-Wunsch. Alam et al. [28] are working since 2008 on various web usage mining approaches and particularly applying Swarm Intelligence. The authors produced the hierarchical sessionization and dropped the weak sessions. This strategy may lead to missing the important patterns and the accuracy of results produced is challengeable. Kundra et al. [67] also applied the PSO for hierarchical sessionization without dropping of weak sessions. The only major issue with the proposed idea for Kundra et al. [67] is the use of two similarity measure while the various similarity measures are available which can handle the both the numeric and categorical nature of the weblog data. Table 2.4 gives the detailed review of existing hierarchical sessionization techniques along with the limitations.

2.12 Findings of the Literature Review

The literature was reviewed in three directions to cover all aspects of web sessionization to establish the significance of web sessionization in the research area and web mining industry and we identified the area where new contributions could be made. The literature review provides the critical evaluation of the different methodologies used in web sessionization as to identify the appropriate methodology for the investigation of the research questions raised in Chapter-1. The researchers are striving in all the web usage mining areas and phases. The most attractive part of web usage mining is to explore the weblog data to find out the hidden knowledge to explore the user traverses motivations. The study of the user behavior is significant, however, a complex job to play with it [96]. In the area of web session similarity, it can be concluded that there is no single suitable measure available for the identification of similar sessions. The researchers are relying on the modified measures for accurate and noise free sessionization. However, industry is still looking for an appropriate web session similarity measure for the correctness and reliability of results. For pattern identification, almost all the web mining techniques have been applied directly and with some modifications. Clustering is the frequently applied technique in all its flavors and producing better results than other web sessionization techniques (Table 2.3, 2.4). Furthermore, hybrid techniques are also being applied to address the sessionization problem. The evolutionary sessionization techniques are emerging research area for web sessionization and delivering promising results. Consequently, the hybrid evolutionary sessionization technique may overcome the limitations of existing sessionization techniques. Followings are the key findings of the literature and these findings would be helpful to address the sessionization in the solution-oriented format.

- The technology revolution has also empowered us to capture the huge amount of web data [8] that is playing a pivotal role in web sessionization research for the useful extraction of previously unknown interesting patterns from this data. The WebKDD is the set of interconnected sequential processes, and web sessionization is important to establish the bridge the WebKDD processes. (Motivation)

- In a number of research studies, it is evident that false results at one stage could nullify the results at the subsequent stages [99]. Consequently, a great deal of research has focused on the sessionization steps [133]. (Motivation)

- The web sessionization must take account of the validity of generated sessions, which entirely depends on upon the correctness [134] and credibility of sessions. This problem is mostly hindering by the proxy server, cache, and firewalls in the client web browser. This area has been surprisingly neglected until recently, as the majority of the literature on web sessionization has focused the pattern discovery phase. (Problem Statement)

- The sessionization problem may fail to obtain noise-free [75] unbiased user profiles [135] with high coverage [122] and precision even though well-known measures such as Euclidean, Cosine, Levenshtein, and Jaccard are prevalent in literature for WUM process at the early learning stages [19, 22, 98, 103]. (Problem Statement)

- The evolving and dynamic nature of WWW leads to enormous challenges for mining web clickstreams for extraction of patterns [95]; user behavior analysis Kotiyal et al. [110]; targeted and focused visualization [89] of coherent sessions. (Research Objective)

- A number of web session similarity measures are found. The main cause of such a huge number is a lack of proper, accurate, and unbiased similarity measure for web data [86]. (Limitation and Research Challenge)

- Almost all the data mining techniques are being tried to discover the hidden patterns from weblog, however, a clustering technique is a most common strategy to cluster the sessions with similar behavior [32, 33]. (Literature Review Conclusion)

- The research investigated that traditional clustering techniques are unable to address their legacy limitations such as number clusters, the center of the cluster, initialization [34, 35]. Furthermore, evolutionary approaches are also facing issues of feature selection, local maxima, efficiency, quality [36], visualization and reliability. (Research Objective)

- While the above literature review provides valuable information regarding the web usage mining, caution needs to be exercised in applying these results to the practical web usage mining area. One should not assume the results obtained from weblog applying various web session techniques are generalized. Limitations of these researches on web sessionization, session performance should not be compared with the individual performance of clustering algorithms.

- Few frameworks have also been proposed to address any one application such as profiling, user behavior analysis, recommended the system, etc., but we found the lack of a complete framework for Sessionization problem.

- Most research involving the experimentally induced information methodology seeks to identify the unbiased, highly accurate, and precise sessions to influence the WebKDD process and therefore the assumption is made that web sessionization is a detrimental process. It may, therefore, be advantageous to also investigate the effects of preprocessing; session similarity; and evolutionary approach as a key framework methodology of high ecological validity. However, few studies have used this methodology, and those that have, have yielded mixed findings. Therefore, future investigation using the hybrid evolutionary methodology would be helpful to better understand the effects of web sessionization.

2.13 Summary

In this chapter, we presented the foundation and research methodology that fulfills the research objectives. We discussed the abstract level of web usage mining life cycle and its utility in the current scenario of World Wide Web. We briefly describe the web usage mining and its effectiveness and necessity. In literature review part of the chapter, a meta-analysis of web sessionization was presented to pinpoint the gaps and limitations in existing literature. Regardless of web sessionization models and techniques, a variety of web session similarity measures are available for session similarity and negating one another. Consequently, there is no single web session similarity measure is available to fill the gap of accuracy, quality, and noise free. Furthermore, when we reviewed the literature on the web sessionization techniques, almost all the web mining techniques have been applied for the pattern discovery.Moreover, when we reviewed the literature, the traditional and flat clustering techniques suffer the few by default defects that debar its application in web sessionization. Ultimately, researchers have shifted to evolutionary approaches to remove the hindrances of traditional clustering. Consequently, based on the review on web sessionization, a complete framework is dire need to cope with the expanding, dynamicity, and scalability issues, the hierarchical sessionization with hybridization of agglomerative and particle swarm optimization along with the accurate and noise free similarity measure.

Chapter 3

A Proposed Framework for Mining Trends

The mounting and dynamic web clickstreams analysis is helpful for user behavior analysis. However, the existing WUM techniques and frameworks are unable to perform the correct, valid and credible sessionization. This chapter proposes a framework that is capable of mining the trends for investigating the user behavior analysis from weblog data. The aim of the proposed framework is to deliver the in-depth weblog visualization through the hierarchical clustering based on an evolutionary approach to address the web sessionization problem.

3.1 An Overview of Web Mining Frameworks

The WUM is a process of converting raw weblogs into useful, actionable, and knowledgeable information in Web Knowledge Discovery in Databases (WebKDD) process. The web mining techniques are applied in a sequence as a coherent set of techniques in the form of the framework as a solution. The taxonomy of web usage mining has been discussed in details in Chapter 2. In most of the web usage research, it has been observed that researchers have partially tackled the issues of web usage mining by addressing the portion of web sessionization issues. We lack the complete methodologies to cover the all issues of web sessionization.

However, the significance of the sub-issues of web usage cannot be denied and the contributions made in this regard are helpful to the industry and academia. In following sections, we are presenting the overview of web usage mining frameworks.

In the early history of the web and its applications, Fayyad et al. [46] iterated the need for a framework to extract the knowledge to assist the user with data mining techniques. The various steps involved in knowledge discovery have been highlighted and explained in detail. According to Peng et al. [136] the knowledge discoveries from the database is an interactive and iterative process. The authors explained the various steps involved in Data Mining Knowledge Discoveries (DMKD) such as data resources, open analysis, and axial analysis (Similarity relationship). For the validation of data mining framework, the data mining techniques are vital and necessary and Peng et al. [136] applied the clustering techniques. Gullo [48] also proposed the framework for pattern identification in data mining and elaborated the properties of frameworks for trend mining in KDD process such as trends must be non-trivial, valid, novel, potentially useful and understandable. If we are interested in mining the trends from weblog data, we have to apply the similar strategy. The interaction among the WebKDD processes is must to deliver the valid and correct solution to the web sessionization.

3.2 Paradigm of Mining Trends

For the last few years, WUM techniques are frequently applied to the user clickstreams data for mining the novel trends and the extraction of hidden knowledge to strengthen the machine learning process to facilitate the web consumers and producers. The research objective of these techniques is to formulate the viable mining frameworks and to model the user traversing behavior. In traditional data mining, WUM is a classical approach to explore and extract the useful patterns from user clickstreams. The crucial steps involved in WUM process are Preprocessing; Sessionization; Pattern Discovery; and Knowledge Visualization [5, 8, 35].

In web usage mining, the trend is an expected user behavior for future traversing based on the past clickstreams. Dueñas-Fernández et al. [137] defined the trend as a topic or event that is above the average over a certain period of time and the impact of the trend on the system is significant. In the area of machine learning, pattern recognition, and data mining, trends are the patterns followed by the user during the website traversing. Whereas the trends are the novel trends which are previously unknown and unseen patterns and marked from weblog clickstreams. There are various data mining techniques in literature that are being applied for the pattern identification from the web usage mining platform on the weblog data. In WebKDD, the pattern extraction process is composed of different steps and phases with different data mining techniques. The set of data mining techniques applied sequentially on the weblog for the mining trends to address the web sessionization problem are best composed in the form of framework. Frameworks address the core issues and deliver the complete methodology for the problem [47] and web mining researchers have proposed different frameworks that may address the overall challenge of web sessionization or frameworks for specific phase of WUM process to address the issues of that phase with complete solution [22, 138–140].

3.3 Proposed Framework for Mining Trends: F_MET

In web usage mining, a proposed F_MET is a conceptual and structured web sessionization solution with specific functionalities to explore the user clickstreams with intended mining objectives (Figure 3.1). The F_MET delivers the complete solution of web sessionization problems and it interrelates the WebKDD processes that work iteratively for the knowledge visualization that was previously unseen in user clickstreams by covering all aspects of the WebKDD process. F_MET takes the raw weblog as input data and delivers the hierarchical sessionization of the weblog as output. The objectives of F_MET have been defined for each step that provides the dynamic solution of the challenges and issues at each step. The major components of F_MET are Preprocessing; Web Session Similarity; and Hierarchical Sessionization. In following sections, we are explaining the components of F_MET.

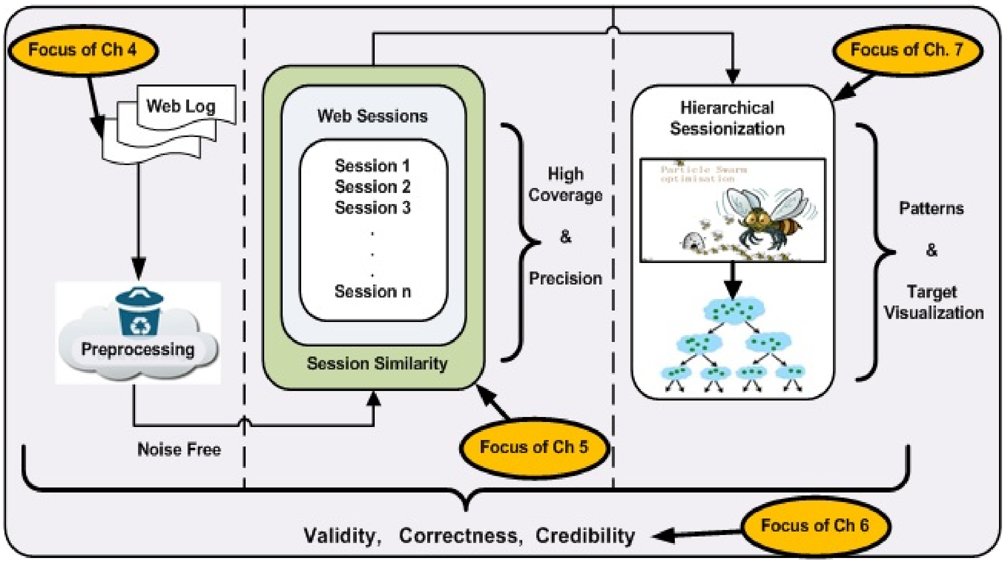


FIGURE 3.1: Framework for Mining Emerging Trends (F_MET)

3.3.1 Preprocessing of Weblog

Preprocessing is the first and one of the most important component of any data-mining framework. Without the proper preprocessing strategy, the ultimate objective of data mining cannot be achieved. Preprocessing prepares the data for next stages of data mining. Weblog data consists of around 80% of raw entries and is filtering out these raw entries are must for sustainable results of web mining process. In Chapter 3, we covered the major aspects of preprocessing in web mining and delivered the various preprocessing algorithms such as data cleansing, session identification, along with the results. The main objective of the preprocessing step is to focus on the bit-ignored phase of data mining and production of noise-free results for web mining process. There are several preprocessing tools available, however, these tool are unable to cater the dynamic nature of the website. Our proposed preprocessing scheme is simple in implementation, delivers accurate, and noise free filtered data for upcoming steps. Another added feature of proposed preprocessing scheme is the session construction. We adopted the Chitraa and Thanamani [141] technique to address the proxy and cache issues.

3.3.2 Web Session Similarity

Peng et al. [142] iterated that axial analysis for finding the similar concepts among the data items is an important step in data mining frameworks. In this regard, we reviewed the web mining literature related to the session similarity measures and discussed the importance of web session similarity measures in the WebKDD process along with the merits and demerits of the existing measures. The web session similarity measures play the vital role in WUM process for the accurate and highly precise results. Web session similarity is core component of F_MET as session (User) similarity is key to most of the web usage mining applications such as user behavior analysis; profiling; recommended system etc. The identification of similar sessions is helpful to cluster the sessions for knowledge discovery and trend mining. There was no proper session similarity metric available in literature and to fill the gap, we proposed the web session similarity measure ST_Index in chapter 5 of this dissertation. The ST_Index incorporates the common web pages along with uncommon web pages between the two sessions for a pair-wise relationship. The ST_Index is based on the argument of Chen et al. [17] that uncommon web pages play significant role to identify the real close relationship among the sessions. The common web pages are computed by assigning the unique UID to the URLs visited by the user after preprocessing weblog. The time spent by the users on the website is also an important factor and was utilized for session similarity in ST_Index. The details of ST_Index are available in chapter 5 along with results and comparison with the existing web session measures.

3.3.3 Hierarchical Web Sessionization

The proposed F_MET delivers end to end solution to the web sessionization issue through state of the art web mining techniques. The first two phases of F_MET have been discussed and implemented in Preprocessing Chapter 4 and Web Session Similarity Chapter 5. In this section, our focus is another important component of F_MET that is Hierarchical Web Sessionization. The input for the hierarchical sessionization is preprocessed weblog data in the form of user based sessions. Each session has been assigned a unique SessionID based on the IP_Host, User_Agent, User OS, and Referrer_Page. This technique of session construction helps to eliminate the Proxy Server, Firewall, and Cache issues for user identification. The other two attributes used for hierarchical sessionization are the set of URLs traversed by a user and time spent by a user in minutes for traversing the website. Another aim of hierarchical sessionization is to cluster the sessions with similar traversing behavior and identify the user groups that share the similar patterns.

Another feature of this proposed scheme is the assignment of distinct URLID to the URL_Resource. This attribute of the weblog is the actual web page traversed by a user in a session. By assigning the URLID in numeric format, the comparison among the session is quite easy and user-friendly. This technique also helps to increase the performance of overall WUM process to defuse the performance issue of Hierarchical Sessionization of the weblog as compared to the partition clustering techniques. The unique URLID is also helpful in finding the common web pages traversed in different sessions to compute the web proximity matrix. We also computed the uncommon web pages among the sessions with each other, as uncommon web pages among respective sessions,do affect the session similarity. This aspect has been ignored in the web session similarity measures research. By introducing this aspect the F_MET has filled the research gap for accurate session similarity.

After computing the SessionIndex and TimeIndex from preprocessed weblog through web proximity matrix, we used the proposed ST_Index web session similarity measure to compute the pair-wise session relationship. The results of web proximity matrix clearly show that single session can have a close relationship with more than one sessions. Such type of issues are beyond the scope of the hierarchical agglomerative clustering algorithm. Furthermore, to enhance the performance and efficiency of hierarchical clustering algorithm with accuracy, evolutionary approaches are more optimized and are frequently applied in a WebKDD process for the accuracy, correctness, and optimized results.

3.3.4 Hierarchical Agglomerative Algorithm

Hierarchical clustering is well-known data mining technique. There are two main types of hierarchical Clustering [143].

- Agglomerative (Bottom-Up)
- Divisive (Top-down)

In Agglomerative approach, we take the each data set as a single cluster. In next step, we apply any suitable similarity measure or distance metric, we pair the single clusters. The output is hierarchical clusters which are grouped based on maximum similar to each other. The agglomerative approach is also known as bottom-up technique. In divisive approach, we take the whole data set as a single cluster and then we recursively divide the single cluster into hierarchical clustering. This approach is also known as top down. Algorithm 3.1 gives a brief description of the agglomerative algorithm [143]. This an adapted algorithm and customized for the hierarchical sessionization.

3.4 Putting F_MET into Work

The proposed framework F_MET is a solution of web sessionization issues with accuracy and precision. F_MET identifies the patterns with high coverage. We tested the F_MET with the first 20 sessions of actual weblog data. The sessions are being constructed from filtered weblog. In each session, total page visited (TPV) and total spent time (TST) are the two key factors that include the total number of pages visited by the user in session and total time spent by a user in a session. On the basis of these two factors, ST_Index based web session similarity is computed for the hierarchical sessionization. The web proximity matrix is computed for the sessions as shown in Table 3.1.

Algorithm 3.1: Standard Hierarchical Agglomerative Clustering Algorithm

```
Data: Set of Objects(sessions) D of $n$ Sessions
Result: Dendrogram DS (Session Paired Clusters)
$D = (S_1, S_2, S_3, \ldots, S_n)$
while do
    for $i = 1 \rightarrow n$ do
        $C_i \longleftarrow S_i$ $C_i$ is $i^{th}$ cluster
    end
    $d \longleftarrow 0$     (d is threshold distance)
    $k \longleftarrow n$     (k is number clusters)
    $S \longleftarrow (C_1, C_2, C_3, \ldots, C_n)$    (S is set of clusters)
    $DS \longleftarrow (d, k, S_i)$
    $Dist \longleftarrow ComputeDistance(S)$   (Euclidean, Cosine, Jaccard, ST_Index, etc.)
    $d = \infty$
    for $i = 1 \rightarrow (k-1)$ do
        for $j = i + 1 \rightarrow k$ do
            if $Dist(i, j) < d$ then
                $d \longleftarrow Dist(i, j)$
                $u \longleftarrow i$
                $v \longleftarrow j$
            end
        end
    end
    $k \longleftarrow k - 1$
    $C(new) \longleftarrow (C_u \cup C_v)$              (Merging of clusters for dendrogram)
    $S \longleftarrow (S \cup C_{(new)} - C_u - C_v)$    (setting of new session from merged clusters)
    $DS \longleftarrow (DS \cup (d, k, S_i))$
    $until\ k = 1$
end
```

Table 3.1: Web Session Similarity ST_Index based Proximity Matrix

SessionID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	1	0.0468	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2		1	0	0	0.0250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3			1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0500
4				1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5					1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6						1	0.0284	0.0350	0.0017	0.0534	0.0025	0.0334	0.0418	0.0635	0.0167	0	0	0.0017	0	0
7							1	0.0075	0.1670	0.0518	0.2500	0.0084	0.0350	0.0184	0.0033	0	0	0.1670	0	0
8								1	0.0175	0.0175	0.0350	0.9750	0.0800	0.0525	0.1000	0	0	0.0175	0	0
9									1	0.0017	0.1250	0.0084	0.0067	0.0033	0.0184	0	0	0.2500	0	0
10										1	0.0025	0.0184	0.0234	0.0367	0.0084	0	0	0.0017	0	0
11											1	0.1250	0.0100	0.0050	0.0275	0	0	0.1250	0	0
12												1	0.0718	0.0501	0.0484	0	0	0.0084	0	0
13													1	0.0601	0.0401	0	0	0.0067	0	0
14														1	0.2510	0.4608	0	0.0330	0	0.0768
15															1	0	0	0.0184	0	0
16																1	0.0501	0	0	0.0518
17																	1	0	0	0
18																		1	0	0
19																			1	0
20																				1

The proximity matrix provides the session-wise relationship for hierarchical sessionization. Each session has ST_Index based proximity value and the maximum value of related sessions (pair) is picked. For next level of hierarchy,the tail of level-1 (pair) is matched with the next best proximity session (pair) and construct the next level of hierarchy with maximum proximity values and so on the hierarchies are defined. Based on the proximity values, the sessions have been paired for hierarchical clustering as:
$(S_1, S_2), (S_3, S_{20}), (S_6, S_{14}), (S_7, S_{11}), (S_8, S_{15}), (S_9, S_{18}), (S_{10}, S_{13}), (S_{14}, S_{16})(S_{16}, S_{17})$.
The different hierarchies of web sessions are given following Table 3.2.

Table 3.2: Web Session Hierarchies

Root Session	Hierarchies	Hierarchy Levels
S_1	$S_1 \rightarrow S_2 \rightarrow S_5$	3
S_3	$S_3 \rightarrow S_{20}$	2
S_6	$S_6 \rightarrow S_{14} \rightarrow S_{16} \rightarrow S_{17}$	4
S_7	$S_7 \rightarrow S_{11} \rightarrow S_{12}$	3
S_8	$S_8 \rightarrow S_{15}$	2
S_9	$S_9 \rightarrow S_{18}$	2
S_{10}	$S_{10} \rightarrow S_{13}$	2
S_{11}	$S_{11} \rightarrow S_{12}$	2

3.5 Comparison of F_MET with Existing Frameworks

In this section, we are presenting the comparison of proposed F_MET with the existing and available web usage mining frameworks. The WebKDD is an iterative and interactive process and its life cycle has been well defined in literature [46, 48]. All the data mining frameworks necessarily follow the data mining life cycle. The web mining is an application and extension of data mining techniques. Consequently, the web mining frameworks are constructed on the basis of data mining frameworks and conform to the data mining process.

3.5.1 Comparison Parameters

The parameters are defined to evaluate the web mining framework as under:

The scalable property of the frameworks is playing important role in the present age of huge and diversified web data. According to Fayyad et al. [46], the frameworks must flexible to handle the high dimensions of weblog data. Another property of web mining frameworks must be the completeness as frameworks essentially follow the complete life cycle of data mining process from data intake to knowledge delivery. The partial adoption of data mining process may address the particular portion rather than a complete problem. The axial analysis is the part of combining the similar objects and concepts in sessionization. It further links the concepts with each other in hierarchical form. Independent concept or trend identification is vital, however, insignificant as a part of the solution. Consequently, we have to link the trends (patterns) for in-depth and enhanced visualization through the frameworks.

There are different techniques and methodologies available in the literature to cope with the challenges of the web at all the stages of WUM process. All the three phases of WUM process are equally important and cannot be left orphan as WUM is an integrated and intra-dependent process. There are very few research in literature that delivers the complete solution to the challenges of the web in the form of frameworks. The plethora of partial and intermediary frameworks is available in the literature for web sessionization. In the following section, we are comparing the F_MET with the existing frameworks. Nasraoui et al. [22] proposed the framework for mining evolving trends in weblog streams on the basis of proposed similarity measure. This is a partial framework as the preprocessing phase of WUM has been totally ignored. The completeness of the proposed framework is not fulfilled, however, the scale issue of web usage mining is resolved. The axial analysis was performed by proposing similarity measure and for knowledge visualization, hierarchical clustering was performed. Whereas, the proposed F_MET is covering its all the phases and fulfilling the completeness property of frameworks.

3.5.2 F_MET Comparison

The authors realized the limitation of their framework and proposed a complete web mining framework for mining user profiles [11]. The proposed framework delivers the solution in five steps and applied the Cosine similarity. The proposed framework is delivering weak axial analysis as the authors are not yet satisfied with their own previous frameworks. Whereas, the proposed F_MET, is based on proposed strong web session similarity measure ST_Index. Another similar research was produced by the Ramkumar et al. [144] based on [11]. Both the frameworks are unable to deliver hierarchical clustering of web sessions as flat clustering techniques are failed to produce the analytical view of the weblog. The axial analysis is a strong key property that must be available in web mining frameworks.Knowledge visualization is essential part of WUM, otherwise, it is of no use. The WUM process is a costly and organizations can not afford it, however, academia may get benefits from such research activities.

Ansari et al. [36] proposed the fuzzy-neural network based framework to mine the overlapping user behavior from the weblog. The Euclidean measure was applied for fuzzy clustering as a scoring function. The added benefit of this proposed framework is the axial analysis through the neural network. Whereas the completeness and scalable properties of frameworks are totally ignored. The beauty of the F_MET is the coverage of all the steps and maintenance the integrity of the web usage mining process. The comparison of few available frameworks are shown in Table 3.3.

Table 3.3: Web Usage Mining Frameworks Comparisons

Authors	**Technique**	**Scalable**	**Complete ness**	**Axial Analysis**
Nasraoui et al. [11]	NICHE	Yes	No	Yes
Ramkumar et al. [144]	HAC	Yes	No	No
Ansari et al. [36]	FuzzyNeuro	No	No	Yes
Hussain and Asghar [12]	HAC	No	Yes	No
Alam et al. [28]	HPSO	Yes	No	No
F_MET	**PSO-HAC**	**Yes**	**Yes**	**Yes**

3.6 Summary

In this chapter, we proposed a framework for mining trends (F_MET) in weblog data to overcome the web sessionization issues at various stages of WUM process. We also discussed the effectiveness of F_MET to address the sessionization issues along with the merits and demerits of existing frameworks. We also explained the components of the proposed framework and execute the dry run to verify its flow, working, and objectivity as a problem solver. The comparison of F_MET with the existing frameworks at abstract level proved its effectiveness. We also presented the chi-square hierarchical sessionization and its published results as initially proposed a solution for web sessionization. However, few limitations were also observed to address the sessionization at full length. Few sessions have more than one best matching ST_Index score and selection of optimized web session pair is difficult with the simple agglomerative hierarchical clustering. As clustering is multi modal optimization problem for intra similarity between the pair of web sessions and only evolutionary approaches can be helpful in this regard. Particle Swarm Optimization and Genetic Algorithms are the two best optimization problem solutions whereas particle swarm is commonly practiced for web sessionization in the literature due to its simplicity in nature with efficient and scalable results. While genetic algorithm suffers the robustness and trained population with a leftover feature that may miss the few interesting patterns. In next, chapter we are interested in implementing the F_MET with particle swarm for hierarchical sessionization as an optimized solution of web sessionization.

Chapter 4

Weblog Preprocessing and Web Session Similarity

The primary objective of WUM process is the identification of interesting patterns and knowledge visualization from the weblog. The researchers have proposed various web mining techniques to achieve the successful WUM process. However, what so ever be the technique, weblog preprocessing is imperative for the valid and effective identification of knowledge discovery. In this chapter, we are presenting the state of the art preprocessing techniques for the noise-free web data for WUM process as rest of the WUM phases solemnly dependent on quality preprocessed web data.

The common limitation among the web session similarity measures is the similarity between the given two sessions (S_i, S_j). How the given two sessions are similar? Is the similarity relation developed between the sessions is accurate? The web sessionization must take account of validity of generated sessions, which entirely depends upon the correctness and credibility of sessions. The sessionization problem may fail to explicitly seek user profiles (behavior) with high coverage and precision even though well-known measures such as Euclidean [23, 28, 41], Cosine [22], Jaccard and Longest Common Sequence [42] are prevalent in literature for WUM process in the early learning stages of WebKDD.

4.1 Significance of Weblog Preprocessing

The preprocessing is a vital step in data mining for quality and noise free results [145]. Mostly, this step is ignored and deliver misleading results at later stages [37].The preprocessing techniques include Data Cleansing; Data Filtering; Path Completion; User Identification; Session Identification; and Session Clustering [14, 38]. There is almost a consensus that all the researchers in literature review have performed Data Cleansing to remove the irrelevant entries from weblog file [39]. Some of the researchers in literature review also carried out Data Filtering techniques. Data Filtering was applied in different flavors [8]. Missing path and missing entries are also completed through applying Path Completion techniques at preprocessing level. We have also seen that User Identification and Session Identification techniques are also applied in some research. Some researchers also have performed Session Clustering at preprocessing level of WUM [13, 72, 146].

By applying, one or more preprocessing techniques, cannot guarantee the reliability of overall results of WUM process. Weblog preprocessing is also necessary for web mining because weblogs are primarily configured to support the server administrators for Operating System errors monitoring and ratification [108, 147]. To prepare the weblog for mining purpose, preprocessing is a mandatory step. Preprocessing phase is a set of interconnected, coherent, and integrated techniques, applied in a sequence to have clear and well-defined results [14]. To cater the issue, we have proposed a complete methodology at the preprocessing level of WUM similar to the framework given in Figure 4.1. The objective of proposed methodology is to drag the most relevant information for the next steps of WUM process and at the same time, improve the quality and structure of information by applying clustering on weblog data based on defined web session similarity metric.

4.2 Weblog the Data Source

The website is launched on the web server and system administrator configures the weblog to capture the website user clickstreams. These weblogs are the primary source of data for web usage process. The weblog contain the user clickstreams in an unstructured format and cannot be made the part of web mining process directly. The different weblogs, their formats, and weblog attributes have been discussed by [33, 72]. A weblog is stored in plain text format (ASCII) and that is a part of the OS rather than a part of web application. Access Log; Agent Log; Error Log; and Referrer Log are commonly available log files on web servers [44]. The Table 2.1 is a generic snapshot of the weblog along with the attributes and their values recorded for users. There are 19 different attributes of weblog files in which user clickstreams are recorded. However, it is the discretion of server administrators to configure the weblog and its attributes to record the user entries. Only the mandatory weblog attributes are activated to capture the user activates on the web. The commonly available weblog attributes are IP, DateTime, Method, HTTP_Protocol, URL, User_Agent, OS, Status_Code, Data Downloaded, and Referrer_Page (Figure 4.1) [3]. Weblog records the user clickstreams in parallel during the user's website traversing without disturbing users. The communication protocol for users is HTTP, which is stateless. Due to the nature of the protocol, weblog records all the objects available on the single web page when the user clicks on that particular web page. The web page may contain the images; audio,

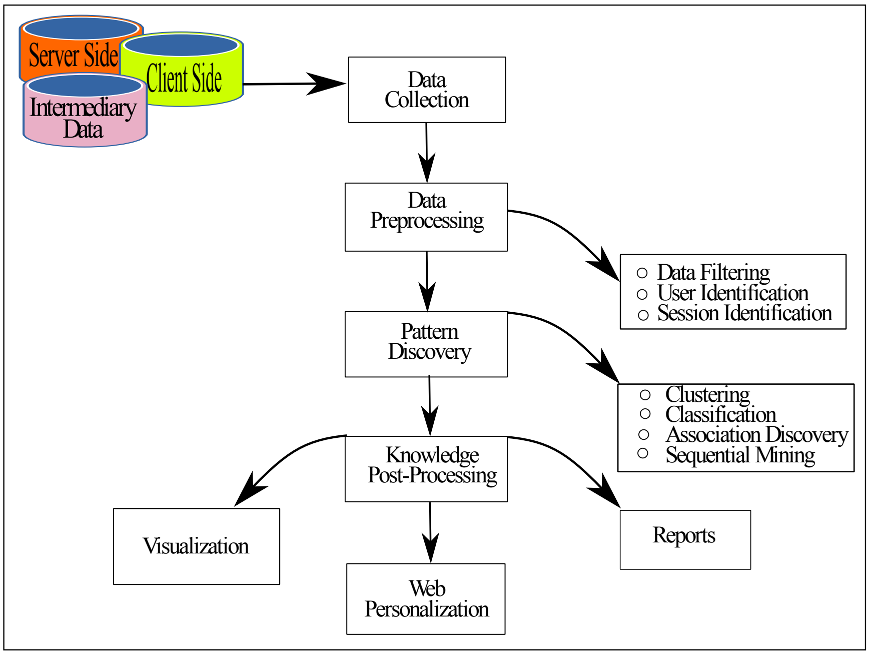


Figure 4.1: Web Usage Mining Process Pierrakos et al. [14]

video and Cascading Style Sheets (CSS). These entries are also recorded in weblog along with actual web URL (Web Page), administrator actions (insert, delete, and update), crawler and robot entries, while the user traversing the website. The capturing of all these entries in weblog against each IP (User) makes the weblog more cumbersome and complex and these entries are irrelevant for mining procedure. With the revolution in secondary storage devices technology (capacity and speed) has empowered us to capture the huge amount of web data in the form of weblog files and requires a complete preprocessing mechanism for analysis of this mega-repository [148].

The Internet users have increased exponentially since its inception in the mid-80s. According to the Internet statistics (2005-2016), the numbers of internet users have been increased from 1 billion to 3.5 billion [149]. Almost half of the world population is using Internet and number of users are increasing around the clock. This fact is also increasing the size of the weblog. Another factor is the growth of web page around 5 billion [21]. These factors are accumulating the web overall and making the weblog more and more raw data prone. The presence of such a huge amount of irrelevant entries in a weblog is a prerequisite for effective and noise-free weblog preprocessing to prepare it for upcoming phases of WUM Process. According to Alam et al. [41] weblogs contain around 80% irrelevant data and that cannot be used for data mining purpose.

4.3 Data Cleansing

Definition 4.1. Given a weblog of n records $L = \{t_1, t_2, t_3, \ldots, t_n\}$ where $n \geq 0$, let $\widetilde{L} = \{t_1, t_2, t_3, \ldots, t_m\}$ where $m \leq n \ \exists \ t(i).URL \neq images \wedge t(i).method = Get \ \wedge t(i).status = 200 \wedge t(i).agent \neq$\{crawler, spider, robot, administrative entries\} be the clean weblog such that $L \subset \widetilde{L}$.

The weblog is a primary raw source consists of noisy data. It requires a proper data cleansing technique to remove the noisy data from the weblog. The definition 4.1 highlights the weblog cleansing methodology.Weblogs are to help to debug of OS errors and to support Administrators, not for data mining purposes [150]. Users accesses and traverse clickstreams are recorded as they come in a sequence but the accesses of a single user cannot be in a sequence necessarily. The weblog records the users clickstreams on the basis of Time factor. However, the clickstreams are distinguished on the basis of IP of the client. The HTTP being stateless protocol, every transaction is recorded separately in weblog file. Thats why; components such as images (jpeg, bmp, gif, tiff, etc), (CSS), and scripting files are recorded as a separate record entity in weblog.

Besides these components, weblog also contains several other log record entries such as administrative actions of the update, insert, or deletes. The crawler, spider, and robot entries are also present in weblogs. As web server cannot recognize between the external and internal user [151]. Similarly, crawler actions are also recorded in the weblog. Therefore, the weblogs contain a number of irrelevant entries, which are needed to be removed. For further cleansing of the weblog, we keep only successful entries with a weblog status code equal to "200". It represents that web page has successfully delivered to users while the others status codes represent the error page and non-delivery of requested page to end user. The further details about the weblog attributes can be checked in [72]. Entries with all other status codes are considered as "Unsuccessful" log entries and are removed. The algorithm for data cleaning of weblog file is given in Algorithm 4.1.

4.4 Weblog Filtering

Definition 4.2. Weblog Filtering: Given a clean and preprocessed weblog $L = \{t_1, t_2, t_3, \ldots, t_n\}$ where $n \neq 0$. The log file consists of different attributes such as IP, Time, URL Visited, Data Downloaded, URL Referrer, Status Code, User Agent, Protocol Used etc. Let A(i) be any attribute of $\widetilde{L}\{t_i\}$. Count A(i), and discord A(i), if A(i) $= 0$

The details about the weblog can be further studied from our previous research [44] and all the three formats of weblog have different log attributes to capture the user

Algorithm 4.1: Weblog Filtering Algorithm

```
Data: Weblog L of $n$ records(transactions)
Result: Processed Weblog
$L = \{IP, DateTime, UserAgent, Method, URLReferrer, Bytes, Status, URL\}$
$L = \{t_1, t_2, t_3, \ldots, t_n\}$where $n \neq 0$
while do
    for $i$=0 to $n$ do
        Read L
        if  t(i).URL $\neq$ (.png, .bmp, .jpg, .avi,.css)
        and t(i).Method = GET
        and t(i).Status = 200
        and t(i).UserAgent $\neq$ (Spider, Robot, Crawler) then
            Record t(i) into Processed Database
        end
    end
    Update Processed Weblog
end
```

clickstreams. This difference is due to the nature and type of server OS. What so ever be the weblog format, the preprocessing is mandatory step to prepare the weblog for mining purpose. Otherwise, the results of WUM process may be mock and unreliable. Each weblog source has different attributes and one framework can not be applied to each format. The data in all the fields of weblog file is rarely available as it is the requirement of OS administrator to configure the weblog as per OS policy. We also cannot use all the fields (attributes of the weblog file) for web mining purpose. Therefore, we are required to remove the unwanted and empty fields. Algorithm 4.2 explains the filtering algorithm such as ”- -” represents the user ID and user password but these two attributes are not enabled for capturing the user data. Such type of attributes are overhead in processing are filtered out at an early stage of weblog preprocessing to make the weblog more accurate and noise free.

4.5 User Identification

Definition 4.3. User Identification: Given a clean weblog $\widetilde{\widetilde{L}} = \{t_1, t_2, t_3, , , t_n\}$ where n $\neq$ 0 if t(i).IP $\neq \emptyset \ \Lambda$ select all distinct t(i).IP into user identification

Algorithm 4.2: Weblog Attribute Removal Algorithm

```
Data: Preprocessed Weblog L of $n$ records (transactions)
Result: Attributes Removed Database
$L = \{IP, DateTime, UserAgent, Method, URLReferrer, Bytes, Status, URL\}$
$L = \{t_1, t_2, t_3, , , t_n\}$ where $n \neq 0$
while do
    for $i = 1 \rightarrow n$ do
        Read L
        Count t(i) in each Attribute
        if $t(i).Attribute = 0$ then
            Drop t(i).Attribute
        end
    end
    Update Weblog Database
end
```

database.

The identification of the distinct and unique users from weblog is an issue of its own kind due to the frequent use of proxy and firewall. The user (customer) to the website (web server) has relationship one to many. A single user can have a number of transactions in weblog file. Each web user is assigned a unique IP Address while connecting to the website through the Internet service provider. This IP Address is stored in weblog as user or client. There are different techniques in literature to find the distinct users of the website. Some researchers use the following combination to identify the users [50, 62].

- IP Address
- Browser used by client
- Operating System (OS)
- Referrer URL

The above-mentioned weblog attributes are significant for the user identification. In our proposed sessionization scheme, we studied the different literature on weblog preprocessing and concluded that the scheme introduced by Pierrakos et al. [14] by utilizing all the four attributes are delivering best results for the user identification [141]. Most of the researchers are applying different heuristics based on these attributes for the user identification. We adopted the scheme of [14, 141] and abstract level of user identification algorithm is defined in following Algorithm 4.3. Just picking the distinct IP from weblog is not enough to identify the true and real website users.

Algorithm 4.3: Weblog Distinct User Identification Algorithm

```
Data: Distinct User Weblog L of $n$ records (transactions)
Result: Session Identification
$L = \{IP, DateTime, UserAgent, Method, URLReferrer, Bytes, Status, URL\}$
$L = \{t_1, t_2, t_3, \ldots, t_n\}$ where $n \neq 0$
while do
    Read L
    for $i$ = 1 to $n$ do
        Select Distinct t(i).IP,t(i).User Agent,t(i).Referrer Page
        Copy r(i) into Weblog Session Database
    end
    Update SessionID Save Weblog Session Database
end
```

4.6 Session Identification

When a user connects to the website and performs, different activities through surfing the website and Server weblog file records these clickstreams. The weblog records the users traverses in various weblog attributes. The common weblog attributes in which user data is stored are IP Host; Date and Time; URLs, Referrer Page; Data Download; User Agent; Status Code; and Method. To the group, the activities of a single user are called a session that is time between login and logout. As long a user is connected to the website, it is the session of that user. In most of the research, 30 minutes timeout was taken as a default session timeout [8, 14, 39, 141, 152].If a user stays more than 30 minutes then it can be divided into episodes of that session. For the session, we took the following parameters (attributes) from the weblog:

- IP Address 110.36.13.223

- Date 06/Dec/2008:14:09:32 +0500
- Time 06/Dec/2008:14:09:32 +0500
- URL visited http://us.mcafee.com/apps/agent/submgr/appsync.asp
- Url_Referer Google.com
- User_Agent (Browser) Mozilla Compatible/2.0 (WinNT; I; NCC/2.0)

The maximum session length was taken 30 minutes and if a user traverses more than 30 minutes than first 30 minutes are taken as a session and so on. This fact produces the multiple sessions or episodes of a session in which time is exceeding 30 minutes. In few research, the maximum time for a session is 30 minutes. If a user spent more than 30 minutes time, it is converted into maximum 30 minutes. The multiple episode heuristics is more appropriate session construction and we adopted this heuristic for session construction in our experiment. Each session represents a user and a basic artifact used for the WUM process. The whole castle of WUM process is based on these sessions for further sessionization steps. During the session construction process, we also count the total pages visited (TPV) in a session and total time spent (TST) by a user in a session. These two entities can further be utilized for sessionization.

4.7 Weblog Preprocessing Results

The effectiveness of weblog preprocessing cannot be denied and overlooked as it works as the foundation of WUM Process. In most of the web mining literature, researchers presented the preprocessing methodologies along with the pattern identifications and pattern analysis techniques. In this section, we have presented the initial preprocessing results based on the above cited preprocessing algorithms at various stages. Table 4.1 is showing the preprocessing of weblog of two actual datasets. There are raw entries about 77% in each dataset and in the presence of such a huge irrelevant entries, the efficacy of web mining techniques is unproductive irrespective of the technique applied. The distinct IP_Host have been taken after the preprocessing and these are the actual distinct users of the website. The

Table 4.1: Preprocessed weblog dataset statistics

Datasets	**Dataset-1**	**Dataset-2**
Total Entries	6032	20408
Filtered Entries (Preprocessed)	1384	4673
% of Raw Data	77.05	77.1
Distinct IP	237	242
IP (Sessions)	201	160
IP+UserAgent+OS (Sessions)	209	165
IP+UserAgent+ReferrerPage (Sessions)	284	678

extraction of the distinct user (Sessions)is a complex phenomenon in weblog and has significance importance in web sessionization in the presence of proxy server, firewall, and cache issues. We managed these issues by incorporating the different heuristics such as IP based; IP, User Agent, and Operating System based and IP, User Agent, Operating System, and Referrer Page based. The most reliable heuristics is IP, User Agent, Operating System, and Referrer Page, that has sufficient literature support and empirical evidence to identify the actual user from the weblog.

4.8 Web Session Similarity

With the advancement of secondary storage devices technology, to capture the huge amount of web data, information and knowledge retrieval from the web has become an attractive and challenging research area. In preprocessing chapter, we concluded that the size and volume of the web are increasing around the clock. The information retrieval and knowledge extraction from huge weblog require the suitable mining procedure [148]. According to Ferrara et al. [91], the web session similarity computation from the weblog data is a useful resource in recommended system; data aggregation; and data analysis.

The web session similarity, computation among the web sessions is although a complex, however, a significant sessionization problem in the web usage mining (WUM) process at the early learning stage of WebKDD. Can we obtain the web sessions with high coverage and precision from preprocessed user clickstreams?

The valid and accurate session construction requires the proper and quality web session similarity metric for enhanced analysis of the web usage mining process to address the web sessionization problem. To achieve the correct and credible sessions, we are introducing a web session similarity (WSS) measure to compute the similarity among the user sessions that must address the sessionization problem at an early stage of WUM learning process. The proposed WSS will be able to statistically compute the significant relationship among the various web sessions for analyzing the user behavior, based on common pages visited by a user in a session along with the uncommon pages accordingly; and time consumed by a user during a session.

4.9 Significance of Web Session Similarity

To access the ongoing worldwide trends, the Internet has emerged as a wonderful tool. The number of clicks on a website and the user's click records determine the user the behavior. Oliner et al. [153] believes that the analysis of user click records is extremely important and is beneficial in many ways. The click records are very advantageous for website management, informational retrieval, fraud detection and web personalization [154]. The user's click history is the key to investigate the user trends , behavior on a specific website [33, 155], website management [156]; website administration; fraud detection; web personalization; information retrieval systems [154]; recommended systems [44] and web data analysis [91]. All the above-mentioned web applications are depending on the production of the valid and credible web session similarity measure that generates close relationship among the sessions to support the web usage mining process accuracy and correctness.

4.10 Limitations of Session Similarity Measures

There are various web session similarity techniques available in the literature for the WebKDD process. Mostly the research community focused the similarity measure as a core for session identification [40]. Nasraoui et al. [22] used Cosine

Measure and criticized the Euclidean and Jaccard Measures. Whereas, Alam et al. [23] applied the Euclidean Measure and clipped the use of Cosine measure for Sessionization. Alam et al. [41] developed a recommended system based on web usage data by applying particle swarm. Alam et al. [41] used various heuristic approaches to minimize the noise in the web log data and claimed the accuracy up to 95%. Alam et al. [41] proposed the similarity measure by combining the Euclidean and Hamming distance measures to overcome the limitations of cosine and Jaccard measures.

In 2016, Sisodia et al. produced the series [86, 96, 122] of research on web sessionization and categorized the similarity measure as a core for capturing the true user behavior. However, authors monopolized the research and used the Cosine Measure in different flavors to produce the results with different clustering algorithms such as K-Medoids, Fuzzy CMeans and Fuzzy C Medoids. Pai et al. [2] also criticized the Euclidean Metric for assigning equal weight to the all the features and is incapable of identifying the accurate sessionization. Pai et al. [2] overcome the limitation of Euclidean session measure such as local maxima by transforming it to Mahanlobis metric. The authors ignored the nature of weblog data in selection of web session measure and the actual problem of sessionization overlooked for the valid, accurate and correct session relation identification among the sessions.

A recent research on the sessionization investigated the effectiveness of clustering to group the similar web pages visited in an order in a session are to be placed in one cluster in an unsupervised manner [97, 98]. The authors combined the techniques of Needleman-Wunsch (NW) and Smith-Waterman (SW) to propose the new similarity measure to overcome the limitations of Euclidean, Manhattan, Levenshtein, Hamming Distance and Longest Common Sequence (LCS). The authors applied the modified form of LCS by taking the size of longest sequence and composed the proposed measure. Duraiswamy and Mayil [104] proposed the same web sessionization technique by utilizing the LCS through dynamic programming without involving the complexity of Needleman-Wunsch (NW) and Smith-Waterman (SW). Duraiswamy and Mayil [104] categorized the user interactions with a website is a key problem and described the click stream data presents the unique patterns and clustering provides the unusual knowledge about the user behavior. Duraiswamy and Mayil [104] proposed the sequence alignment for session clustering.

In 2011, Azimpour-Kivi and Azmi [123] criticized all the well-known measures and defined that measure plays a key role in session clustering. Azimpour-Kivi and Azmi [123] applied the sequence alignment for Sessionization. Bianco et al. [124] discussed the threshold mechanism for session clustering and claimed that the threshold value is significant to understand the user behavior statistically and it optimizes the performance of session identification techniques in terms of speed and precision. Bianco et al. [124, 125] applied the Poisson and correlation measure to identify the session. Bianco et al. [124] produces another research on web sessionization and further criticized the threshold base techniques Huidrom and Bagoria [81] and proposed the technique that works without requiring a prior threshold value. Huidrom and Bagoria [81] compared the threshold based techniques and suggested clustering based session generation approaches must be used rather than threshold base approaches. The authors are unable to give the clear mechanism for session clustering. The positive point is that Li [103] also criticized the traditional measures used for web session clustering and proposed URL visited and time-based similarity measure to reduce the shortcomings of prevalent measures. It can be concluded that web session similarity measures play a prominent role in the web usage mining process to address the web sessionization issue. However, at the same time, we lack the proper, efficient, and accurate session metric for WUM. The researchers have not only applied the traditional measures but also have tried the proposed measures to fill the gap. Consequently, the web usage mining still lacks the session metric like Euclidean, Cosine, and Jaccard on which consumer can rely for efficient and accurate results.

4.11 Web Session Similarity Measure (Initial Proposed Chi-Square)

In previous section and literature review chapter, we come up with research findings and gaps of existing web session similarity measures. The web session similarity is a primary and utmost artifact to articulate the sessionization problem and researchers have tried almost all the renowned measures to cope with the sessionization problem. However, researchers were failed to come up with a proper and accurate measure to address the problem and still, industry is waiting for the viable solution. Keeping in view the importance and complexity of sessionization issue, and failure of traditional similarity metrics to compute the close relationship among the sessions, can we have a session metric that could deliver the accurate and correct results for WUM process?

One of the reasons for the failure of traditional similarity measures for weblog data is the nature of weblog data. While the weblog data is of mixed type, numerical and categorical. The web pages traversed in a session by a user is of categorical while the total time spent by a user in a session is numerical. To compute the relationship among the sessions from mixed data,we adapted a Chi-square independence test for web session similarity [157, 158]. A Chi-square is a statistical approach to finding the closeness and relativeness between the objects. The reason to apply, the Chi-square to compute the similarity between the pair of web sessions was the nature of weblog data and to evaluate the existing measures for performance, precision and recall. The Chi-square Eq 4.1 is given below:

$$\chi^2 = \frac{(P_1 * T_2 - P_2 * T_1)^2 (P_1 + P_2 + T_1 + T_2)}{(P_1 + P_2) * (P_1 + T_1) * (P_2 + T_2) * (T_1 + T_2)} \tag{4.1}$$

Where P_1 is the total number of page visited by a user in time T_1 in session S_1 and P_2 is the total number of page visited by a user in time T_2 in session S_2. Similarly, the Chi-square values are calculated for every session with the other sessions and now two sessions can be joined by applying hypothesis.

$$H_0 = S_1 \text{ and } S_2 \text{ are independent}$$

$$H_1 = S_2 \text{ and } S_2 \text{ are dependent}$$

The Chi-square value is computed between two sessions $\{(S_1, S_2), (S_1, S_3), \ldots, \ldots, (S_1, S_n)\}$ in succession. The pair that holds the maximum Chi value exhibits a maximum tendency of the similar session. The marked pair is interrelated and dependent on each other and will not participate for the next iteration of finding similar sessions. The minimum value can also be taken into account for dependency test by taking them highly related. The selection of hypothesis has entirely based the nature of data and test values while the expert has also the right of hypothesis selection. The Chi-square based web sessionization algorithm is given below in Algorithm 4.4.

Algorithm 4.4: Chi-Square based Web Sessionization Algorithm (adapted)

Data: Distinct User Weblog L of n records (transactions)

Result: Chi Square Similar Sessions

```
$L = \{IP, DateTime, UserAgent, Method, URL\_Referrer, Bytes, Status, URL\}$
$L = \{t_1, t_2, t_3, \ldots, t_n\}$ where $n \neq 0$
while do
    for $i$ = level + 1 → Total Sessions $n$ do
        Read L
        Max Chi = 0
        for $j$ = $i$ + 1 → Total Sessions $n$ do
            $R_1 = i$
            $R_2 = j$
            Calculate Chi Square $(R_1, R_2)$
            if Chi Square > Max Chi then
                Max Chi = Chi Square
                $R_1 = i$
                $R_2 = j$
                Avg Value $(R_1, R_2)$
            end
        end
    end
    Update Web Session Database
end
```

4.12 Experimental Results of Chi Square based Sessionization

For the experimental work, we took the weblog data of two different university websites. The details of experimental work are as under: The Weblog 1 contains the total user clickstreams of 60302 of four days and Weblog 2 contains the total user traverses of 65536 in one day. After preprocessing phase, sessionization step was performed to create the user sessions. From weblog 1, we obtained the 1738 unique sessions and from Weblog 2, we obtained 1987 sessions. The results of Weblog 2 are being presented in the following section.

Table 4.2: Chi Square Values and Related Sessions

Session 1	**Session 2**	**Chi-Square Values**
1	17	23.440
2	5	18.110
3	19	9.500
4	6	12.810
7	10	19.460
8	20	16.690
9	15	21.470
11	13	10.140
12	14	13.370
16	18	17.000
21	27	13.860
22	23	15.280
24	29	16.160
25	28	11.260
26	30	11.250
31	32	22.490
33	34	9.910
36	37	10.000

On the basis of the proposed algorithm 4.4, we computed the web session similarity among the sessions. Table 4.2 is the snapshot of Chi-square values among the sessions. The proposed Chi-square algorithm was executed on the same dataset with same parameters just replacing the similarity metrics. The similarity measures used for the experiment are Euclidean Distance; Cosine; Jaccard; Angular;

and Canberra. The comparison of results of different similarity measures was performed on the basis of precision and recall metrics. The parametric values of precision and recall are computed. The True Positive (TN), True Negative (TN), False Positive (FP) and False Negative (FN) for each similarity measure are calculated and shown in Figure 4.2 and 4.3.The precision [0,100] and recall [0, 100] are marked along the y-axis and similarity measures are marked along the x-axis (Figure 4.2). The precision and recall measures analysis showed the remarkable results for all measures. The average precision and recall remained around 60. These results guided us that any one measure can be used for Sessionization and no need for the proposed session similarity measure. The results validated the claims of Ibrahimov et al. [157], Chen and Chen [158] that the Chi-square is showing steady performance at recall test over the Cosine and Jaccard measures. The slightly improvement in Chi- based results is due to the nature of Chi-square metric.

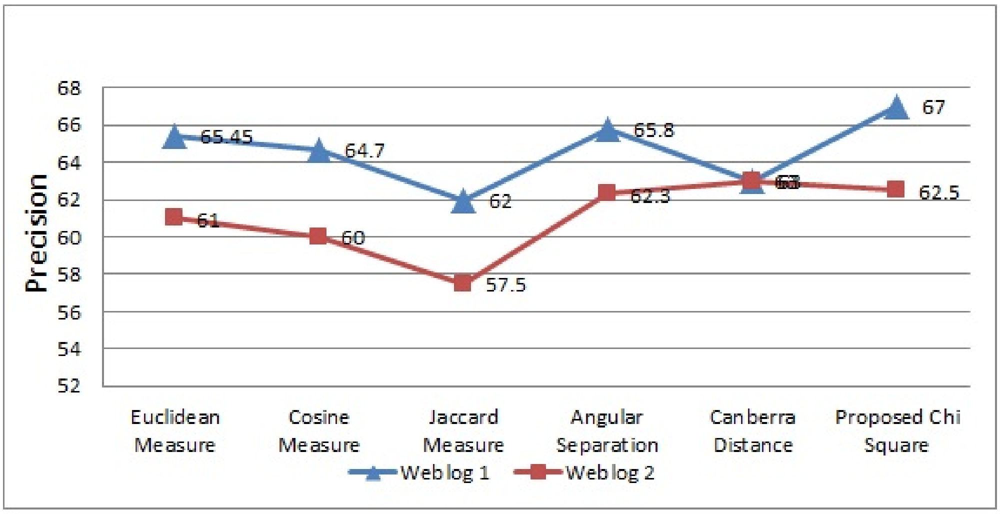


FIGURE 4.2: Precision for Web Session Similarity Measures

The Chi-square is initially proposed similarity measure to evaluate the other existing measures, even though Chi-Square measure is performing steadily at precision and recall level on both weblog 1 and weblogs 2 datasets but still it is not fulfilling the basic requirement of generation of accurate, valid and correct session relationship among sessions with high precision and coverage. Like other measures, Chi-square is also suffering the few by birth limitations such as all the sessions are

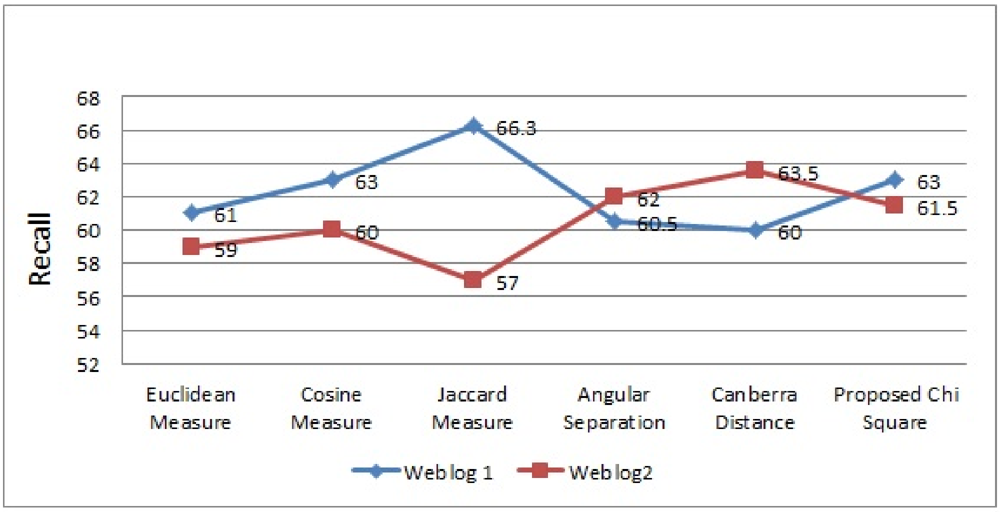


FIGURE 4.3: Recall for Web Session Similarity Measures

not explicitly independent while Chi-square requires independent sessions. The two given sessions, share the same number of pool (web pages) and not explicitly independent. Secondly, Chi-square requires a minimum threshold of web pages in a session such as at least 10 or more web page in each session. For fewer numbers of web pages in a session produces mock results and even dropped from session similarity computation. This is a big hurdle in the accuracy of session similarity and results cannot be produced with high precision and coverage. The Chi-square is unable to produce the accurate, valid, and correct relationship among the sessions. Consequently, we required a web session similarity measure to address the sessionization problem and produce the valid results to overcome the limitations of the existing session similarity metrics.

4.13 A proposed Web Session Similarity Measure: ST_Index

In the previous section, we presented the results of initially adapted Chi-square session similarity measure along with traditional measures. The Chi-square measure produced better results than the other measures. However, Chi-square was unable to produce the valid and correct session relationship among the sessions.

The Chi-square measure is depending on a number of web pages present in a session. Higher the number of web pages in a session produces better results and less number of web pages is producing invalid session relation among the sessions. Another limitation of Chi-square measure is computation complexity, which affects the efficiency of the measure as a whole. Chi-Square calculated the closeness and relatedness among the sessions and incorporated the hypothesis that higher the Chi-square value, higher the relationship between the paired sessions. The factor of relatedness was based on web pages traversed in session in given session time. However, like other distance measures for web sessionization, Chi-Square was also failed to convince for real relationship among the sessions as users are traversing the same website and a number of users have the common traversing patterns.

We have also concluded from literature review chapter, that similarity measure is a focal point of weblog Sessionization. As a Euclidean family of measures is frequently applied in web Sessionization and widely criticized by the research community due to mismatch with the data type and assigning equal weight to all the sessions irrespective of web pages traversed in different sessions. Cosine and Jaccard similarity measures have also failed to provide the accurate and correct session relationship among the sessions due to the approach, as both consider the related and unrelated sessions as a different entity while both the sessions may be linked syntactically rather than attribute-wise [40]. The majority of researchers proposed their own measures to overcome the limitations of the frequently used measures.

Due to the criticality of the session similarity issue in web sessionization, the appropriate session similarity measure is vital and keeping in view the limitations of existing measures, we introduce the new session similarity measure ST_Index based on Session Index (SI) and Time Index (TI). The synopsis of proposed similarity measure is to cater not only the shared web pages and shared time between the two sessions; in fact, it also assigns the weight to the unshared pages with respect to the sessions each other as users have the same pool of web pages to traverse them with different objectives [17]. In following steps, the session similarity is being computed between the two sessions with synthetic data of ten sessions along with a number of web pages visited and time consumed in each session by users. A structure of website and set of 10 sessions along with web pages traversed and time spent in each session. We adapted the idea of [159] to represent the website in a structural form and labeled the URLs of weblog in numeric format. This approach is useful in representing the user web session in graphical form and to find out the pair-wise relationship among the sessions.

In this proposed web session similarity measure, we are considering the two key features from user clickstreams records, URL visited by the user in a session, and time utilized by the user in the session. Furthermore, we are interested in assigning weight to uncommon URLs between the two sessions for similarity as an uncommon URL factor various among the different user session. The proposed web session similarity measure is defined below:

Definition 4.4. Clickstreams: Given a weblog of n records $L = \{t_1, t_2, \ldots, t_n\}$ where the i^{th}-transaction t_i is defined as user transactions tuples containing User IP; Client Identifier; User Name; Time Stamp; Request Method; URL Resource; HTTP Protocol; Status Code; Data Transferred; Referrer URL; User Agent. Let $S = \{S_1, S_2, S_3, \ldots\ , S_k\}$ be the collection of k sessions where each $S_j \subset L$ containing transactions of j^{th} session and CS_j be the collection of clickstreams in j^{th} session is defined as $CS_j = \{C_{j1}, C_{j2}, C_{j3}, \ldots\ , C_{jm}\}$ be the user clickstreams where C_{jl} be the l^{th} click of j^{th} session. In fact $CS_j = \{F_{j1}, F_{j2}, F_{j3}, \ldots\ , F_{jp}\}$ where F_{jp} is the user clickstreams features extracted from the transactions related to the j^{th} session.

The user clickstreams history is recorded in weblog and each user transaction has various attributes and predefined parameter such as User IP; Time Stamp; URLs; Method; Code; Data Transferred; etc.

Definition 4.5. Web User Session: is defined as $S_j = \{t_{j_1}, t_{j_2}, t_{j_3}, \ldots\ , t_{j_k}\}$ where each S_j be the j^{th} user session and user have browsed the n number of web pages $\{p_1, p_2, p_3, \ldots, p_n\}$ in time (minutes) $\{t_1, t_2, t_3, \ldots, t_n\}$.

Definition 4.6. Common Session Index (CSI): For the i^{th} session S_i and j^{th} session S_j is defined as Eq 4.2.

$$CSI(S_i, S_j) = \frac{||S_i \cap S_j||}{||\widetilde{S}_i \cup \widetilde{S}_j|| + ||S_i \cup S_j||} \tag{4.2}$$

Where $||\widetilde{S}_i \cup \widetilde{S}_j||$ be the total number of uncommon URLs (pages) traversed in the i^{th} and j^{th} sessions with respect to each other. By assigning weight to uncommon URL/pages helps to address the precision and high coverage between the sessions.The Common Session Index is a modified form of Jaccard Measure. We have added the uncommon web pages weight for the granularity of the proposed measure along with minimum time utilized by the sessions. If we remove the weight of uncommon web pages,it is Jaccard web session measure. For further improvement in proposed measure, we are adding the time factor of sessions. We can claim that our proposed measure is time and uncommon pages based and different from Jaccard on the basis of these two factors.

Definition 4.7. Common Time Index (CTI): For the i^{th} session S_i and j^{th} session S_j, we defined time index as Eq 4.3.

$$CTI(S_i, S_j) = \frac{min(S_i(time), S_j(time))}{max(S_i(time), S_j(time))} \tag{4.3}$$

Where $S_i(time)$ be the total time spent in i^{th} session for traversing the website and $S_j(time)$ be the total time utilized in j^{th} session for surfing the website. As we are interested to find out the similarity between the two sessions, both the sessions S_i and S_j have spent minimum common time while traversing common URLs. Although, both the sessions are traversing the URLs with different objectives and spending different amount of time. It is assumed that the common pages visited in both the sessions deliver same contents and addressing the same objectives.

Definition 4.8. Web Session Similarity: We defined the web session similarity between the i^{th} and j^{th} session as in Eq 4.4.

$$ST\_Index(S_i, S_j) = CSI(i, j) * CTI(i, j) \tag{4.4}$$

Lemma 4.9. ***Similar Sessions:*** *The two given sessions are similar, if both the sessions share some similar traversing pattern such as similar web page in given time and have a tendency to exhibit the similar user objective and traversing behavior. The similarity function $f : S \times S \to C$ for the given two sessions based on some predefined threshold is defined as Eq 4.5.*

$$\phi(S_i, S_j) = \sum_{|t|>0} S_i(C_i) * S_j(C_j) \geq \widetilde{\phi} \tag{4.5}$$

Proof: *The threshold $\widetilde{\phi}$ is dynamic and varies among the various similar sessions. The proximity table is used for the computation of similarity score of each session with other sessions. After computing proximity values of sessions in a table, we adopt only the highest value among the sessions. For example, we have to compute the similarity among the three sessions ($S_i, S_j$and S_k) and we are interested to find out the most similar pair. We compute the proximity values among the sessions (S_i, S_j) and (S_j, S_k). If (S_j, S_k) has a higher proximity value than (S_i, S_j). Then it can be concluded that (S_j, S_k) pair exhibit more similar traversing pattern and pair has stronger similar relation than the rest of web sessions.*

Lemma 4.10. ***Change in Behaviour (CB)*** *For the given two S_i and S_j, the behavioral changes in website traversing are depicted as follows Eq 4.6*

$$CB = || \prod_{i}^{n} P_i | P_i \in (S_i \cup S_j) \wedge P_i \notin (S_i \cap S_j) || \tag{4.6}$$

Where P_i be collection of web pages traversed in S_i and S_j.

Proof: *The change in traversing behavior of among different sessions is significant to study the user browsing pattern. The web pages are traversed by the either session but not by the both. This is twist in behavior as the shared web pages indicate the similar behavior while the uncommon portion indicates the change in behavior of diversion in behavior of two sessions.*

Now in the following section, we are validating the abstract model of ST_Index against the prototype data. We took the 20 web sessions along with the total web pages visited and total time spent in each session. We suppose that the data is preprocessed and sessions have been constructed as per the mechanism defined in preprocessing chapter. The structure of the website is also shown to validate the results either the similar sessions are following the website structure or not. The web pages traversed in each session are given along with the time utilized by the user in each session. The structure and data table of the prototype is shown in Figure 4.4 and Table 4.3.

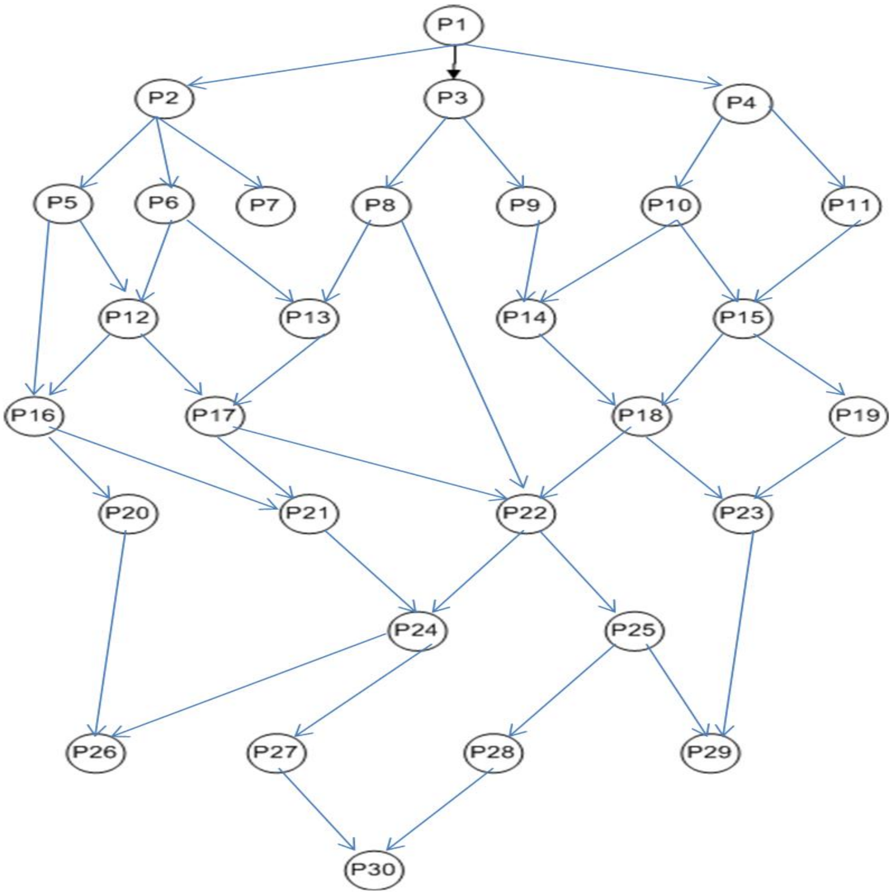


FIGURE 4.4: The General Website Structure with P1 as index page and showing the links of web pages with each other.

4.13.1 Common Pages Calculation

The major ingredient for web session similarity is common pages visited between the two given sessions. Number of researchers have used this notion rather than the counting of pages traversed in a session. Especially the research in which, the researchers opted for longest common subsequence for the calculation of similarity

Table 4.3: Web pages visited by users in different web sessions along with Total Page Visited (TPV) and Total Spent Time (TST) by users in sessions

SID	**Web Pages**	**TPV**	**TST**
S1	P1,P2,P6,P13,P17,P21,P24,P27	8	15
S2	P1,P3,P8,P13,P17,P21,P24,P27,P30,P38,P47	11	23
S3	P1,P2,P6,P13,P17,P21,P24,P26,P34	9	11
S4	P1,P2,P6,P13,P19,P23,P29	7	20
S5	P1,P3,P9,P14,P18,P22,P25,P28,P30,P41	10	19
S6	P1,P2,P6,P13,P17,P21	6	8
S7	P1,P4,P11,P15,P18,P22,P24,P27,P30,P41,P45,P54	12	9
S8	P1,P2,P7,P8,P22,P25,P28,P30,P41,P51,P55,P62,P64,P66,P69	15	12
S9	P1,P3,P9,P14,P18,P23,P29,P37,P41	9	21
S10	P1,P4,P10,P15,P18,P22,P24,P26,P31,P45,P53,P55,P62	13	13

among the sessions applied the common web pages. In Table 4.4, the common web pages are computed among the sessions.

Table 4.4: Web Session Common Pages

Common TPV	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
S1	**8**	6	7	4	1	6	3	2	1	2
S2		**11**	5	2	3	4	4	3	2	2
S3			**9**	4	1	6	2	2	1	3
S4				**7**	1	4	1	2	3	1
S5					**10**	1	5	6	6	3
S6						**6**	1	2	1	1
S7							**12**	4	3	7
S8								**15**	2	4
S9									**9**	2
S10										**13**

The second important segment is the computation of uncommon web pages among the sessions with relevant to each other. In following Table 4.5, the uncommon web pages are computed.

4.13.2 Common Session Index

The session index of each session with every other session is calculated on the basis of common and uncommon web pages traversed in both the sessions S_1 and S_2 as explained in Eq 4.2. The assigning weights to the uncommon page with

Table 4.5: Web Session Uncommon Pages

Uncommon TPV	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
S1	**8**	2	1	4	7	2	5	6	7	6
S2	5	**11**	6	9	8	7	7	8	9	9
S3	2	4	**9**	5	8	3	7	7	8	6
S4	3	5	3	**7**	6	3	6	5	4	6
S5	9	7	9	9	**10**	9	5	4	4	7
S6	0	2	0	2	5	**6**	5	4	5	5
S7	9	8	10	11	7	11	**12**	8	9	5
S8	13	12	13	13	9	13	11	**15**	13	11
S9	8	7	8	6	3	8	6	7	**9**	7
S10	11	11	10	12	10	12	6	9	11	**13**

each other is playing significant role to identify the true and granular relationship among the web sessions. In Table 4.6, we computed the proximity values for Session Index from the common URLs and uncommon URLs visited by the users among sessions. Otherwise all the users are visiting the same pool of web pages and similarity computation may be biased and most of users(sessions) have same similarity.

Table 4.6: Web Proximity Matrix (URL Visited)

CSI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
S1	1	0.3000	0.5385	0.2222	0.0303	0.6000	0.0968	0.0500	0.0323	0.0556
S2		1	0.2000	0.0667	0.0909	0.1818	0.1176	0.0698	0.0588	0.0476
S3			1	0.2000	0.0286	0.5000	0.0556	0.0476	0.0303	0.0857
S4				1	0.0323	0.2857	0.0286	0.0526	0.1304	0.0270
S5					1	0.0345	0.1724	0.1875	0.3000	0.0811
S6						1	0.0303	0.0556	0.0370	0.0286
S7							1	0.0952	0.0909	0.2414
S8								1	0.0476	0.0909
S9									1	0.0526
S10										1

4.13.3 Common Time Index

Time is another important attribute for web sessionization and plays vital role in web session similarity measures. We picked the minimum time(minutes) between the two sessions as at least minimum time is shared between the sessions. The assigning the weight to minimum time is logically sharing the minimum time for identifying the session similarity. We computed the Common TimeIndex(CTI) for the two sessions, by taking the minimum time from the both the sessions along with the maximum time utilized in both the sessions. The details of CTI are given in Eq 4.3 and Table 4.7.

Table 4.7: Common Session Time Index

CTI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
S1	1	0.6522	0.7330	0.7500	0.7895	0.5333	0.6000	0.8000	0.7143	0.8667
S2		1	0.4783	0.8696	0.8261	0.3478	0.3913	0.5217	0.9130	0.5652
S3			1	0.5500	0.5789	0.7273	0.8182	0.9167	0.5238	0.8462
S4				1	0.9500	0.4000	0.4500	0.6000	0.9524	0.6500
S5					1	0.4211	0.4737	0.6316	0.9048	0.6842
S6						1	0.8889	0.6667	0.3810	0.6154
S7							1	0.7500	0.4286	0.6923
S8								1	0.5714	0.9231
S9									1	0.6190
S10										1

4.13.4 Web Session Similarity (ST_Index)

The final session similarity among the sessions is computed on the basis of Session Index and Time Index in Eq 4.4 by just taking the product of both the indices. The web session similarity score for matching the sessions is computed in normalized form and results shown in Table 4.8. The algorithm for the computation of similarity among the web sessions based on ST_Index is shown in Algorithm 4.5.

Table 4.8: Web Session Similarity Computation (ST_Index)

STI	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
S1	1	0.1957	0.3947	0.1667	0.0239	0.3200	0.0581	0.0400	0.0231	0.0482
S2		1	0.0957	0.0580	0.0751	0.0632	0.0460	0.0364	0.0537	0.0269
S3			1	0.1100	0.0166	0.3637	0.0455	0.0436	0.0159	0.0725
S4				1	0.0307	0.1143	0.0129	0.0316	0.1242	0.0176
S5					1	0.0145	0.0817	0.1184	0.2714	0.0555
S6						1	0.0269	0.0371	0.0141	0.0176
S7							1	0.0714	0.0390	0.1671
S8								1	0.0272	0.0839
S9									1	0.0326
S10										1

4.14 Experimental Results of ST_Index

The proposed web session similarity measure ST_Index is being evaluated against the available three datasets along with the other measures such as Chi-square, Euclidean; Jaccard; Cosine; Angular; and Canberra Distance measures. One dataset was shared by [31] and marked as Dataset-1 and other two datasets are of local universities(marked as Dataset-2 and Dataset-3). The details of datasets are in Table 4.9. The aim of this experiment is to evaluate the ST_Index for the accuracy; quality; scalability; and noise free sessionization and to deliver the unbiased and focused results for upcoming steps of WUM process. The accuracy of sessions generated is being evaluated against the actual web pages traversed in each session and for quality assessment, we compared the ST_Index results with well-established measures. The scalability is being evaluated by testing on different sizes of datasets and with a different number of sessions constructed with time and space attributes. To obtain the noise and distortion free results of web sessionization, we performed the preprocessing step by applying various cleansing technique. The weblogs are the part of the web server and are in textual format.

Table 4.9: Datasets along with initial preprocessing, sessions and ST_Index (Pair wise similarity) results

Datasets	**Dataset-1**	**Dataset-2**	**Dataset-3**
Total Entries	6032	20408	668619
Filtered Entries (Preprocessed)	1384	4673	88152
Distinct IP	237	242	12560
IP (Sessions)	201	160	9896
IP+User Agent+OS (Sessions)	209	165	11449
IP+User Agent+Referrer Page	284	678	14481
ST_Index (Pair wise Similarity)	254(127)	624(312)	12602(6301)
Total Pages(After Processing)	134	474	4085

The weblog was imported into Oracle Database. Preprocessing was performed to remove the irrelevant entries and removed the 77% to 87% irrelevant entries. The variation in data filtering is due to the structure of websites. The filtered data was then converted into web user sessions. There are different heuristics available in the literature for web sessionization and we applied the scheme of [141] for sessionization.

Table 4.10: Pair-wise Web Session Similarity, No. of missed sessions and % showing missed sessions measure-wise

Similarity Measures	**Dataset-1**			**Dataset-2**			**Dataset-3**		
	Session Pair	Missed Sessions	% of Missed Sessions	Session Pair	Missed Sessions	% of Missed Sessions	Session Pair	Missed Sessions	% of Missed Sessions
Euclidean Measure	214	70	32.71	556	122	21.94	12532	1949	15.55
Cosine Measure	224	60	26.79	588	90	15.31	12540	1941	15.48
Jaccard Measure	230	54	23.48	592	86	14.53	12484	1997	16.00
Angular Separation	226	58	25.66	586	92	15.70	12547	1934	15.41
Canberra Distance	228	56	24.56	568	110	19.37	12562	1919	15.28
Chi Square	248	36	14.52	580	98	16.90	12574	1907	15.17
ST_Index	**254**	**30**	**11.81**	**624**	**54**	**8.65**	**12602**	**1879**	**14.91**

One way is to construct the IP based sessions. We also constructed the sessions based on IP and it was observed that very few sessions are constructed due to proxy and cache issue. Next, we also constructed the sessions based on IP, User Agent, and Operating System; the results were not up to mark as most of the users are using same User Agent (User Browser) and OS. For more accurate results, the attribute of Referrer Page was also used for session construction along with IP, User Agent, and OS. These heuristics generated better results comparatively. We also retained only those sessions, where page traversed by users are more than 1. In few cases, it was observed that only one page was hit by the user and such type of users are ineffective for the analysis of user behavior.

After the construction of web sessions, we implemented the algorithm to compute the ST_Index. The details of ST_Index algorithms are given in Algorithm 4.5. To verify the results of ST_Index, we also implemented the other well-known measures (Table 4.10, 4.11). The proposed ST_Index generated better results than the Euclidean, Jaccard, Cosine, Canberra, Angular, and initially proposed Chi-Square measures. ST_Index produced remarkable results of over Euclidean Family of measures as Euclidean measures compute the distance between sessions rather than computing similarity among the sessions. In this regard, ST_Index outclassed the Euclidean Family of measures. Jaccard and Cosine and Chi-Square measures produced results on based web pages traversed among sessions. However, only common URLs (Pages) are insufficient for session similarity. ST_Index weighted the common URLs among the sessions along with the respective uncommon web pages in the form of SessionIndex (SI). ST_Index also applied another, important attribute Time spent by users in a session. ST_Index computed the TimeIndex (TI). Alam et al. [41] applied the time factor along with the data downloaded with pages traversed. Such type of attributes may help in Euclidean measures but insignificant to the vector-based similarity measures. These features give an edge to ST_Index over the other competitive measures.

The analysis of sessions generated showed that web session similarity computed through Euclidean measures, most of the time interlink the sessions which have nothing in common between two sessions in question. The results generated

Algorithm 4.5: ST_Index Web Proximity Matrix

```
Data: Weblog Session Database S of $n$ records(transactions)
Result: ST_Index Web Proximity Matrix
$L = \{SessionID, S_1, S_2, TPV, TST, SessionIndex, TimeIndex, ST\ Index\}$
where $S_1$ and $S_2$ are two sessions, Total Page Visited (TST), Total Spent Time (TST)
while do
    for $i$ = level +1 → TotalSessions do
        Read L
        SessionIndex = 0
        TimeIndex = 0
        ST_Index = 0
        $S_1 = i$
        for $j$ = i+1 → TotalSessions do
            $S_2 = j$
            Compute SessionIndex $(S_1, S_2, TPV)$
            Compute TimeIndex $(S_1, S_2, TST)$
            Compute ST_Index (SessionIndex, TimeIndex)
        end
        Update Database
        $S_1 = i$
        $S_2 = j$
    end
    Generate Proximity Matrix
    Update WebSession Database
end
```

through such measures may deliver the biased results and these results are fatal to whole WUM process. Cosine, Jaccard, and Chi-Square are producing weak results, in some sessions, their matching is accurate, and in some sessions, matching is irrelevant. Such type of mismatched behavior is vulnerable for the WebKDD and WUM particularly. The placement of each session at its right place is baseline requirement of ST_Index. However, some of the sessions do not take part in the run with other sessions due to a number of factors such as less number of web pages visited in sessions and time consumed by the user in session is negligible.

In Table 4.10, we computed the session pair as the identification of true relation between the two session is worthy to understand the traversing pattern between them. The Missed sessions are also being computed on the basis of actual data. The missed sessions may be due to scoring values. We pick the session pair with maximum scoring value with first occurrence. It may be possible more than two session pair have same score. The percentage of missed sessions is also computed and it can be seen that ST_Index has less number of missed sessions as compared to the other measures. In Figure 4.5, the same comparison is shown of other session similarity measures with ST_Index and ST_Index is displaying significant less number of misplaced sessions.

For further validation of results, we construct the confusion matrix based on the TN (True Negative) (session is active and participated in similarity computation), FP (False Positive) (session is active and did not participate in similarity computation), TP (True Positive) (session is inactive and did not participate in similarity computation) and FN (False Negative) (session is inactive and participated in similarity computation) for ST_Index and other measures from the filter data of weblog (Table 4.11). The accuracy, validity, and correctness for sessions similarity is cornerstone and guarantee for the success of WUM process overall. In this regard, we tested the ST_Index along with other measures of the Precision, Recall, F-Measure, and Accuracy. ST_Index is showing better results as compared to the other measures with all the three datasets.

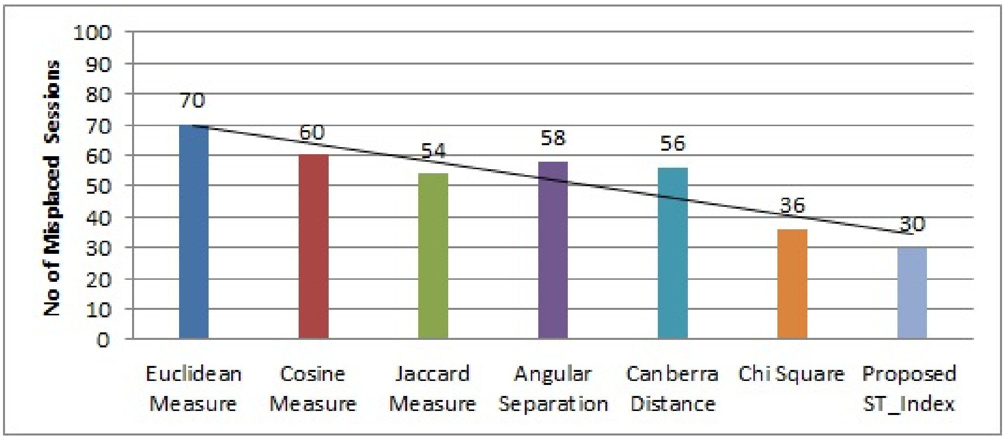


Figure 4.5: Comparison of Missed Sessions of ST_Index and Other Measures

Table 4.11: Performance Evaluation of different Similarity Measures

Similarity Measures	Datasets	TN	FP	FN	TP	Precision	Recall	F-Measure	Accuracy
ST_Index	**Dataset-1**	**254**	**30**	**0**	**36**	**0.8944**	**0.8759**	**0.885**	**0.9063**
	Dataset-2	**624**	**54**	**0**	**82**	**0.9204**	**0.8839**	**0.9017**	**0.9289**
	Dataset-3	**12602**	**1879**	**0**	**2664**	**0.8702**	**0.8255**	**0.8473**	**0.8904**
Chi-Square	Dataset-1	248	36	0	36	0.8732	0.8732	0.8732	0.8875
	Dataset-2	580	98	0	82	0.8555	0.8761	0.8657	0.8711
	Dataset-3	12574	1907	0	2664	0.8683	0.8252	0.8462	0.8888
Canberra Distance	Dataset-1	228	56	0	36	0.8028	0.8636	0.8321	0.825
	Dataset-2	568	110	0	82	0.8378	0.8738	0.8554	0.8553
	Dataset-3	12562	1919	0	2664	0.8675	0.825	0.8457	0.8881
Angular Separation	Dataset-1	226	58	0	36	0.7958	0.8626	0.8278	0.8188
	Dataset-2	586	92	0	82	0.8643	0.8772	0.8707	0.8789
	Dataset-3	12547	1934	0	2664	0.8664	0.8249	0.8451	0.8872
Jaccard Measure	Dataset-1	230	54	0	36	0.8099	0.8647	0.8364	0.8313
	Dataset-2	592	86	0	82	0.8732	0.8783	0.8757	0.8868
	Dataset-3	12484	1997	0	2664	0.8621	0.8241	0.8427	0.8835
Cosine Measure	Dataset-1	224	60	0	36	0.7887	0.8615	0.8235	0.8125
	Dataset-2	588	90	0	82	0.8673	0.8776	0.8724	0.8816
	Dataset-3	12540	1941	0	2664	0.866	0.8248	0.8449	0.8868
Euclidean Measure	Dataset-1	214	70	0	36	0.7535	0.856	0.8015	0.7813
	Dataset-2	556	122	0	82	0.8201	0.8715	0.845	0.8395
	Dataset-3	12532	1949	0	2664	0.8654	0.8247	0.8446	0.8863

4.15 Summary

In this chapter, we iterated that preprocessing is a first major step for the WUM process and without the proper and effective preprocessing, the correctness of the results is impossible. Mostly, researcher ignored the importance of weblog preprocessing and performed the web usage mining. In web session similarity section, we proposed the web session similarity measure ST_Index to overcome the limitations of existing web session measures. ST_Index utilizes the theme of common web pages traversed in sessions along with the weight assigned to mismatch web pages in relevant sessions. Secondly, time index has been used by picking the minimum time in both the sessions. ST_Index computes the similarity rather than a distance among the sessions. ST_Index results are validated against the actual datasets. The proposed measure ST_Index has been evaluated against the well-known web session similarity measures. The performance of ST_Index is better in all the metrics such as Precision; Recall; F-Measure; and Accuracy. This indicates that use of such a realistic approach like ST_Index is going to a valuable addition in the field of Web Session Similarity Measures.

Chapter 5

Results and Evaluation

The F_MET is a complete working model of web usage mining in web mining knowledge discovery in Databases (WebKDD). The F_MET is being implemented by incorporating the various state of the art data mining techniques stepwise according to web mining phases. In this chapter, we are implementing the F_MET in the practical domain through the particle swarm optimization for hierarchical sessionization. The core working of the F_MET is based on the proposed ST_Index, the web session similarity measure. We also implemented the chi-square web session similarity measure for hierarchical sessionization.

5.1 Hierarchical Sessionization

Hawwash and Nasraoui [5] generated the hierarchical sessionization by applying GA-based Hierarchical Unsupervised Niche Clustering (HUNC) Algorithm for user profiling. The fitness function applied for HUNC is $f_i = \frac{\sum_{j=1}^{N_i} W_{ij}}{\sigma_i^2}$ where f_i is density of profile i, W_{ij} weight, σ_i^2 variance and N_i be the total number sessions. The proposed framework works efficiently by delivering accurate clusters with improved visualization. The GA suffers from the fitness function for sessions and in the case of updating the user profiles, the database scan can further affect the efficiency of the WebKDD process. On the other hand, Chakraborty and Bandyopadhyay [120] applied a Fast Optimal Global Sequence Alignment Algorithm (FOGSAA) technique to address the scalability and overlapping user behavior. Sequence alignment was used to overcome the matching hurdle of Needleman-Wunsch.[28] are working since 2008 on various web usage mining approaches and particularly applying Swarm Intelligence. The authors produced the hierarchical sessionization and dropped the weak sessions. This strategy may lead to missing the important patterns and the accuracy of results produced are challengeable. Kundra et al. [67] also applied the PSO for hierarchical sessionization without dropping the weak sessions. The only major issue with the proposed idea of Kundra et al. [67] is the use of two similarity measures while the various similarity measures are available which can handle both the numeric and the categorical nature of the weblog data. In following sections, we are applying simple hierarchical sessionization based on chi square and PSO based hierarchical sessionization.

5.2 Chi-Square based Hierarchical Sessionization

Hierarchical sessionization is an important aspect of proposed F_MET and there are various techniques available in the literature as discussed in literature review chapter. For the empirical validation of hierarchical sessions, we performed the Chi-Square based hierarchical sessionization [12]. After preprocessing step, we constructed the sessions and calculated the chi-square values based on the parameters of a number of web pages and a session time in each session. The chi-square value of each session is computed with every other session and the highest chi-square value shows the strong correlation between these two sessions. If more than one sessions have the same higher value, than the first occurrence is considered a more appropriate pair of related sessions. This is the first level hierarchy. We also computed the average of the most related pairs, for the calculation of next hierarchy level and for the height of related session in the dendrogram. We applied the following proposed algorithm 5.1 for chi based hierarchical sessionization of the weblog. The results of Chi-Square based hierarchical sessionization were compared with the existing hierarchical clustering algorithms of WUM.

Algorithm 5.1: Chi Square Web Sessionization

```
Data: Weblog Session Database S of $n$ records (transactions)
Result: Chi Square Similar Sessions
$L = \{SessionID, TPV, TST, R_1, R_2, AvgValue, Level\}$
where $R_1$ is not related to $R_2$, Total Page Visited (TST), Total Spent Time (TST)
while do
    for $i = level + 1 \rightarrow TotalSessions$ do
        Read L
        Max Chi = 0
        $R_1 = i$ for $j = i + 1 \rightarrow TotalSessions$ do
            $R_2 = j$
            Calculate Chi Square $(R_1, R_2)$
            if $ChiSquare > MaxChi$ then
                MaxChi = ChiSquare
                $R_1 = i$
                $R_2 = j$
                $AvgValue(R_1, R_2)$
            end
        end
    end
    Update WebSession Database
end
```

5.3 Genetic Algorithm based Web Sessionization

In 1970, Genetic Algorithm (GA) was introduced based on the principles of genetics. GA incorporates the biological principle of reproduction (genes) of population. GA has wide range of applications in data mining and extensively applied for the solution of complex problems. The major evaluation of GA is survival of the fittest based on the fitness or objective function. During the search process of GA, a dataset (population) is selected against the predefined fitness function according to the given environment.The fitness function helps to eliminate the weaker population at an early stage of the GA learning. The GA operators "Crossover" and "Mutation" further helps to improve the scoring function of GA at number of iterations. Finally the results are evaluated for the optimized results.

The web sessionization is complex problem and can be addressed through the GA. The population size is the number of web sessions after the processed and filtered data. The population (sessions) are initialized as per the URLs (web pages) traversed in a session. For the selection of sessions that will take part in the next stage of GA is based on the fitness function ST_Index. We select 5 max value pair from ST_Index results and take the average as fitness function. After the selection process, we have to apply crossover and mutation. since we are not applying the crossover operator as we are not converting session into binary form. We are just applying the mutation operator. for the mutation process, we have alloted a unique URLID to the web pages during the sessionization step. We select a randomly URLID and replace it randomly in the pool of selected web sessions in such a way that the number of web pages remained unchanged in sessions. However, mutation will effect the fitness value on each iteration. At the evaluation step, we select the sessions with maximum ST_index value. The stopping criteria for the GA is the occurrence of super fit (max valued pair of sessions).

For the comparison of GA and PSO, we keep the few external parameters same such as size (the number of sessions), fitness function (ST_Index) and URLID. The GA is implemented through Matlab 10. The results of GA are presented in Table 5.1.

5.4 Particle Swarm Optimization

The use of swarm model to the real scenario is helpful to overcome the limitations of traditional web mining approaches and guarantee the performance and accuracy of results of the WebKDD process. For the last few years, swarm model is producing the promising results in data mining and particularly in the WebKDD process to enhance the reliability and dependability of web-based applications. The swarm model is being applied in various web mining techniques such as clustering, classification, feature selection and outlier detection. Eberhart and Kennedy [160] proposed the swarm-inspired meta-heuristic approach Particle Swarm Optimization (PSO) through incorporating the social and cognitive behavior of swarms. PSO models the computational problem by introducing the swarm intelligence in meta-heuristic way. Due to these added features, PSO has become the ultimate choice of the research community in evolutionary approaches to problem-solving. A real life problem can be modeled through incorporating the swarm intelligence flocks and each individual item is called swarm agent (particles). These particles move under certain constraints and rules in an environment linking with each other and communication with each other to take part as a solution particle by exhibiting the swarm intelligence. While moving towards the target, particles reveal the local and global social behavior for the communication and cooperation with each other in a decentralized environment. Furthermore, modeling of the problem to swarm intelligence is quite simple in implementation, which makes the PSO ultimate choice to solve the complex computational problem in an optimized way.

PSO adapts this behavior of birds and searches for the best solution vector in the search space. A single solution is called particle. Each particle has a fitness value that is evaluated by the function to be minimized, and each particle has a velocity that directs the movement of the particles. The particles move through the search space by following the optimum particles.

The algorithm is initialized with particles at random positions, and then it explores the search space to find better solutions. In each iteration, each particle adjusts its velocity and position to follow two best solutions. There are three main factors, which are causing movement and controlling the particles Eq. 5.1.

$$\begin{cases} Cognitive factor, (pBest - x_i) \\ Social factor, (gBest - x_i) \\ Self - organizing factor, (y_i - x_i) \end{cases} \tag{5.1}$$

The first is the cognitive part, where the particle follows its own best solution found so far. This is the solution that produces the lowest cost (has the highest fitness). This value is called pBest (particle best or local best). The other best value is the current best solution of the swarm, i.e., the best solution by any particle in the swarm. This value is called gBest (global best). Then, each particle adjusts its velocity and position with the following equations Eq. 5.2 & 5.4:

$$v_i = c_o * v + c_1 * rand() * (pBest - x_i) + c_2 * rand() * (gBest - x_i) \tag{5.2}$$

$$x_i = x + v_i \tag{5.3}$$

v is the current velocity, v_i the new velocity, x the current position, x_i the new position, pBest and gBest as stated above, rand( ) is even distributed random numbers in the interval $\{0, 1\}$, and c_1 and c_2 are acceleration coefficients. Where c_1 is the factor that influences the cognitive behavior, i.e., how much the particle will follow its own best solution and c_2 is the factor for social behavior, i.e., how much the particle will follow the swarm's best solution. The pseudocode of Particle Swarm Optimization is given in Algorithm 5.2.

Algorithm 5.2: Standard Particle Swarm Optimization

```
Data: Particles(Clusters); Optimize Function; Swarm Size; Problem Dimension
Result: The Best Fitness Value gBest
for Each Particle do
    Initialize Particles
end
while do
    for Each Particle do
        Calculate fitness Value pBest
        if Fitness Value >pBest then
            Set Current Value = pBest
        end
    end
    Choose the Particle with the Best Fitness Value of gBest
    for Each Particle do
        Calculate fitness Velocity
        Update Particle Position
    end
    While Maximum Iteration or Minimum Error Criteria is not Attained
end
```

5.5 A Proposed Particle Swarm Optimization based Hierarchical Clustering Algorithm (PSO-HAC)

The proposed particle swarm optimization based hierarchical clustering algorithm is simply working like agglomerative hierarchical clustering in an optimized way. PSO-HAC takes all the sessions (particles) clusters and merges them into pairs based on ST_Index criteria. The merging of sessions (clusters) continues until a complete web session hierarchy is achieved in an iterative mode. The sessions adjust their best position during the iterations for an optimized solution. Both the hierarchical and partitioning clustering algorithms suffer initialization, local maxima of particles and efficiency deficiencies by default. The proposed PSO-HAC is the combination of swarm particles optimization and agglomerative algorithm to overcome the above-cited issues in an efficient and optimized way.

The proposed PSO-HAC has three main components like the PSO such as initialization of particles, social and cognitive learning process through the movement (velocity) of sessions and updating the new positions of sessions.

Initialization: The input for the PSO-HAC is preprocessed and filtered database where the sessions have been created. Each session is assigned a distinct SessionID and there are n number of total sessions in session database. For the PSO-HAC, we marked the particles as sessions and initialized with the total page visited (TPV) and total time spent in a session (TST) in each session in swarm search space. These are the local best positions of sessions. The number of iterations for the PSO-HAC is total number of sessions n in database.

Fitness Function: In this experiment, we are applying the proposed similarity measure ST_Index as fitness function. We created a separate session database in which we calculated the proximity values of each session with every other session based on the parameters defined in ST_Index. We compare the sessions and select the pair of sessions which have maximum session ST_Index (proximity value).

Session Position: The particles never die in swarm search space and just change their positions by adapting the change in position vectors. The sessions are initialized on the basis of two parameters TPV and TST. The change in position of sessions is controlled by the time vector ($x_i = x_i[TST]+1$). During each iteration, we select the winning pair($pBest$) of sessions in a separate database with maximum proximity value(ST_Index). The leading session from the pair is removed from the search space for the next round(iteration). Again proximity values are calculated and best pair is updated in database. We are not directly applying cognitive and social factors. We are applying the self-organizing factor of standard PSO.

Construction of Hierarchies: We arrange the winning sessions and select the *gBest* pair of sessions and joined the sessions with head to tail method through the connect function. If pair of session is not connected with the other sessions, then its hierarchy level is marked 1. Similarly, if pair of session is being connected with the other pairs (sessions) successively, the hierarchy level is increased. The starting session is marked as root session in hierarchy table.

5.6 Experiments

In this section, we are carrying out the experiments to evaluate the proposed framework for mining trends (F_MET) and how it is effective along with the proposed PSO-HAC based Hierarchical Sessionization. The summary of the experiments is as under:

- The experiments are being conducted based on three datasets, whose detail has been discussed in Chapter 5.
- Chapter 6, we explained the components of proposed (F_MET) and performed comparison with available web usage mining frameworks.
- The experiments are being supported by the various clustering analysis measures to reveal the performance and defects of proposed session clustering algorithm. The performance measures are page coherence (PC), F-Measure, Accuracy, Precision, and Recall.

Algorithm 5.3: Particle Swarm Optimization based Hierarchical Sessionization

```
Data: Web Sessions SessionID, TPV TST
Data: $S = \{S_1, S_2, S_3, \ldots, S_n\}$where n $\neq 0$
Data: Web Sessions as Swarm Particles, pBest, gBest
Result: The set of gBest Particles(Clusters)
for Each Particle do
    Initialize Particles with TPV, TST
end
while do
    for $i = 1 \rightarrow MaxSessions$ do
        for $j = i + 1 \rightarrow MaxSessions$ do
            $ST\_Index(S_i, S_j) = SessionIndex * TimeIndex$
        end
        Pick the Web Sessions Pair with Max(ST_Index)
        Update the Positions of Web Sessions
        $x_i$= $x_i(TST) + 1$
    end
    Update the pBest, gBest Databases
    Update the Web Sessions
end
Construction of Session Hierarchies
for $i = 1 \rightarrow MaxSessions$ do
    Connect head to tail of the sessions from winning gBest sessions.
    Marked the leading session as RootSession.
    level=level + 1
    Update the Hierarchy Table
end
```

- The proposed F_MET results were compared with the standard heirarchical agglomerative clustering (HAC).
- The clustering technique was compared with the [28] works. Their research is closely related to our proposed scheme.

5.6.1 Resource Datasets

We are applying the proposed F_MET on the three databases. The proposed framework F_MET can be applied on any weblog data for the extraction of user patterns and knowledge discovery. The details about the datasets have already been discussed in Chapter 4. The proposed F_MET is not being developed as a tool, however the F_MET prototype was developed. The weblog datasets have been converted from text to MS Excel database where we labeled the weblog attributes. For the F_MET implementation, we further transformed the MS Excel database into Oracle 10g database. The F_MET procedures and algorithms are being developed through PL SQL. The machine used for prototype is core i7 with 32GB memory.

5.6.2 Comparison Metrics

In this section, we are explaining the metric (measures) used for the performance as VC (Visit Coherence) or PC (Page Coherence); Accuracy; Coverage; and evaluation of the proposed particle swarm optimization based hierarchical clustering algorithm. These measures are commonly practiced in web usage mining to ensure the accuracy and quality of the web session clustering algorithm and results produced by the PSO-HAC.The measures used are also discussed in [77, 161]. The VC measure checks the right placement of the web pages visited by the users in various sessions. It is very important that session clustering must be coherent of the page visited in that sessions. It has been observed that mostly the distance base similarity measures displace the actual page traversed the user sessions. The PC measure will ensure that how the proposed web session clustering algorithm will place the clusters with a coherent set of web pages. Consequently, while developing the cluster hierarchy, it is essential the right placement of the web pages in right cluster. The PC will further help us to investigate the trends (Patterns) and knowledge learning process from WebKDD.

Visit-coherence is utilized to evaluate the quality of the clusters (navigation patterns) produced during the off-line phase. Furthermore, visit-coherence quantifies a session intrinsic coherence. As in the Page Gather system, the basic assumption here is that the coherence hypotheses hold for every session. To evaluate the visit-coherence, we split the dataset into two halves after the pretreatment phase. The clustering task is applied on the first half dataset and the recommendation engine is employed on the second half dataset to create recommendations. Visit-coherence is then evaluated based on the recommendations. The second half of the dataset is known as evaluation dataset. In this study, parameter is defined to measure the number of Web pages in every session i that belongs to a navigation pattern (cluster) found for that session as in Eq. 5.4.

$$\beta_i = \frac{\{p \in S_i | p \in C_i\}}{N_i} \tag{5.4}$$

where p is a page, S_i is an i_{th} session, C_i is the cluster representing i, and N_i is the number of pages in an i_{th} session. The average value for overall N sessions in the evaluation part of dataset is shown in Eq. 5.5.

$$\alpha = \frac{\sum_{i=1}^{N_s} \beta_i}{N_s} \tag{5.5}$$

Where α is percentage of the visit-coherence and N_s total sessions.

For further analysis of hierarchical sessionization, we applied the three more quality measures Accuracy; Coverage; and F-Measure. The web session clustering revolves around the web pages traversed in web sessions. The user behavior analysis is based on these web sessions and common interest among the sessions is extracted from the common web pages. We are applying these metrics to evaluate the performance of these sessions, in following paragraph. An accuracy is a number of relevant Web pages retrieved and divided by the total number of Web pages in cluster paired. The Accuracy for session clustering is defined in Eq. 5.6.

$$Accuracy = \frac{|S_i \cap S_j|}{\sum_{i,j=1}^{n}(S_i, S_j)} \tag{5.6}$$

Where $S_i \cap S_j$ are the common web page traversed in both the sessions. Another evaluation parameter, coverage is defined in Eq. 5.8. The coverage is the ratio between the numbers of relevant Web pages retrieved and the total number of Web pages that actually belongs to the user session. On the other hand, coverage measures the ability to manage the sessions as a whole.

$$Coverage = \frac{S_i \cap S_j}{\sum_{i,j=1}^{n}(\widetilde{S}_i, \widetilde{S}_j)} \tag{5.7}$$

The F1-measure attains its maximum value when both accuracy and coverage are

TABLE 5.1: Comparison of Proposed PSO-HAC, HPSO and Genetic Algorithm on the basis of performance metric

Technique	Datasets	VC	Coverage	Accuracy	F1-Measure
Proposed PSO-HAC	**Dataset-1**	**0.8944**	**0.7890**	**0.8350**	**0.8113**
	Dataset-2	**0.8204**	**0.7039**	**0.8517**	**0.7707**
	Dataset-3	**0.8702**	**0.7255**	**0.7973**	**0.7597**
HPSO (Alam et al., 2014)	Dataset-1	0.8732	0.7732	0.7532	0.7630
	Dataset-2	0.8555	0.7661	0.7457	0.7557
	Dataset-3	0.8683	0.7252	0.7262	0.7257
Genetic Algorithm	Dataset-1	0.8237	0.7802	0.7232	0.7530
	Dataset-2	0.8155	0.7061	0.7257	0.7857
	Dataset-3	0.8300	0.6914	0.7390	0.7765

maximized.

$$F1 = \frac{(2 * Accuracy \times Coverage)}{(Accuracy + Coverage)} \tag{5.8}$$

We applied these parameters on all the three datasets to evaluate the performance of web session clustering. We also applied these parameters on the [28] for fair comparison of our proposed techniques with it. The results of the parameters are shown in Table 5.1.

5.6.3 Hypothesis Testing (t *test*)

In this section we are interested to compare the significance of PSO-HAC with standard genetic algorithm (GA) statistically. We are applying student t *test* to check the significance and effectiveness of PSO-HAC and GA. Since both PSO and GA are evolutionary heuristics approaches and are widely applied in industry and academia. During the null hypothesis testing H_0, we set the null hypothesis as ***"There is nothing significance difference between PSO-HAC and GA"*** with respect to the web sessionization problem. The alternative hypothesis H_1 is defined as: ***"The proposed technique has better significance (accuracy wise) with respect to standard GA on basis of*** α***=0.05."*** If null hypothesis H_0 is accepted during the student t *test*, this means that both PSO and GA have same level of significance and can be applied both for web sessionization. If hypothesis is rejected, it will be clear indication that PSO has better efficacy over GA in this particular problem of web sessionization.

For *t test*, we used the Microsoft Excel 2013. PSO and GA based results results were generated by applying ST_Index as fitness function for PSO and for GA, we took the average of 5 max value ST_Index paired as fitness function. For data analysis, we applied unequal variance as for both the techniques PSO and GA, we don't know the variance. According to null hypothesis, both the techniques have same level of significance, therefore we kept the hypothesized mean difference 0. The significance level α is marked as 0.05 (5%) as a probability of rejecting null hypothesis. We tested the *t test* on the sample from the $Dataset - 1$. The details of outcome of the tests are given below in Table 5.2.

Table 5.2: *t test* parametric comparison of PSO-HAC and GA

Parameters	**PSO-HAC**	**GA**
Mean	0.8765	0.7945
Variance	0.8279	5.2657
Sample Size	50	50
$t - Stat$	2.8927	
$P(T <= t)two - tail$	0.0065	
$t\ Critical\ two - tail$	2.0280	

Since the $P(T <= t)two - tail$ is less than the α value and hypothesis H_0 is rejected that there is no significance difference between PSO-HAC and GA.

5.6.4 Results and Evaluation

We applied proposed PSO-HAC algorithm to produce the hierarchical sessionization. The core of PSO-HAC is our proposed web session similarity measure ST_Index. The dendrogram represents the hierarchy of clusters as shown in Figure 5.1. The dendrogram was obtained through the Matlab 10. Due to space and compactness of full dendrogram of experiment, we are just showing the 30 sessions in dendrogram. The dendrogram shows the relationships among the sessions based on proposed ST_Index. We are presenting the dendrogram of one only experiment due to shortage of space. In Figure 5.2, the number of web pages present in web sessions is shown. In few sessions, the number of web pages traversed is more than 50. In one session, web pages visited are 388. The presence of such a high number of web pages in a session affects the accuracy and coverage issues. However, on

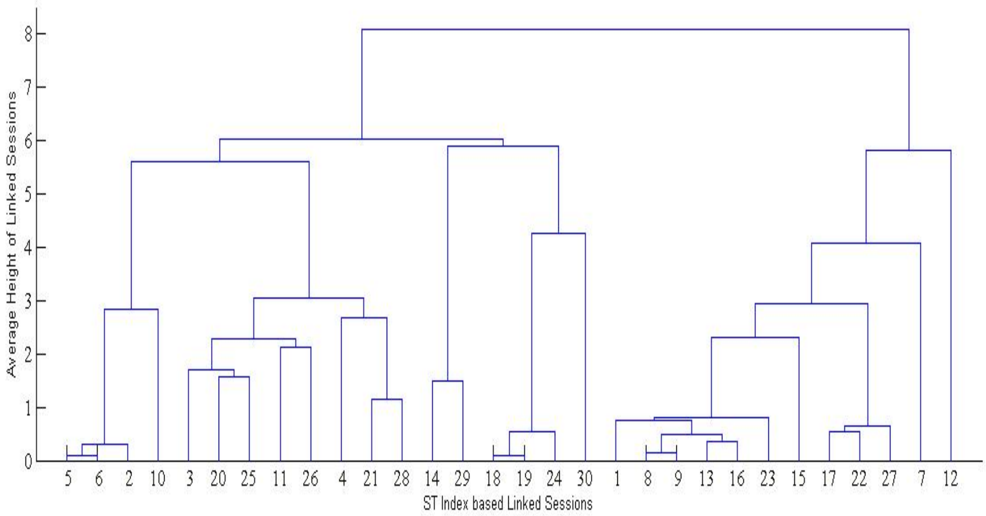


Figure 5.1: The Dendrogram of Web Session Clusters based on Proposed PSO-HAC

the analysis of the weblog, it indicates that the session construction at preprocessing level is a complex phenomenon. We applied the various heuristics approaches to identify the sessions as defined in preprocessing chapter. However, the true

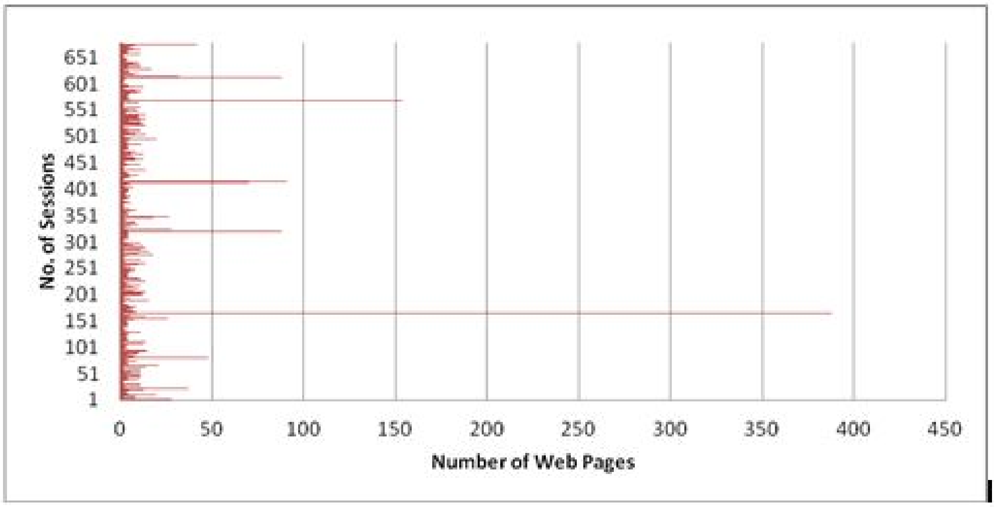


Figure 5.2: No. of Pages traversed in Each Session

user identification from the weblog is complex and affects the overall web usage mining process. This artifact is being ignored in most of the research. This issue was not properly managed by [28] that causes the accuracy and coverage of web sessionization process. Another factor that is affecting the performance of web usage mining process is the time spent in a session by a user. In Figure 5.3, we are showing the time spent in each of the session used in our experiment. In most of

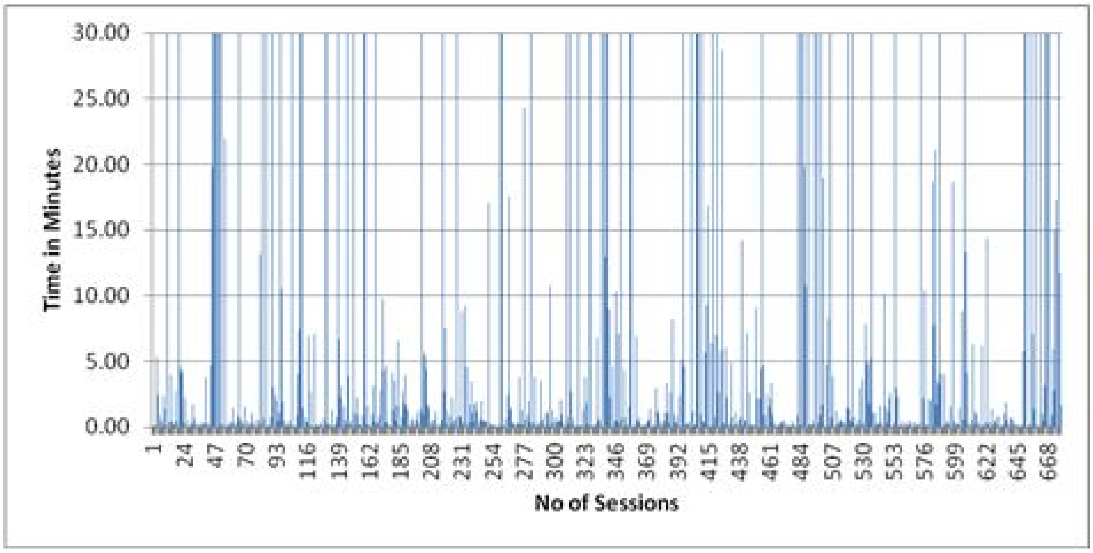
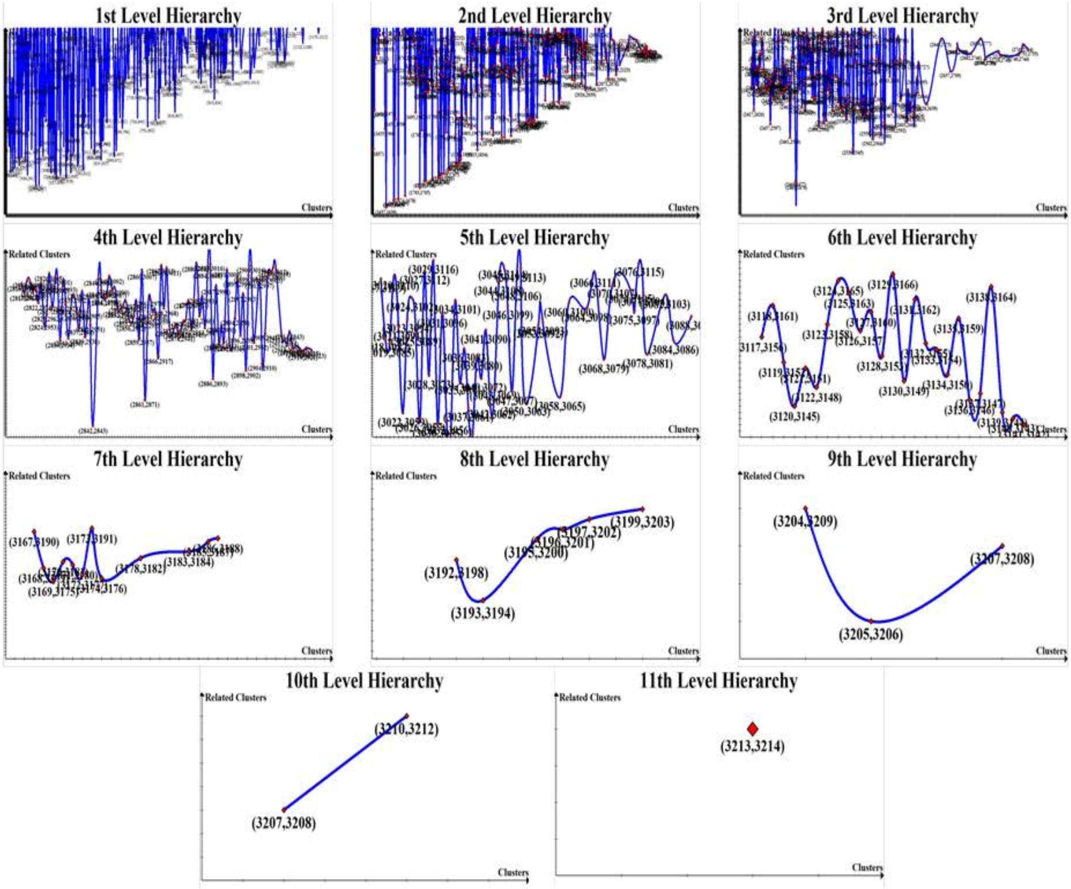


Figure 5.3: Session-Wise Time Spent by Users

the research, researchers used the 30 minutes heuristics for sessions, and we also applied the same heuristics of 30 minutes. However, more than 30 minutes time spent was further converted into multiple sessions. Paying the due attention to this factor, the overall improvement in quality results can be achieved in the web usage mining process. The average time for this particular dataset is 5 minutes for an average of 6 pages visited by the user in a session. Our average time used in a session and an average number of pages visited in sessions are due to the noise filtering at preprocessing level. For the effective and targeted web browsing, the average time and page visited are 5 10 minutes and 5-10 pages visit. This notion can vary from the category of the website to website. The entertaining websites and educational websites have a different level of usage and different average time utilization can be observed.



1st Level Hierarchy, 2nd Level Hierarchy, 3rd Level Hierarchy, 4th Level Hierarchy, 5th Level Hierarchy, 6th Level Hierarchy, 7th Level Hierarchy, 8th Level Hierarchy, 9th Level Hierarchy, 10th Level Hierarchy, 11th Level Hierarchy

Figure 5.4: Chi square based hierarchical sessionization

We extended the experiment with different similarity measures to check the efficacy of various well-known similarity measures as shown in Figure 5.5. Most of these measures are available in different data mining tools and are applied for the data analysis. It has been observed that for small datasets (Sessions), the most of the measures are generating good results. However, for increased number of data values (Sessions), the proposed ST_Index is showing increased accuracy. The measures used for hierarchical sessionization are frequently used in web usage mining for the pattern identifications.

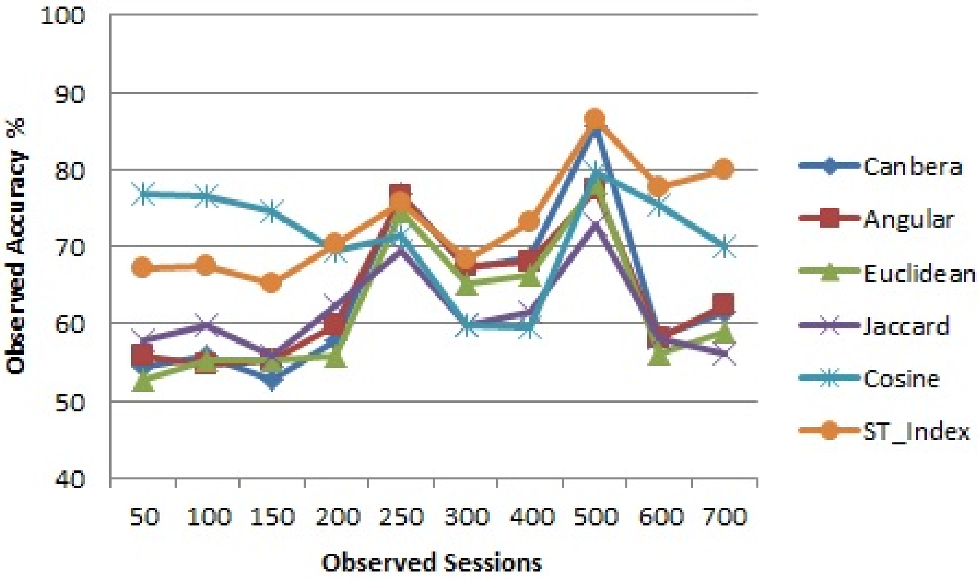


Figure 5.5: Accuracy of proposed PSO-HAC with Different Similarity Measures

Alam et al., (2014) applied the HPSO based on Euclidean and Hemming Distance for hierarchical sessionization. In their previous research [23, 41] applied the Euclidean Measure for hierarchical sessionization, however, the Euclidean measure has been failed to produce the accurate and quality results. The authors himself changed the measure and applied the Euclidean measure with a hamming measure to improve the performance of PSO. The authors also added the downloaded data parameter from the weblog. The addition of this parameter is another overhead on performance and scalability of hierarchical sessionization. Today, the dynamic websites are the requirement of the web-based applications and data downloaded attribute may vary from user to user and can lead inaccurate results. We also compared the results of our proposed algorithm with the Alam et al. [28] HPSO, Hussain and Asghar [12] Chi-HAC, and standard hierarchical agglomerative clustering algorithm (HAC). The accuracy and precision are the two parameters on which compared the proposed algorithm PSO-HAC with other algorithms. The precision and accuracy of algorithms is computed from the attributes such as TP, TN, FP, and FN as shown in Table 4.10. The precision and accuracy were computed on the different chunks of clusters such as 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. Figure 5.6 is representing the precision of PSO-HAC with the HPSO. In HPSO, Alam et al. [28] used the attributes 4, 8, 12, 16, 20, and 24 and explained only the 4 weblog attributes. The selection of attributes for HPSO is complex and performance overhead as the HPSO is unable to cater the issue of dynamic behavior of websites. PSO-HAC is steady as the number of clusters are increasing. PSO-HAC outperformed the even Chi-HAC and simple HAC. The different levels of hierarchies are shown in Figure 5.4. The Chi-HAC and HAC have their own limitations as both the algorithms are unable to address the optimized results. However, both the algorithms have built the clusters hierarchy. Accuracy is another parameter to evaluate the clusters and are also applying the accuracy parameter to evaluate the proposed PSO-HAC along with HPSO, Chi-HAC, and HAC. In Figure 5.7, the results of accuracy of PSO-HAC shows that with the increasing number of clusters we get better the accuracy. The accuracy of HPSO is again weblog attribute dependent. If the number of attributes are increased, the performance and accuracy will decrease especially by adding the unproductive attributes such as Status Code, Method, HTTP Protocol, data download, etc. In PSO-HAC, we applied the only most suitable and relevant weblog attributes such as IP, Time, URLs(pages visited). The rest of weblog attributes are overheads and unproductive for the optimized clustering. PSO-HAC also outperformed the Chi-HAC (Figure 5.4) and HAC. Both the algorithms are showing poor accuracy as the graph in Figure 5.7 can be analyzed.

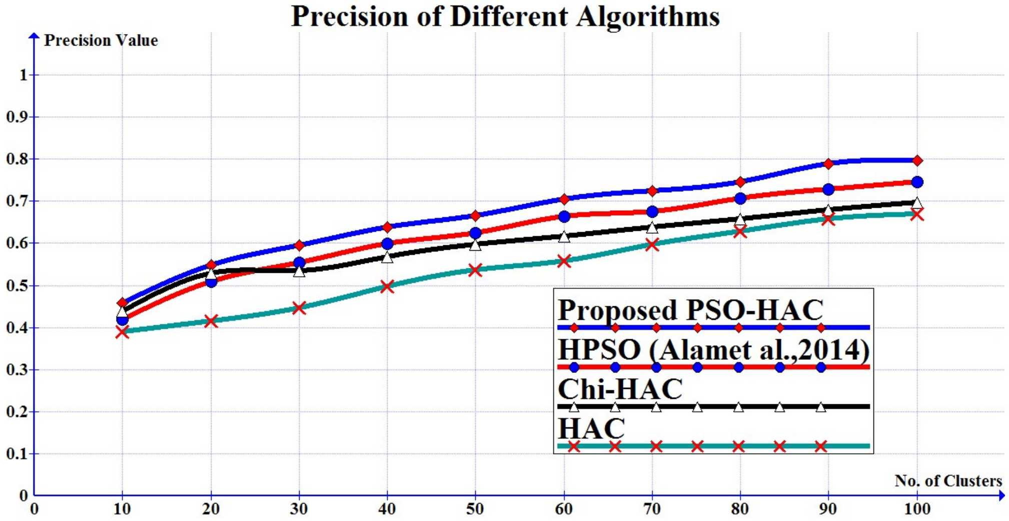


Figure 5.6: Precision Comparison of PSO-HAC with other Techniques

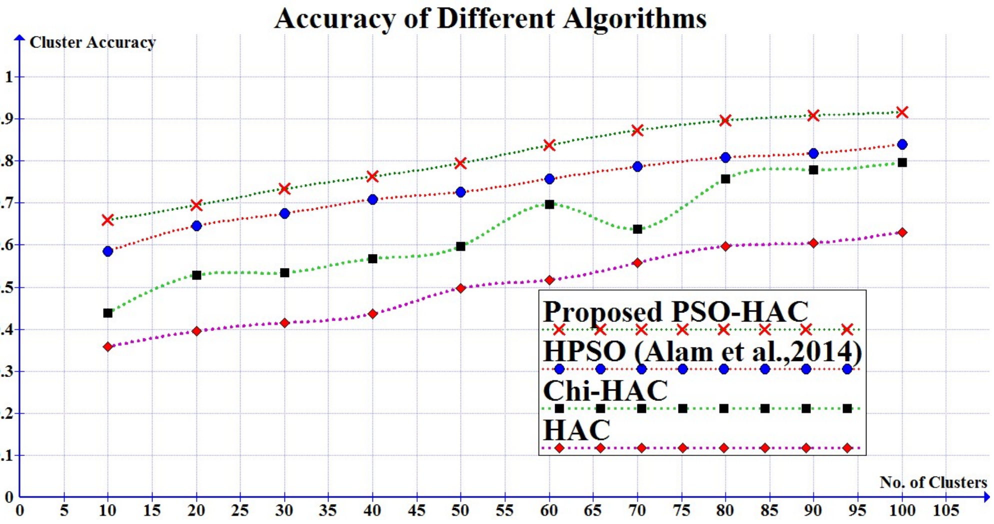


FIGURE 5.7: Accuracy Comparison of PSO-HAC with other Techniques

5.7 Knowledge Visualization

The knowledge visualization is an important step and phase of the WebKDD process. In our proposed framework F_MET, knowledge visualization is an important component and without the knowledge visualization, the web usage mining process is incomplete. For knowledge visualization, we adopted the hierarchical sessionization from its first level to onward until the matching criteria of sessions are finished. We link the sessions through an iterative process and find the local best from particle swarm optimization. The linking criteria are ST_Index. We find out the linkage of sessions (particles) at first level and then successively find the next best matching sessions. In each iteration, we change the position and velocity of sessions and find out the set of best matching values. We adopted the decreasing of iterations as the matching set grows. In Table 5.3, we present the knowledge analysis from the weblog. In this experiment, we took the weblog having 678 web sessions after the preprocessing and filtering. In Root Session, there is starting session and next, path gives the linkage. The session S_{68} is linked with session S_{111}, and so on, based on ST_Index scoring. In the level, column gives the total number of linked sessions in a row that how many users are interested in that specific theme. This is a complete one rule and it gives knowledge about one common subject. The experiments ware performed on the university weblog and rules generated show the users interests and preferences in university. In first rule,

Table 5.3: Knowledge Visualization from Weblog

ROOTSESSION	PATH	LEVEL
58	58 → 111 → 173 → 217 → 265 → 334 → 586	7
98	98 → 101 → 117 → 228 → 372 → 441 → 627	7
17	17 → 137 → 275 → 419 → 446 → 506	6
25	25 → 86 → 87 → 253 → 418 → 515	6
27	27 → 276 → 393 → 436 → 443 → 632	6
50	50 → 188 → 219 → 281 → 442 → 514	6
68	68 → 239 → 316 → 318 → 319 → 489	6
101	101 → 117 → 228 → 372 → 441 → 627	6
102	102 → 183 → 212 → 271 → 359 → 531	6
111	111 → 173 → 217 → 265 → 334 → 586	6
5	5 → 66 → 259 → 291 → 589	5
8	8 → 126 → 200 → 346 → 591	5
16	16 → 264 → 269 → 435 → 641	5
47	47 → 109 → 113 → 438 → 467	5

the users are seeking information about the Admission and in the second rule; the users are looking for Training offered by university. Hierarchical clustering algorithm, itself is best data analysis tool and particle swarm optimization further adds the best pair and finds the relationship in an optimized way. Each row represents the most closely related sessions group and identifies the group with similar traversing objectives and patterns. In one of our experiment, we get total 678 sessions. Our proposed particle swarm hierarchical cluster paired the global best of sessions based on ST_Index and produced the best intra-linked sessions to generate the business rules. In Figure 5.8, we show the different number of rules generated along with the paired level. Out of 678 sessions, 438 sessions have been

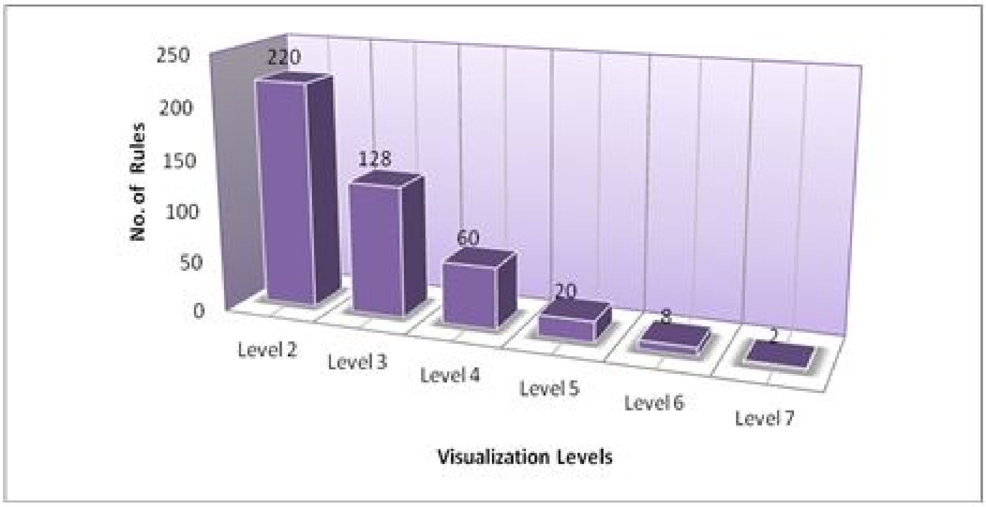


Figure 5.8: Visualization Levels and Corresponding Rules

so for utilized in knowledge discovery. In level 2, we get 220 rules. These rules have different traversing themes and only two sessions are paired. These are single paired sessions and show different traversing objective. In Level 3, three sessions are linked and it generated 128 rules. This shows that, there are 128 rules where three users have similar traversing behavior. In level 4, 4 sessions are linked and out of 678 sessions, 60 rules are generated and these 60 groups have 60 different searching objectives with a group of 4 sessions. In level 7, only two rules are visualized. These are a maximum number of linked session rule. In each rule, 7 sessions having a similar traversing taste.

5.7.1 Ground Truth (Expert Review)

There was no benchmark trends (patterns) data is available for the pattern identified from WUM process. To find the ground truth of the generated patterns is a complex job. To evaluate the effectiveness and accuracy of generated trends with the help of proposed PSO-HAC based on proposed web session similarity measure ST_Index, we took the services of two experts from the two local universities. Both the experts are maintaining the university websites. Expert A is a web administrator for the last 8 years in university, while the Expert B has more than 10 year of experience of web administration.

Process: Before the review process, we conducted the separate meeting with the experts and explained the whole procedure to them. We also explained them the research objectives and working of F_MET. During the meeting, experts raised the various questions regarding the working of F_MET. When experts were well aware of the F_MET, we delivered the data to them. The dataset includes the set of rules (trends) generated in the form of hierarchies and session database in MS Excel. The experts were given two different datasets for evaluation. The experts were given two-week time to give their expert opinion.

Feedback: After two weeks, expert A and B gave their suggestions and opinion. Overall, both the experts were satisfied with the effectiveness and accuracy of results produced by the F_MET. Expert A explained the different hierarchies and interconnection of hierarchies. Expert B explained the trends generated and appreciated the accuracy of F_MET. Expert B pointed out some missing referrer links. The missing link was due to the use of cache and sometimes the weblog entries are missing or broken. This is affecting the accuracy and validity of trends. The few results of experts and rules generated by the F_MET are shown and discussed in Conclusion Chapter section 8.4 Research Recommendations and Claims.

5.8 Summary

This chapter summarizes the results produced by the proposed framework F_MET. The proposed F_MET delivers the end to end solution to web sessionization problem. The major component of F_MET was the implementation of Particle Swarm Optimization algorithm to produce the hierarchical sessionization. The ST_Index measure was used as PSO scoring function in F_MET. The results of F_MET are based on the three different datasets of three different resources. The outcome of F_MET was the production of trends. The results are compared with the published results of Alam et al. [28].

Chapter 6

Conclusion

In this dissertation, we address the web sessionization problem and highlight its effectiveness in the web based applications. The growth of the web and big data has made the WebKDD process more challenging and demanding. The trends and knowledge discovery in weblog have attracted a great deal of research and new techniques have been introduced to deliver the smart solution to the web sessionization. The traditional web mining approaches and techniques are no longer delivering the elevated results and the cutting-edge revolution in technology has gathered a lot of user clickstreams. Only the leading-edge web mining techniques can cope with the sessionization issues at present day for the accurate and optimized sessionization.

6.1 Introduction

In Chapter I, we highlighted the Sessionization problem in details. We also synopsis and composed the overview of the web usage mining and its effectiveness in user clickstreams. The trend identification in a weblog is a challenging and complex phenomenon. The research motivation paved the path for the research objectives and research scope in the field of web usage mining. The problem statement highlighted the web sessionization problem that is being faced in web usage mining and is essential to be tackled in the form of framework to produce the complete solution. The abstract level of proposed framework F_MET was discussed as a solution of web sessionization. At the end of the chapter, we also elaborated the significance of web usage mining in industry and academia. Both Internet and its users are growing day by day; consequently, web mining tasks and challenges are increasing. How the new web mining approaches can be incorporated in developing and designing the web-based applications to provide secure; reliable and smooth functionality to its users.

For the last two decades, the research community is striving hard to tackle the challenges of knowledge identification from exponentially growing web data by applying web mining techniques and also working hard to strike out the approaches as well. In this thesis, we tried to update the web mining importance; utility; approaches and application. We also investigated the research gaps and issues for future research.

In chapter III, a meta-analysis of web sessionization was presented to pinpoint the gaps and limitations in existing literature. Regardless of web sessionization models and techniques, a variety of web session similarity measures are available for session similarity and negating one another. Consequently, there is no single web session similarity measure is available to fill the gap of accuracy, quality, and noise free. Furthermore, when we reviewed the literature on the web sessionization techniques, almost all the web mining techniques have been applied for the pattern discovery. However researchers are indecisive over the single choice, but clustering is widespread in all its forms. This is clear indication to address the sessionization problem in full swing, clustering is the ultimate choice. Moreover, when we reviewed the literature, the traditional and flat clustering techniques suffer the few by default defects that debar its application in web sessionization. Ultimately, researchers have shifted to evolutionary approaches to remove the hindrances of traditional clustering. Consequently, based on the review on web sessionization, a complete framework is dire need to cope with the expanding, dynamicity, and scalability issues of web sessionization.

In Chapter IV, we iterated that preprocessing is a first major step for the WUM process and without the proper and effective preprocessing, the correctness of the results is impossible. Mostly, researcher ignored the importance of weblog preprocessing and performed the web usage mining. We also discussed in details the various user identification heuristics and adopted the best option for user identification in the presence of proxy server and firewall. The true user identification is a challenging and complex job. We identified the users, with all the possible grouping such as IP; IP and User_Agent; and IP, User_Agent and Referrer_Page. The last approach is more appropriate and suitable in the current scenario for web sessionization. We also discussed in details the session construction mechanism with time constraint and applied the 30 minutes heuristic.

In Chapter V, we proposed the web session similarity measure ST_Index to overcome the limitations of existing web session measures. ST_Index utilized the theme of common web pages traversed in sessions along with the weight assigned to mismatch web pages in relevant sessions. Secondly, time index was used by picking the minimum time in both the sessions. ST_Index computed the similarity rather than a distance among the sessions. ST_Index results are validated against the actual datasets. The proposed measure ST_Index was evaluated against the well-known web session similarity measures. The performance of ST_Index is remarkable in all the metrics such as Precision; Recall; F-Measure; and Accuracy. This indicates that use of such a realistic approach like ST_Index will be a benchmark in the field of Web Session Similarity Measure and the results produced by ST_Index are the solid ground for authentic results of WUM process especially the Pattern Discovery and Knowledge visualization phases. The observed limitation of ST_Index is the generation of more than one similar relationship among the sessions. This limitation may be overcome by implementing the ST_Index through any evolutionary approach for pattern identification.

In Chapter VI, we proposed a framework for mining emerging trends (F_MET) in weblog data to overcome the web sessionization issues at various stages of WUM process. We discussed the effectiveness of F_MET to address the sessionization issues along with the merits and demerits of existing frameworks. We explained the components of the proposed framework and execute the dry run to verify its flow, working, and objectivity as a problem solver. The comparison of F_MET with the existing frameworks at abstract level proved its effectiveness. We presented the chi-square hierarchical sessionization and its published results as initially proposed a solution for web sessionization. However, few limitations were observed to address the sessionization at full length. Few sessions have more than one best matching ST_Index score and selection of optimized web session pair is difficult with the simple agglomerative hierarchical clustering.

As clustering is multi-modal optimization problem for intra similarity between the pair of web sessions and only evolutionary approaches can be helpful in this regard. Particle Swarm and Genetic Algorithms are the two best optimization problem solutions whereas particle swarm is commonly practiced for web sessionization in literature due to its simplicity in implementation with efficient and scalable results. While genetic algorithm suffers the robustness and trained population with a leftover feature that may miss the few interesting patterns. In next, chapter we are interested in implementing the F_MET with particle swarm for hierarchical sessionization as an optimized solution of web sessionization.

6.2 Summary of Contributions

In the following section, we are summarizing the research contributions of this dissertation. The comprehensive literature review was conducted that covered the all major aspects of web sessionization such as web session similarity measures; web session techniques; and hierarchical sessionization. The literature review helped us to compose the web sessionization problem statement. Introduced the preprocessing algorithms that helped us to prepare the noise free weblog data for upcoming web usage mining phases. Introduced the web session similarity measure, ST_Index, to overcome the limitations of existing web session similarity measures. The proposed framework F_MET was introduced as a complete solution to address the web sessionization problem. The proposed F_MET followed the complete life cycle of data mining and incorporated the web usage mining techniques to produce the accurate and quality results with high coverage. The proposed F_MET was tested and evaluated against the particle swarm optimization based hierarchical sessionization. The proposed algorithm was based on the web session similarity measure ST_Index and produced the quality results.

6.3 Research Limitations

In this research work, we have observed the following limitations. At the preprocessing level, we used the server weblog. The integration of weblog files was not performed. One reason is the availability of such a weblogs that is from the multiple web server of the same website such as Google, Facebook, YouTube, and Amazon etc. Integration technique at preprocessing assembles the different weblogs of the same period and can deliver optimized results. Secondly, the most of the available weblogs are incomplete due to cache issues. During the user traversing, some of the web pages traversed have missing links at server weblog. Number of techniques are available at preprocessing to complete the web server weblogs. In this research, the issues of missing links were not handled. Web session similarity is an important segment for the identification of similar groups from weblog and in this dissertation, one of the major contributions is the introduction of web session similarity measure ST_Index. The proposed measures work based on common, uncommon and time factor attributes of the weblog. However, we did not cater the sequence of web pages traversed by a user in a session and time spent by each user on traversed web pages. The research focus was to identify the users or group of users with similar traversing behavior rather than the computation of single web pages importance. Such type of heuristics can be obtained through the number of hit count of various web pages.

6.4 Research Recommendations and Claims

The web usage mining is playing active role to disseminate the proper and accurate information to the web users. In this regard, our proposed framework F_MET is offering the best solution to the sessionization issue. In following section, we are presenting claims and recommendations. The weblog preprocessing is an important and mandatory phase of web usage mining and without the proper weblog preprocessing, the noise cannot be removed, and weblog cannot be used in WebKDD process. Consequently, the quality sessionization cannot be achieved. Moreover, the use of preprocessing tools are also failed to deliver noise free preprocessing due to the dynamic nature of websites and advancements of web designing technology. One of the weblog dataset was shared by [31] and weblog dataset was prepared by a tool WumPrep 2003. The WumPrep removed 63% irrelevant entries from weblog, while our proposed preprocessing methodology removed 77% irrelevant entries from the same dataset. Most of the researchers claim that around 80% irrelevant entries exists in a raw weblog. The noise free weblog data ensures the accurate, valid and correct results from the rest of web usage mining process.

The session generation with high coverage and precision is a crucial step in weblog sessionization process due to proxy server and firewall. However, the proper use of heuristics approach based on the weblog attributes (IP, Time Date, User Agent, User OS, Referrer Page) is useful for session generation. The experiment was performed with different weblog attributes to generate the web session on all the three available datasets. Dimitrijevic et al. [31] generated the 436 sessions through WumPrep tool while we adopted the Chitraa and Thanamani [141] technique to overcome the proxy server and firewall issues we obtained the 284 sessions. One reason for generating lesser number of sessions as compared to 436 is proper data filtering. Mostly the session construction is based on IP attribute only. This heuristics leads to the poor coverage. Whereas the precision is achieved through the proper data filtering techniques. The high coverage of session construction is shown in Table 4.1.

In addition to that the identification of real and convincing relationship (similarity) among the sessions further enhances the granularity of precision and coverage parameters. In this regard, we proposed a web session similarity measure ST_Index based on the common and uncommon web pages among the sessions along with the time shared factor. The notion of uncommon web pages do affect the session similarity measure [17] and in our proposed scheme, we enhanced this argument.

We prepared the test bad that supported us to test the various web session similarity measures such as Euclidean, Cosine, Jaccard etc. The coverage parameter out performed the other measures as the notion of uncommon web pages narrow down the similarity and enhanced the accuracy as all the users are traversing the same websites with different motives. All the other tested measures compute the similarity on the basis of number of web pages and common web pages. The proposed ST_Index, performed better on all the accuracy measures such as Precision, Recall, F-Measures and Accuracy. The ST_Index improved the results overall 5 to 10 % as shown Table 4.10.

Given the dynamic nature of the web, the F_MET is delivering the accurate, correct, and quality trends. These trends are used for the analysis of user behavior which are helpful in many ways such as customer relationship management; personalization; recommended system; and website management. In Chapter 7, Table 5.3, we computed the web business rules (Trends) from weblog such as $47 \rightarrow 109 \rightarrow 113 \rightarrow 438 \rightarrow 467$, This rule shows that Session 47 and Session 109 are the most relevant and have high degree of closeness as compare to the rest of weblog sessions. Session 109 is closely related to Session 113, and so on up to level 5. This rule shows that users are interested in Training, that is being offered by the university. The other interesting aspect of this rule is the different paths followed by the users to reach at the "Training link". These users are following the the training trend. Different rules deliver different user trends.

These trends can be utilized in various web applications. These rules are helpful in most of the web usage mining applications. The websites can be managed and traced that web pages traversed rarely and can be removed from the website. Similarly, the frequently traversed web pages can be optimized both contents and resources wise. The rules can be utilized to gain the business edge as well. The customers relationship management is helpful in many ways such as business promotions, product analysis. Moreover, the web personalization can also be achieved through these rules. The online purchasers can be offered the similar trends. The online theft and fraudulent activities can also be gauged by expanding this research further in future.

6.5 Future Guidelines

Web usage mining is an active research area and researchers in this area are carrying out a lot of research. The objective of the researchers is to deliver the practical solutions and improvement of information retrieval systems. This dissertation is also an effort to produce the practical solution of web sessionization issue. However, the improvements are never ending task in research and following are the few future guidelines that would be helpful to further enhance the research.

Preprocessing is a mandatory and crucial phase of web usage mining. Without the proper and accurate preprocessing, the web usage mining process cannot deliver the promising and optimized results. In the future, the weblog integration may be performed for the identification of more precise results. To complete the broken links of weblog will be done through path completion techniques. In this context, the website structure will be used. The analysis of available weblog will also be helpful to recover the broken links of the website. Both the techniques will resolve the path completion to manage the cache issue. The introduction of such type of techniques at preprocessing level will equip us to capture the strong results overall in the WebKDD process. The introduction of sequence based web session similarity measure along with the time spent on each page by the user will be helpful to study the user behavior more closely. The accurate computation of true relation among the sessions (users) will strengthen the web usage mining techniques such as clustering; classification; and outlier detection. In ST_Index, we are applying the common and uncommon web page artifacts along with the minimum time-shared between the sessions.

Bibliography

[1] G. Xu, "Web mining techniques for recommendation and personalization," *Ph.D Thesis 2008: The School of Computer Science & Mathematics, Faculty of Health, Engineering & Science, Victoria University, Australia*, 2008.

[2] D. Pai, A. Sharang, M. M. Yadagiri, and S. Agrawal, "Modelling visit similarity using click-stream data: A supervised approach," in *International Conference on Web Information Systems Engineering-WISE 2014*. Springer, 2014, pp. 135–145.

[3] M. Mohammadpourzarandi and R. Tamimi, "The application of web usage mining in e commerce security," *International Journal of Information Science and Management (IJISM)*, pp. 17–23, 2013.

[4] T. Hussain and S. Asghar, "Evaluation of similarity measures for categorical data," *The Nucleus, The journal Pakistan Atomic Energy Commission, Pakistan*, vol. 50, no. 4, pp. 387–394, 2013.

[5] B. Hawwash and O. Nasraoui, "Mining and tracking evolving web user trends from large web server logs," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 3, no. 2, pp. 106–125, 2010.

[6] A. Jha, M. Dave, and S. Madan, "A review on the study and analysis of big data using data mining techniques," *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol. 6, no. 3, pp. 94–102, 2016.

[7] A. Mohebi, S. Aghabozorgi, T. Ying Wah, T. Herawan, and R. Yahyapour, "Iterative big data clustering algorithms: a review," *Software: Practice and Experience, Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2341*, vol. 46, no. 1, pp. 107–129, 2016.

[8] F. M. H. Fernandez and R. Ponnusamy, "Data preprocessing and cleansing in web log on ontology for enhanced decision making," *Indian Journal of Science and Technology*, vol. 9, no. 10, pp. 1–10, 2016.

[9] K. Sharma, G. Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on.* IEEE, 2011, pp. 399–403.

[10] C. Lin and T. Hong, "A survey of fuzzy web mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 3, pp. 190–199, 2013.

[11] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, "A web usage mining framework for mining evolving user profiles in dynamic web sites," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 202–215, 2008.

[12] T. Hussain and S. Asghar, "Chi-square based hierarchical agglomerative clustering for web sessionization," *Journal of the National Science Foundation of Sri Lanka*, vol. 44, no. 2, pp. 211–222, 2016.

[13] E. Manohar and D. S. Punithavathani, "Effective preprocessing and knowledge discovery in web usage mining," *Middle-East Journal of Scientific Research*, vol. 23, no. 10, pp. 2433–2439, 2015.

[14] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, "Web usage mining as a tool for personalization: A survey," *User modeling and user-adapted interaction, Kluwer Academic Publisher, Netherlands*, vol. 13, no. 4, pp. 311–372, 2003.

[15] E. Saka and O. Nasraoui, “A recommender system based on the collaborative behavior of bird flocks,” in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on.* IEEE, 2010, pp. 1–10.

[16] M. Chaurasia and C. Satsangi, “Enhancing proxy server cache management using log analysis and recommendations,” *International Journal of Computer Applications*, vol. 113, no. 2, pp. 9–14, 2015.

[17] L. Chen, S. S. Bhowmick, and W. Nejdl, “Cowes: Web user clustering based on evolutionary web sessions,” *Data and Knowledge Engineering, Elsevier*, vol. 68, no. 10, pp. 867–885, 2009.

[18] M. Aldekhail, “Application and significance of web usage mining in the 21st century: a literature review,” *International Journal of Computer Theory and Engineering, IACSIT Press*, vol. 8, no. 1, pp. 41–47, 2016.

[19] M. A. Bayir and I. H. Toroslu, “Link based session reconstruction: finding all maximal paths,” *International Journal on Advanced Science Engineering Information Technology,arXiv preprint arXiv:1307.1927*, vol. 1, no. 1, pp. 1–3, 2013.

[20] M. A. Bayir, I. H. Toroslu, A. Cosar, and G. Fidan, “Discovering more accurate frequent web usage patterns,” *International Journal on Advanced Science Engineering Information Technology,arXiv preprint arXiv:0804.1409*, pp. 1–19, 2008.

[21] M. d. Kunder, “www.worldwidewebsize.com,” *The size of the World Wide Web (The Internet), Accessed on: December 8, 2016*, 2016.

[22] O. Nasraoui, C. Rojas, and C. Cardona, “A framework for mining evolving trends in web data streams using dynamic learning and retrospective validation,” *Journal of Computer Networks, Elsevier*, vol. 50, no. 10, pp. 1488–1512, 2006.

[23] S. Alam, G. Dobbie, and P. Riddle, “Particle swarm optimization based clustering of web usage data,” in *Proceedings of the 2008 IEEE/WIC/ACM*

International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. IEEE, 2008, pp. 451–454.

[24] I. Mele, "Web usage mining and its applications to web search and recommendation," *Ph.D Thesis, Sapienza University of Rome, 2014*, 2014.

[25] Wikipedia, "History of the world wide web - wikipedia, the free encyclopedia," *https://en.wikipedia.org/w/index.php?title=Historyof the World Wide Web = 773665610, Accessed on: January 8,2017*, 2017.

[26] G. De Francisci Morales, "Big data and the web: algorithms for data intensive scalable computing," *PhD Thesis, IMT Institute for Advanced Studies, Lucca,Italy*, 2012.

[27] V. A. F. A. Daniel A. Menasc, "Customer behavior models," in *Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning. 2000.* Prentice Hall, 2000, ch. 2, pp. 41–53.

[28] S. Alam, G. Dobbie, Y. S. Koh, and P. Riddle, "Web usage mining based recommender systems using implicit heterogeneous data," *Web Intelligence and Agent Systems: An International Journal*, vol. 12, no. 4, pp. 389–409, 2014.

[29] D. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method," *Journal of Applied Computing and Informatics*, vol. 12, no. 1, pp. 90–108, 2016.

[30] L. Chen, S. S. Bhowmick, and L.-T. Chia, "Fracture mining: mining frequently and concurrently mutating structures from historical xml documents," *Data and Knowledge Engineering, Elsevier*, vol. 59, no. 2, pp. 320–347, 2006.

[31] M. Dimitrijevic, Z. Bosnjak, and S. Subotica, "Discovering interesting association rules in the web log usage data," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 5, no. 1, pp. 191–207, 2010.

[32] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PLoS ONE, https://doi.org/10.1371/journal.pone.0144059*, vol. 10, no. 12, pp. 1–20, 2015.

[33] J. Vellingiri, S. Kaliraj, S. Satheeshkumar, and T. Parthiban, "A novel approach for user navigation pattern discovery and analysis for web usage mining," *Journal of Computer Science*, vol. 11, no. 2, pp. 372–382, 2015.

[34] A. Ahmad and S. Hashmi, "K-harmonic means type clustering algorithm for mixed datasets," *Journal of Applied Soft Computing,Elsevier*, vol. 48, no. 2016, pp. 39–49, 2016.

[35] P. Suthar and B. Oza, "A survey of web usage mining techniques," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5073–5076, 2015.

[36] Z. Ansari, M. F. Azeem, A. V. Babu, and W. Ahmed, "A fuzzy clustering based approach for mining usage profiles from web log data," *International Journal on Advanced Science Engineering Information Technology*, vol. 9, no. 6, pp. 70–79, 2015.

[37] S. Garcia, J. Luengo, and F. Herrera, "Data preprocessing in data mining," *Springer Book Series (Intelligent Systems Reference Library), DOI:10.1007/978-3-319-10247-4, ISBN: 9783319102474 (online)*, 2015.

[38] C. Halatsis and D. Petrilis, *Combining SOMs and Ontologies for Effective Web Site Mining.* Book Title (Self Organizing Maps-Applications and Novel Algorithm Design), Publisher: INTECH Open Access, 2011.

[39] P. Kherwa and J. Nigam, "Data preprocessing: A milestone of web usage mining," *International Journal Of Engineering Science And Innovative Technology (IJESIT)*, vol. 4, no. 2, pp. 281–289, 2015.

[40] E. Saka, "Swarm intelligence for clustering dynamic data sets for web usage mining and personalization," *Ph.D Thesis 2011, The University of Louisville, USA*, 2011.

[41] S. Alam, G. Dobbie, Y. S. Koh, and P. Riddle, "Clustering heterogeneous web usage data using hierarchical particle swarm optimization," in *Swarm Intelligence (SIS), 2013 IEEE Symposium on.* IEEE, 2013, pp. 147–154.

[42] P. Loyola, P. E. Roman, and J. D. Velasquez, "Predicting web user behavior using learning-based ant colony optimization," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 5, pp. 889–897, 2012.

[43] D. Arotaritei and S. Mitra, "Web mining: a survey in the fuzzy framework," *Fuzzy Sets and Systems,Elsevier*, vol. 148, no. 1, pp. 5–19, 2004.

[44] T. Hussain, S. Asghar, and N. Masood, "Web usage mining: A survey on preprocessing of web log file," in *Information and Emerging Technologies (ICIET), 2010, IEEE International Conference on.* IEEE, 2010, pp. 1–6.

[45] J. Ranjan, D. Goyal, and S. Ahson, "Data mining techniques for better decisions in human resource management systems," *International Journal of Business Information Systems*, vol. 3, no. 5, pp. 464–481, 2008.

[46] U. Fayyad, G. Piatetsky Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine, www.aaai.org*, vol. 17, no. 3, pp. 37–43, 1996.

[47] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth *et al.*, "Knowledge discovery and data mining: Towards a unifying framework." in *Proceedings Knowledge Discovery in Databases (KDD-96), AAAI (www.aaai.org).* KDD-96, 1996, pp. 82–88.

[48] F. Gullo, "From patterns in data to knowledge discovery: What data mining can do," *Physics Procedia, Elsevier*, vol. 62, pp. 18–22, 2015.

[49] J. Han, M. Kamber, and J. Pei, "Mining frequent patterns, associations, and correlations," *Book Chapter: Data Mining: Concepts and Techniques. 2nd Ed, San Francisco, USA: Morgan Kaufmann Publishers*, pp. 227–283, 2006.

[50] G. Raju, P. Satyanarayana, and L. Patnaik, "Knowledge discovery from web usage data: Extraction and applications of sequential and clustering

patterns–a survey," *International Journal of Innovative Computing, Information and Control*, vol. 4, no. 2, pp. 381–389, 2008.

[51] O. Nasraoui and C. Petenes, "Combining web usage mining and fuzzy inference for website personalization," in *Proceedings of the Fifth WEBKDD workshop; Webmining as a Premise to Effective and Intelligent Web Applications (WebKDD 2003), in conjunction with ACM SIGKDD conference, Washington, DC, USA.*, 2003, pp. 37–46.

[52] P. E. Roman, R. F. Dell, J. D. Velasquez, and P. S. Loyola, "Identifying user sessions from web server logs with integer programming," *Intelligent Data Analysis, IOS Press Content Library*, vol. 18, no. 1, pp. 43–61, 2014.

[53] G. M. Thomaz, A. A. Biz, E. M. Bettoni, L. Mendes-Filho, and D. Buhalis, "Content mining framework in social media: A fifa world cup 2014 case analysis," *Information & Management*, vol. 54, no. 6, pp. 786–801, 2017.

[54] F. Johnson and S. K. Gupta, "Web content mining techniques: a survey," *International Journal of Computer Applications*, vol. 47, no. 11, pp. 44–50, 2012.

[55] S. Fernando, S. Perera *et al.*, "Empirical analysis of data mining techniques for social network websites," *Compusoft*, vol. 3, no. 2, pp. 582–592, 2014.

[56] A. Kumar and R. K. Singh, "A study on web structure mining," *International Research Journal of Engineering and Technology (IRJET)*, vol. 04, no. 1, pp. 715–720, 2017.

[57] Z. Abdullah and A. Hamdan, "Comparative study of universities? web structure mining," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 10, pp. 2237–2243, 2015.

[58] P. Kumari, P. Ranout, A. Sharma, and P. Sharma, "Web mining-concept, classification and major research issues: A review," *Asian J. Adv. Basic Sci*, vol. 4, no. 2, pp. 41–44, 2016.

[59] N. Kaur and H. Aggarwal, "Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement," *International Journal of Information Technology*, vol. 10, no. 1, pp. 99–110, 2018.

[60] 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*, vol. 26, pp. 1–37, 2017.

[61] P. Lopes and B. Roy, "Dynamic recommendation system using web usage mining for e-commerce users," *Procedia Computer Science*, vol. 45, pp. 60–69, 2015.

[62] K. Suneetha and R. Krishnamoorthi, "Classification of web log data to identify interested users using decision trees," in *Ubiquitous Computing and Communication Journal Special Issue for The International Conference on Computing, Communications and Information Technology Applications (CCITA-2010).* IEEE, 2009, pp. 1–7.

[63] M. H. A. Wahab, M. N. H. Mohd, H. F. Hanafi, and M. F. M. Mohsin, "Data pre-processing on web server logs for generalized association rules mining algorithm," *World Academy of Science, Engineering and Technology*, vol. 48, no. 1, pp. 190–197, 2008.

[64] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[65] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, pp. 120–127, 2012.

[66] V. Dixit and S. Bhatia, "Refinement of clusters based on dissimilarity measures," *Int J Multidiscip Res Adv Eng (IJMRAE)*, vol. 6, no. 1, pp. 33–54, 2014.

[67] K. Kundra, U. Kaur, and D. Singh, "Efficient web log mining and navigational prediction with ehpso and scaled markov model," in *Computational*

Intelligence in Data Mining-Volume 3. Springer, New Delhi, 2015, pp. 529–543.

[68] D. Kavyasrujana and B. C. Rao, "Hierarchical clustering for sentence extraction using cosine similarity measure," in *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1.* Springer, 2015, pp. 185–191.

[69] P. Kumar, B. S. Raju, and P. R. Krishna, "A new similarity metric for sequential data," *Exploring Advances in Interdisciplinary Data Mining and Analytics: New Trends: New Trends*, pp. 16–32, 2011.

[70] W. Wang and O. R. Zaane, "Clustering web sessions by sequence alignment," *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pp. 394–398, 2002.

[71] S. Kumbharkar and Y. Gurav, "Canberra distance based web structure analysis for improving user navigation," *IJRIT International Journal of Research in Information Technology*, vol. 2, no. 8, pp. 16–32, 2011.

[72] T. Hussain, S. Asghar, and N. Masood, "Hierarchical sessionization at preprocessing level of wum based on swarm intelligence," in *Emerging Technologies (ICET), 2010 6th International IEEE Conference on.* IEEE, 2010, pp. 21–26.

[73] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *European Semantic Web Conference.* Springer, 2007, pp. 624–639.

[74] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *Journal of ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[75] X. Yu, M. Li, K. A. Kim, J. Chung, and K. H. Ryu, "Emerging pattern-based clustering of web users utilizing a simple page-linked graph," *Sustainability,doi:10.3390/su8030239*, vol. 8, no. 3, pp. 1–18, 2016.

[76] Z. A. Ansari, S. A. Sattar, and A. V. Babu, "A fuzzy neural network based framework to discover user access patterns from web log data," *Advances in Data Analysis and Classification, Springer*, vol. 11, no. 38, pp. 1–28, 2015.

[77] R. Forsati, A. Keikha, and M. Shamsfard, "An improved bee colony optimization algorithm with an application to document clustering," *Neurocomputing, Elsevier*, vol. 159, no. 1, pp. 9–26, 2015.

[78] P. Berkhin, "A survey of clustering data mining techniques," *Grouping multidimensional data*, pp. 25–71, 2006.

[79] S. Mehrotra and S. Kohli, *Application of Clustering for Improving Search Result of a Website.* Springer, 2016, pp. 349–356.

[80] R. Kaur and S. Kaur, "A review: Techniques for clustering of web usage mining," *International Journal of Science and Research (IJSR) ISSN (Online)*, pp. 2319–7064, 2012.

[81] N. Huidrom and N. Bagoria, "Clustering techniques for the identification of web user session," *International Journal of Scientific and Research Publications*, vol. 3, no. 1, pp. 1–4, 2013.

[82] Y. Hamasuna, Y. Endo, and S. Miyamoto, "On agglomerative hierarchical clustering using clusterwise tolerance based pairwise constraints," *JACIII*, vol. 16, no. 1, pp. 174–179, 2012.

[83] K. Katariya and R. Aluvalu, "Agglomerative clustering in web usage mining: A survey," *International Journal of Computer Applications*, vol. 89, no. 8, 2014.

[84] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Journal of Expert Systems with Applications*, vol. 42, no. 5, pp. 2785–2797, 2015.

[85] R. M. Stefan, "Cluster type methodologies for grouping data," *Procedia Economics and Finance*, vol. 15, no. 1, pp. 357–362, 2014.

[86] D. S. Sisodia, S. Verma, and O. P. Vyas, "Augmented intuitive dissimilarity metric for clustering of web user sessions," *Journal of Information Science,DOI: 10.1177/0165551516648259 jis.sagepub.com*, vol. 43, no. 4, pp. 1–12, 2016.

[87] D. A. Menascé and V. A. Almeida, "Challenges in scaling e-business sites," in *Int. CMG Conference*, 2000, pp. 329–336.

[88] M. H. A. Elhiber and A. Abraham, "Access patterns in web log data: A review," *Journal of Network and Innovative Computing*, vol. 1, no. 2013, pp. 348–355, 2013.

[89] T. Gopalakrishnann, P. Sengottuvelan, and A. Bharathi, "Two level clustering of web log files to enhance the quality of user data," *Int J AdvEngg Tech (April-June)*, vol. 7, no. 2, pp. 1138–1144, 2016.

[90] C. J. Carmona, S. Ramirez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. Garcia, "Web usage mining to improve the design of an e-commerce website: Orolivesur. com," *Expert Systems with Applications, Elsevier*, vol. 39, no. 12, pp. 11 243–11 249, 2012.

[91] A. Ferrara, L. Genta, and S. Montanelli, "Similarity recognition in the web of data," in *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014)*. Citeseer, 2014, pp. 263–268.

[92] R. Geng and J. Tian, "Improving web navigation usability by comparing actual and anticipated usage," *Human-Machine Systems, IEEE Transactions*, vol. 45, no. 1, pp. 84–94, 2015.

[93] D. Anupama and S. D. Gowda, "Clustering of web user sessions to maintain occurrence of sequence in navigation pattern," *Procedia Computer Science, Elsevier*, vol. 58, no. 2015, pp. 558–564, 2015.

[94] S. k. Jain and V. S.A.T.I., "Tree based techniques for web access patterns?a survey," *International Journal of Application or Innovation in Engineering and Management (IJAIEM)*, vol. 5, no. 2, pp. 25–28, 2016.

[95] X. Xu, J. Lu, and W. Wang, “Incremental hierarchical clustering of stochastic pattern-based symbolic data,” in *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.* Springer, 2016, pp. 156–167.

[96] D. S. Sisodia, S. Verma, and O. P. Vyas, “Performance evaluation of an augmented session dissimilarity matrix of web user sessions using relational fuzzy c-means clustering,” *International Journal of Applied Engineering Research*, vol. 11, no. 9, pp. 6497–6503, 2016.

[97] V.-T. Luu, G. Forestier, F. Fondement, and P.-A. Muller, “Web site audience segmentation using hybrid alignment techniques,” in *Book:Trends and Applications in Knowledge Discovery and Data Mining, DOI:10.1007/978-3-319-25660-3-3.* Springer, 2015, pp. 29–40.

[98] V. T. Luu, M. Ripken, G. Forestier, F. Fondement, and P. A. Muller, “Using glocal event alignment for comparing sequences of significantly different lengths,” in *Book: Machine Learning and Data Mining in Pattern Recognition, DOI:10.1007/978-3-319-41920-6-5,.* Springer, 2016, pp. 58–72.

[99] V. Dixit and S. K. Bhatia, “Refinement and evaluation of web session cluster quality,” *International Journal of System Assurance Engineering and Management*, vol. 6, no. 4, pp. 373–389, 2015.

[100] V. S. Dixit, S. K. Bhatia, and V. Singh, “Evaluation of web session cluster quality based on access-time dissimilarity and evolutionary algorithms,” in *International Conference on Computational Science and Its Applications,Springer.* Springer, 2014, pp. 297–310.

[101] R. Forsati, A. Moayedikia, and M. Shamsfard, “An effective web page recommender using binary data clustering,” *Information Retrieval Journal, Springer*, vol. 18, no. 3, pp. 167–214, 2015.

[102] Y. Han and K. Xia, “Data preprocessing method based on user characteristic of interests for web log mining,” in *Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2014 Fourth International Conference on.* IEEE, 2014, pp. 867–872.

[103] C. Li, "Research on web session clustering," *Journal of Software*, vol. 4, no. 5, pp. 460–468, 2009.

[104] K. Duraiswamy and V. V. Mayil, "Similarity matrix based session clustering by sequence alignment using dynamic programming," *Journal of Computer and Information Science*, vol. 1, no. 3, pp. 66–72, 2008.

[105] M. Kumar and M. Meenu, "A survey on pattern discovery of web usage mining," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 3, no. 1, pp. 379–385, 2017.

[106] M. R. Sundari, Y. Srinivas, and P. P. Reddy, "A review on pattern discovery techniques of web usage mining," *International Journal of Engineering Research and Applications*, vol. 4, no. 9, pp. 131–136, 2014.

[107] R. S. Rao and J. Arora, "A survey on methods used in web usage mining," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 5, pp. 2627–2631, 2017.

[108] G. Shivaprasad, N. S. Reddy, and U. D. Acharya, "Knowledge discovery from web usage data: An efficient implementation of web log preprocessing techniques," *International Journal of Computer Applications*, vol. 111, no. 13, pp. 27–32, 2015.

[109] R. Mishra, P. Kumar, and B. Bhasker, "A web recommendation system considering sequential information," *Journal of Decision Support Systems, Elsevier*, vol. 75, pp. 1–10, 2015.

[110] B. Kotiyal, A. Kumar, B. Pant, and R. Goudar, "Classification technique for improving user access on web log data," in *Intelligent Computing, Networking, and Informatics.* Springer, New Delhi, 2014, pp. 1089–1097.

[111] M. A. Bayir, I. H. Toroslu, and A. Cosar, "A new approach for reactive web usage data processing," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on.* IEEE, 2006, pp. 44–44.

[112] L. Ying', "Fuzzy-clustering web based on mining," *Journal of Multimedia,Scimago Journal & Country Rank (SJR), Finland*, vol. 9, no. 1, pp. 123–129, 2014.

[113] L. Ying, "Web user interest-based clustering method," *International Journal of Digital Content Technology and its Applications(JDCTA),doi:10.4156/jdcta.*, vol. 7, no. 6, pp. 337–345, 2013.

[114] M. A. Awad and I. Khalil, "Prediction of user's web-browsing behavior: Application of markov model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1131–1142, 2012.

[115] G. Shivaprasad, N. S. Reddy, U. D. Acharya, and P. K. Aithal, "Neuro-fuzzy based hybrid model for web usage mining," *Procedia Computer Science*, vol. 54, pp. 327–334, 2015.

[116] S. S. Patil and H. P. Khandagale, "Enhancing web navigation usability using web usage mining techniques," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 6, pp. 2828–2834, 2016.

[117] S. Malarvizhi and B. Sathiyabhama, "Frequent pagesets from web log by enhanced weighted association rule mining," *Cluster Computing*, vol. 19, no. 1, pp. 269–277, 2016.

[118] Y. Dong, Y. Zhuang, K. Chen, and X. Tai, "A hierarchical clustering algorithm based on fuzzy graph connectedness," *Fuzzy Sets and Systems,Elsevier*, vol. 157, no. 13, pp. 1760–1774, 2006.

[119] J. Ghorpade-Aher and R. Bagdiya, "A review on clustering web data using pso," *International Journal of Computer Applications*, vol. 108, no. 6, pp. 31–36, 2014.

[120] A. Chakraborty and S. Bandyopadhyay, "Clustering of web sessions by fogsaa," in *Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in.* IEEE, 2013, pp. 282–287.

[121] B. H. Kumar, L. Vibha, and K. Venugopal, “Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order markov model,” in *Region 10 Symposium (TENSYMP), 2016.* IEEE, 2016, pp. 1–6.

[122] D. S. Sisodia, S. Verma, and O. P. Vyas, “Quantitative evaluation of web user session dissimilarity measures using medoids based relational fuzzy clustering,” *Indian Journal of Science and Technology*, vol. 9, no. 28, pp. 1–9, 2016.

[123] M. Azimpour-Kivi and R. Azmi, “A webpage similarity measure for web sessions clustering using sequence alignment,” in *Artificial Intelligence and Signal Processing (AISP), 2011 International Symposium on.* IEEE, 2011, pp. 20–24.

[124] A. Bianco, G. Mardente, M. Mellia, M. Munafo, and L. Muscariello, “Web user session characterization via clustering techniques,” in *Global Telecommunications Conference, 2005. GLOBECOM’05. IEEE*, 2005, pp. 1102–1107.

[125] A. Bianco, G. Mardente, M. Mellia, M. Munafò, and L. Muscariello, “Web user-session inference by means of clustering techniques,” *IEEE/ACM Transactions On Networking*, vol. 17, no. 2, pp. 405–416, 2009.

[126] C. Nasa and Suman, “Evaluation of different classification techniques for web data,” *International Journal of Computer Applications*, vol. 52, no. 9, pp. 34–40, 2012.

[127] J. Chand, A. S. Chauhan, and A. K. Shrivastava, “Review on classification of web log data using cart algorithm,” *International Journal of Computer Applications*, vol. 80, no. 17, pp. 41–43, 2013.

[128] M. Udantha, S. Ranathunga, and G. Dias, “Modelling website user behaviors by combining the em and dbscan algorithms,” in *2016 Moratuwa Engineering Research Conference (MERCon, Engineering Research Unit (ERU) of the University of Moratuwa, Sri Lanka).* IEEE, 2016, pp. 168–173.

[129] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial bee colony (abc) algorithm," *Applied soft computing, Elsevier*, vol. 11, no. 1, pp. 652–657, 2011.

[130] H. Lu and T. T. S. Nguyen, "Experimental investigation of pso based web user session clustering," in *Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of.* IEEE, 2009, pp. 647–652.

[131] C. Dou and J. Lin, "Improved particle swarm optimization based on genetic algorithm," in *Software Engineering and Knowledge Engineering: Theory and Practice.* Springer, Berlin, Heidelberg, 2012, pp. 149–153.

[132] B. Rajdeepa and D. Sumatht, "Web structure minging for users based on a hyvrid ga/pso approach," *Journal of Theoretical and Applied Information Technology*, vol. 70, no. 3, pp. 573–578, 2014.

[133] M. Srivastava, R. Garg, and P. Mishra, "Analysis of data extraction and data cleaning in web usage mining," in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering and Technology (ICARCSET 2015).* ACM, 2015, pp. 13–18.

[134] V. Vidyapriya and V. Pushpa, "Identifying web users from weblogs using classification algorithms," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 7, pp. 13 722–13 728, 2016.

[135] N. Algiriyage, S. Jayasena, and G. Dias, "Web user profiling using hierarchical clustering with improved similarity measure," in *Moratuwa Engineering Research Conference (MERCon), Sri Lanka.* IEEE, 2015, pp. 295–300.

[136] Y. Peng, G. Kou, Y. Shi, and Z. Chen, "A descriptive framework for the field of data mining and knowledge discovery," *International Journal of Information Technology & Decision Making*, vol. 7, no. 04, pp. 639–682, 2008.

[137] R. Dueñas-Fernández, J. D. Velásquez, and G. L?Huillier, "Detecting trends on the web: A multidisciplinary approach," *Information Fusion, Elsevier*, vol. 20, no. 1, pp. 129–135, 2014.

[138] J. Gao, W. Fan, J. Han, and P. S. Yu, "A general framework for mining concept drifting data streams with skewed distributions," in *Proceedings of the 2007 SIAM International Conference on Data Mining.* SIAM, 2007, pp. 3–14.

[139] F. Roohi, "Neuro fuzzy approach to data clustering: A framework for analysis," *European Scientific Journal*, vol. 9, no. 9, pp. 183–192, 2013.

[140] P. Verma and N. Kesswani, "Web usage mining framework for data cleaning and ip address identification," *International Journal of advanced studies in Computer Science and Engineering IJASCSE*, vol. 3, no. 8, pp. 39–43, 2014.

[141] V. Chitraa and A. S. Thanamani, "A novel technique for sessions identification in web usage mining preprocessing," *International Journal of Computer Applications*, vol. 34, no. 9, pp. 23–27, 2011.

[142] Y. Peng, G. Kou, Y. Shi, and Z. Chen, "A descriptive framework for the field of data mining and knowledge discovery," *International Journal of Information Technology and Decision Making*, vol. 7, no. 04, pp. 639–682, 2008.

[143] M. W. Berry and M. Browne, *Book,Lecture notes in data mining.* World Scientific Publisher, ISBN: 978-981-4478-05-2 (ebook), 2006.

[144] S. Ramkumar, G. Emayavaramban, and A. Elakkiya, "Performance evaluation of mobile sensor network," *Journal of Applied Engineering (JOAE)*, vol. 2, no. 8, pp. 151–155, 2015.

[145] J. Mehra and R. Thakur, "An effective method for web log preprocessing and page access frequency using web usage mining," *International Journal of Applied Engineering Research*, vol. 13, no. 2, pp. 1227–1232, 2018.

[146] T. Pamutha, S. Chimphlee, C. Kimpan, and P. Sanguansat, "Data preprocessing on web server log files for mining users access patterns," *International Journal of Research and Reviews in Wireless Communications (IJRRWC)*, vol. 2, no. 2, pp. 92–98, 2012.

[147] B. Bakariya, K. K. Mohbey, and G. Thakur, "An inclusive survey on data preprocessing methods used in web usage mining," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012).* Springer, 2012, pp. 407–416.

[148] J.-C. Ou, C.-H. Lee, and M.-S. Chen, "Efficient algorithms for incremental web log mining with dynamic thresholds," *The VLDB Journal*, vol. 17, no. 4, pp. 827–845, 2008.

[149] T. Kroeger, "https://www.statista.com/topics/1145/internet-usage-worldwide," *Statista Inc. Accessed on: Jan 1, 2017*, 2008.

[150] N. Khasawneh and C.-C. Chan, "Active user-based and ontology-based web log data preprocessing for web usage mining," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on.* IEEE, 2006, pp. 325–328.

[151] Z. Pabarskaite, "Implementing advanced cleaning and end-user interpretability technologies in web log mining," in *Information Technology Interfaces, 2002. ITI 2002. Proceedings of the 24th International Conference on.* IEEE, 2002, pp. 109–113.

[152] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, pp. 12–23, 2000.

[153] A. Oliner, A. Ganapathi, and W. Xu, "Advances and challenges in log analysis," *Journal of Communications of the ACM*, vol. 55, no. 2, pp. 55–61, 2012.

[154] A. Sote and S. Pande, "Web page clustering using self-organizing map," *A Monthly Journal of Computer Science and Information Technology*, vol. 4, no. 1, pp. 78–84, 2015.

[155] H. Wang, C. Yang, and H. Zeng, "Design and implementation of a web usage mining model based on upgrowth and preflxspan," *Journal of Communications of the IIMA, available: http://scholarworks.lib.csusb.edu*, vol. 6, no. 2, pp. 68–84, 2015.

[156] R. Vaarandi, "A data clustering algorithm for mining patterns from event logs," in *Proceedings of the 3rd IEEE Workshop on IP Operations & Management (IPOM 2003) (IEEE Cat. No.03EX764).* IEEE, 2003, pp. 119–126.

[157] O. Ibrahimov, I. Sethi, and N. Dimitrova, "The performance analysis of a chi-square similarity measure for topic related clustering of noisy transcripts," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on.* IEEE, 2002, pp. 285–288.

[158] Y.-T. Chen and M. C. Chen, "Using chi-square statistics to measure similarities for text categorization," *Expert systems with applications, Elsevier*, vol. 38, no. 4, pp. 3085–3090, 2011.

[159] F. Hadzic and M. Hecker, "Alternative approach to tree-structured web log representation and mining," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01.* IEEE, 2011, pp. 235–242.

[160] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on.* IEEE, 1995, pp. 39–43.

[161] M. Jalali, N. Mustapha, M. N. Sulaiman, and A. Mamat, "Webpum: A web-based recommendation system to predict user future movements," *Expert Systems with Applications, Elsvier*, vol. 37, no. 9, pp. 6201–6212, 2010.