Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

# Ranker Information Gain Preprocessed Mutal Reinforcement Clustering For Mining Web Access Logs

**Nagan B. K. Mathan* and C. Chandrasekar**

Department of Computer Science, Dravidian University, kuppam, Andhra Pradesh,
Email: mathan_nagan@yahoo.com M. 99946 34644
Department of Computer Science, Periyar University, Salem, Tamil Nadu,
Email: ccsekar@gmail.com

## ABSTRACT

Web usage mining investigates the navigation patterns in web access logs and obtains the most relevant user information. This helps to introduce different strategies for several web-related applications, to name a few are, file sharing, online file editing and distributed social network. Many research works has been conducted in this area, therefore improving the clustering quality and accuracy of navigation patterns being generated. The current work demonstrates clustering of user sessions according to the mutual reinforcement clustering based on the domain-specific and similar interests of web users by proposing a ranker information gain preprocessing model to extract dimensionality reduced user sessions. The method is called as Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering (RIGP-MRC) for efficient mining of web access logs. To start with, a preprocessing model using Ranker Information Gain is applied to the input web log files to extract dimensionality reduced user sessions. Next, Mutual Reinforcement Clustering algorithm cluster the user sessions. Finally, within each user session, domain specific and similar interests for selecting specific URL are generated based on Mutual Reinforcement Clustering. The selected specific URL is then ranked according to the domain and similar interests of web users for summarizing session clusters into user profiles. Results of large scale evaluation demonstrate that RIGP-MRC is more effective than previous approaches for finding domain-specific and similar interests web users. More importantly, the RIGP-MRC method significantly improves the precision rate with minimum computational time and computational overhead over the state-of-the-art methods.

**KEYWORDS**: Web access logs, Ranker Information Gain, Mutual Reinforcement Clustering, Domain-specific, Similar interests, web users

## *Corresponding author

## B. K. Mathan Nagan

Research Scholar,

Department of Computer Science,

Dravidian University, kuppam,

Andhra Pradesh,

Email: mathan_nagan@yahoo.com M. 99946 34644

## INTRODUCTION

Customer Relationship Management (CRM) uses data from inside and outside the business establishments to provide an understanding of its customers via customer profile management. An up-to-date understanding of the customer's preferences, requirements and niches allow the business establishments to profit, i.e., through cross selling or selling the produces correlated to the ones that the customer needs to buy. Hence, authentic and genuine knowledge regarding the customers' preferences and necessity forms the basis for effective CRM. The advancement of the internet besides the popularity of the web has gained a large deal of attention between the research persons to web usage mining.

In Search Goal Shift [1], initially based on the behavioral characteristics of searchers in the search goal shift processes, all the queries that were submitted in the search goal shift processes were extracted from search engine logs. This was performed with the help of machine learning mechanism. With result, queries were constructingthe search goal shift graph. Finally, a random walk algorithm was applied to the resultant queries to extract the query recommendations in search goal shift graph. However, certain other aspects may also influence the navigation patterns. To address this issue, in this work, two main aspects, like, domain-specific knowledge and specific interests of each user is analyzed based on the mutual reinforcement principle.

Web sessions clustering using hybrid sequence alignment measure (HSAM) [2] performed user sessions clustering prevailing to uneven lengths. The objective of HSAM remained in identifying the access patterns by applying a distance method to group user sessions. Besides, the hybrid distance measure exploits information regarding the access path to identify the distance between sessions without changing the sequence in which web pages were visited.

In addition, two different validity indices namely, Jaccard Index and Davies–Bouldin validity indices were used to measure the clustering quality and accuracy. As a result, optimal number of clusters were said to be formed and was found to be encouraging in terms of several navigation patterns being obtained. Despite improving the clustering quality and accuracy, the computational overhead and computational time involved in accessing several navigation patterns were not concentrated. To address this issue, in this work, Ranker Information Gain Preprocessing model is investigated that uses the entropy and measures Information Gain for each attributes with respect to the class variable, therefore, minimizing the computational overhead and computational time involved during preprocessing, reflecting the overall process.

Modified Cognitive Style Analysis (Modified CSA) [3] involved a three-fold design. They were, detecting relationship among cognitive styles and user navigation behavioral patterns. Second

design included the clustering techniques exploited to group users of specific cognitive style via measures derived from cognitive style analysis test. Finally, Modified CSA evaluated the use of navigation content metrics to identify the groups of users that possess similar navigation pattern. Results revealed that clustering algorithms were efficient for CSA with a specific cognitive style, therefore improving the optimum users, rather than applying a rule-based method based on specific thresholds. Navigation behavior based on the linear and non-linear patterns were analyzed in an efficient manner.

In this paper, we develop and evaluate a new method to solve above said two problems for web usage mining. The concept of ranker information gain is introduced to preprocess the given input datasets with the objective of extracting relevant attributes and then show how the results can be used for further processing. This is a novel method, since instead of relying on assumptions about the user profiles being mined according to several mining concepts it estimates the mutual reinforcement principle to cluster according to similar interests and domain-specific knowledge. The main contributions of this work is described as below

- To achieve precision with web log files during web usage mining for mining user profiles, propose a new method called Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering (RIGP-MRC). In such a method, the essential data is said to be sent for further processing whereas the other web attributes in web log files retain in the web log files as they are not of relevance. It reduces the time taken for navigate access patterns of user profiles. This is performed by the application of ranker information gain preprocessing model.
- To provide a method for minimizing the computational overhead for navigating access patterns based on the mutual reinforcement clustering to support strong rate of precision and accuracy, which by generating two different separate entities for similar interests and domain-specific knowledge, therefore minimizing the burden of computational overhead.
- The web usage mining for accessing web navigation patterns indicates that the proposed method involves minimum time for accessing the patterns and also rationalizes the performance by tangible executions. The result shows that the proposed method attains desirable rate of accuracy with minimum computational time and overhead.

The paper is ordered as follows. In Section 2, an overview of web usage mining methods and algorithms is presented. In Section 3, Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering (RIGP-MRC) for efficient mining of web access logs is investigated. In Section 4 experimental settings is presented. Section 5 provides a detailed discussion with the aid of graph. Consequently, we conclude the paper in Section 6.

## RELATED WORKS

Extracting the common characteristics of web users is said to be performed by several experts so that the common characteristics of anomalous behaviors can be learnt. In [4], a novel frequent episode mining algorithm was designed to identify various patterns based on the characteristics of their query volume time series. The results indicated interested patterns being mined in an efficient manner. A comprehensive survey of semantic data web mining was presented in [5] for constructing content-based recommender systems.

While ensuring market research information, the content generated by the user on the web poses several challenges with respect to systematic analysis, differences and unique characteristics of various social media channels and so on. In [6], reports on the determination of such particularities were presented and their impact on text preprocessing and opinion mining were examined. Clustering of data streams for web mining was designed in [7] among micro-clusters via a shared density graph, therefore improving clustering quality. A guest editorial on advances in web services was presented in [8]. Rough set theory and dominance principle were applied in [9] to yield crucial set of effectively consistent information.

With increased usage of mobile devices, online markets are used by consumers for effective purchasing.Data flood refers to that the consumers are in a state of confusion with several options to be select amidst their favorite products and therefore makes the decision making a cumbersome process.

One solution is recommender systems that assists consumers in identifying items of interest and also provide with additional items. In [10], privacy enhanced matrix factorization was investigated with Local Differential Privacy (LDP) to evaluate the recommendation accuracy. Despite improvement observed in recommendation accuracy, the time consumed was not concentrated. To address this issue, Most Interesting Pattern-based Parallel FP-growth algorithm [11] was designed that not only improved the execution time but also improving the support factor.

Certain challenges and opportunities in web mining with respect to crowd sourcing were analyzed in [12]. In [13], the effect of task complexity was analyzed using machine learning models with the objective of obtaining higher quality of results. With the exponential opinions available on the websites, tourists are often provided with wide range of information. Due to the high availability of information, it becomes extremely cumbersome to avail the information for decision making. In [14], a fuzzy aspect based opinion classification system was designed to extract aspects from user opinions and performed accurate classification, henceenhances classification accuracy.

One of the very popular active researches in natural language processing is the analysis of sentiments or simply sentiment analysis. Sentiment analysis deals with structured and unstructured data. With these two types of data, people's opinions are detected and then extracted in several resources of subjectivity, namely, product reviews, blogs, social networks, etc. In [15], yet another classification-based approach was presented to extract explicit classification. Co-occurrences and sequence patterns were mined for diagnosing cancer [16] using disparities. By applying disparities among various patient groups, diagnosis accuracy was said to be improved.

Sites for web-based shopping are exponentially increasing nowadays. There, business establishments are anxiously thinking about their client purchase pattern. Among cash and merchandise, internet shopping is considered to be one of the most familiar methods for powerful exchange which is completed by the end clients without the needs of energy spam. In [17], high-recommendation web-based sites were dissected with the aid of an integrated collection mechanism and a swarm-based improvement system. In [18], process mining based service composition approach was presented using process mining algorithm with the objective of enhancing service composition adaptiveness and efficiency. Four classification algorithms namely, Support Vector Machines, Random Forest, Statistical and Logical analysis of data and Logistic Classifier was applied in [19] to improve classification accuracy for identifying e-commerce. An automatic recommendation for online users using two tier architecture was presented in [20] with the advantage of improving user intuition.

## METHODOLOGY

The framework for our Web usage mining and an overview to the rest of this paper is summarized in Figure 1, which starts with the preprocessing of Web log files, clustering the user sessions, summarizing session clusters into user profiles.

**Figure 1Ranker Information Gain-Preprocessed - MRC**

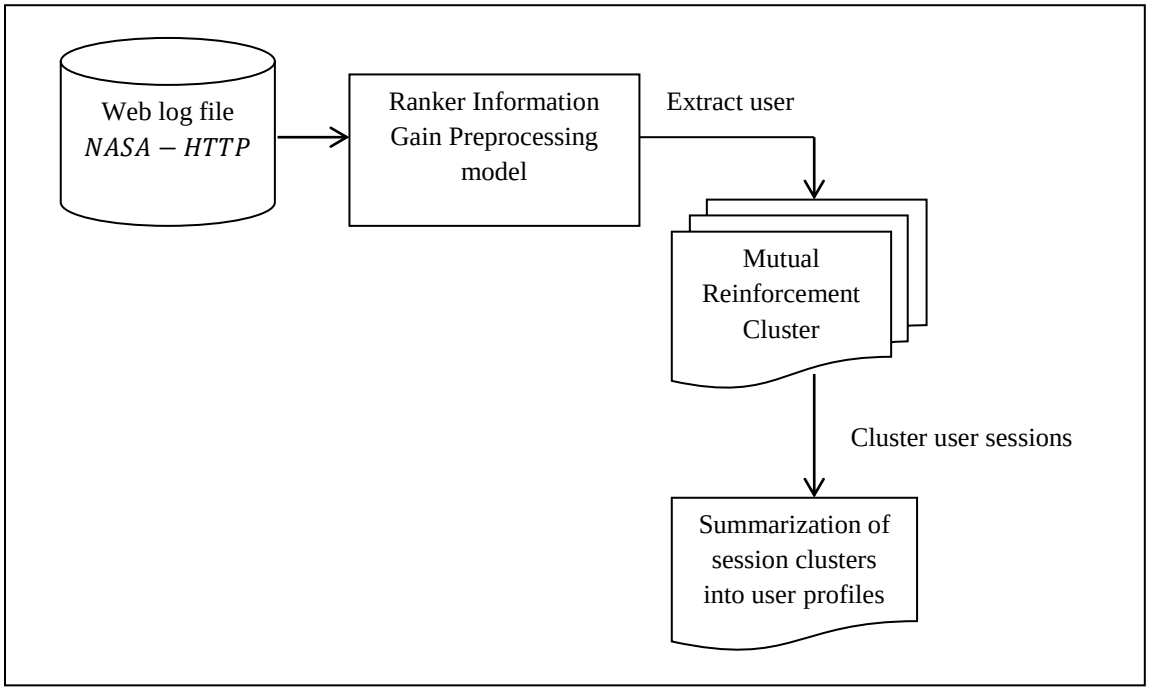



Figure 1 shows the block diagram of Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering (RIGP-MRC) method for efficient mining of web access logs. This is followed by clustering results to acquire Web user profiles and ends with tracking profile evolution.The automatic identification of user profiles consists of regularly mining updated user access log files and is outlined in the following steps:

a. Preprocessing web log file using Ranker Information Gain to extract dimensionality reduced user sessions
b. Cluster user sessions using Mutual Reinforcement
c. Summarizing cluster user sessionsaccording to domain-specific and similar interests web users

The elaborate description of the design of Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering method is given below.

**a. Ranker Information Gain Preprocessing model**

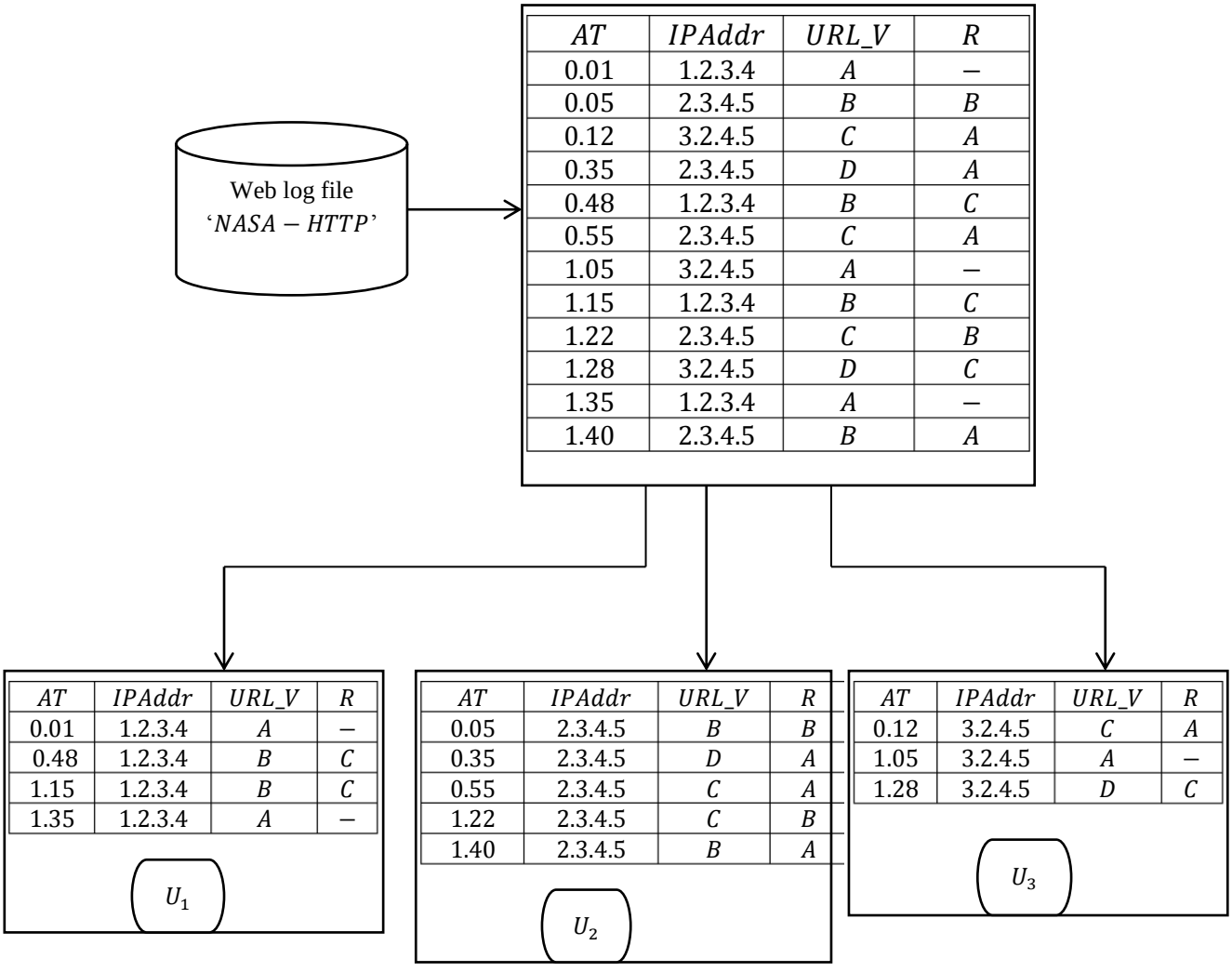
Data preprocessing refers to the process of web log file parsing, data cleaning and filtering. The most imperative steps of data mining is data preprocessing. Here, data are prepared for mining with the aid of a given input data set or web log file for each user session. User session is said to be established when the web user makes the first request to the Web server. Alternatively, the user session is said to be completed upon a period of inactive time from the web user, with the inactive time referred to as the session timeout.

With the inclusion of both active and inactive time, substantial time is said to be consumed during preprocessing or extracting user sessions. There are several techniques utilized in distance measures [1] to identify most notable features or attributes of a predictive problem. Several features or aspects were discovered but measures were not taken to identify how effective those factors were accurately predicting and extracting user sessions with minimum time.

To address this issue, Ranker Information Gain Preprocessing (RIGP) model is utilized in this work that performs dimension reduction to capture the most relevant attributes for further processing. Here, the attributes in web log file refers to each log entry (i.e., access time '$AT$', IP address '$IPAddr$', URL viewed '$URL\_V$', Referrer '$R$' and so on). A sample block diagram representing the attributes in web log files with the resultant extract of user for different sessions with the aid of RIGP model is provided in figure 2.

**Figure 2**

**IP address and URL viewed in web log files**



The RIGP model exploits entropy and measures the Information Gain '$IG$' for each attributes with respect to the class variable, resulting in a ranking of each attributes with a range between '0' and '1'. Attributes contributing more possess higher entropy and chosen for further processing. Attributes contributing less possess lower entropy are eliminated.

Given a web log file '$f$' containing attributes '$Attr = A_1, A_2, \dots, A_n$', then the objective of RIGP model is to generate smaller set of attributes '$A_{i1}, A_{i2}, \dots, A_{ik}$'. Here '$k$' refers to the length of the dimensionality reduced attributes with '$k \leq n$'. Then, with '$Attr$' representing the set of all attributes and '$Tr$' representing the set of all training examples, '$value\ (r, a)$' with '$r \in Tr$' defines the value of a specific example for attribute '$a \in Attr$' and '$H$' represent the entropy. Then, the information gain for an attribute is mathematically formulated as given below.

$$IG(Tr, a) = H(Tr) - \sum_{v \in values\ (a)} \left( \frac{r \in Tr \mid val(r,a)=v}{Tr} \right) * (r \in Tr \mid val(r, a) = v) \qquad (1)$$

From the above equation (1), the information '$IG$' to extract relevant user sessions is obtained using the training samples '$Tr$' and the attributes involved '$a$'. Along with the URL viewed '$URL_V$', Ranker Information Gain Preprocessing model also considers referrer '$R$' to extract users sessions as lists, thus saving on both computational overhead and computational time involved during preprocessing (i.e. minimizing the computational overhead and computational time for extracting user session).

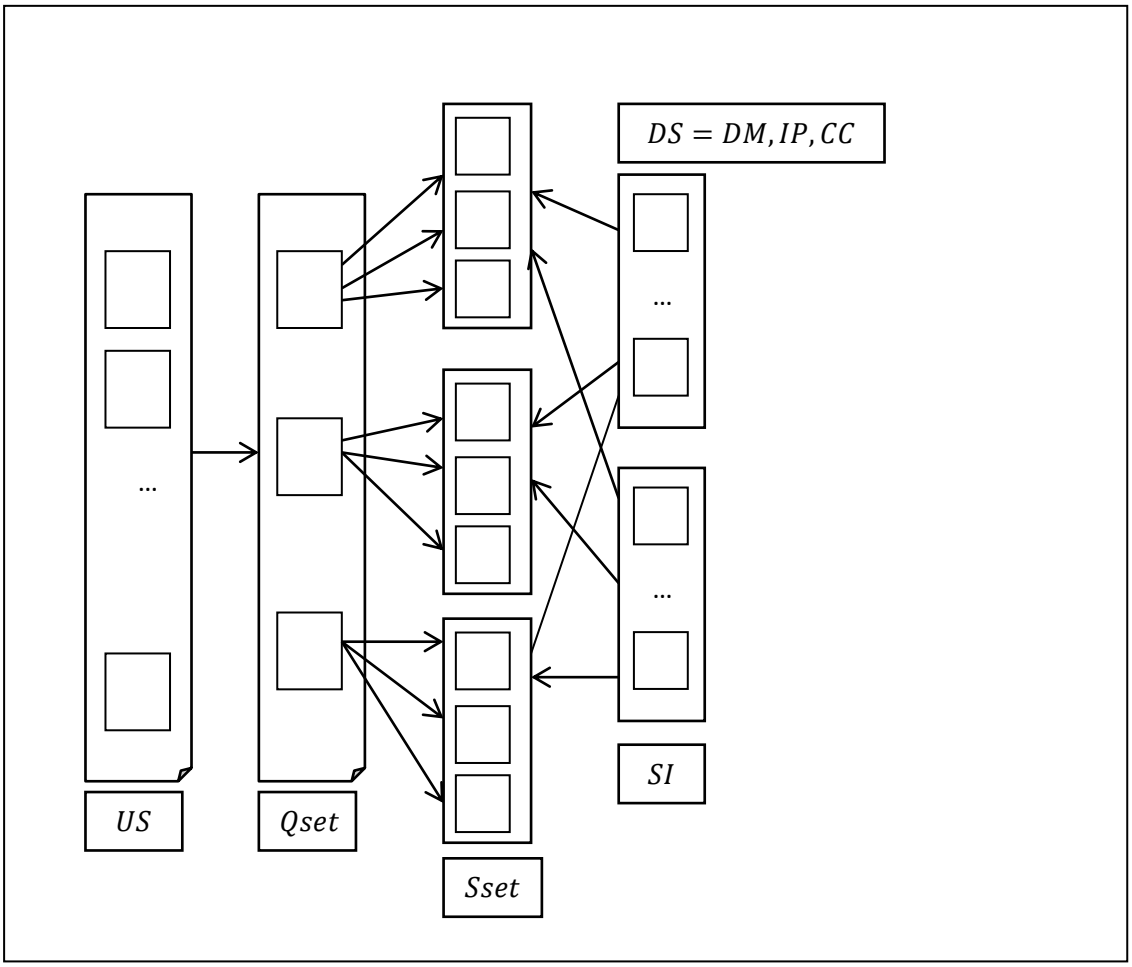
**b. Mutual Reinforcement Clustering model**

With the extracted user sessions, the next step in web usage mining to cluster user sessions. In this work, Mutual Reinforcement Clustering model is applied to the extracted user sessions with the objective of partitioning of user sessions into optimal clusters. The main outline of the Mutual Reinforcement Clustering (MRC) algorithm is elaborated as given below. The purpose that we use MRC algorithm instead of other clustering algorithms is that unlike most other algorithms, MRC algorithm can handle both domain aspect with similar interests and automatically partitions the user sessions into optimal clusters.

Hence, the purpose of using MRC remains in clustering the user sessions based on the domain-specific and similar interests simultaneously. The mutual reinforcement principle for web usage mining that underlies our approach to solve the problem with higher true positive rate. A URL is said to be selected then if the resultant retrieval is obtained both from similar interest '$SI$' and domain-specific '$DS$' web users. For example, domain specific includes data mining '$DM$', image processing '$IP$' and cloud computing '$CC$' respectively. The design of MRC includes three entities, namely, user session, '$US$', request '$Q$', response '$S$' and their relationships. The relationship

between these three entities with respect to similar interest and domain specific web users is shown in the figure 3 given below.

**Figure 3Domain-specific and similar interests-based bipartite**

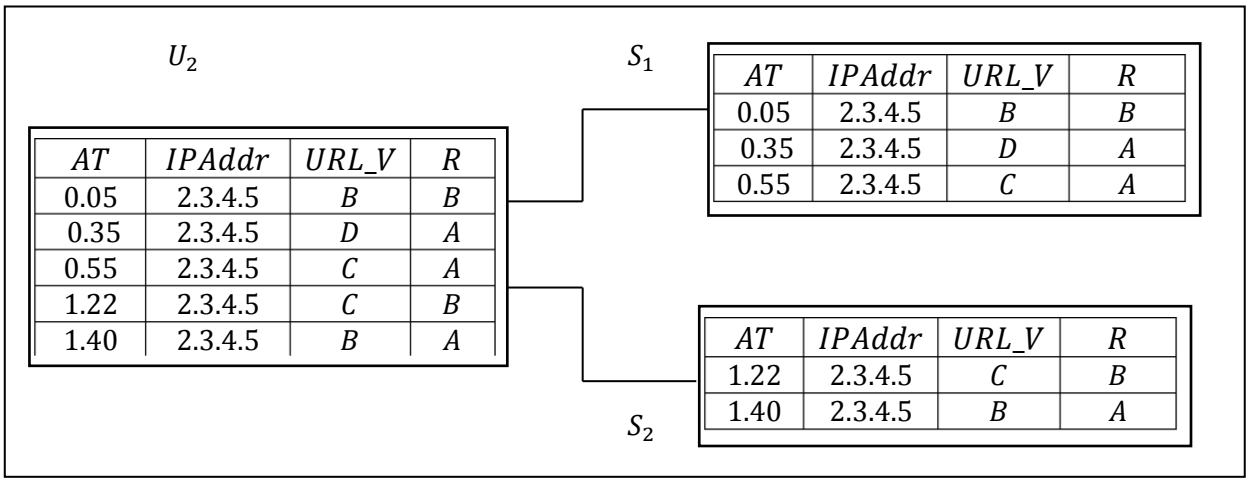



The mathematical representation of domain-specific and similar interests-based bipartite graph is mathematically formulated as given below.

$$BG = [US, Q, S, M_{USQ}, M_{USS}, M_{QS}] \quad (2)$$

From the above equation (2), the bipartite graph '$BG$' is evolved based on the user session '$US$', request made '$Q$' to extract the user profile, response obtained '$S$', '$M_{USQ}$' matrix representing the pair-wise edges between user session and request, '$M_{USS}$' matrix representing the pair-wise edges (i.e. considering domain knowledge or domain-specific and similar interests) between user session and response and '$M_{QS}$' matrix representing the pair-wise edges (i.e. considering domain knowledge or domain-specific and similar interests) between request and response. The resultant cluster user sessions is shown in figure 4 as given below.

**Figure 4Resultant cluster user sessions**




As shown in the above figure, by applying the mutual reinforcement principle for clustering the each user according to the domain-specific and similar interest pattern, the corresponding cluster user sessions are obtained. The equivalent pseudo code representation of Mutual Reinforcement Clustering is given below.

**Algorithm 1 Mutual Reinforcement Clustering algorithm**

| |
|---|
| **Input**: User Sessions ' $US = US_1, US_2, \dots, US_n$ ', web log file ' $f$ ', attributes ' $Attr = A_1, A_2, \dots, A_n$ ', set of all training examples '$Tr$' |
| **Output**: Cluster User Sessions '$CUS$' |
| 1: **Begin**<br>2: **For** each web log file '$f$' with user sessions '$US$' and training examples '$Tr$'<br>3: Measure information gain for an attribute using (1)<br>4: Express domain-specific and similar interests-based bipartite graph using (2)<br>5: Return (Cluster User Sessions according to domain-specific and similar interest pattern)<br>6: **End for**<br>7: **End** |

As given in the above Mutual Reinforcement Clustering algorithm, with training examples for corresponding web file given as input, the objective of the algorithm remains in clustering the user sessions according to the domain and similar interests with higher rate of accuracy. This is said to be achieved by initially performing the preprocessing by applying the information gain for the corresponding attributes. Next, domain-specific and similar interests-based bipartite graph is derived for returning the Cluster User Sessions with higher rate of accuracy.

**c. Summarizing cluster user sessions into user profiles**

Finally, in this section, with the cluster user sessions, summarization of user profiles according to the domain-specific knowledge and similar interests are obtained. This in turn helps in mining user profiles found to be relevant for Customer Relationship Management (CRM). Here, we start with the weight matrix for bipartite graph '$G = (DS, SI, M)$'. Here, each matrix '$M$' is represented by a row and column vector with row vector denoting the domain-specific knowledge '$DS$' and column vector denoting the similar interests '$SI$' respectively. According to the mutual reinforcement principle denoted in the above section, summarization of user profiles is performed via four simultaneous equations governing the similar interests '$y_{US}^{SI}$' and domain-specific knowledge '$y_{US}^{DS}$' for a specific user session '$US$' and the corresponding similar interests '$y_{SI}$' and the domain-specific knowledge '$y_{DS}$' generated. The four mathematical equations are given below.

$$y_{US}^{SI} = M_{(US)(SI)} y_{SI} \quad (3)$$

$$y_{SI} = M_{(US)(SI)} y_{US}^{SI} + y^{SI} \quad (4)$$

$$y_{US}^{DS} = M_{(US)(D)} y_{DS} \quad (5)$$

$$y_{DS} = M_{(US)(DS)} y_{US}^{DS} + y^{DS} \quad (6)$$

With the resultant values obtained from the above equation (3), (4), (5) and (6), varied user profile summarizes a group of users with similar access activities (i.e. according to domain-specific and similar interests) and consists of their pages or URLs being viewed, inquiring and inquired links and so on. The pseudo code representation of Mutual Reinforced- User Profile is as given below.

**Algorithm 2 Mutual Reinforced-User Profile**

| |
|---|
| **Input**: Cluster User Sessions '$CUS = US_1, US_2, \ldots, US_n$', bipartite graph '$G = (DS, SI, M)$' |
| **Output**: User Profiles |
| 1: **Begin**<br>2: **For** each Cluster User Sessions '$CUS$' with bipartite graph '$G = (DS, SI, M)$'<br>3: Generate simultaneous equations governing similar interests using (3)<br>4: Generate resultant similar interests using (4)<br>5: Generate simultaneous equations governing domain-specific knowledge using (5)<br>6: Generate resultant domain-specific knowledge using (6)<br>7: **End for**<br>8: **End** |

As given in the above Mutual Reinforced-User Profile algorithm, for each Cluster User Sessions with bipartite graph, the objective of the algorithm remains in tracking the user profiles

based on the domain-specific and similar interests present in the web log files. To start with, for each training samples, simultaneous equations governing similar interests and domain-specific knowledge is obtained. Followed by which the actual user profiles based on the similar interests and domain-specific knowledge is obtained. This in turn aids in discovering the patterns of web users from web data and therefore limits the discovered user profiles regarding certain subject or class of products.

## EXPERIMENTAL SETTINGS

In this section, the results of a series of experiments carried out to evaluate the effectiveness of proposed method and compare with other state-of-the-art methods are presented. Ranker Information Gain-Preprocessed Mutual Reinforcement Clustering (RIGP-MRC) for efficient mining of web access logsuses JAVA platform for the experimental work. The RIGP-MRC method uses the web server log files for the experimental work.

Web server log files consist of repository of information regarding web browsing performed by the users in the internet community. Mining on this data collected from web server log files can provide information to be of highly valuable that includes the web access patterns of users. When we apply clustering techniques in web usage mining environment, the web users access patterns with corresponding user profiles are said to be predicted in an efficient manner. The data used in this work contains web server maintaining a log that includes history of web page requests made by several users at different time intervals. Several information or attributes like, user's IP address, user name, requested web page, HTTP response code and number of bytes transferred are included.

In this work, NASA data set is used that is a standard dataset used for experimental purpose. The dataset is taken from NASA Kennedy space centres' www server in Florida (http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html) [2] and is available for free download. It consists of more than 10,00,000 entries. The log has the data collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days and the size of the file is 114 MB.

Each entry of the server log file is parsed to extract the required attributes such as the IP address/user name, date/time, requested web page, HTTP response code and number of bytes transferred. In order to evaluate the performance of the RIGP-MRC method, certain metrics are introduced to measure the web user profile analysis and compared with the existing methods namely, Web sessions clustering using Search Goal Shift [1] and hybrid sequence alignment measure (HSAM) [2]. The performance metrics including computational time, computation overhead and precision involved in accessing several navigation patterns are measured in the following sections.

## DISCUSSION S

The RIGP-MRC method compares with the existing work such as Web sessions clustering using Search Goal Shift [1] and hybrid sequence alignment measure (HSAM) [2]. The method, RIGP-MRC is experimented on factors such as computational time, computational overhead and precision.

In order to evaluate the web usage mining framework for mining user profiles, three groups of test are performed on the NASA Kennedy space center'sdata set. The first group of tests is on the effect of number of users to the computational time. The second group of test shows the effect of number of users to the computational overhead. Finally, the effect of precision to the number of users is revealed.

**1. Computational time in accessing navigation patterns**

The computational time refers to the time taken to access different navigation patterns at different time periods with respect to several users. Navigation pattern accessing time is referred as the time taken to access several navigation patterns with respect to different users. In this work, the navigation pattern accessing time is measured according to the time taken for retrieving corresponding similar interest users and time consumed for navigating the pattern according to the domain specific knowledge.

$$CT = \sum_{i=1}^{n} U_i * [Time\ (y_{DS}) + Time(y_{SI})] \quad (7)$$

From the above equation (7), the computational time '$CT$' refers to the time taken to obtain user profiles with domain specific '$Time\ [y_{DS}]$' and similar interest users '$Time\ [y_{SI}]$' with respect to the users considered for experimentation '$U_i$'. It is measured in terms of milliseconds (ms). The sample calculations for measuring computational time are evaluated as given below followed by which, the graphical representation is provided.

**Sample calculations**

- **Proposed RIGP-MRC**: The time consumed for obtaining domain specific user profiles was found to be '$0.025ms$', time consumed for obtaining similar interest user profiles was found to be '$0.013ms$', then the overall computational time involved is measured as given below.

$$CT = 15 * [0.025ms + 0.013ms] = 0.57ms$$

- **Search Goal Shift**: The time consumed for obtaining domain specific user profiles was found to be '$0.028ms$', time consumed for obtaining similar interest user profiles was found to be '$0.015ms$', then the overall computational time involved is measured as given below.

$$CT = 15 * [0.028ms + 0.015ms] = 0.645ms$$

- **HSAM**: The time consumed for obtaining domain specific user profiles was found to be '$0.032ms$, time consumed for obtaining similar interest user profiles was found to be '$0.018ms$', then the overall computational time involved is measured as given below.

$$CT = 15 * [0.032ms + 0.018ms] = 0.75ms$$

**Figure 5: Computational time**

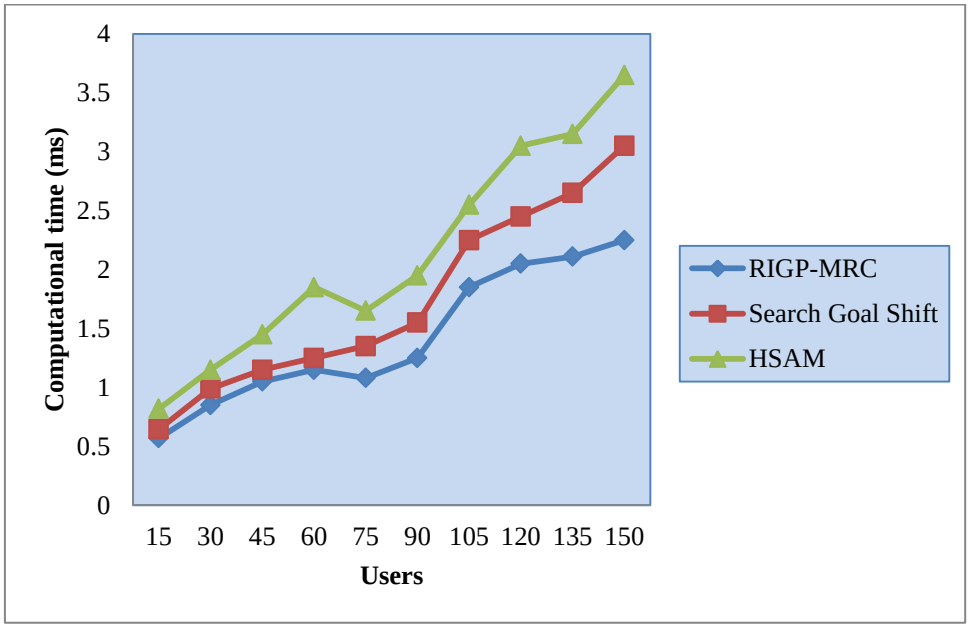



Figure 5 given above shows the convergence plot of computational time with respect to different number of users. Here, the computational time refers to the time consumed for generating user profiles with similar interests and corresponding domain-specific knowledge. With the increase in the number of users the time for obtaining user profiles with similar interests and domain-specific also increases. As a result, the overall computational time for navigating different access patterns of users also increases. Also a linear increase is found to be observed. Besides, from the sample calculation provided above, with '15' users considered, the time consumed for user profiles using RIGP-MRC method with similar interests was found to be '$0.025ms$' and domain-specific knowledge was found to be '$0.013ms$' and therefore the overall computational time was found to be '$0.57ms$'. In a similar manner, the time consumed for user profiles using Search Goal Shift method with similar interests was found to be '$0.028ms$' and domain-specific knowledge was found to be '$0.015ms$' and therefore the overall computational time was found to be '$0.645ms$'. Finally, the time consumed for user profiles using HSAM method with similar interests was found to be '$0.032ms$' and domain-specific knowledge was found to be '$0.018ms$' and therefore the overall computational time was found to be '$0.75ms$'. From the sample calculations, it is inferred that the overall computational time was found to be reduced by applying RIGP-MRC method. This is because by applying Ranker Information Gain Preprocessing model, dimension reduction is said to

be performed with the given input dataset. This in turn extracts the most relevant attributes for further processing. With the most relevant attributes extracted, similar interest with domain-specific user profiles are said to be extracted in a minimum time. As a result, the computational time using RIGP-MRC method is reduced by 16% compared to Search Goal Shift and 32% compared to HSAM.

2. **Computation overhead**

In this section, the performance of various computational overhead measurements for obtaining user profiles is analyzed. Two state-of-the-art web usage mining methods are analyzed, including, Search Goal Shift [1] and HSAM [2] method. The computational overhead '($CO$)' measures the memory required to perform web usage mining for mining user profiles. It is measured in terms of kilobytes (KB) and mathematically expressed as given below.

$$CO = \sum_{i=1}^{n} U_i * [Mem(y_{DS}) + Mem\ (y_{SI})] \quad (8)$$

From above equation (8), the computational overhead '$CO$' is measured with respect to the memory required for mining user profiles based on similar interests '$Mem(y_{DS})$' and domain-specific knowledge '$Mem\ (y_{SI})$' respectively. The sample calculations for measuring computational overhead are evaluated as given below followed by the graphical representation is provided.

**Sample calculations**

- **Proposed RIGP-MRC**: With the memory required for mining user profiles based on similar interests being '$13KB$' and domain-specific knowledge being '$15KB$', the overall computational overhead involved in navigating access patterns for several user profiles is given below.

$$CO = 15 * [13KB + 15KB] = 420KB$$

- **Search Goal Shift**: With the memory required for mining user profiles based on similar interests being '$18KB$' and domain-specific knowledge being '$23KB$', the overall computational overhead involved in navigating access patterns for several user profiles is given below.

$$CO = 15 * [18KB + 23KB] = 615KB$$

- **HSAM**: With the memory required for mining user profiles based on similar interests being '$25KB$' and domain-specific knowledge being '$31KB$', the overall computational overhead involved in navigating access patterns for several user profiles is given below.

$$CO = 15 * [25KB + 31KB] = 840KB$$

**Figure 6: Computational overhead**

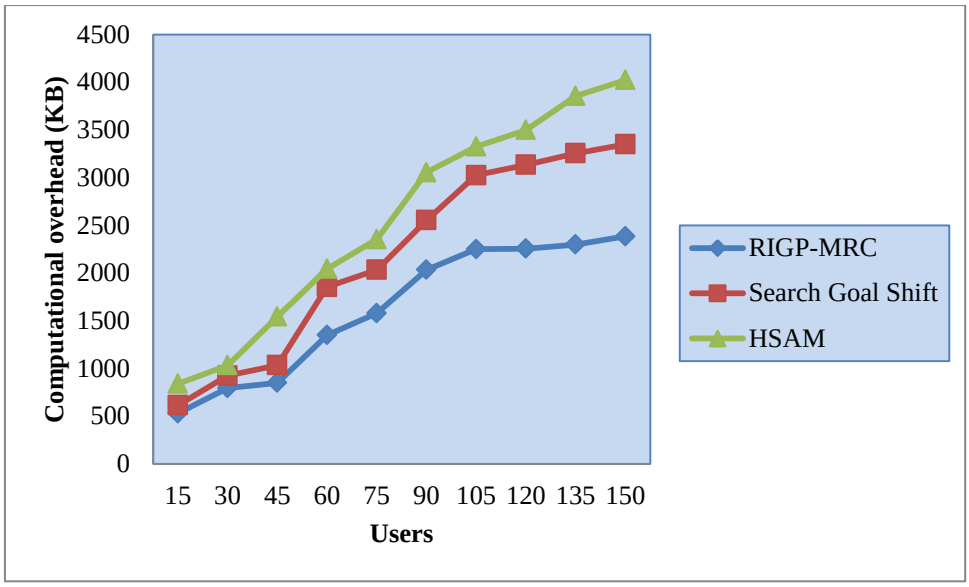



Figure 6 given above shows the comparison analysis of computational overhead with respect to different web users in the range of 15 – 150, using three methods, namely, RIGP-MRC, Search Goal Shift [1] and HSAM [2]. While considering '15' web users for carrying out the simulation work, RIGP-MRC utilized '$420KB$', whereas Search Goal Shift[1] and HSAM[2] utilized '$615KB$' and '$840KB$' respectively. As a result, computational overhead using proposed RIGP-MRC method is lower when compared to the other existing works[1,2]. However, by increasing the number of web users, the URL to be searched gets increased and hence the computational overhead also increases. But comparatively, the computational overhead of web usage mining for mining user profiles using proposed RIGP-MRC method is lower. This is because of the application of Mutual Reinforcement Clustering algorithm. By applying, Mutual Reinforcement Clustering algorithm with the NASA dataset given as input, only the relevant attributes is extracted using the Ranker Information Gain preprocessing model. Due to this, the attributes considered to be irrelevant are not considered for further processing. Hence, while extracting user profiles with specific interest and domain-specific knowledge, only the most relevant user profiles for Customer Relationship Management are mined. This in turn extracts the corresponding user profiles based on the bipartite graph. By incorporating bipartite graph in RIGP-MRC method, efficient web mining is said to take place. This assists for RIGP-MRC method in efficient mining of user profiles by training the data via deep Mutual Reinforcement principle. This in turn helps in reducing the computational overhead in a significant manner. Thus, RIGP-MRC method minimizes the computational overhead by 23% and 35% when compared to Search Goal Shift[1] and HSAM[2] respectively.

**3. Precision**

Precision refers to the relevant item retrieved. In this work, precision refers to the fraction of all the web user profiles retrieved that are relevant, considering domain-specific and similar interest profiles. Precision is mathematically represented as given below.

$$P = \sum_{i=1}^{n} \frac{U_i}{n} * 100 \quad (9)$$

From the above equation (9), the rate of precision '$P$' is measured as the ratio of web user profiles retrieved based on similar interests and domain-specific '$U_i$' retrieved that are relevant to the overall web users '$n$' considered for experimentation. The sample calculations for measuring computational overhead are evaluated as given below followed by the graphical representation is provided.

**Sample calculations**

- **Proposed RIGP-MRC**: With '15' number of web users considered for experimentation and '12' web profiles retrieved that are relevant according to domain-specific and similar interests, the precision is as given below.

$$P = \left[\frac{12}{15}\right] * 100 \; = \; 80\%$$

- **Search Goal Shift**: With '15' number of web users considered for experimentation and '10' web profiles retrieved that are relevant according to domain-specific and similar interests, the precision is as given below.

$$P = \left[\frac{10}{15}\right] * 100 \; = \; 66.66\%$$

- **HSAM**: With '15' number of web users considered for experimentation and '9' web profiles retrieved that are relevant according to domain-specific and similar interests, the precision is as given below.

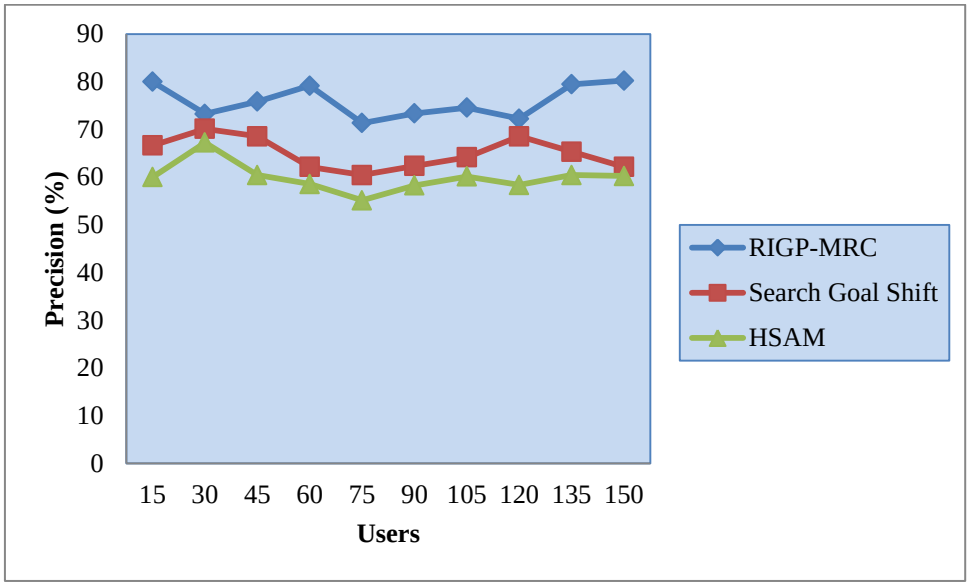
$$P = \left[\frac{9}{15}\right] * 100 \; = 60\%$$

As provided in the above sample calculations, with '15' web users considered for experimentation, '12' web profiles were retrieved that were found to be relevant using RIGP-MRC method, '10' web profiles were retrieved that were found to be relevant using Search Goal Shift and '9' web profiles were retrieved that were found to be relevant using HSAM respectively. Based on this resultant values, the graphical representation for rate of precision are provided below.

Figure 7 depicts the rate of precision average of three different methods providing web usage mining. It is found that the rate of precision of all the three methods decreases as the number of web users increases. When the web users are high there are more web users waiting in the queue for

placing their request in search of either web site or doing any social activities via web server, and this decreases the rate of precision. Search Goal Shift and HSAM provided minimum precision when number of web users increases than compared to RIGP-MRC.

**Figure 7: Precision**




The proposed method has maintained constant rate of precision throughout the simulated scenarios because while summarizing the user profiles based on the cluster user session, domain-specific knowledge and similar interests are considered separately according to the mutual reinforcement principle. It shows that based on the Mutual Reinforced-User Profile algorithm when the number of web users increases, rate of precision also gradually increased. But in case of Search Goal Shift and HSAM, when the number of web users increases precision rate is not gradually decreased, rather than it decreases abruptly. So that compared to existing Search Goal Shift [1] and HSAM [2], RIGP-MRC method improves the precision rate by 17% and 27% respectively. This result shows that RIGP-MRC method has an ability to sustain application performance even for large number of web users ensuring scalability.

## CONCLUSION

The mining of web for extracting user profiles is highly important for web site developers. The results obtained by extracting user profiles are found to be used in several areas, such as electronic commerce, security and crime investigation, electronic business, digital library and so on. In this study, the effect of mutual reinforcement clustering on Web usage mining is studied. By introducing the mutual reinforcement clustering, web usage mining algorithms are performed in terms of specific interests and domain-specific knowledge instead of web page addresses. The proposed method first performed preprocessing by applying Ranker Information Gain where relevant attributes for further processing were extracted with minimal computational time and overhead. With

the resultant preprocessed attributes, mutual reinforcement principle is applied for clustering user profiles according to specific interests and domain-specific knowledge of corresponding web users. The effectiveness of RIGP-MRC method is measured in terms of computational time, computational overhead and precision and compared against with state of the art works. With the simulations conducted for RIGP-MRC method, it is illustrative that the computational time is found to be less compared to state-of-the-art works. Besides, the simulation results demonstrate that the RIGP-MRC method provides better performance by minimizing the computational overhead and improving the rate of precision when compared to state-of-the-art works.

**REFERENCES**


1. Chao Ma, Bin Zhang, "A New Query Recommendation Method Supporting Exploratory Search Based on Search Goal Shift Graphs", IEEE Transactions on Knowledge and Data Engineering, 1 Nov. 2018; 30(11).
2. G. Poornalatha, S. RaghavendraPrakash, "Web sessions clustering using hybrid sequence alignment measure (HSAM)", Social Network Analysis and Mining, Springer, Jun 2013
3. Marios Belk, EfiPapatheocharous, PanagiotisGermanakos, George Samaras, "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques", The Journal of Systems and Software, Elsevier, Dec 2013; 86(12).
4. Hongyuan Cui, Jiajun Yang, Ying Liu, Zheng Zheng, Kaichao Wu, "Data Mining-based DNS Log Analysis", Annals of Data Science, Springer, Jan 2015
5. PetarRistoski, Heiko Paulheim, "Semantic Web in data mining and knowledge discovery: Acomprehensive survey", Web Semantics: Science, Services and Agentson the World Wide Web, Elsevier, Jan 2016
6. Gerald Petz, MichałKarpowicz, Harald Furschus, Andreas Auinger, Vaclav Stritesky,Andreas Holzinger, "Computational approaches for mining user's opinionson the Web 2.0", Information Processing and Management, Elsevier, Aug 2014
7. Michael Hahsler, and Matthew Bolanos, "Clustering Data Streams Based on SharedDensity between Micro-Clusters", IEEE Transactions on Knowledge and Data Engineering 1 June 2016; 28(6).
8. John A. Miller, Hong Zhu, and Jia Zhang, "Guest Editorial:Advances in Web Services Research",IEEE Transactions on Services Computing 1 Jan.-Feb. 2017; 10(1).
9. Couto, Ayrton Benedito Gaia do, Gomes, Luiz Flavio Autran Monteiro, "Multi-criteria web mining with DRSA", Information Technology and Quantitative Management, Elsevier, Mar 2016

10. Hyejin Shin, Sungwook Kim, Junbum Shin, Xiaokui Xiao, "Privacy Enhanced Matrix Factorization for Recommendation with Local Differential Privacy", IEEE Transactions on Knowledge and Data Engineering 1 Sept. 2018; 30(9).
11. Dilip Singh Sisodia, Vijay Khandal, Riya Singhal, "Fast prediction of web user browsing behaviours using most interesting patterns", Journal of Information Science, Nov 2016
12. Ujwal Gadiraju, Demartini, Ricardo Kawase and Stefan Dietze, "Human beyond theMachine: Challenges and Opportunities of Microtask Crowdsourcing", IEEE Intelligent Systems, IEEE Computer Society, Dec 2015
13. Ujwal Gadiraju, Gianluca Demartini Ricardo Kawase, Stefan Dietze, "Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection", Springer, Jun 2018
14. Muhammad Afzaal, Muhammad Usman, A. C. M. Fong, Simon Fong, and Yan Zhuang, "Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews", Advances in Fuzzy Systems, Sep 2016
15. Farek Lazhar, "Mining hidden opinions from objective sentences", International Journal of Data Mining, Modelling and Management, Jun 2018; 10(2).
16. Yu Wang, Wei Hou, Fusheng Wang, "Mining co-occurrence and sequence patterns from cancer diagnoses in New York State", PLOS ONE | https://doi.org/10.1371/journal.pone.0194407 April 26, 2018
17. S. K. Lakshmanaprabu,K. Shankar, Deepak Gupta, Ashish Khanna, Joel J. P. C. Rodrigues, Plácido R. Pinheiro, and Victor Hugo C. de Albuquerque, "Ranking Analysis for Online Customer Reviews of Products UsingOpinion Mining with Clustering", Complexity, The Hindawi, Jun 2018
18. Chengxi Huang, HongmingCai, Yulai Li, Jiawei Du, Fenglin Bu, and Lihong Jiang, "A Process Mining Based Service Composition Approach forMobile Information Systems", Mobile Information Systems, Jan 2017
19. Gianpiero Bianchi, Renato Bruni, and Francesco Scalfati, "Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms", Mathematical Problems in Engineering, Aug 2018
20. Ms. Dipa Dixit, Jayant Gadge, "Automatic Recommendation for Online Users Using Web Usage Mining", International Journal of Managing Information Technology (IJMIT) August 2010; 2(3).