User Identification, Classification and Recommendation in Web Usage Mining – An Approach for Personalized Web Mining

Priyanga P [1], Dr. Naveen N C [2]

[1] Assistant Professor, CSE, VTU/K.S. Institute Of Technology,
Bengaluru, Karnataka / 560078, India

[2] Professor, CSE, VTU/K.S. Institute Of Technology,
Bengaluru, Karnataka / 560078, India

Abstract

In recent years, Web Analytics (WA) is turning out to be an emerging research topic due to the extensive advancements in the techniques that aid in accessing the web contents, which millions of people have shared on the web. The information that has connection to the theme being searched may not be recognized always, if the personalization system operates in accordance with the usage-dependent outcomes alone. In this research work a new method is introduced for Personalized Web Search system, wherein, the users are enabled to have access to the relevant web pages as per their choice from the URL list. The first stage of this research deals with Semantic Web Personalization, which provides a merging between the content semantics as well as the usage data that are stated as ontology terms. This system supports the computation of the navigational patterns that are semantically improvised, so that constructive recommendations can be successfully engendered. It can be perceived that no other systems excluding the semantic web personalization system described here is employed in non-semantic web sites. The second stage of the work is to assist in augmenting the quality of the recommendations depending on the structure lying beneath the website. Finally, the testing is achieved through the utilization of a prolonged database link. The analysis of variation that exists among the different classes of parameters is made later, when the privacy is formulated using the memory usage and the period of execution.

Keywords: Web analytics; Semantic Web; Personalization

1. Introduction

The World Wide Web (WWW) refers to a massive and eminent repository for documents [1]. Currently, the total count of the online websites has exceeded 150 millions. The World Wide Web has emerged as the most influencing arena, where majority of the valuable information can be incorporated, published or retrieved [2] [3]. Researches pertaining to the analysis of Web data are posed with severe issues, since the Web has its own associated traits like considerably larger size, nature of being dynamic as well as diverse and not well-structured. At this juncture, the Quality of Service (QOS) is of much concern. A web developer should be familiar with the user's actual activity, envisage the pages that capture the user's attention a lot and offer the user with the desired Web pages through the understanding of the navigational patterns of the user for guaranteeing enhanced quality of service, in addition to rising the number of user's clicks on a certain website [4] [5]. All these steps are together known as Web Personalization. As per the history, a number of names that include information harvesting, data mining, information discovery, data archeology and knowledge extraction were used to denote the meaningful data extraction. Etzioni was the one, who coined the term web mining that has association with the withdrawal of information from the Web [6] [7].

Web Mining is a research domain, which deals with the withdrawal of more fascinating and beneficial patterns that are inherent within the World Wide Web (WWW). Normally, Web mining falls into three categories and they are the Web structure mining, the Web usage mining and the Web content mining. The mining process with which the significant information from the document contents in Web can be extracted is known as Web content mining [8] [9]. On the other hand, the Web usage mining lets the user patterns of interest to be explored from the Web, so that the user patterns can be made familiar and the applications relying on the Web can be rendered with better service [10]. The user's personal characteristics and the browsing actions pertaining to a specific website will be recorded from the onset of browsing. The organization of Web can be deemed as a Web graph, in which the Web pages constitute the nodes and the edges that provide a link between two relevant pages serve as the hyperlinks. Of all the three Web mining kinds, Web usage mining (WUM) along with Personalization make up the attractive research domains. In addition, the contents that reside in a Web page too have a tree structure in accordance with the XML and HTML tags contained in a page. As a consequence, the process involved in the finding of Web structure information can be thought as the Web structure mining [11].

WUM supports plenty of applications like, system enhancements, personalization, business intelligence and site alterations. This work is mainly concentrated on personalization through Web Usage Mining due to the following reasons [12]. As a usual case, the users hunt the necessary information on the Internet via making a move between the pages in line with the Web links. Hence, this requirement of information gain induces the need of several surfing patterns [13] [14]. Personalization of content that is conveyed to the user is of more importance because of the vast advancements in the systems. In the past, personalization has been subjected to enormous modifications. Yet, its primary objective that involves offering the user with their requirements or envisaging the user's requisite before they demand is left unaltered. Normally, personalization implies the ability of introducing changes to the services, products and information or otherwise, it alters the information depending on the service or the product [15] [16]. This extensive domain encloses additional systems like customization as well as adaptive websites and the recommender systems as well. The various key elements included in Web Personalization are the establishment of a match between the pages and the users, modeling among the users and so forth.

The final issue is related to the means through which each and every user can be kept in contentment. Personalization can satisfy this requirement. However, it is not that very much easy because ample techniques that may vary between the uncomplicated database views and software agents as well as collaborative filtering algorithms are engaged in Personalization [17] [18].

The Web Personalization is assumed to succeed, if the group of people engaged in personalization considerably progresses the technology. Moreover, web personalization has grown to be a basic requisite in most of the bulkier information services as well as the commercial spots in a very few years [19]. The idea behind personalization allows the contents in the site to be modified in a semi-automatic or automatic sense, so that the site can be personalized and a particular set of users are enabled to have control over the contents in the site as per their wish. Other advantages granted through the Personalization of the Web takes account of the formation of fresh index pages, development of the intended advertisements, product encouragements and user recommendations [20].

2. Related Work

Mamoun A. Awad and Issa Khalil [21] have made an effort, which aids in forecasting a group of Web pages that that the user is going to search next via the information regarding the pages viewed before. For achieving Web prediction, they have made an investigation on the Markov model as well as the all-Kth Markov model. Additionally, they have put forth a novel modified Markov model for getting rid of the problem associated with scalability in the total count of the paths. They have also dealt with an innovative structure for handling the two-tier prediction problem. This structure has made use of the training examples and the classifiers produced to form an example classifier (EC). It was evident from the results that the improvement in the duration spent for prediction was as good as the improvement in accuracy. The results also reveal that the accuracy was found to increase, when the higher orders of all-Kth model was utilized.

Tomas Arce *et al.* [22] have suggested that it is unavoidable to incorporate the user's action, while the web is redesigned, for ensuring more effective service. The user's activities on the Web can be made available any time through the maintenance of a Web log file, which notes the details of the user's visits in a partial manner. There method was a heuristic one that relies on simulated annealing to tackle the issues related with the formation of sessions. Their scheme has brought about a decrement in the processing period for about 166 times than the time spent in the integer-programming model. The heuristic solution achieved with their approach was found to discover fresh optimum values as well.

Nishad Deshpande *et al.* [23] have suggested an e-commerce approach for Intellectual Property Rights. The quick progression in e-commerce has lead Internet to be a more significant means for promoting advertisements. The innovations in the technological background of Internet have supported the advertisers to capture the minds of a specific group of targeted viewers, which was really infeasible through the conventional media. Hence, the advertisers are very much benefitted through the Internet to attract the audience depending on several decisive factors. But, Intellectual Property Rights (IPR) assists in safeguarding the technologies through numerous ways. The patents become necessary, when safety has to be ensured at times the business methodologies and the progression in information technology are merged. Patent rights permit a person not to practice an already claimed invention for a short duration until the expiry of patent's life. Practice refers to manufacturing, employing or selling a invention. Patents frequently serve as the basis for exposing the technology once or more than once and in addition, play a major role as the significant techno-legal documents that enclose inestimable number of data kinds having connection with technology or its enrichments. Here, they have employed patents to know the way the business has got affected using ICT developments. They have also dealt in detail about the ICT mechanism with which the internet dependent targeted advertising was accomplished.

Athanasios Papagelis *et al* [24] have presented a technique, which depicts the activity of the user as a top-down methodology during the communication with the Web. The random-surfer model that acts as the central piece of Google's Page Rank may be stated as the most familiar instance of this technique. The fraction of the number of web surfers, who are examining a particular web page, can be assessed in this model through the utilization of the linking arrangement of the Web. They have also presented a bottom-up scheme for making an investigation on the web dynamics in relation to the Web data, which the end users have browsed, gathered, tagged and arranged partially. A wide range of experiments was carried out for illustrating the merits and the demerits of the proposed scheme along with the attributes of user created web data, which have association with the quality as well as the quantity. Moreover, the experimental outcomes were compared against those of the conventional approaches. At the end of their paper, they have given a discussion about the mode of incorporating their scheme with PageRank. With this integration, it is possible to produce a novel kind of page-ranking algorithm that unites the user's interest and the link analysis exclusively.

Limeng Cui *et al.* [25] have come up with an idea that the single-class SVM can be exploited in news recommendation systems. The testing of this system employs the Dot Net software. The Web pages from Sogou Labs were initially pre-processed. Each and every single Web page contains its own intrinsic domain and for all the domains, the construction of single-class SVM models was achieved. As the next step, the user interest models for the entire number of users have been built according to the investigation on their search histories. A comparison is made between the user interested model and every single domain model to locate the domains on which the user shows increased interest. At last, the single-class SVM model is realized through the utilization of the Web pages associated with the domains and the search history of the users. The single-class SVM enables the computation of the Web pages that are highly related to the user's need and the user is recommended with those Web pages. In this algorithm, the single-class SVM makes the computation of the likeness that exists among the web pages and the user's interest. On the other hand, the accuracy of the outcomes can be enriched further with the application of the hierarchical model. The results reveal the better functioning of this algorithm.

While summing up this section, it is clear that the proposed research methodology aims to attain highly successful Personalized Web mining through the consideration of processes like detection and categorization of the account user as well as ensuring the quality of the recommendations.

3. Research Methodology

The proposed work aims to achieve the following.

- To gather and pre-process the data available in the web in an effectual manner.
- To efficiently find the users on the Web in accordance with their content and Web utilization
- To make a classification of the web pages in relation to the user's choice.
- To become accustomed to the users' varying interests.
- To examine the influence of personalization on the access to Information Retrieval.
- To recommend information to the users in connection with their behavioral patterns.

Majority of the research works found in the literature has executed the identification process for making use of several classifier techniques to categorize the objects of Web Usage Mining using the sequential decision procedure. However, only a few works have been proposed for performing the non-stationary sequential decision process and most of them have resulted in poor model directionality. Naive Bayesian Classification algorithm has been employed to classify the interested users in the currently available systems. The models that have been derived lack improved precision and accuracy values. All these shortcomings can be tackled with the utilization of sequential pattern mining, which is capable of performing non-stationary sequential decision processes that support in achieving enhanced accuracy levels.

4. Web Usage Mining and Personalization

This paper attempts to attain an enriched accuracy level through the exploitation of recommendation technique in Personalized Web Mining. The steps involved in the presented system include pre-processing, information extraction, classification, recommendation or knowledge discovery and output evaluation in order. The preprocessing provides the input data in a more appropriate way. Collaborative filtering is employed in the information extraction phase to guesstimate the user's requisite with the help of their preferences. The content based recommendation allows the classification phase to be conducted and the sequential pattern mining is applied, while the pre-processed data is passed into the knowledge-based discovery step. The output file is yielded in the ultimate step and subsequently, the measurement of how far the output is accurate is carried out. Figure 1 portrays the architecture associated with the proposed scheme.

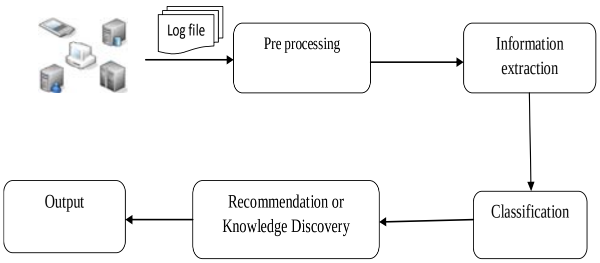


Figure 1: Architecture of the Proposed Work

4.1 Log File

Log files indicate a set of files, which the Web server manages to list down the activities that have already took place in the Web. Log files are made available from browser application, email and OS. The server admin alone has a frequent access to this file type.

Log files offer the details of the Web user such as, the path involved in navigation, the location of the Web user and the number of times the specific website has been visited. In a highly complex system, the improper functioning can be eliminated using the Log files. The log files make a correlation among the servers and trace the most prominent issues that exist among various systems demanding a single and significant solution to mend the entire number of issues. Assume a network $N = \{s_i \setminus i = 1,2,3.....n\}$. Each and every single system s_i possesses log files $LF = \{lf_1, lf_2......lf_j......lf_m\}$ where $j \in m$. The log files are grouped in the database $D = \{LF_1, LF_2......LF_p \setminus p > m\}$. These log files are then subjected to pre-processing for yielding a more dependable data mining process.

```
Input: log file
Output: cleaned log file
Step 1: Begin
        Read records in log file
Step 2:  For each record
         Read (status code, method)
Step 3:  If (status code= ‘40*’ and method= ’**’)
                Then, remove that status field
                Get IP address and URL link
Step 4: If (suffix_URL_link= {.gif, .jpg, .css, .av}
                        && request= “implicit”)
                Then, remove that URL_link
         Else
                Save IP address and URL_link
        End if
      End if
Step 5: If (status code! = ‘40*’ and method= ’**’)
                Then, Get IP address and URL link
        If (suffix_URL_link= {.gif, .jpg, .css, .av}
                        && request= “implicit”)
                Then, remove that URL_link
        Else
                Save IP address and URL_link
        End if
      End if
      Next record
  End for
End
```

Figure 2: Pseudo code for data cleaning

4.2 Preprocessing

Data pre-processing in Web Usage Mining is more often highly complicated. The preprocessing step aspires to grant a consistent, integrated and structural data source for use in the fore coming processes in the personalizing phenomenon. With data pre-processing, the data gets transformed into a format for enabling uncomplicated and efficient data processing. The ultimate function of pre-processing is to make a choice of the standardized data from the actual log files, in order to make it ready for applying the user navigation pattern discovery algorithm. In preprocessing, data cleaning and user identification are the two processes that are conducted.

4.2.1 Data cleaning

In the preprocessing stage, the superfluous references associated with the objects embedded and undesired fields are got rid. Few sound or style files and some graphics that have no connection with the Web page contents may be present in a web page. One such instance is the advertisement that describes about the weight losing procedure in the online shopping page. When a web document is downloaded in HTML format, these unnecessary objects also get downloaded and saved in the log file without the user confirmation for downloading from the Web page. Data cleaning plays a major role in eradicating such unrelated data. The HTTP status code field will be deemed as not required, when the status code does not succeed and the server then returns a three-digit status code. The status codes fall into three types, namely, Server Error (SOD Series), Redirect (300 Series), Success (200 Series) and Failure (400 Series). The most familiar failure codes are 403, 401 and 404 and they respectively imply the request that is forbidden to a confined subdirectory, authentication failure and the undetected file. Such irrelevant and unsuccessful entries are futile for making analysis and hence, swept away from the log files. Figure 2 elucidates the pseudo code for data cleaning.

4.2.2 User identification

Once the data cleaning in the log file has got completed, the identification of the user either through the IP address or the cookies has to be carried out. In this work, the IP address is employed to identify the user. The IP address as well as the browser from all the web pages in the log file is observed. If two entries have the dissimilar IP addresses, the user is believed to be a new one. On the other hand, if the IP addresses are found to be the same, the matching process will be executed with a user of the browser. A match between two browsers indicates the user as old. Otherwise, the user is considered as a new one and the total count of the users is incremented. Figure 3 reveals the pseudo code for identifying the user with the help of the IP address.

```
Input: cleaned log file
Output: Unique Users file
Step 1: Begin
          Initialize IP_List=0; Users List=0; Browser
          List=0; No-Of-users=0;
          Read Record From cleaned log file
  Step 2: For each page in log file
          Read Record.IP address and Record. Browser
  Step 3: If Record.IP address is not in IP_List
                  Then, add Record.IP address in to
                  IP_List
                  Add Record. Browser in to Browser List
                  No-Of-users++
                  Add new user in to User List.
             Else
                  If (Record.IPaddress is present in
                  IP_List and Record. Browser not in
                  Browser List)
                          Then, No-Of-users++
                          Add new user in to User List.
                  End If
          End for
End
```

Figure 3: Pseudo code for user identification

4.3 Information Extraction

The process, which assists in extracting the related information that resides in the documents, is called as information extraction. With information extraction, a document is examined to disclose the previously determined actions, elements or their connections. In Web mining, it points to the extraction of Web page contents according to the user's wish. Collaborative filtering makes use of the user's log file as well as the particulars in the input log file to achieve the extraction of information.

4.3.1 Collaborative Filtering

During the identification of the user, the information regarding the number of times the same user as well as the distinct users visits a browser is found. The collaborative filters use these details to operate. A user's preference over a certain item can be computed with the following equation.

$$pref(a,p) = r_a^{-} + \frac{\sum_{p\in N} sim(a,b) * (r_{b,p} - r_b^{-})}{\sum_{p\in N} sim(a,b)}$$

Where, $pref(a,p)$ stands for the user's preference on page p; r_a^{-} specifies the user's interest averaged over a particular page p; $sim(a,p)$ offers the correlation existing among the user a b; $r_{b,p}$ represents the user rating b for the item p and r_b^{-} points to the mean rating of user b. For a user a, the mean interest specifies the user's average interest over all the pages he/she visits.

$$r_a^{-} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where, N indicates the total count of the pages that the user a visits. Likewise, the mean interest of user b is given by the following expression.

$$r_b^{-} = \frac{1}{N}\sum_{i=1}^{N} X_j$$

Where, N points to the number of pages that the user b visits. Pearson correlation coefficient allows the computation of correlation coefficient that exists among two users. The linkage between the variables can be found with the help of the correlation technique. Pearson's correlation coefficient $sim(a,p)$ determines the degree of association between two variables as follows:

$$sim(a,p)=\frac{\sum_{p\in N}(r_{a,p}-\bar{r}_a)(r_{b,p}-\bar{r}_b)}{\sqrt{\sum_{p\in N}(r_{a,p}-\bar{r}_a)^2\sum_{p\in N}(r_{b,p}-\bar{r}_b)^2}}$$

Pearson coefficient offers the weighted relationship, which is decided with the user's interest on a specific page. Here, the weight is nothing but the coefficient that is calculated using the Pearson correlation coefficient. Collaborative filter renders each and every single user's preferences over a certain page. The Pearson coefficient that forms a portion of the collaborative filter gives the information pertaining to various users' preferences on a single page. All these outputs are used in content based filtering to make a Web classification.

4.4 Classification

The classification based on the content based recommendation is performed, immediately after the information extraction phase has got ended up. Classification can be assumed to belong to the category of supervised learning problem, wherein, the classifier is trained using a set of labeled data that can be later applied to future examples. The classification of the Web page can be achieved using the outcomes of the collaborative filter.

4.4.1 Content Based Recommendation

The content based recommendation systems concentrate more on the traits of the items and examines the report about the items for detecting the item, which the user prefers much. The likeness among the items can be assessed using the identical traits in those items. The pages with high ratings and low ratings are respectively categorized as highly recommended pages and lowly recommended pages. Each and every item in a content-based system should have its own constructed profile, which can be a single or multiple records that characterizes the features of an item. For a definite user, the collaborative filter allows the measurement of the preferences that are related to a specific page. These preferences are necessary to classify the pages into two types. The recommender system helps a new user to access the desired page using the recommendation it provides depending on the user's interest and the categorized groups. The identical nature that is present among the user's interest and the entries in the classified portion is computed with an overlapping coefficient, represented as O.

$$O=2\frac{|D\cap Q|}{\min(|D||Q|)}$$

Where, D denotes the pages with high user's interest and Q indicate the pages residing in the classified list A. If D is found to be the subset of Q or the reverse, then the overlap coefficient takes a value of one. If $O=1$, the user is believed to search the pages coming under the classified list A. Figure 4 concisely presents the pseudo code for content base recommendation.

```
Input: categorized log file
Output: log file based on user's interest
    Step 1: Begin
      Get user input $(i_1, i_2 ...... i_n)$
      Categorized data $(c_1, c_2 ...... c_n)$ //calculate
      overlap coefficient to find the similarity
      between I and C
    Step 2: $sim(i_j, c_j) = 2\frac{|i_j \cap c_j|}{\min(|i_j||c_j|)}$
              $O = sim(i_j, c_j)$
    Step 3: If $((i_j \subset c_j) || (c_j \subset i_j))$
              Then $O = 1$
           End if
       If $O == 1$
              Then information retrieval
       Else
              Denied
      End if
   Get next input data
End
```

Figure 4: Pseudo code for content based recommendation

4.5 Knowledge Based Recommendation

Knowledge discovery with respect to databases represents the process with which the meaningful information can be unveiled from a group of data. Knowledge discovery encompasses the processes that are dealt with preparing as well as the choosing of data, refining the data, including the known information with the data sets and inferring the precise solutions from the gained outcomes. Basically, the objective is to separate the knowledge at high-level from the lower level. KDD is a process involving multiple disciplines like the storing or making an access to the data, scaling the algorithms for yielding huge data sets and for deducing the results. The Knowledge-based recommender systems make use of unambiguous knowledge, which are related to the group of items, user's interest and principles of recommendation. Plenty of knowledge based discovery algorithms are available, but the one used in this work is the sequential pattern mining.

4.5.1 Sequential Pattern Mining using prefix span GSP

The approach to discover the related patterns that exist among the data, whose values are in a sequence form, is called as sequential pattern mining. Generalized Sequential Pattern algorithm (GSP) refers to a priori algorithm that performs mining. The term 'priori algorithm' implies a level-wise paradigm and it supports finding the items that are repeatedly occurring from the log file. It also points to the sites occurring only once in the database. Next, the items that do not occur so often are eliminated. Thus, the end result will have the recommended pages alone as per the user needs. The GSP initiates its working with the recommended pages. If the user makes numerous accesses to a specific Web site, the frequency of the respective page undergoes increment at every time of the site access. This will be categorized as the frequent item list. A non-frequent item list will be created, if the user does not have access to any of the web sites more than one time. Web mining takes the frequent item list alone in to account ad neglects the non-frequent item list.

Assume $I = \{i_1, i_2 .....i_n\}$ as the set containing the entire number of items. . An item set X is defined in such a way that it is alphabetically sorted I. Next, a sequence $s = \{X_1, X_2 ....X_s\}$ is defined and it points to a list of item sets that are arranged in ascending form with regard to the frequency of occurrence. Later, the target sequential data base $SDB = \{s_1, s_2 ....s_n\}$ is defined. From these sequences, two frequent sequences α and β are created. Frequent sequences refer to the sequences rendering large assistance, when compared against the user mentioned sub support. In our methodology, the user specified sub support is indicated as min_sup and its takes a preset value of 0.5. Consider $\alpha = \{X_1, X_2 ....X_n\}$. Two subsequences are produced from α . One subsequence serves as the prefix $prefix(\alpha, X_p) = \{X_1, X_2 ...X_{j-1} .....X_n\}$ and the other functions as the postfix $postfix(\alpha, X_p) = \{X_j, X_{j+1} ....X_p\}$. Here, $X_n \subset X_j$ and $1 \le j \le p$. At the condition j=ϕ , the prefix and the postfix will possess their value as null. The subsequences that resemble that of α are produced forβ also. Both α and β appear as projection and gets converted to the targeted output later. Figure 5 reveals the algorithm for generating the projection.

```
Input: log file with user recommended pages
Output: mined log file
    Step1: Begin
        Given $SDB = \{s_1, s_2 .......s_n\}$ , Where
        $s_i = \{X_1, X_2 ........X_n\}$
                Set min_sup=0.5
        For all s
    Step 2: Scan $X_j$ to find frequent items such that
            $1 \le j \le n$
        Find the support for each frequent item
        Generate SDB for $X_j$
    Step 3: Scan SDB ( $X_j$ ) to find frequent sequence.
        Find support Generate SDB ( $X_{j1}$) and scan to
        find frequent sequence
    Step 4: If (frequent sequence with a single item
                !=ϕ)
        Continue extraction of frequent sequence
        containing < $X_j$ > as prefix
        Until no frequency sequence with single item
        exist
      Else
    Step 5: Terminate extraction of frequent sequence
            containing < $X_j$ > as prefix
            Extract frequent sequences containing
            < $X_{j+1}$ >, < $X_{j+2}$ >, < $X_{j+3}$ > as prefix
        End if
                Output= {frequent sequences}
End
```

Figure 5: Algorithm for projection generation in prefix span GSP

5. Result

5.1. Data collection

In our research, the effectiveness of three ranking schemes is being compared depending on the pure similarity, plain Page Rank and weighted (personalized) URL Rank. The data set used here does not have a wider view, when comparing with the web portals. The visitors form two major groups, namely, the researchers and the students. Hence, further examination like the association rules are formulated to discover the following: (i) the visits made to access the course materials, (ii) the visits to numerous publications and the information regarding the researchers. Online processing of the contents has restricted us from utilizing the other publicly available web log sets. This is due to the fact that the information was gathered for years before and the sites we desire to visit may be sometimes unavailable. Further, the web logs of renowned web sites or portals that would support our experimentation in an efficient manner are deemed as private and the owners do not reveal them.

5.2. Evaluating Indicators

This section demonstrates the results of experimentation obtained after developing an algorithm for web mining of sequence patterns using a newly formulated measure. The synthetic URL datasets were utilized to perform the comparison between the proposed web-mining algorithm and the existing scheme.

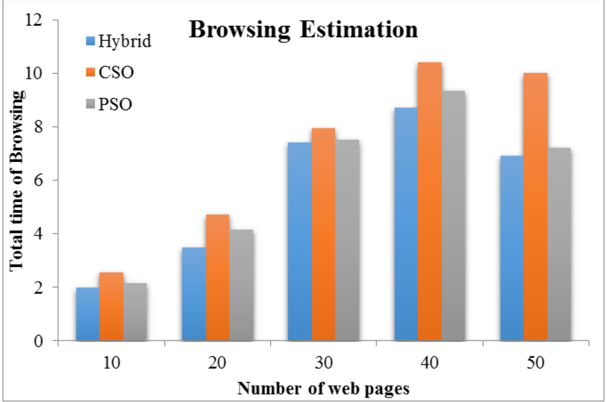
5.3. Evaluation Metrics

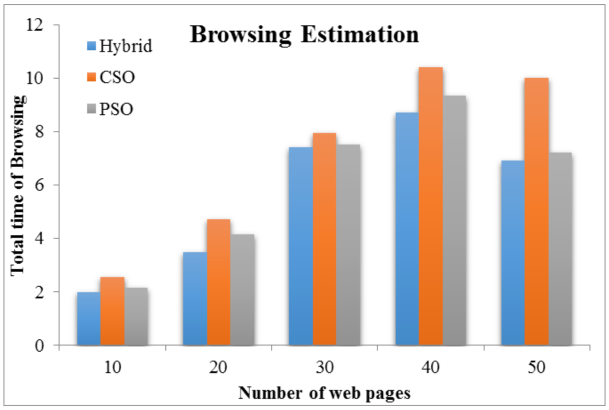
The gathering of the web logs files have spanned a period of over 6 months (01/01/10 – 30/12/14). Later to pre-processing, the size of the entire web logs was roughly 500 hits together with a set of over 50.000 dissimilar and unidentified user sessions on 2000 web pages. The session was carried out with a unique IP and time limit constraints. The maximum time between the successive hits from the same user has been preset as 20 minutes. Three evaluation metrics are employed to assess the effectiveness of the proposed event r-tree miner, which serves as an efficient scheme to mine the sequential patterns from the spatio-temporal event dataset. This paper introduces a framework to accomplish Semantic Enhancement for Web Personalization. This web personalization framework merges the content semantics and the users' navigational patterns based on ontology. Ontology is used to describe about the content as well as the usage of the web site. The accuracy associated with personalization is 85%, while for the random search it is just 69.6 %. It can also be seen that the personalization accuracy averages to 85.7%. This improvement in accuracy is dedicated to the use of user interested recommendations of the web page rather than making a random recommendation. Because the personalization in the interested domains is done in view of the user-selected domain, the accuracy is superior to the random recommendation in our experiment. Figure 2 to Figure 5 shows how the personalization accuracy of the interested domains that rely on random selection is compared against the accuracy produced in our personalization scheme.

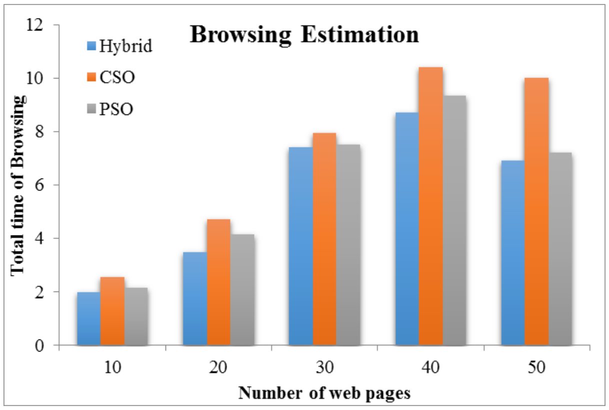
5.3. Experimental Design

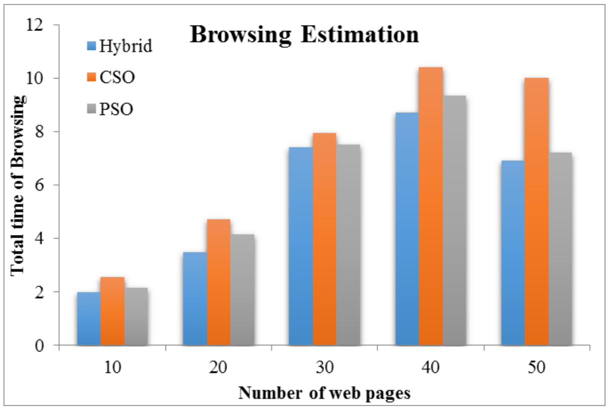
Table 1:

Session	Total records	Total users	Total Time
01-07-2014 - 29-07-2014	606	12	13:00
01-08-2014 - 30-08-2014	797	20	18:00
01-09-2014 - 30-09-2014	970	32	09:00
02-10-2014 - 30-10-2014	854	25	12:00
02-11-2014 - 30-11-2014	528	12	08:00









6. Conclusion and Future work

This paper has presented a novel and reusable software engine to induce a faster design of the applications, which are related to the personalized recommendation systems. When a user makes a search on the Web, the Web personalization engine offers the user with a list of desired domains along with a collection of personalized pages. Thus, the user is rendered the ability to switch between various interested domains. The importance of using web mining schemes in the web personalization domain has also been discussed in this paper. The future work of this research work is to bring about the enhancement in quality and to let an extension of this model to be developed. Improvements can be achieved, if the Web mining techniques are applied to Web applications involving Web contents, its exploitation as well as the structure. Investigating the architectures and the algorithms that permit mining of contents, usage and structure from diverse sources are found to be more valuable and may guide to the next generation of intelligent Web applications. The duration spent on searching the Web will be greatly reduced with this approach because the user is allowed to concentrate on the desired domains alone rather than going behind unrelated domains.

Acknowledgments

The Authors would like to thank VGST (Vision Group on Science and Technology), Government of Karnataka, India for providing infrastructure facilities through the K-FIST Level I project at KSIT, CSE R&D Department.

References

[1] M. Malarvizhi and S. A. Sahaaya Arul Mary,” Preprocessing using Web Log Data for Finding Frequent Patterns using Weighted Association Rule Mining Technique”, European Journal of Scientific Research, Vol. 74, No.4, pp. 617-633, 2012.
[2] Guangcan Yu, Chuanlong Xia, XingyueGuo,” Research on Web Data Mining and its Application in Electronic Commerce”, In proceeding of: Computational Intelligence and Software Engineering, pp. 1-3, 2009.
[3] JozefKapusta, Michal Munk, Martin Drlík, “User Session Identification Using reference Length”, DIVAI 2012 - 9th International Scientific Conference on Distance Learning in Applied Informatics, pp. 175-184, 2012.
[4] Ali Mirza Mahmood and Mrithyumjaya Rao Kuppa, “A novel pruning approach using expert knowledge for data-specific pruning”, Engineering with Computers, Springer, Vol.28, pp.21–30, 2011.
[5] Ling Chen, Sourav S. Bhowmick , Wolfgang Nejdl, “COWES: Web user clustering based on evolutionary web sessions”, Data & Knowledge Engineering, Vol. 68, pp. 867–885, 2009.
[6] Yongjian Fu Ming-Yi Shih,” A Framework for Personal Web Usage Mining”, In International Conference on Internet Computing IC’, pp. 595 -600, 2002.
[7] BalajiPadmanabhan ,Zhiqiang Zheng , Steven O. Kimbrough, ”Personalization from Incomplete Data: What You Don't Know Can Hurt “ In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 154-163, 2001.
[8] MagdaliniEirinaki and Michalis Vazirgiannis,” Web mining for web personalization”, Journal ACM Transactions on Internet Technology (TOIT), Vol. 3 No 1, pp.1-27, 2003.

[9] Kavita D. Satokar S. Z. Gawali, “Web Personalization Using Web Mining”, International Journal of Engineering Science and Technology, Vol. 2, No. 3, pp. 307-311, 2010.

[10] Jonathan P. Bowen and Silvia Filippini-Fantoni, “Personalization and the Web from a Museum Perspective”, in D. Bearman& J. Trant (Eds.), In Proc. MW2003: Museums and the Web Charlotte, USA, 19-22, 2003.

[11] Przemyslaw Kazienko1 and MaciejKiewra,” ROSA–Multi-agent System for Web Services Personalization”, AWIC , LNAI 2663, Springer-Verlag Berlin Heidelbergk, pp. 297–306, 2003.

[12] Ivan Marcialis and Emanuela De Vita,” SEARCHY: An Agent to Personalize Search Results”, The Third International Conference on Internet and Web Applications and Services, 2008

[13] Fang Liu, Clement Yu and WeiyiMeng,” Personalized Web Search for Improving Retrieval Effectiveness”, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, January 2004

[14] DimitriosPierrakos, Georgios Paliouras, Christos Papatheodorou, Constantine D. Spyropoulos, “KOINOTITES: A Web Usage Mining Tool for Personalization “, User Modeling and User-Adapted Interaction, Vol.13, No. 4, 2003, pp. 311-372.

[15] RajniPamnani, PramilaChawan, “Web Usage Mining: A Research Area in Web Mining”, Knowledge Discovery and Data Mining, WKDD 2009, Second International Workshopon, 2009.

[16] D. Backman and J.Rubbin, “Web log analysis: Finding a recipe for success”, Journal of Information Science, Vol. 26 No. 6 pp. 399-411, 2000.

[17] RenátaIváncsy, IstvánVajk,” Frequent Pattern Mining in Web Log Data”, ActaPolytechnicaHungarica Vol. 3, No. 1, 2006.

[18] Navin Kumar Tyagi, A.K. Solanki & Sanjay Tyagi, “An Algorithmic Approach to Data Preprocessing in Web Usage Mining”, International Journal of Information Technology and Knowledge Management , Volume 2, No. 2, pp. 279-283, 2010.

[19] J. Srivastava, P. Desikan, and V. Kumar, “Web Mining: Accomplishments and Future Directions,” In Proceeding. US Nat’l Science Foundation Workshop on Next-Generation Data Mining (NGDM), Nat’l Science Foundation, 2002.

[20] Dipa Dixit, MrJayantGadge,” Automatic Recommendation for Online Users Using Web Usage Mining”, International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, 2010

[21] Mamoun A. Awad and Issa Khalil, “Prediction of User’s Web-Browsing Behavior: Application of Markov Model”, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 42, No. 4, pp. 1113-1142, 2012.

[22] Tomás Arce, Pablo E. Román, Juan Velásquez and VíctorParada “Identifying web sessions with simulated annealing”, Expert Systems with Applications, Vol. 41, No. 2, pp. 1593–1600, 2014.

[23] Nishad Deshpande, Shabib Ahmed and AlokKhode “Web Based Targeted Advertising: A Study Based on Patent Information”, Procedia Economics and Finance, Vol. 11, pp. 522 – 535, 2014.

[24] Athanasios Papagelis and Christos Zaroliagis “A Collaborative Decentralized Approach to Web Search”, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 42, No. 5, pp. 1271-1290, 2012

[25] Limeng Cui and Yong Shi “A Method Based on One-Class SVM for News Recommendation”, Procedia Computer Science, Vol. 31, pp. 281-290, 2014.