Session identification techniques used in web usage mining

A systematic mapping of scholarly literature

Bahjat Fatima, Huma Ramzan and Sohail Asghar
Department of Computer Science,
COMSATS Institute of Information Technology, Islamabad, Pakistan

Received 16 August 2015
Revised 25 March 2016
Accepted 12 May 2016

Abstract

Purpose – The purpose of this paper is to critically analyze the state-of-the-art session identification techniques used in web usage mining (WUM) process in terms of their limitations, features, and methodologies.

Design/methodology/approach – In this research, systematic literature review has been conducted using review protocol approach. The methodology consisted of a comprehensive search for relevant literature over the period of 2005-2015, using four online database repositories (i.e. IEEE, Springer, ACM Digital Library, and ScienceDirect).

Findings – The findings revealed that this research area is still immature and existing literature lacks the critical review of recent session identification techniques used in WUM process.

Originality/value – The contribution of this study is to provide a structured overview of the research developments, to critically review the existing session identification techniques, highlight their limitations and associated challenges and identify areas where further improvements are required so as to complement the performance of existing techniques.

Keywords Clustering, Pre-processing, Session identification, Sessionization, Web log file, Web usage mining

Paper type Literature review

Introduction

World Wide Web has experienced tremendous growth since its inception. It has evolved from a static content repository to a place where users can contribute, interact, and collaborate. Now-a-days, e-business has become a standard operating procedure for the vast majority of companies. Consequently, the structure of the web is rapidly evolving from a loose collection of websites into organized market places, through which millions of visitors pass daily. However, over time, the competition has intensified and now personalized services are considered as an integral part of e-business. Personalization is making it easier and more pleasant for users to surf the web and find what they want (Iváncsy and Vajk, 2006). The user's experience can be personalized by tailoring web pages to their individual preferences, offering services of their interest, and letting them bypass irrelevant content (Kumar and Rukmani, 2010; Wu *et al.*, 2015).

Recent advancements in technology have made it possible to record users' browsing behaviors through web log files (Stenmark, 2008; Jansen, 2006; Stenmark and Jadaan, 2007; Limam *et al.*, 2010). As shown in the Figure 1, web usage mining (WUM) aims at discovering interesting patterns from web usage data (Rahaman *et al.*, 2014; Siddiqui and Aljahdali, 2013). Due to large amount of irrelevant information present in the web log file, it cannot be directly used in the WUM process. Therefore, data pre-processing is regarded as an essential part to improve the efficiency and quality of the later stages of the WUM process (Domenech and Lorenzo, 2007).

Online Information Review
Vol. 40 No. 7, 2016
pp. 1033-1053
© Emerald Group Publishing Limited
1468-4527
DOI 10.1108/OIR-08-2015-0274

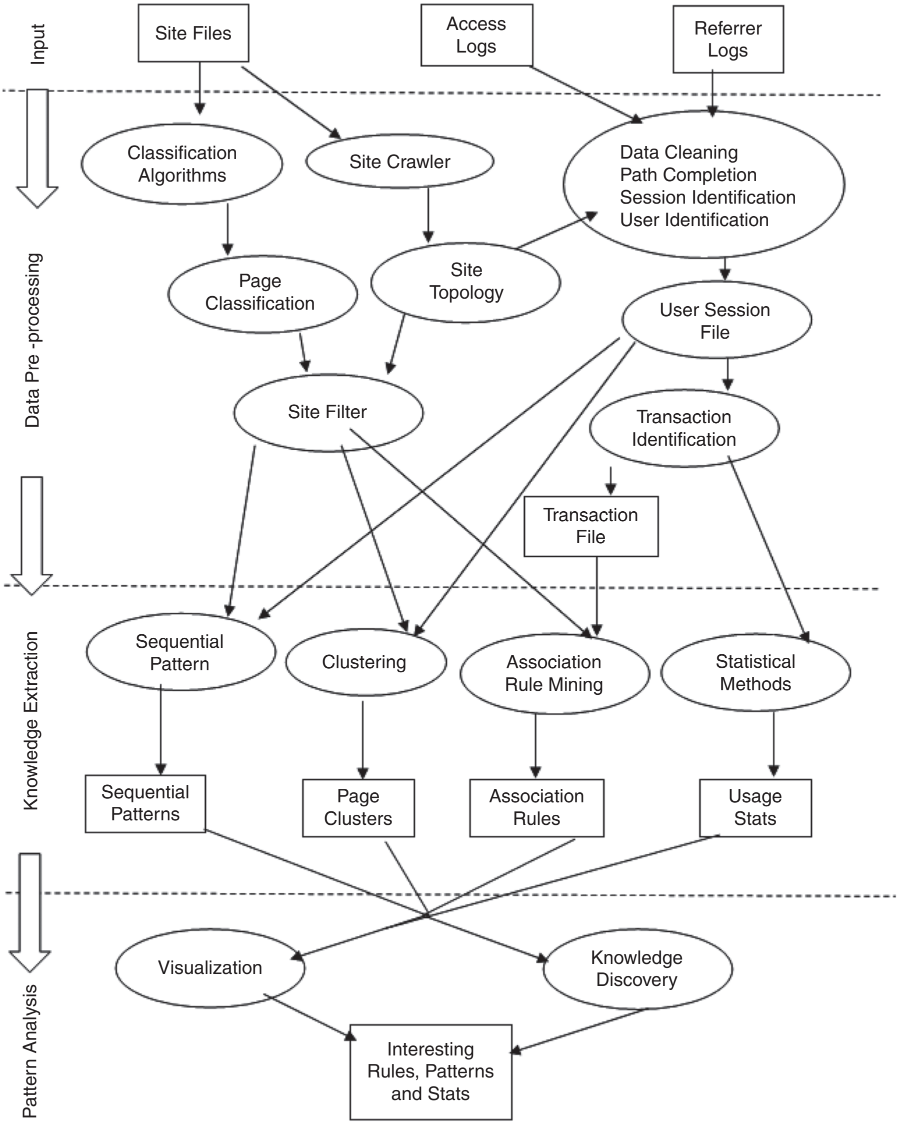
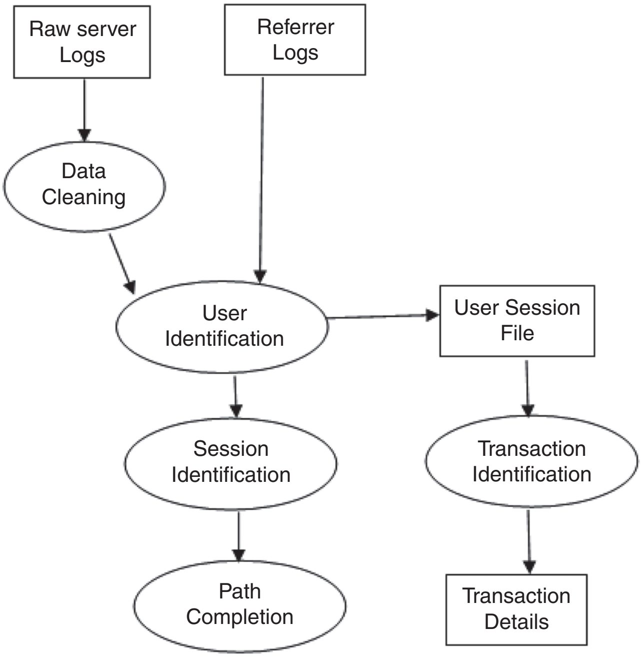


Figure 1.
Web usage mining (WUM) process

Source: Sharma (2008)

Data pre-processing includes data cleaning, user identification, session identification, path completion, and transaction identification. As shown in Figure 2, the goal of the data pre-processing step is to transform the raw web log data into a set of user profiles. Each such profile captures a sequence or a set of URLs representing a user session.

Session identification is considered as an important yet most difficult step and if not handled properly will directly affect the subsequent stages of WUM process. Session identification is the process of segmenting the user activity log of each user into



Source: Sharma (2008)

Figure 2.
Data Pre-processing phase

sessions, each representing a single visit to the site (Asadianfam and Mohammadi, 2014). The goal of a sessionization heuristic is to reconstruct, from the clickstream data, the actual sequence of actions performed by one user during one visit to the site (Gopalakrishnan *et al.*, 2014).

In the recent years, several techniques for session identification have been proposed by the researchers (Sindhuja *et al.*, 2014). However, the existing literature lacks the critical review of recent session identification techniques used in WUM process. In this paper, we will critically review the state-of-the-art session identification techniques used in WUM process in terms of their features, methodologies, performance, and limitations. The study intends to provide a summarized view of the research work done in the domain of session identification, highlight their limitations and associated challenges, and identify areas where further improvements are required so as to complement the performance of existing techniques.

The rest of the paper is organized as follows. Next section describes the research method used in this review. The section after that discusses the existing session identification techniques used in WUM process, followed by the critical evaluation of the techniques. The subsequent section includes discussion. The section after that contains the summary and findings, and the final section concludes the paper.

Research method

In this systematic literature review, we have used the review protocol approach (Brereton *et al.*, 2007) with some modifications, which has been shown in Figure 3.

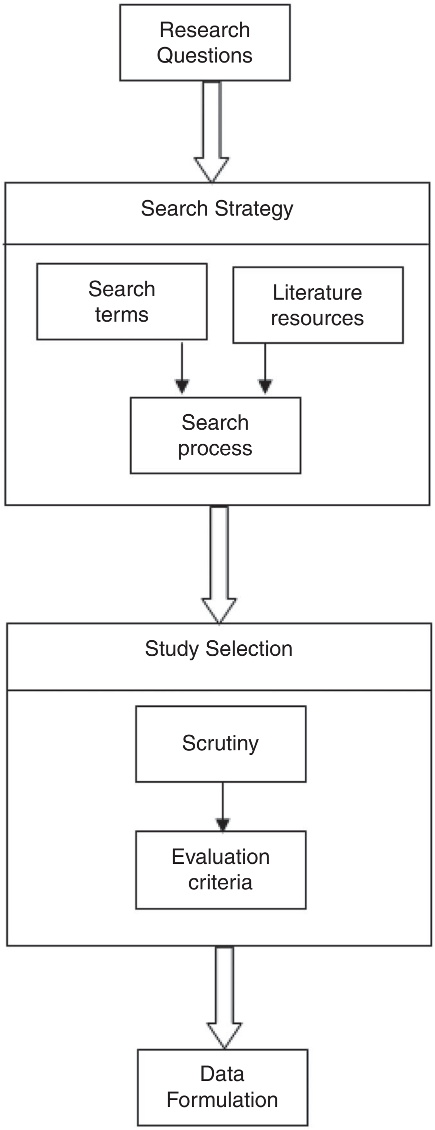


Figure 3.
Review protocol approach

Source: Brereton *et al*. (2007)

Research questions

To achieve our objective, the research questions were formulated as follows:

RQ1. What are the descriptions (e.g. underlying methodology, working, and results) of existing session identification techniques?

RQ2. What are the research contributions of existing session identification techniques?

RQ3. What are the limitations of existing session identification techniques?

Search strategy

Search terms. Keywords from relevant papers were used as search terms, mainly – Clustering, Data Pre-processing, Session identification, Sessionization, Web log file, and Web usage mining.

Literature resources. In this study, four online database repositories (i.e. IEEE, Springer, ACM Digital Library, and ScienceDirect) and Google were used to conduct searches for published journal papers, and conference proceedings.

Search process. In the first step, a thorough search was conducted using in the online resources and a set of 35 prospective papers was obtained. In the second step, additional three relevant papers were identified with the help of reference lists of the previous papers (identified in the first step).

Study selection

Scrutiny. In order to eliminate duplicate and irrelevant papers, the 42 prospective papers gathered from the last step were scrutinized. Only the papers published during the period 2005-2015 were kept.

Evaluation criteria. Finally, the remaining 33 prospective papers were evaluated to check whether they are capable to answer our research questions. As a result, 29 papers were selected.

Data formulation

The contents of 29 selected papers were further analyzed in detail to obtain the answers to our research questions.

Literature review of existing session identification techniques

In this section, we will discuss and analyze different session identification techniques used in WUM process.

One approach for session identification is to use time-oriented heuristics to reconstruct user sessions. In 2006, Zhou *et al* proposed a time-oriented algorithm which is based on a time threshold value for user session identification. Each user request in a session is marked with a URL and the requested time. The threshold time is set for 30 min for each user session after which a new session is created. The difference between the request time of the successor and current entry determines the duration of a URL. If there is no successor URL, the duration is determined by finding an average of the current session duration. This technique helps predicting the otherwise costly process of user navigation patterns to be done in advance.

Castellano *et al.* (2007) developed a tool for the pre-processing of the web log file called log data pre-processor (LODAP) that extracts user sessions from the requests stored in the log file of a website. LODAP performs complete data pre-processing, including data cleaning, data filtering, and session identification along with the summary of each step of the processor. By using a time-based method, data structuration module groups the remaining requests as user sessions. Experiment results revealed the effectiveness of the LODAP tool in identifying significant user sessions.

Sisodia *et al.* (2015) evaluated the performance of ensemble-based learners in order to differentiate web robot sessions from human sessions. Web server logs contains user visit entries in sequential order and after some prepressing on this log data, session identification algorithm is used to divide multiuser visits into individual sessions based on time threshold heuristic. Sessionization log analyzer extracts relevant features from the identified sessions and ensemble classifier labels them as robot or human session by comparing the session IP. The authors evaluated these ensemble learners using F1 measure, recall, and precision. However, by using a combination of certain feature extraction techniques performance of these robot detections could be significantly improved.

Contrary to the previous approaches, Peng and Zhao (2010) argued that session identification algorithms based on uniform threshold values do not consider the different user behaviors, i.e. time spent by different users on a particular web page is not the same. Consequently, during the session identification, there are good chances that a long session can be falsely divided into the next session. To address this problem, authors suggested that for session identification an algorithm based on average threshold value should be used rather than uniform threshold value. The proposed scheme is a two-step average threshold-based algorithm. In the first step, average threshold value is calculated and adjusted dynamically when a web page requests records to add in the current session. In this way, only the requests with relatively large interval time need to be calculated. In the second step, previously identified sessions are rechecked to remove the errors. Experimental evaluation revealed the effectiveness of the proposed algorithm in identifying long sessions with increased accuracy.

Similar to the previous approach, Xinhua and Qiong (2011) presented a dynamic time-out based session identification algorithm. At the start of the algorithm, the time-out is determined for an individual web page using the statistical results, in combination with the degree of importance of a page. Once the procedure of session identification starts, time-out is dynamically altered. The results of the experiments showed that the proposed algorithm can perform better than traditional time-based algorithms used for session identification.

Dinuca and Ciobanu (2012) identified the potential errors caused by using fixed values based algorithms and proposed a new method to improve the process of session identification. The modified algorithm uses the average time of visiting web pages to indicate the end of a session. Visit time is calculated for each page visited by the user by using the consecutive timestamp values. Average visit duration of a particular page is the average of all the times spent on that page. Both algorithms, classic and modified were implemented in Java. Experimental results showed that both algorithms have the same running time, but the modified algorithm has a greater number of sessions than classic algorithm.

Chitraa and Thanamani (2011) presented an optimal algorithm that can be used to convert web log entries into user sessions without considering the site topology. The authors also proposed a new session identification technique in which sessions are identified by constructing a matrix. In the matrix, web pages are represented as columns and users along with their sessions are represented as rows. By using the user's traversal data, browsing time and weights are calculated. These weights are stored in matrix cells and each row is considered as a user's individual session traversal. If the weight is 100, the next value is stored in the next row as a new session. The scheme was implemented in Matlab and experimental evaluation revealed the effectiveness of the proposed sessionization technique in constructing sessions accurately.

In 2013, Chitra combined time-oriented heuristics with graphs and proposed a new session reconstruction algorithm which is based on the timestamp of the visited pages.

The method called MAG creates a graph containing vertices. These nodes represent the visited web pages with calculated weights and the links represent the navigation between the different web pages. Associated weights reflect the page's importance. This could be done using the user ID of the visiting users who have traversed through the different web pages using page navigations provided on a website. Session construction is done by traversing through this graph. A MAG structure will be built per user session after weighting each web page according to browsing time. Graph traversals model the user navigation activities on the website. From the experimental results, it is obvious that the proposed method helps in identifying the session in short span of time with extra accuracy.

As an alternate for time-oriented heuristics, Robert *et al.* (Roman *et al.*, 2008) proposed a novel algorithm for session reconstruction using integer programming. This technique groups the requests coming from the same IP address using log registers. But unlike heuristics, it constructs all the user sessions at the same time. Each constructed session represents an ordered list of log registers; integer program uses a binary variable that has value 1 if log register is assigned to it and 0 otherwise provided each register can be used once in a session. This technique constructed better sessions. The time-based heuristics, on the other hand, are substantially faster but less accurate as compared to integer programming.

Another approach for session identification is to use structure-oriented heuristic to reconstruct the user sessions. Based on the fact that user and session identification are two major steps in data pre-processing for WUM, Khasawneh and Chan (2006) presented a new pre-processing technique. The proposed scheme includes data cleaning, user identification, and session identification. Users are identified by using a fast active user-based identification algorithm with time complexity O(n). To do so, the algorithm uses a set of records visited by a user during that period of time, IP address, date, and time of visit. An ontology-based algorithm is utilized for session identification which is based on the features extracted from the pages and structure of a website.

Heydari *et al.* (2009) presented a combination of the statistical analysis and graphical method for session reconstruction, which is an influential way to evaluate website usage. This method considers the client side which was not taken into account previously. The algorithm reconstructs user session by applying statistical analysis, using browsing time and graph method for mapping user navigation. The website's structure is divided into weighted vertices representing web pages and user navigation path as traversal on it, edge represents link between web pages, and weight of node represents browsing time of web page.

Arumugam and Suguna (2009) proposed an optimal technique for user session identification using two-way hashed tables to correctly map the user navigational history. The motivation behind their work is the fact that traversing the tree structure for backward references is a very time-consuming task. When backward references are made, the proposed algorithm generates a complete path based on access history list (AHL) within optimal time. AHL stores the identified user sessions. It uses the IP address, intersession and intrasession timeouts, direct link between referred pages, and backward reference analysis without searching the whole tree representing the server pages. Server web logs are represented by an array list and server pages are represented by hash table. To represent user-accessed page sequences, a two-way hashed data structure is used to store AHL. Primary hash table stores user sessions and contain the references to the secondary hash table which actually hold the complete navigation path. In the AHL, only the visited pages are stored, which saves searching time. Whenever a reference is made, a page is directly searched from the history list. If it is not available, a new session is started. Pages that are referred from other server can directly be accessed rather than starting the search from the root server. The experimental results have shown correct session identification within an optimal time than maximal forward and reference length methods.

Yuankang *et al.* (Fang and Huang, 2010) presented a session identification algorithm, which uses frame page and page access threshold for identifying individual user sessions. The algorithm starts with the user identification after which frames are filtered for each web page. According to the content of the page the page threshold is set and all web structure and user's session sets were identified by this threshold. In the first step of the proposed technique, the users are identified and the frame page is filtered. Next the contents of multiple pages are combined into an overall web structure forming the actual user session. In comparison to fixed threshold traditional methods, the proposed algorithm can identify longer sessions. This technique produces more accurate and effective sessions than the traditional techniques for session identification.

Ting *et al.* (Kimble and Kudenko, 2005) proposed a novel WUM approach called UBB mining, which detects the patterns that deviate from the common usage patterns. UBB mining is based on two algorithms; one is the segmentation in which the pre-processed visitor route data is segmented into multiple fragments and the second is UBB discovery in which the identified segments from first step are compared with predefined expected routes. These unexpected differences help a web designer to review, redesign, or improve the website according to the user behavior.

Unlike previous work, Dohare *et al.* (2012) proposed a hybrid two-phased session identification scheme. The proposed algorithm is both time oriented and structure oriented. In the first phase, user request sequences are made using an overall time spent by user on the page and the session duration. Each user has a page request sequence represented by an adjacency matrix. In the second phase, candidate sessions are generated by processing user request sequence. This algorithm is better than previously developed both time-oriented and structure-oriented heuristics since each page must be connected to some other page through a hyperlink, this technique does not allow non-related pages in a session. The sessions reconstructed through this technique are shorter and therefore easier to process than those generated by previous algorithms.

Several works in the domain of WUM, leverage clustering techniques with the purpose of characterizing user's navigation behavior (Nichele and Becker, 2006). Clustering is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters (Sudhamathy, 2011). Murray *et al.* (2006) claimed that the global threshold value is not appropriate for all web users, and as an alternative, proposed a user-centered variant of hierarchical agglomerative clustering approach to find the session threshold. The algorithm first loop through the intervals between user queries, the intervals are first ordered in ascending order, then for each one quotient of the interval divided by the standard deviation for short intervals are determined. This ratio is then maximized for the longer intervals. The threshold is then set between this maximum valued interval and the preceding one. The algorithm aims to find the intervals that are significantly longer and have greater impact on the variance.

One of the short comings of traditional threshold-based techniques is the selection of the correct threshold value and its sensitivity, which may affect the number of correctly identified user sessions. Bianco *et al.* (2009) used a mix of hierarchical agglomerative clustering and partition approach for session identification in a three-step process. First clustering is done using partition approach, followed by the aggregation of clusters using agglomerative algorithm and finally the identified clusters are fine-tuned using partitional algorithm. This approach is able to aggregate a set of user sessions generated from a single source. However, due to a large number of clicks present in a session, it is difficult for clustering techniques to identify a user within short time span.

Alam *et al.* (2008) proposed a particle swarm optimization (PSO)-based web session clustering technique that uses time and browsing sequence dimensions for clustering web usage sessions. For PSO, the sessions were taken as particles and Euclidean distance was used to measure the distance between sessions. Evaluation results showed that the proposed scheme performs better than the benchmark K-means clustering algorithm.

Haiyan *et al.* (Lu and Nguyen, 2009) proposed another sequence clustering technique based on PSO. For each user, the algorithm maintains two sequences consisting of common and unique items, respectively. For sequence clustering, similarity measure is defined as a ratio of common items and unique items present in those sequences. Similarity is measured as the order of occurrence of the items in these sequences. This algorithm performs better than K-means.

In order to determine the interesting user patterns, Jyoti *et al.* (2009) presented another WUM scheme to identify individual user navigational behavior on the web. These patterns can then be used to predict the next sequence of pages accessed by the user. The paper presents a novel data pre-processing approach for finding user sessions from the server log. After this, the important sessions are clustered together using rough session clustering techniques based on maximum pages visited. The technique uses a Prediction Prefetching Engine (PPE) that resides on the server and processes the user's past patterns. Using RST, only those sessions are clustered in which the user spent his quality time, and hence it narrows the web log. This decreases the WUM pre-processing time. Experimental results showed that the algorithm outperforms the traditional clustering techniques in order to derive groups of users which exhibit similar access patterns.

Hussain *et al.* (2010) proposed a comprehensive hierarchal cluster-based pre-processing methodology for WUM. It covers all the pre-processing steps from data cleaning to session clustering and converts the categorical web log data into numerical data. During a session identification phase, user sessions are identified by calculating the End Time and Start Time of user accesses. User sessions are then converted into session vectors. Unlike Alam's work, the authors used a combination of three different similarity measures, i.e., Angular Separation, Canberra Distance, and Spearman Distance to group the sessions. During a hierarchal session-based clustering phase, PSO algorithm is applied to obtain the set of winning sessions and in the second step agglomerative algorithm is applied for hierarchal sessionization of user sessions. Experimental evaluation showed that the proposed scheme provides more structured information about the user sessions than previous approaches. It also revealed the importance of selection of proper similarity measure for session clustering.

Chordia and Adhiya (2011) proposed another clustering technique to group user sessions using server log, based on the user access similarity patterns. The author has put forward a local and global alignment method to group sessions using dynamic programming. Web access sequences are grouped together by sequence alignment technique. The sequence alignment method is used as a measure of similarity for user's navigational patterns.

Huidrom and Bagoria (2013) argued that the threshold-based heuristics used for session identification need a priori definition of the threshold value. But this method is subtle to small changes in the threshold. To overcome the limitations of the threshold-based mechanism, the author has given a robust and effective clustering method for user session identification that does not require any prior information. The proposed algorithm uses the best features of hierarchical and partitional clustering techniques. In the first step, an initial clustering is obtained by applying K-mean partitional algorithm, after that, a hierarchical agglomerative clustering technique is applied to get a final number of clusters, finally the partitional algorithm is applied again to get definition of fine number of clusters. The experimental results have shown much better performance than the threshold-based techniques.

Chang-bin and Li (2010) proposed a collaborative filtering-based web log data pre-processing algorithm. The collaborative filtering algorithm is based on the idea that users can be categorized according to their interests. The algorithm employs K-nearest neighbor classifier to calculate the similarities between former user and target user. Adjusted cosine similarity, cosine similarity, and person correlation coefficient methods are mainly used for measurement. In order to assist the decision making, items are assigned ratings based on the measure of users' similarity and weighted value approximation of nearest neighbor. Experimental evaluation revealed the effectiveness of the proposed algorithm in user identification, even with the missing log entries.

To overcome the limitations of time-oriented and structure-oriented heuristics, link-based algorithms are developed. Smart Miner is the first such framework for WUM, devised by Bayir *et al.* (2009) as a part of their web analytics service. The motivation behind Smart Miner is to capture more realistic user behavior by producing more accurate user sessions. Unlike previous approaches, in Smart Miner, sessions are a set of paths traversed in the web graph that corresponds to users' navigations between web pages. Smart Miner employs a new algorithm Smart-SRA that views the session construction process as a web graph problem and identifies the maximal paths traversed. Smart-SRA utilizes timestamp ordering and topology rules and works in two stages. Initially, the short candidate sessions are generated by using time-oriented heuristic rules. In the second stage, maximal sub-sessions are generated by using page stay time rule. Here the topology rule is also used, which guarantees to prevent the session from backward movements. Proposed framework was implemented as software that works on a distributed architecture. Experimental evaluation is done on both simulated and real data. Results showed that the maximal frequent patterns identified by the proposed framework are 30 percent more accurate than the ones discovered by previous session construction methods.

In 2013, to overcome the limitations of Smart-SRA, Bayir and Toroslu (2013) proposed a new technique called Complete Session Reconstruction Algorithm (C-SRA). C-SRA is an extension of the Smart-SRA, which produces link-based sessions with complete set of maximal paths. C-SRA is also a two-stage session reconstruction algorithm. In the first stage, C-SRA works the same as Smart-SRA. In the second stage, C-SRA models the session reconstruction problem as maximal paths in a vertex sequence problem and produces all maximal navigation path sequences from the candidate sessions generated in the first stage. Experimental evaluation is done by replacing the Smart-SRA with C-SRA in WUM component and applying it to the same simulator data previously used for Smart-SRA. Results showed 20-25 percent more accuracy as compared to the previous approaches.

Tiknaz (2013) extended Bayir's work and proposed an improved link-based limited session reconstruction scheme for mining web usage data. The motivation behind limited session reconstruction algorithm was that both the previous link-based heuristics Smart-SRA and C-SRA have some limitations, i.e., Smart-SRA generates less false sessions but misses some of the correct sessions, which leads to low recall value while C-SRA captures most of the sessions but generates false sessions as well which leads to low precision value. The new algorithm LSRA attempts to increase the precision of C-SRA by utilizing a limit function and three different algorithms. The underlying models are time limited, node limited, and a hybrid model, which uses both node and time limitation while constructing sequences. The first stage of the algorithm is the same as the other link-based heuristics Smart-SRA and C-SRA. In the second stage, it puts some limitations while generating sub-sessions by using a website's web topology graph. A limit function is used to compare parameters with threshold values and the parameters are changed with the applied model. Experimental evaluation results showed that the proposed algorithm gives better accuracy on simulated data as compared to other heuristics. While on real data, its accuracy is better than C-SRA, but not as good as Smart-SRA.

Patel and Parmar (2014) proposed a new session identification technique which combines both the time-oriented and structure-oriented heuristics, i.e., maximum forward reference length and reference length model. Based on the page requests, the length model generates length of reference from two consecutive requests made by the same user. Using the length request, a cut-off time is defined which classifies the web page as either content page or navigation page. An initial time-out is calculated by combining the statistical results and an importance degree of the page. Later on session time-out is dynamically adjusted based on the average time spent by user on the website. The proposed cut-off algorithm is compared with both traditional time-out based and dynamic time-out based session identification algorithms. Evaluation results revealed the effectiveness of the proposed algorithm in constructing user sessions more accurately.

Cooley *et al.* (2013) proposed a session segmentation technique which divides the sessions into meaningful transactions. Using these segmented transactions, association rules are developed by WEBMINER system. These transactions are identified using reference length approach and maximal forward reference approach. Reference length approach performs well both on real and artificial data while maximal forward reference approach does not produce good results on real data.

Critical evaluation

In this section, we will sum up all the session identification techniques we have reviewed above, in the form of a table. Table I highlights the classification of techniques based on its heuristic type (i.e. whether it is time-oriented, structure-oriented, or link-based), the WUM process steps it covers (e.g. data cleaning, user identification, session identification, path completion, and transaction identification), the underlying methodology, whether or not the proposed scheme is evaluated, and what are the limitations.

Discussion

Several techniques have been proposed by the research community to perform session identification in WUM process. Sessionization heuristics can be broadly classified as: time-oriented (Zhou *et al.*, 2006; Castellano *et al.*, 2007; Peng and Zhao, 2010; Xinhua and

Table I.
Critical evaluation of session identification techniques

References	Heuristic type	WUM process steps	Methodology	Evaluation (Y/N)	Limitations
Zhou *et al.* (2006)	Time oriented	Session identification	Fixed time-out based threshold value (i.e. 30 min)	Y	The algorithm works inefficiently when there are a number of URL's in the session
Castellano *et al.* (2007)	Time oriented	Data cleaning, session identification and data filtering	Fixed time-based structuration module	Y	No classification performed
Peng and Zhao (2010)	Time oriented	Session identification	Average threshold-based algorithm	Y	The comprehensive web log file is required, missing log entries to be handled
Xinhua and Qiong (2011)	Time oriented	Session identification	Dynamic time-out based threshold value	Y	Factors like loading time of modules in a web page, content size of web pages and busy communication line are not considered
Dinuca and Ciobanu (2012)	Time oriented	Session identification	Average time-based clickstream analysis	Y	IP sharing; sessions of two different users could get conflicted
Chitraa and Thanamani (2011)	Weight matrix	Data cleaning, user identification and session identification	User's traversal data	Y	Only few aspects are considered while identifying the users
Chitra (2013)	Time oriented	Session identification	Mixed Ancestral Graph (MAG) technique	Y	A comprehensive web log is needed in order to generate the MAG graph with complete user traversals in order to correctly regenerate the user session
Roman *et al.* (2008)	Integer programming	Session identification	Log registers	Y	Increased solution time
Khasawneh and Chan (2006)	Structure oriented	Data cleaning, user identification and session identification	User based and ontology based	Y	The sessions are not grouped further to reduce the complexities of subsequent web usage mining stages
Heydari *et al.* (2009)	Graph	Session identification	Statistical analysis and graph method considering client data	Y	The assumption that factors like switching between tabs and hitting backward or forward button are not worthy as they do not create a new navigation path, is not justified. These factors can affect the browsing time and hence should not be ignored

(*continued*)

References	Heuristic type	WUM process steps	Methodology	Evaluation (Y/N)	Limitations
Arumugam and Suguna (2009)	Hash table	Session identification	Access history list (AHL)	Y	Implementation complexity
Fang and Huang (2010)	Structure oriented	User identification, session identification	Frame page and page access threshold	Y	n/a
Dohare *et al.* (2012)	Hybrid	Session identification	Requests adjacency matrix, time-oriented and structure-oriented heuristics	Y	The time-based methods are not accurate as the user might get involved in other activities after opening the web pages
Murray *et al.* (2006)	Clustering	Session identification	Hierarchical agglomerative clustering	Y	There may be more high values and the first high value might not correctly identify the actual first high value
Alam *et al.* (2008)	Clustering	Session identification	Particle swarm optimization (PSO) and Euclidean distance	Y	The results should also be compared to some other PSO-based clustering technique as well, since K-Mean and PSO are different in nature
Lu and Nguyen (2009)	Clustering	Session identification	Particle swarm optimization (PSO)	Y	The performance decreases when the content and the order of occurrence are considered equally
Jyoti *et al.* (2009)	Clustering	Session identification	Prediction Prefetching Engine (PPE) and rough set clustering (RST)	Y	n/a
Hussain *et al.* (2010)	Clustering	Data cleaning, user identification and session identification	Particle swarm optimization (PSO) and three similarity measures (i.e. Angular Separation, Canberra Distance and Spearman Distance)	Y	n/a
Chordia and Adhiya (2011)	Clustering		Sequence alignment method and dynamic programming	Y	The algorithm is computationally intensive

(*continued*)

Table I.

References	Heuristic type	WUM process steps	Methodology	Evaluation (Y/N)	Limitations
Huidrom and Bagoria (2013)	Clustering	Session identification	K-mean partitional algorithm and hierarchical agglomerative clustering technique	Y	In order to improve the quality of identified sessions, a second identification is needed to remove the errors
Chang-bin and Li (2010)	Collaborative filtering	Session identification	K-nearest neighbor classifier, adjusted cosine similarity, cosine similarity and person correlation coefficient methods	Y	The scarcity of users' data and the real-time of nearest neighbor inquiry
Bayir *et al.* (2009)	Link-based	Session identification	Web graph, timestamp ordering and topology rules	Y	The usage of rules forbids the algorithm from finding some of the correct sessions as well and it still cannot capture particular user behaviors when user navigation is more complex
Bayir and Toroslu (2013)	Link-based	Session identification	Maximal paths in a vertex sequence (MPVS)	Y	Although complete-SRA captures more correct sessions as compared to the smart-SRA, but it gives low precision value
Tiknaz (2013)	Link-based	Session identification	Web topology graph, limit function and three different algorithms (time limited, node limited and a hybrid model)	Y	On real data its accuracy is better than C-SRA but not as good as smart-SRA
Patel and Parmar (2014)	Hybrid	Session identification	Time- oriented and structure-oriented heuristics	Y	n/a

Table I.

Qiong, 2011; Dinuca and Ciobanu, 2012) and structure-oriented (Khasawneh and Chan, 2006; Fang and Huang, 2010). One approach for session identification is to use clustering techniques for grouping user sessions (Murray *et al.*, 2006; Alam *et al.*, 2008; Lu and Nguyen, 2009; Jyoti *et al.*, 2009; Hussain *et al.*, 2010; Chordia and Adhiya, 2011; Huidrom and Bagoria, 2013; Chang-bin and Li, 2010). Later, link-based sessionization heuristics were proposed as an alternate for time-oriented and structure-oriented heuristics (Bayir *et al.*, 2009; Bayir and Toroslu, 2013; Tiknaz, 2013). In addition, few hybrid techniques have been proposed that uses both time-oriented and structure-oriented heuristics (Dohare *et al.*, 2012; Patel and Parmar, 2014). Moreover, some researchers have combined conventional sessionization heuristics with other techniques, for example, matrix, graphs, hash tables, and statistical analysis (Chitraa and Thanamani, 2011; Chitra, 2013; Roman *et al.*, 2008; Heydari *et al.*, 2009; Arumugam and Suguna, 2009).

Time-oriented heuristics can be further classified as fixed time-out and dynamic time-out. A time-oriented algorithm presented in Zhou *et al.* (2006) uses fixed time-out threshold value (i.e. 30 min) for user session identification. This technique helps predicting the otherwise costly process of user navigation patterns to be done in advance. However, this algorithm works inefficiently when there are a number of URLs in the session. LODAP (Castellano *et al.*, 2007) is a data pre-processing tool which performs data cleaning, data filtering, and session identification along with the summary of each step of the processor. However, if LODAP was able to perform some classification, it would be an effective tool for data pre-processing.

Considering the fact that fixed time-out threshold values do not take into account the different user behaviors, it has been suggested that dynamic time-out threshold values should be used rather than fixed time-out threshold values. A two-step average threshold-based algorithm is proposed by Peng and Zhao (2010), where average threshold value is calculated and adjusted dynamically. However, there are some limitations, i.e., the complete web log is required for the proper working of a session identification algorithm so missing log entries need to be handled. In Xinhua and Qiong (2011), another dynamic time-out based session identification algorithm is proposed. However, the factors like loading time of modules in a web page, content size of web pages, and busy communication line are not considered. In Dinuca and Ciobanu (2012), a modified version of fixed time-out-based session identification algorithm is presented; however, there is a limitation of IP sharing, sessions of two different users could get conflicted.

Another approach for session identification is to use structure-oriented heuristic to reconstruct the user sessions. In Khasawneh and Chan (2006), an ontology-based algorithm is utilized for session identification, which is based on the features extracted from the pages and structure of a website. However, the sessions are not grouped further to reduce the complexities of subsequent WUM stages. In Fang and Huang (2010), another session identification algorithm is presented, which can identify longer sessions with increased accuracy.

Clustering techniques are also used for session identification with the purpose of characterizing user's navigation behavior. Considering the fact that the global threshold value is not appropriate for all web users, a user-centered variant of hierarchical agglomerative clustering approach is proposed in Murray *et al.* (2006), which can be used to find the session threshold. However, in case of maximum appearance of non-reasonable values, there may be more high values and the first high value might not correctly identify the actual first high value. Considering the fact that threshold-based heuristics are subtle to small changes in the threshold value, an effective clustering method is proposed in Huidrom and Bagoria (2013) that does not require any prior information. The proposed algorithm uses the best features of hierarchical and partitional clustering techniques. The experimental results have shown much better performance than the threshold-based techniques. However, to improve the quality of identified sessions, a second identification is needed in order to remove the errors.

Using PSO different web session clustering techniques are proposed in Alam *et al.* (2008), Lu and Nguyen (2009), and Hussain *et al.* (2010). In Alam *et al.* (2008), sessions are taken as particles and Euclidean distance is used to measure the distance between sessions. However, the results should also be compared to some other PSO-based clustering technique as well. In Lu and Nguyen (2009), algorithm maintains two sequences for each user, consisting of common items and unique items. This algorithm performs better than K-means. However, the performance decreases when the content and the order of occurrence are considered equally. In addition, PSO-based sequence clustering algorithms are sensitive to the selection of centroids of K-clusters, i.e., which should be the existing sequences in the data sets. In Hussain *et al.* (2010), PSO algorithm is applied to obtain the set of winning sessions and in the second step, agglomerative algorithm is applied for hierarchal sessionization of user sessions. This scheme provides more structured information about the user sessions than previous approaches. It also revealed the importance of selection of proper similarity measure for session clustering.

In order to determine the interesting user patterns, Jyoti *et al.* (2009) presented a WUM scheme to identify individual user navigational behavior on the web. Experimental results showed that the algorithm outperforms the traditional clustering techniques in order to derive groups of users which exhibit similar access patterns. In Chordia and Adhiya (2011), a local and global alignment method has put forward group sessions using dynamic programming. Although, the algorithm is computationally very intensive, but experiments showed that it performs better than other similarity measures to find similarity between different user navigation patterns to form clusters.

To overcome the limitations of time-oriented and structure-oriented heuristics, link-based algorithms are developed. Smart Miner (Bayir *et al.*, 2009) is the first such framework, which employs a new algorithm Smart-SRA that views the session construction process as a web graph problem and identifies the maximal paths traversed. However, there is a limitation, although the usage of rules forbids the algorithm from generating too many sessions and reduces the total number of false sessions produced, but at the same time, it forbids the algorithm from finding some of the correct sessions as well. Consequently, it still cannot capture particular user behaviors when user navigation is more complex due to its greedy nature. To overcome the limitations of Smart-SRA, C-SRA (Bayir and Toroslu, 2013) was proposed, which produces link-based sessions with complete set of maximal paths. Although Complete-SRA captures more correct sessions as compared to the Smart-SRA, but it gives low precision value due to the number of false sessions. In Tiknaz (2013), an improved link-based limited session reconstruction method is proposed, which attempts to increase the precision of C-SRA by utilizing a limit function and three different algorithms. Proposed algorithm gives better accuracy on simulated data as compared to other heuristics. While on real data, its accuracy is better than C-SRA but not as good as Smart-SRA.

A hybrid two-phased session identification scheme is proposed in Dohare *et al.* (2012), which uses both time-oriented and structure-oriented heuristics. However, the time-based methods are not accurate as the user might get involved in other activities after opening the web pages. Another hybrid session identification technique is presented in Patel and Parmar (2014), which combines the both maximum forward reference length and reference length model. The proposed cut-off algorithm constructs user sessions more accurately than fixed time-out-based and dynamic time-out based session identification algorithms.

A new session identification technique is presented in Chitraa and Thanamani (2011), where sessions are identified by constructing a weight matrix of web pages and users along with their sessions. However, the user identification process still needs to be improved as only few aspects are considered while identifying the users. In Chitra (2013), time-oriented heuristics are combined with MAG graphs and a new session reconstruction algorithm is proposed, which is based on the timestamp of the visited pages. However, a comprehensive web log is needed in order to generate the MAG graph with complete user traversals in order to correctly regenerate the user session. A combination of the statistical analysis and graphical method is presented in Heydari *et al.* (2009) for session reconstruction, which considers the client side that was not taken into account previously. Here, the assumption that factors like switching between tabs and hitting backward or forward button are not worthy as they do not create a new navigation path, is not justified. These factors can affect the browsing time and hence should not be ignored. An optimal session identification technique is proposed in Arumugam and Suguna (2009), which uses two-way hashed tables to correctly map the user navigational history. It gives correct session identification within an optimal time as compared to maximal forward and reference length methods.

A novel algorithm for session reconstruction is proposed in Roman *et al.* (2008), which uses integer programming to group the requests coming from the same IP address through log registers. Unlike heuristics, it constructs all the user sessions at the same time with increased accuracy. However, the ability of the integer program to construct all sessions together does come at the cost of increased solution time (i.e. it is substantially slower as compared to time-based heuristics). A collaborative filtering-based web log data pre-processing algorithm is proposed in Chang-bin and Li (2010), which employs K-nearest neighbor classifier to calculate the similarities between former user and target user. However, there are some limitations such as the scarcity of users' data and the real-time of nearest neighbor inquiry.

Summary and findings

The findings revealed that this research area is still immature and existing literature lacks the critical review of recent session identification techniques used in WUM process. In this section, we will summarize our research findings in the form of a table. Table II highlights the challenges and limitations of the existing approaches and identify areas where further improvements are required so as to complement the performance of existing techniques.

Conclusion

Data pre-processing is an essential part to improve the efficiency and quality of the later stages of WUM process. In this paper, we have critically reviewed the state-of-the-art session identification techniques used in WUM process that have been proposed

Table II. Summary and findings

Time oriented	Fixed time-out threshold values do not take into account the different user behaviors and it has been suggested that dynamic time-out threshold values should be used rather than fixed time-out threshold values. While using dynamic time-out threshold values, issues like missing log entries and IP sharing needs to be addressed. Moreover, factors like loading time of modules in a web page, content size of web pages and busy communication line should also be considered
Structure oriented	Structure-oriented heuristics should be extended to group sessions so that it can reduce the complexities of subsequent steps in WUM process
Clustering	Clustering techniques which are user centric and do not require any priori information are more effective. Selection of proper similarity measure is important for session clustering
Link based	Some researchers have suggested different link-based heuristics and experimental results revealed that the idea of viewing the session construction process as a web graph problem is actually very effective
Hybrid	Extensive research is being carried out on combining conventional sessionization heuristics with other techniques, for example matrix, graphs, hash tables, statistical analysis. Recently, hybrid approach that uses a combination of time-oriented and structure-oriented heuristics has received interest and is likely to gain more attention in the future

over the period of 2005-2015. It was discovered that despite the enormous research, existing techniques are limited in some identification aspects. Session identification is an important yet most difficult step and if not handled properly will directly affect the subsequent stages of WUM process. Hence, more robust techniques for session identification are required. We strongly believe that this review paper will help in improving the existing session identification techniques and it will accelerate the research efforts in the domain of WUM.

References

Alam, S., Dobbie, G. and Riddle, P. (2008), "Particle swarm optimization based clustering of web usage data", *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 451-454, available at: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4740819 (accessed April 26, 2015).

Arumugam, G. and Suguna, S. (2009), "Optimal algorithms for generation of user session sequences using server side web user logs", *2009 International Conference on Network and Service Security*, pp. 1-6.

Asadianfam, S. and Mohammadi, M. (2014), "Identify navigational patterns of web users", *International Journal of Computer-Aided Technologies (IJCAx)*, Vol. 1 No. 1, pp. 1-8.

Bayir, M. and Toroslu, I. (2013), "Link based session reconstruction: finding all maximal paths", available at: http://arxiv.org/abs/1307.1927 (accessed April 26, 2015).

Bayir, M.A., Toroslu, I.H., Cosar, A. and Fidan, G. (2009), "Smart miner: a new framework for mining large scale web usage data", *Proceedings of the 18th International Conference on World Wide Web*, pp. 161-170, available at: http://dl.acm.org/citation.cfm?id=1526732

Bianco, A., Mardente, G., Mellia, M., Munafò, M. and Muscariello, L. (2009), "Web user-session inference by means of clustering techniques", *IEEE/ACM Transactions on Networking (TON)*, Vol. 17 No. 2, pp. 405-416.

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M. and Khalil, M. (2007), "Lessons from applying the systematic literature review process within the software engineering domain", *Journal of Systems and Software*, Vol. 80 No. 4, pp. 571-583.

Castellano, G., Fanelli, A.M. and Torsello, M.A. (2007), "LODAP: a log data preprocessor for mining web browsing patterns", *Proceedings of the 6th Conference on 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, World Scientific and Engineering Academy and Society (WSEAS)*, pp. 12-17.

Chang-bin, J.C.J. and Li, C.L.C. (2010), "Web Log Data Preprocessing Based on Collaborative Filtering", *2010 Second International Workshop on Education Technology and Computer Science (ETCS), 2*, pp. 118-121.

Chitra, S. (2013), "A novel Preprocessing Mixed Ancestral Graph technique for session construction", *International Conference on Computer Communication and Informatics*, pp. 1-7.

Chitraa, V. and Thanamani, A. (2011), "A novel technique for sessions identification in web usage mining preprocessing", *International Journal of Computer Applications*, Vol. 34 No. 9, pp. 24-28.

Chordia, B.S. and Adhiya, K.P. (2011), "Grouping web access sequences using sequence alignment", Vol. 2 No. 3, pp. 308-314.

Cooley, R., Mobasher, B. and Srivastava, J. (2013), "Data preparation for mining World Wide Web browsing patterns", *Knowledge and Information Systems*, Vol. 1 No. 1, pp. 5-32, available at: http://link.springer.com/10.1007/BF03325089 (accessed March 23, 2016).

Dinuca, C.E. and Ciobanu, D. (2012), "Improving the session identification using the mean time", *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 6 No. 2, pp. 265-272, available at: http://naun.org/main/NAUN/ijmmas/17-705.pdf (accessed March 12, 2015).

Dohare, M.P.S., Arya, P. and Bajpai, A. (2012), "Novel web usage mining for web mining techniques", *International. Journal of Emerging Technology and Advanced Engineering*, Vol. 2 No. 1, pp. 253-262, available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.414.529 (accessed March 15, 2015).

Domenech, J.M. and Lorenzo, J. (2007), "A tool for web usage mining", *International Conference on Intelligent Data Engineering and Automated Learning*, Springer Berlin Heidelberg, December, pp. 695-704.

Fang, Y. and Huang, Z. (2010), "A session identification algorithm based on frame page and page threshold", *Proceedings – 2010 3rd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2010*, 6, pp. 645-647.

Gopalakrishnan, T., Kavya, M. and Gowthami, V.S. (2014), "Advanced preprocessing techniques used in web mining: a study", *International Journal of Computer Applications*, Vol. 101 No. 13, pp. 16-20, available at: www.ijcaonline.org/archives/volume101/number13/17747-8822 (accessed April 27, 2015).

Heydari, M., Helal, R.a. and Ghauth, K.I. (2009), "A graph-based web usage mining method considering client side data", *2009 International Conference on Electrical Engineering and Informatics, August 1*, pp. 147-153.

Huidrom, N. and Bagoria, N. (2013), "Clustering techniques for the identification of web user session", Vol. 3 No. 1, pp. 1-8.

Hussain, T., Asghar, S. and Masood, N. (2010), "Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence", *2010 6th International Conference on Emerging Technologies (ICET)*, pp. 21-26.

Iváncsy, R. and Vajk, I. (2006), "Frequent pattern mining in web log data", *Acta Polytechnica Hungarica*, Vol. 3 No. 1, pp. 77-90.

Jansen, B.J. (2006), "Search log analysis: what it is, what's been done, how to do it" *Library & Information Science Research*, Vol. 28 No. 3, pp. 407-432.

Jyoti, J. *et al.* (2009), "A novel approach for clustering web user sessions using RST", *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pp. 10-12.

Khasawneh, N. and Chan, C. (2006), "Active user-based and ontology-based web log data preprocessing for web usage mining", *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pp. 325-328, available at: http://dl.acm.org/citation.cfm?id=1248823.1249128 (accessed April 26, 2015).

Kimble, C. and Kudenko, D. (2005), "UBB mining: finding unexpected browsing behaviour in clickstream data to improve a web site's design", *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 179-185, available at: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1517840 (accessed March 23, 2016).

Kumar, B. and Rukmani, K. (2010), "Implementation of web usage mining using APRIORI and FP growth algorithms", *International Journal of Advanced Networking and Applications* , Vol. 1 No. 6, pp. 400-404, available at: http://ijana.in/papers/6.11.pdf (accessed May 10, 2015).

Limam, L. *et al.* (2010), "Extracting user interests from search query logs: a clustering approach", *2010 Workshops on Database and Expert Systems Applications, IEEE*, pp. 5-9, available at: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5591973 (accessed August 3, 2015).

Lu, H. and Nguyen, T.T.S. (2009), "Experimental investigation of PSO based web user session clustering", *SoCPaR 2009 – Soft Computing and Pattern Recognition*, pp. 647-652.

Murray, G.C., Lin, J. and Chowdhury, A. (2006), "Identification of user sessions with hierarchical agglomerative clustering", *ASIS&T Annual Meeting*, *November*, pp. 1-4.

Nichele, C.M. and Becker, K. (2006), "Clustering web sessions by levels of page similarity", in Ng, W.-K. *et al.* (Eds), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, pp. 346-350.

Patel, P. and Parmar, M. (2014), "Improve Heuristics for user session identification through web server log in web usage mining", *International Journal of Computer Science and Information Technologies*, Vol. 5 No. 3, pp. 3562-3565, available at: http://ijcsit.com/docs/Volume 5/vol5issue03/ijcsit20140503201.pdf (accessed April 26, 2015).

Peng, Z. and Zhao, M. (2010), "Session identification algorithm for web log mining", *2010 International Conference on Management and Service Science, IEEE*, pp. 1-4.

Rahaman, A., Sait, W. and Meyappan, T. (2014), "Data preprocessing and transformation technique to generate pattern from the web log", *International Conference on Computer Science and Information Systems (ICSIS' 2014)*, pp. 17-18.

Roman, P.E., Velasquez, J.D. and Dell, R.F. (2008), "Web user session reconstruction using integer programming", *2008 IEEE/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society*, Vol. 1, pp. 385-388.

Sharma, A. (2008), "Web usage mining: data preprocessing, pattern discovery and pattern analysis on the RIT web data", available at: http://www.cs.rit.edu/~aps2177/APS_MSProject_Report_REV3.pdf (accessed May 28, 2015).

Siddiqui, A. and Aljahdali, S. (2013), "Web mining techniques in e-commerce applications", available at: http://arxiv.org/abs/1311.7388 (accessed May 28, 2015).

Sindhuja, R., Sengottuvelan, P. and Lokeshkumar, R. (2014), "A survey on preprocessing of web log file in web usage mining to improve the quality of data", *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4 No. 8, pp. 2250-2459, available at: www.ijetae.com/files/Volume4Issue8/IJETAE_0814_35.pdf (accessed May 10, 2015).

Sisodia, D.S., Verma, S. and Vyas, O.P. (2015), "Agglomerative approach for identification and elimination of web robots from web server logs to extract knowledge about actual visitors", *Journal of Data Analysis and Information Processing*, Vol. 3 No. 1, pp. 1-10, available at: www.scirp.org/journal/PaperInformation.aspx?PaperID=55973abstract (accessed March 23, 2016).

Stenmark, D. (2008), "Identifying clusters of user behavior in intranet search engine log files", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 14, pp. 2232-2243, available at: http://doi.wiley.com/10.1002/asi.20931 (accessed August 3, 2015).

Stenmark, D. and Jadaan, T. (2007), "Intranet users' information-seeking behaviour: an analysis of longitudinal search log data", *Proceedings of the American Society for Information Science and Technology*, Vol. 43 No. 1, pp. 1-19, available at: www.researchgate.net/publication/28805742_Intranet_Users_Information-Seeking_Behaviour_An_Analysis_of_Longitudinal_Search_Log_Data (accessed July 16, 2015).

Sudhamathy, G. (2011), "Web log clustering approaches – a survey", *International Journal on Computer Science and Engineering*, Vol. 3 No. 7, pp. 2896-2903.

Tiknaz, B. (2013), "Link based limited session reconstruction method for mining web usage data", available at: http://etd.lib.metu.edu.tr/upload/12616384/index.pdf (accessed April 26, 2015).

Wu, D., Zhang, G. and Lu, J. (2015), "A fuzzy preference tree-based recommender system for personalized business-to-business e-services", *IEEE Transactions on Fuzzy Systems*, Vol. 23 No. 1, pp. 29-43, available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6783754 (accessed April 5, 2015).

Xinhua, H. and Qiong, W. (2011), "Dynamic timeout-based a session identification algorithm", *2011 International Conference on Electric Information and Control Engineering, IEEE*, pp. 346-349, available at: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5777587 (accessed March 15, 2015).

Zhou, B., Hui, S. and Fong, A. (2006), "An effective approach for periodic web personalization", *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pp. 284-292.

About the authors

Bahjat Fatima holds a Bachelor's Degree in Software Engineering from the University of Engineering and Technology (UET), Taxila, Pakistan and currently she is enrolled in MS Computer Science at the COMSATS Institute of Information Technology, Islamabad, Pakistan. Bahjat Fatima is the corresponding author and can be contacted at: bahjatfatima.se14@gmail.com

Huma Ramzan holds a BSc in Software Engineering from the University of Engineering and Technology (UET), Taxila, Pakistan and currently she is enrolled in MS Computer Science at the COMSATS Institute of Information Technology, Islamabad, Pakistan.

Sohail Asghar is currently working as the Chief Technologist at the COMSATS Institute of Information Technology, Islamabad, Pakistan. In 2006, he received his PhD from Faculty of Information Technology at the Monash University, Melbourne, Australia. He has also served as a Program Committee Member of numerous international conferences and has more than 100 publications in international journals as well as conference proceedings.