# A Novel Approach to Analyse User Satisfaction Level On Web pages using Ontologies

**Dr. S. Chitra**

*Associate Professor, Department of Computer Science, Government Arts College (Autonomous), Coimbatore – 641 018, Tamilnadu, India*

---

**Abstract** - *Web access log analysis is to analyze the patterns of web site usage and the features of user's behavior. The proposed method constructs sessions as a Directed Acyclic Graph which contains pages with calculated weights. This will help site administrators to find the interesting pages for users and to redesign their web pages. After Session Construction a web usage analysis is used for finding the correlation between consumer emotions and buying behaviors. A semantic web usage mining technique is proposed for finding web access patterns from the annotated web usage logs. It includes consumer emotions and behaviors via self-reporting and behavioral tracking. To signify the real-time temporal concepts and requested resource attributes of periodic pattern based web access activities fuzzy logic is used. The consumer emotions and behaviors are integrated into a Personal Web Usage Lattice which represents the web access activities. From this we create Personal Web usage Ontology which facilitates semantic web applications. But the limitation is less efficient in terms of accuracy and user satisfaction level. So, in this manuscript an innovative technique is introduced which is called Optimum Session Interval based Particle Swarm Optimization(OSIPSO). This technique is used to find the optimum session interval. Additionally, an associative classification is used to enhance the level of accuracy. Associative classification is a combination of associative rule mining and classification rule mining. An experimental result shows that the proposed work achieves high accuracy and highly efficient in terms of user satisfaction level.*

***Key Words***: Session Construction, Directed Acyclic Graph (DAG), Robots Cleaning, Emotion and behavior profiling, ontology generation, semantic web, Particle swarm optimization, Associative classification

## 1. INTRODUCTION

This work is partitioned into 2 phases, namely

Phase I : Preprocessing of logs

Phase II : Analysis of user satisfaction level of the web pages.

Phase I : Preprocessing of logs

A web access log is a time series record of user's requests each of which is sent to a web server whenever a user sent a request. Web usage mining extracts regularities of user access behaviour as patterns, which are defined by combinations, orders or structures of the pages accessed by the internet. Web usage mining consists of three main steps:

- Data Preprocessing
- Knowledge Extraction
- Analysis of Extracted Results

Preprocessing is a significant step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase.

Graph and traversal are extensively used to model a number of classes of real world problems. For example, the structure of Web site can be modelled as a graph in which the vertices represent Web pages, and the edges correspond to hyperlinks between the pages [1]. Mining using graphs turns out to be a centre of interest. Traversals on the graphs are the models of User navigations on the Web site [2]. Once a graph and its traversals are specified, important information can be discovered. This paper provides a new version to the previous works by considering weights attached to the vertices of graph. Such vertex weight may reflect the importance of vertex. For example, each Web page may have different consequence which reflects the value of its contents.

Phase II : Analysis of user satisfaction level on the web pages.

Web usage mining is an automatic detection of patterns in clickstreams and related data collected as a result of user relations with one or more Web sites. The main intent of web usage mining is to examine the behavioral patterns and profiles of users interacting with a web site. The discovered patterns are generally characterized as collection of pages, objects or resources which are regularly accessed by groups of users with common interests. Human emotions are a significant factor of human behaviors in web mining analysis [3]. The relationships between consumer emotions and their buying behaviors have been well recognized [4] [5].

The self-report is used to integrate emotions into a personalized consumer profile and web access patterns. At the end of every web access request, users are asked to record the changes in their emotional state. This information is used to determine the emotional influence of the accessed resources for the users. Web usage mining [6] is one of the promising approaches to capture the users access pattern and to find frequent user access patterns by using web usage logs. A semantic web usage mining is an approach to correlate each requested webpage with one or more ontological entities to better understand the pattern of web navigation.

Previous web usage mining techniques [7] [8] focused on mining general access patterns, which have happened regularly within the entire duration of all access sessions. But the semantic web usage mining technique mines periodic access patterns, that frequently occurred in a specific period. By utilizing the periodic web access patterns of a user we easily analyze the frequently used resources at a particular time. Furthermore, the ontology will gather personal information on web access behaviours and habits and also the emotional influence of the accesses resources. The proposed method aims at mining semantics from semantically enriched web usage logs robotically and creates personalized web usage ontology for the Semantic Web. In the proposed work, Optimum Session Interval based on Particle Swarm Optimization(OSIPSO) is introduced to identify the optimum session interval. Particle swarm optimization is a technique which is used to discover the search space of a given problem to discover the settings or parameters essential to maximize a particular objective.

## 2. PREVIOUS RESEARCH

### Phase I:

Various commercially available web server log analysis tools are not designed for high traffic web servers and provide less relationship analysis of data among accessed files which is essential to fully utilize the data gathered in the server logs [9]. The statistical analysis introduces a set of parameters to describe user's access behaviors. With those parameters it becomes easy for administrators to define concrete goals for organizing their web sites and improve the sites according to the goals. But the drawback in this analysis is that the results are independent from page to page. Since user's behavior is expected to be different dependent on length of browsing time, the calculation of accurate browsing time is more important [10].

A labeled graph is a tuple $G = (V, E, \varphi)$, where $V$ is the set of vertices, $E$ is the set of edges and $\varphi: V \rightarrow L$ is a labeling function with $L$ a finite set of labels [9]. For an edge $(u, v) \in E, u$ is the parent of $v$ and $v$ is the child of $u$. If there is a set of vertices $\{u_1, \ldots, u_n\} \subseteq$ V such that $(u_1, u_2) \in E, \ldots, (u_{n-1}, u_n) \in E$, $\{u_1, \ldots, u_n\}$ is called a path, $u_1$ is an ancestor of un and $u_n$ is a descendant of $u_1$. There is a cycle in the graph if a path can be found from a vertex to itself. An edge $(u, v) \in E$ of the graph is said to be a transitive edge if besides the edge $(u, v)$, there also exists another path from $u$ to $v$ in $G$. A labeled DAG is a labeled graph without cycles. Let $D = \{D_1, \ldots, D_n\}$ be a set of labeled DAGs and $\epsilon \geq 0$ be an absolute frequency threshold. DIGDAG algorithm specifies that a DAG P is a frequent embedded sub-DAG of D if it is embedded in at least $\epsilon$ DAGs of D.

Duration time is the time that a user spends on reading a page in a session. Let $P_i$ and $P_{i+1}$ are two adjacent pages in a session. The timestamp field of $P_i$ is Ti, and of $P_{i+1}$ is Ti+1. Suppose $T_3$ is the loading time of $P_i$, and $T_4$ is the loading time ancillary files. By subtracting the time required for loading $P_i$ and the ancillary files from the time difference between the requests of $P_i$ and that of $P_{i+1}$, the duration time of $P_i$ can be calculated [11].

### Phase II:

Many of the ontology generation techniques have been investigated for extracting user access patterns.. These techniques predominantly focus on generating concept hierarchy for creating ontology.

Jian Pei et.al suggested mining access pattern from the weblogs [8]. A web access pattern tree is a new data structure is suggested for mining access patterns from weblogs. This tree accumulates compressed critical information for access pattern mining and also assists the development of new algorithms for mining access patterns in huge set of log pieces.

Gerd Stumme et.al proposed semantic web mining for analyzing the records of web usage [12]. The main motivation of the semantic web is to develop the current web by machine-processable information in order to facilitate for semantic-based tools supporting the human user.

Amalia Todirascu et.al design a prototype of a system for querying the web in natural language [13]. The semantic resources are used to filter the search and a data-driven methodology is adopted for resource acquisition.

Philipp Cimiano et.al proposed a new method that is based on the postulation that verbs pose strong selectional limitations on their arguments [14]. The conceptual hierarchy is then built on the basis of the inclusion relations between the extensions of the selectional limitations of all the verbs, whereas the verbs themselves presents intensional descriptions for each concept. After that, to formalize the design interms and this method is used to attain a concept hierarchy for the tourism domain out of texts.

Quan Thanh Tho et.al suggested a framework which is called Fuzzy Ontology Generation. This framework consists of the subsequent steps: Fuzzy Formal method Analysis, Fuzzy Conceptual Clustering, Fuzzy Ontology creation, and Semantic Representation adaptation [15]. This framework has also been used to create Machine Service Ontology for Semantic Ontology in this research.

Dai and Mobasher [16] used domain ontology to improve web usage mining for conventional web usage logs, but the mapping from requested URLs to ontological entities lacks reliability, particularly for dynamic websites. A framework for semantic enrichment of web usage logs by mapping each requested URL to one or more concepts from the ontology of the underlying website [17]. From the semantic enhanced weblogs, the specific user interests are clustered and pertain association rule mining to the semantically improved weblogs.

By using the ontology editor ontologies can be created manually. For ontology creation process, we use the combination of knowledge acquisition with machine learning techniques. For creating ontology, many methods are investigated. These methods contains Natural Language Processing (NLP) techniques [13], association rule mining [18], hierarchical clustering. On the other hand, these techniques focus primarily on creating concept hierarchies from text documents or relational databases.

## 3. PREPROCESSING

Data Cleaning, user identification, sessions construction are the steps in preprocessing.

### 3.1 Data Cleaning

The log format used in this method is Extended Common Log Format with the fields as follows: "ipaddress, username,password,date/timestamp, url, version ,status-code, bytes-sent, referrer-url, user-agent".

The removal process includes elimination of irrelevant records as follows:

- If the status code of all record is fewer than 200 and better than 299 then those records are eradicated.
- The cs-stem-url field is verified for its extension filename. If the filename has gif, jpg, JPEG, CSS, and so on they are eradicated.
- The records which request robots.txt are eradicated and if the time taken is incredibly little like less than 2 seconds are considered as automated programs traversal and they are also eradicated [8].
- All the records which have the name "robots.txt" in the requested resource name (URL) are recognized and straightly eradicated.

### 3.2 User Identification

Unique users are identified as follows:

- If two records has dissimilar IP address they are differentiated as two different users else if both IP address are similar then User agent field is verified.
- If the browser and operating system information in user agent field is dissimilar in two records then they are recognized as different users else if both are identical then referrer url field is checked.
- If URL in the referrer URL field in present record is not accessed before or if url field is blank then it is considered as a new user.

### 3.3 Session Identification

A user session is defined as a sequence of requests made by a single user over a certain navigation period and a user may have a single or multiple sessions during a period of time. The objective of session identification is to segregate the page accesses of each user into individual sessions. Reconstruction of precise user sessions from server access logs is a difficult task because the access log protocol (HTTP protocol) is status less and connectionless. There are two simple methods for session identification. One is based on total session time and other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [12] to 24 hours [19] at the same time as default time is 30 minutes by R. Cooley [11]. The second method depends on page stay time which is calculated with the difference between two timestamps. If it goes over 10 minutes the second entry is understood as a new session. The third method based on navigation of users through web pages. But this is accomplished by using site topology which is not used in our method.

## 4. SESSION DAG CONSTRUCTION

DAG construction phase has following tasks.

### 4.1 Calculation of Browsing Time

Real Browsing time is very difficult to calculate since it depends on network transfer rate, user's actions, and computer specifications and so on. Browsing Time and Request Time recorded in log are abbreviated as $BT$ and $RT$. Browsing time $BT_p$ of page 'p' is equal to the period of time with the time difference between the $RT_p$ of the request which include '$p$' as a reference and another $RT$ of the request which include '$p$' as a requested page. In the log record one of the fields is bytes_sent which is the size of the web page. '$c$' is the data transfer rate. So the real browsing time is assumed as

$$BT_p = BT_p' - \text{bytes\_sent} / c$$

where $BT_p'$ is the difference between reference and request page of 'p'.

### 4.2 Calculation of Weight of Pages

The second task in this method is to fix minimum and maximum browsing time for each page as $BT_{min}$ and $BT_{max}$ is used to calculate the weighing function which is to be used as a label in the graph. They are assumed by the administrators. The next step is to discretise the browsing time and given to each page as the weight which denotes the length of browsing time. Weighting function is calculated as follows

$Wt(p, BT_p) = 0$ when $BT_p \neq$ null and $BT_p < BT_{min}$

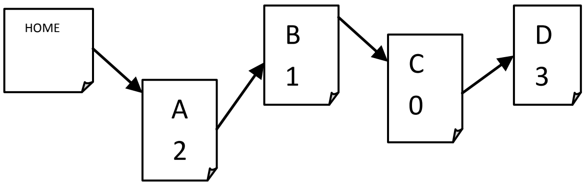
$Wt(p, BT_p) = 1$ when $BT_p \neq$ null and

$BT_{min} \leq BT_p \leq BT_{max}$

$Wt(p, BT_p) = 2$ when $BT_p \neq$ null and $BT_{max} < BT_p$ $Wt(p, BT_p) = 3$ when $BT_p =$ null

### 4.3 DAG Construction

Directed Acyclic Graph (DAG) is a tuple of Vertex, Edge and a label. After weighting all pages based on the browsing time a DAG structure is built for each user session. Vertex is labeled by a page and its weight. Each vertex is represented by a set of page and it's weight as (p, wt (p, BTp). Edge connects reference page to request page for each request. Edges show users page transition and only one direction is considered. In this method DAGs which give user session information for mining is constructed. The advantage over other graph methods is use of numerical values like browsing time is considered. A simplest form of the weighting function is used depending on the browsing time which is longer or shorter than the threshold. The threshold is based on the content of the page.




### 4.4 Pattern Extraction

Once a graph and its traversals are specified, valuable information can be retrieved through graph mining. Normally they are in the form of patterns. Frequent patterns which are sub traversals occurred in a large ratio are considered for analysis. To discover DAG's i.e., sub graphs DIGDAG mining algorithm is used which derive closed frequent sets. It replaces closed frequent DAG mining problem with the problem of closed frequent item-set mining on edges with the restriction that all the labels of the vertices in a DAG must be distinct. By the reconstruction of DAG structures from the mined closed frequent edge set, closed frequent DAG's are obtained. DIGDAG extracts the embedded DAGs based on not only on parent-child relationship but also ancestor-descendant relationship of vertices. The input for DIGDAG are the user session DAG set and the minimum support $\epsilon(\geq 0)$ as inputs. Access patterns are obtained as frequent DAGs.

## 5. ANALYSING USER SATISFACTION ON WEB PAGES

A Personal Web Usage Ontologies is created of individual users by using the semantically enriched web usage logs. This method follows the four steps. Create personal Web Usage Lattice, Creation of Global Web Usage Lattice, Generating Web usage Ontolgy and Geneartion of Personal Web Usage Ontologies.

### 5.1 Web Usage Lattice Creation

The web access session of user is defined as $S = < (URL_1, t_1), (URL_2, t_2), \ldots (URL_n, t_n) >$ is a succession of $URL_i$ with timestamp $t_i$. The user interest level of particular URL is nothing but time spent for a specific URL. The estimation of duration of a specific URL is $d_i = (t_{i+1} - t_i)$. Every

$URL_i$ in the user access session is related to set of resource attributes $M_{ri} \subseteq M_r$ for denoting the the semantics of the content in specific URL. The user access session is treated as a sequence of sets of resource attributes $M_{ri}$ instead of a sequence of individual $URL_i$ and it is denoted as $S =< (M_{r1}, t_1, d_i), (M_{r2}, t_2, d_2), .. (M_{rn}, t_n, d_n) >$. The estimation of level of interest is depends on the total duration for each resource attribute $m_k \in M_r$ during the user access session.

$$d(S, m_k) = \sum_{i=1}^{n} \alpha_{ki} d_i, \quad where$$

$$\alpha_{ki} = \begin{cases} 1, & if\ m_k \in M_{ri} \\ 0, & otherwise \end{cases}$$

$K = (G, M_p, M_r, I)$ is a fuzzy periodic Web Usage Context, where G denotes set of user access sessions, $M_p$ is set of periodic attributes, $M_r$is set of resource attributes, I is a fuzzy set of the domain to denote the associations between user access sessions and attributes. Each fuzzy relation $R(g, m) \in I$ is repreented by a membership value $\mu(g, m) \in [0,1]$ where,

$$\mu(g, m) = \begin{cases} \mu_p(g, m), & if\ m \in M_p \\ \mu_r(g, m), & if\ m \in M_r \end{cases}$$

Each user access session $g \in G$ can also be represented as a fuzzy set on the domain.

For a periodic attribute $m_p \in M_p$, the membership value $\mu_p(g, m_p)$ in a user access session $g \in G$ can be computed using the period of g. The membership function is defined as,

$$\mu_p(g, m_p) = \max_{t \in p(g)} \{\mu_p(t, m_p)\}, where\ \mu_p(t, m_p)$$

For a resource attribute, the membership value $\mu_r(g, m_r)$ in a user access session $g \in G$ can be computed using the total duration of $m_r$. The membership function is defined as,

$$\mu_r(g, m_r) = \begin{cases} 0, & if\ z(g, m_r) < \frac{1}{2} Z(m_r) \\ \frac{2z(g, m_r)}{Z(m_r)} - 1, & if\ \frac{1}{2} Z(m_r) \leq z(g, m_r) \leq Z(m_r) \\ 1, & if\ z(g, m_r) > Z(m_r), \end{cases}$$

Where $z(g, m_r) = \frac{d(g, m_r)}{te(g) - ts(g)} e^r\ and$

$$Z(m_r) = \frac{\sum_{g_k \in G} d(g_k, m_r)}{\sum_{g_k \in G} (te(g_k) - ts(g_k))}$$

$Z(m_r)$ is nothing but the proportion of the total duration of accessing the resource in all web access sessions of the user, which denotes the user's global interest of the resource. $Z(g, m_r)$ is the proportion of the duration of accessing the resource within the user access session g, weighted by an emotional influence factor $e_r$ derived from the consequent $\Delta E$ using one of the following rules:

Rule 1: -E: This is the baseline situation where $e_r = 1$; ignoring emotional influence.

Rule 2: $+E_1$ . If $\Delta E < 3$ then $e_r = 0.5$; otherwise $e_r = 1$; we only repress resources with negative emotional influence.

Rule 3: $+E_2$. $e_r$ is derived by using the formula, $e_r = 0.1\Delta E + 0.7$, thus the larger value is assigned to $e_r$ with increasing $\Delta E$, i.e., $e_r$ =0.8,0.96,1,1.1,1.2

$z(g, m_r)$ denotes the user local interest and emotional influence of the resource. For a given web usage context $K = (G, M_p, M_r, I)$ the set of common to user access sessions are defined as $A \subseteq G$ and the set of user access sessions are defined as $B \subseteq M_p$. The fuzzy support of set of attributes is defined as,

$$Sup(B) = \frac{\sum_{g \in B^*} (\mu_p(g) \times \mu_r(g))}{|G|},$$

The fuzzy confidence of $v(B)$ is defined as,

$$Conf(v(B)) = prob\left((B \cap M_r) \middle| (B \cap M_p)\right) = \frac{Sup(B)}{Sup(B \cap M_p)}$$

To create the global web usage lattice, the set of selected periodic attributes $M_p$ and resource attributes $M_r$ for all users. $W_G = \{B_k\}$ represents the set of all web access activities. $|W_G|$ is the total number of global web access activities. There should be a total of $\binom{a}{i} \times \binom{b}{j}$ global web access activities with i periodic attributes and j resource attributes, where $\frac{a}{i}$ and $\frac{b}{j}$ represent the number of combinations. Each web access activity has direct sub activities. The direct sub activity relationships are used to create a Global Web Usage Lattice.

### 5.2 Generation of Ontology

Ontology includes a taxonomy with a set of inference rules. The expression of taxonomy is a set of domain concepts and the associations among them. The Global Web Usage Ontology can be created by using class and hierarchy mapping, and property mapping. A set of activity classes is initiated by the class and hierarchy mapping and the activity class hierarchy is build based on the Global Web Usage lattice. Web ontology language has more advantages which does not change the original ontology and sustains incremental ontology creation and easily update the activity class hierarchy.The personal web usage ontology is generated by combining the personal web usage lattice of a user with the global web usage ontology by using the concept of instance mapping. Instance mapping creates a set of activity instances from the corresponding activity classes in the global web usage ontology and an activity instance hierarchy.

Finally, the user satisfaction level is determined. The satisfaction for the overall web personalization is defined as,

$$Satisfaction = \frac{\sum_{PR_i \in PR_a} satisfaction(PR_i)}{|PR_a|}$$

Where, $PR_a$ is a subset of personalized resources. $PR_i$ is set of personalized resources. The evaluation of satisfaction is used to measure how well the user is interested in the personalized resources.

### 5.3 Optimum Session Interval based on Particle Swarm Optimization (OSIPSO)

In the previous method the accuracy is less in terms of user satisfaction level. So, in this article Optimum Session Interval based on Particle Swarm Optimization (OSIPSO) is introduced. The session interval chosen in this method is very significant for improving the accuracy. The optimum session interval is identified by using particle swarm optimization algorithm. The particle swarm optimizations is a computational method which optimizes a problem by continuously trying to enhance a candidate solution with regard to a given measure of quality. In every iteration process, each candidate solution is calculated by the objective function being optimized, deciding the fitness of that solution. Every particle preserves its position, composed of the candidate solution and its evaluated fitness, and its velocity. Furthermore, it considers the best fitness value it has accomplished thus far during the process of the algorithm, referred to as the individual best fitness, and the candidate solution that achieved this fitness, referred to as the individual best position. At last, the PSO algorithm maintains the best fitness value accomplished among all particles in the swarm, called the global best fitness, and the candidate solution that achieved this fitness, called the global best position or global best candidate solution.

The PSO algorithm includes three major steps:

1. Compute the fitness of every particle
2. Update individual and global best fitnesses and positions
3. For every particle update velocity and position

Algorithm 1: Optimum Session Interval based Particle Swarm Optimization (OSIPSO) algorithm

1. Initialize N number of particles in the swarm, each particle having a position and velocity . Let be the best known position of particle i and gbest is the best known position of the entire swarm
2. Initialize the particle's position
3. For each particle i=1,2,…N
4. Calculate fitness value for every particle
5. //For fitness calculation satisfaction is taken from the previous algorithm
6. $Fitness = \frac{Satisfaction}{Number\ of\ user\ sessions \times Number\ of\ particles}$ // Computatuion of fitness
7. If fitness value is better than the best fitness value (pBest)
8. Set current value as the new pBest
9. Until a termination criterion is met
10. Select the particle with best fitness value of all particles as the gbest
11. For every particle
12. // Calculation of particle velocity
13. $V_i(t+1) = wv_i(t) + c_1r_1[\hat{x}_i(t) - x_i(t)] + c_2r_2[g(t) - x_i(t)]$ //where, the index of the particle is represented by i, is the velocity of particle i at

time t, is the position of particle i at time t, parameters w, c1, and c2 are coefficients

14. Update particle position
15. $x_i(t+1) = x_i(t) + v$
16. Until some stopping condition is met

### 5.4 Associative Classification Technique

In this work, the combination of both associative and classification rule mining, which are applicable for practical applications, are used. By combining these two techniques a new method is proposed which is called associative classification to improve the accuracy.

In classification there are two phases: (1) training data (2) testing data

The training phase contains a periodic association access pattern of the users with support and confidence values. So, in the training data, already we have the set of emotional behavoiur of particular resources for the users in every sessions. The fuzzy support and confidence denote the quality of such periodic association access patterns. From this we obtain the number of users interested in particular resources in a particular session. In the testing phase, a test data is given for identifying the accuracy level. In this phase, the obtained periodic association access pattern is matched with training data and evaluate the accuracy.

Algorithm 2:

Input:Input dataset

Output: No. of Users interested in particular resource at a particular session

// Training data

1. Generate periodic web access pattern
2. $Sup(B) = \frac{\sum_{g\in B^*}(\mu_p(g)\times\mu_r(g))}{|G|}$ // Fuzzy support value
3. $Conf(v(B)) = prob\left((B\cap M_r)\middle|(B\cap M_p)\right) = \frac{Sup(B)}{Sup(B\cap M_p)}$ // Fuzzy confidence value
4. Evaluate number of users interested in particular resource at a particular session

// Testing data

5. Form test data
6. Generate periodic web access pattern
7. Computation of support value
8. Computation of confidence value

Match with the training data and evaluate the accuracy

## 6. EXPERIMENTAL RESULTS

Phase I:

The experimental results is evaluated for the existing web usage ontology generation (WUOG) approach and the proposed Optimum Session Interval based Particle Swarm Optimization(OSIPSO) including associative classification. The raw web server log data were acquired from a web forum at Nanyang Technological University, Singapore. The web forum contains seven main topics and 57 subtopics. Access data of the top 50 users were used in the experiments.

Table-I: The Processes and Results of Data Preprocessing in Web Usage Mining

| Number of records in raw web log | Number of records after data cleaning | Number of users | Number of session construction using DAG |
|---|---|---|---|
| 747890 | 112783 | 55052 | 57245 |

Table 1 show that after data cleaning, the number of log data diminished from 747890 to 112783.

Four samples from the same web forum are obtained to evaluate the cleaning phase. From Figure-1 it is confirmed that the unwanted and irrelevant records are cleaned.

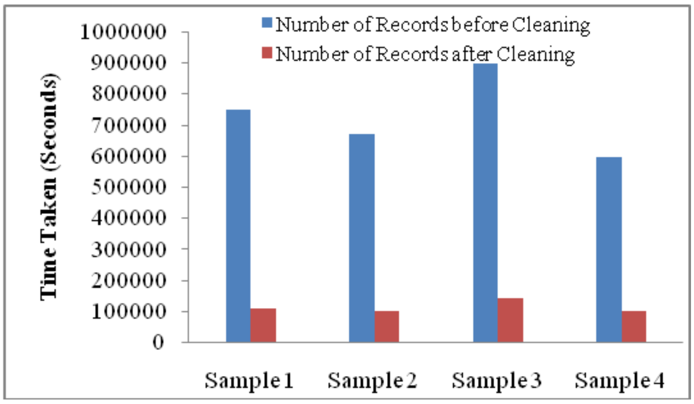



Figure-1: Data Cleaning of Sample Records

Table-II: User Session Identification by using Directed Acyclic Graph (DAG)

| IP Address | User id | Session id | Path Completed |
|---|---|---|---|
| 116.128.56.89 | 1 | 1 | 16-17-18-17-18-19-20 |
| 116.128.56.89 | 1 | 2 | 25-26-30-35 |

From Table-II, it can be observed that using the Directed Acyclic Graph (DAG) the user session is identified correctly. Finally, on the basis of user identification's results, 57245 sessions have been recognized by a threshold of 30 minutes and path completion.

Phase II: Satisfaction level

Satisfaction level evaluates how probably it is that a user is involved in one of the personalized resources in the period-supported sessions.

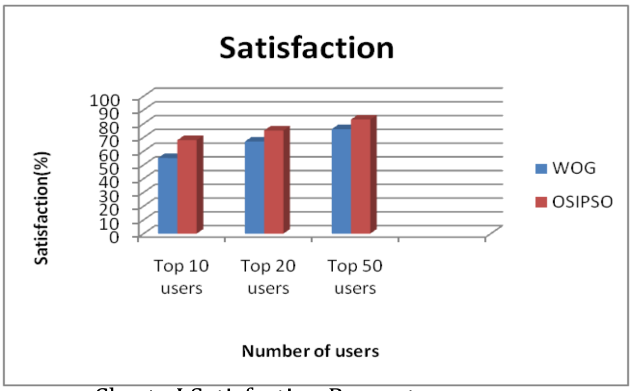



Chart –I Satisfaction Percentage

The satisfaction level is measured for the existing web usage ontology generation (WOG) approach and the proposed Optimum Session Interval based Particle Swarm Optimization(OSIPSO) including associative classification. Compared to the existing WOG method in the proposed method achieves high satisfaction level.

Accuracy

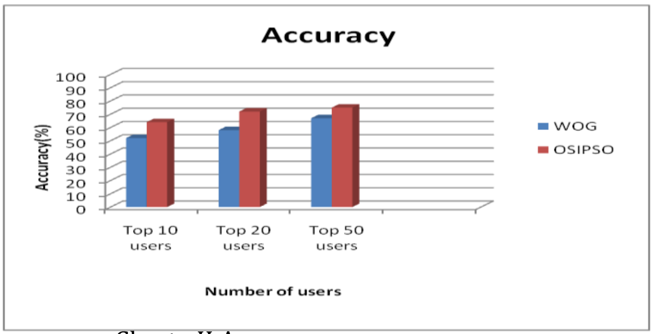



Chart –II Accuracy

The accuracy is measured for the existing web usage ontology generation (WOG) approach and the proposed Optimum Session Interval based Particle Swarm Optimization(OSIPSO) including associative classification. Furthermore an associative classification is used to evaluate the classifying accuracy. Compared to the existing WOG method in the proposed method achieves high accuracy.

## 7. CONCLUSIONS

Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. For this along with statistical analysis data mining techniques is to be applied in web log data. In this paper, proposed a method to analyze web logs in detail by constructing sessions as Directed Acyclic graphs Web site administrators follow the results and improve their web sites more easily. From the experimental results it is obvious that the proposed method successfully cleans the web log data and helps in identifying the user session.

A new method is presented for automatic generation of Personal Web Usage Ontology of periodic access patterns from web usage logs that have been semantically developed with information on emotional influence. From this the consumer web access behavior and emotional influence web resources is captured. This Personal Web Usage Ontology is utilized by software agents to give Semantic Web services. Furthermore, the user queries are processed by search engines and rerank the search results based on the user behavior which is obtained from Personal Web Usage Ontology. his method finds the consumer emotions in every session interval. If the session interval changes the accuracy level also changes. So, in order to find optimum session interval, Optimum Session Interval based on Particle Swarm Optimization(OSIPSO) is proposed. So, by using Particle Swarm Optimization algorithm the best session interval is to be chosen. Furthermore, the user queries are processed by search engines and rerank the search results based on the user behavior which is obtained from Personal Web Usage Ontology.

**REFERENCES**

[1] Seong Dae Lee, Hyu Chan Park, “Mining Weighted Frequent Patterns from Path Traversals on Weighted Graph “, IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4, April 2007.

[2] WANG Tong,HE Pi-Lian”Find Duration Time Maximal Frequent Traversal Sequence on Web Sites”, IEEE International Conference On Control and Automation , 2007.

[3] R.J. Dolan, “Emotion, Cognition, and Behavior,” Science, vol. 298, no. 5596, pp. 1191-1194, 2002.

[4] P. Weinberg and W. Gottwald, “Impulsive Consumer Buying as a Result of Emotion,” J. Business Research, vol. 10, no. 1, pp. 43-57, 1982.

[5] F. Piron, “A Comparison of Emotional Reactions Experienced by Planned, Unplanned and Impulse Purchasers,” Advances in Consumer Research, vol. 20, nos. 341-344, pp. 199-204, 1993.

[6] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” ACM SIGKDD Explorations, vol. 1, no. 2, pp. 12- 23, 2000.

[7] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, “Effective Personalization Based on Association Rule Discovery from Web Usage Data,” Proc. Third Int’l Workshop Web Information and Data Management, pp. 9-15, 2001.

[8] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, “Mining Access Patterns Efficiently from Web Logs,” Proc. Fourth Pacific-Asia Conf. Knowledge Discovery and Data Mining, Current Issues and New Applications, pp. 396-407, 2000.

[9] Cooley, R., Mobasher, B., and Srivastava, J. (1999). “Data preparation for mining World Wide Web browsing patterns”, Knowledge and Information Systems, 1999.

[10] Koichiro Mihara, Masahiro Terabe and Kazuo Hashimoto,” A Novel web usage mining method Mining and Clustering of DAG Access Patterns Considering Page Browsing Time”,2008

[11] Cooley, R., Mobasher, B., and Srivastava, J., “Web mining:Information and Pattern Discovery on the World Wide Web,”,International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[12] G. Stumme, B. Berendt, and A. Hotho, “Usage Mining for and on the Semantic Web,” Proc. NSF Workshop Next Generation Data Mining, pp. 77-86, Nov. 2002.

[13] A. Todirascu, F. De Beuvron, D. Gaˆlea, and F. Rousselot, “Using Description Logics for Ontology Extraction,” Proc. First Workshop Ontology Learning at the 14th European Conf. Artificial Intelligence, Aug. 2000.

[14] P. Cimiano, S. Staab, and J. Tane, “Deriving Concept Hierarchies from Text by Smooth Formal Concept Analysis,” Proc. GI Workshop Lehren-Lernen-Wissen-Adaptivit, pp. 72-79, 2003.

[15] T.T. Quan, S.C. Hui, A.C.M. Fong, and T.H. Cao, “Automatic Fuzzy Ontology Generation for Semantic Web,” IEEE Trans.Knowledge and Data Eng., vol. 18, no. 6, pp. 842-856, June 2006.

[16] H. Dai and B. Mobasher, “Using Ontologies to Discover Domain- Level Web Usage Profiles,” Proc. Second Semantic Web Mining Workshop at ECML/PKDD, 2002.

[17] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez, “Conceptual User Tracking,” Proc. First Int’l Atlantic Web Intelligence Conf., pp. 155-164, May 2003.

[18] A. Maedche and S. Staab, “Discovering Conceptual Relations from Text,” Proc. 14th European Conf. Artificial Intelligence, pp. 321-325, Aug. 2000.

[19] Spilipoulou M.and Mobasher B, Berendt B. “A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis,” INFORMS Journal on Computing Spring, 2003.