Web Usage Pattern Discovery using Contextual Factor: Credibility

Ms. Ravita Mishra
(Department of Information Technology, RAIT College, MUMBAI University, Mumbai -400 706

ABSTRACT

Here we present a new technique of pattern discovery technique based on analysis of contextual factors. The approach is based on Web classification algorithm and task classification algorithm. In order to find the discovered pattern, here we present a preprocessing and web page classification technique based on the Cs-Uri-Stem field of log file, in which we classifies the pages index as well as content pages. The result of classified web page via index or content will be used for further task classification algorithm. The task classification will be achieved by connecting cs-uri-stem and cs-uri-query field of log file. The Experimental result shows the cluster of user (session) and their task like casual and careful user. At the end result Analysis of contextual factor search interest and difficulty, credibility, page frequency and browser dependence is observed. Using this approach we can find other contextual factors easily and improve the performance of system. The result of finding is useful for adaptive services as well as site modification of web sites. Main task of this system is observation of the web log file manually and recommendations for the modification of the sites.

***Keywords* -** Contextual factors, Credibility, Task classification, Web usage mining, Web page classification**.**

I. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. Web usage mining provides the support for the web site design, providing personalization server and other business making decision.

It is important to analyze the user's web usage log for developing any personalized web services. The available web usage log has several difficulties and usage log show uncertain pattern. Pattern discovery is challenging task in web usage mining using contextual parameter analysis this techniques pattern will be simpler and used for on line help, suggestion given to new site developer , site improvement as well as modifying the structure of current web site. With the help of long term experiments user can easily find which contextual factor affect the web usage log and why usage log show uncertain pattern [1]. The proposed approach of pattern discovery technique uses the web page classification algorithm can classify the user's current task and time spent on the web page and clustering algorithms are used to cluster the users IP address and identify the user type. The result of experiment is different contextual factor via User interest, Complexity, Difficulty, Task type etc. These factors are influential factor and it is used in a natural web environment for developing personalization services. In addition pattern discovery is also capable to distinguish which contextual factors are useful for further analysis or not and also calculate the credibility of web page it will be achieved by the observing the user-agent and cs-uri-stem field of log file. If pages are credible then this result will be used for site modification etc.

In our paper, we present an algorithm for pre processing of web log file and a new approach for clustering of user session and user identification based on classification technique of IP address. For the clustering of user session we create a unique identification number for each user. This number is used for further access time calculation and classification of user's page as well as task. And it is based on connectivity between referrer and URI pages and we propose a formula for calculating the access time spent on web page and classification of web pages [2]. The results represent that our approach can improve the quality of clustering for user session creation and contextual factors calculation in web usage mining systems. These results can be use for predicting user's next request in the huge web sites, and site improvements, modification of existing sites and guidelines for launching new site. The rest of this paper is organized as follows: In section 2, we review some researches that advance in pattern discovery techniques. Section 3 describes the design consideration and implementation for the pattern discovery technique. Results are shown in section 4. Finally, section 5 summarizes the paper and introduces future work.

II. RELATED WORK

Recently, several Web Usage Mining algorithms have been proposed to discover the user's navigational pattern. In the following we review some of the most significant pattern d discovery techniques and algorithm in web usage mining that can be compared with our system.

Pattern discovery using classification and clustering is one of the earliest methods used in web usage mining [2]. In the following we review some of the most significant pattern discovery system systems and algorithm in web usage mining area that can be compared with our system. Pattern discovery using SOM was one of the earliest clustering methods to be used in Web usage mining by kobra, Amin[6]. Kohenan self organizing map(SOM) is clustering technique to discover the users navigational pattern SOM also represents clustering concept by grouping similar data together.

It reduces the data dimension and displays similarities among the data once the data have been entered into the system, the network of artificial neurons is trained by providing information about input. SOM does not require target vector and classify the training data without any sequential pattern [6]. SOM provides the data visualization techniques which help to understand high dimensional data by reducing the dimension of data to a map. It is simple do not need to involve in complex mathematical relation. It does not depend on pattern length and it is able to extract pattern of any length. The result could help web site owners to attract new customers, retain current customer, improve cross marketing /sales, effectiveness of promotional campaigns, tracking leaving customers, etc.

Pattern discovery presented in paper is based on contextual factors analysis. The core element of this system is web page classification and task classification algorithm and contextual parameter evaluation. Web page classification algorithm take users session and access time as input and classify the web pages into two categories content page and index page[3]. Task classification algorithm is another important technique applied in pattern discovery. User's web task is classified into two main groups 1.Casual 2. Careful In casual searching the user wants to find the precise and credible information. In careful searching the credibility and accuracy of the search results are not important. Task classification algorithm classifies the user's current task into two categories [2]. The frequently visited URLs will be considered as an indicator URL. Applying the Log file user can easily classify the task. If URL contain more credible link and paid user then the assigned task is careful task. If URL contains no credible link then the assigned task is casual user, and different contextual parameter will be analyzed and their percentage calculated.

2.1 Contextual Factors: The proposed architecture of pattern discovery is based on contextual factors. Contextual factors are subjective assessments of Contextual factors include subjective assessments about contents, situational factors, a user's individual characteristics, and so on (ghiyuk, chio, seo, 2009).Here several contextual factors are analysed. They are:

2.1.1 Search Interest**:** Search interest is one of the main important contextual factors. This factor is used to find the how much time user spent on the web page.

2.2.2 Difficulty**:** The difficulty factor tells us difficulty of the content displayed. Downloading and uploading problem, find the net connection and http status code.

2.2.3 Page frequency**:** In page frequency calculation user find the frequently accessed pages and number of time that page is executed. The total frequently accessed page and infrequently accessed pages are counted and then finalizing the how many time same page is accessing.

2.2.4 Credibility**:** Assessing the credibility of web pages is therefore becoming an increasingly important aspect of information with the difficulty of assessing web sites credibility manifests itself in several problematic phenomena. For instance, providing account information to malicious sites masquerading as authentic ones, as in phishing attacks, results in the loss of billions of dollars annually despite the integration of phishing toolbars into mainstream browsers. The presence of misleading, questionable, and factually incorrect information on the web is yet another source of concern. Credibility can be categorized into four types [5].

1. Presumed credibility is based on general assumptions in the users' mind (e.g., the trustworthiness of domain identifiers like *.gov*).
2. Surface credibility is derived from inspection of a site, .is often based on a first impression that a user has of a site, and is often influenced by how professional the site's design appears.
3. Earned credibility refers to trust established over time, and is often influenced by a site's ease of use and its ability to consistently provide trustworthy information.
4. Reputed credibility refers to third party opinions of the site, such as any certificates or awards the site has won.

2.2 Page Selection: Hand labeling Web pages for credibility is a time-consuming process. A credible webpage is one whose information one can accept as the truth without needing to look elsewhere. If one can accept information on a page as true at face value, then the page is credible. If one need to go elsewhere to check the validity of the information on the page, then it is less credible [5]. Page selection will be done two ways on page feature and off page feature.

2.2.1 On-Page Features Selection: On-Page features are present on a page but are difficult or time-consuming for a person to quantify or attend to. The techniques used in on page feature selection are spelling errors, advertising and domain type.

2.2.2 Off-Page Features Selection**:** Off-page features require the user to leave the target page and look elsewhere for supplementary data. Awards, PageRank and sharing are used in off page feature selection [5].

III. DESIGN CONSIDERATION AND IMPLEMENTATION

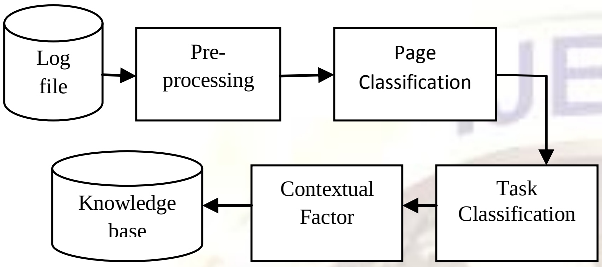


Fig.1 Web Usage Pattern Discovery Techniques

3.1 Log File: Data sets consisting of web log records for 5048 users are collected from De Paul University website. Web log is unprocessed text file which is recorded from the IIS Web Server. Web log consist of 17 attributes with the data values in the form of records.

3.2 Pre-Processing: Generally, several pre-processing tasks need to be done before performing web mining algorithms on the Web server logs. Data pre-processing a web usage mining model (Web-Log pre-processing) aims to reformat the original web logs to identify user's access sessions. The Web server usually registers all users' access activities of the website as Web server logs [3]. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc. Data pre-processing is done using following steps.

3.2.1 Data Cleansing: Irrelevant records are eliminated during data cleansing. Since target of web usage mining is to get traversal pattern, following two kinds of records are unnecessary and should be removed .The records having filenames suffixes of GIF, JPEG, CSS and so on, which can be found in cs-uri-stem field of record and by examining the status field of every record in the web log, the record with status code over 299 and under 200 are removed [7]**.**

Algorithm steps: DataPreparation

- Start
  Check for data available in server log
- If raw data is available goto step 4 else goto step 2
- Cleaning data by removing gap, .jpg , .gif or sound file.
- Execute UserIdentification
- Execute SessionIdentification.
- Divide the session in transaction with a certain duration
- If any data available goto step 4 else goto step 9
- Exit

3.2.2 User and Session Identification: The task of user and session identification is to find out the different user sessions from the original web access log. A referrer-based method is used for identifying sessions. The different IP addresses distinguish different users.

Algorithm steps: UserIdentificaton

- Start
- Take data from cleaned HTTP log file.
- while any data is available do
  i. converting ip address to domain name by reverse DNS lookup.
  ii. Sending cookies to identify user
  iii. Busting cache to prevent use of cache.
  iv. Referring URL.
- Exit

Algorithm steps: SessionIdentificaton

- Start
- Take time of the first log entries.
- Calculate the threshold time from the starting time.
- If threshold >30 min session change
- else same session
- Exit

*3.2.3 ConTent Retrieval***:** Content Retrieval retrieves content from users query request i.e. cs-uri-query. Eg:Query:http/1www.cs.depaul.edu/courses/syllabus.asp?course=323-21 603&q=3&y=2002&id=671. Retrieve the content like /courses/syllabus.asp which helps in fast searching of unvisited pages by other users which are similar to user's interest.

3.2.4 Path Completion**:** Path Completion should be used acquiring the complete user access path. The incomplete access path of every user session is recognized based on user session identification [7]. If in a start of user session, Referrer as well URI has data value, delete value of Referrer by adding '-'. Web log pre-processing helps in removal of unwanted click-streams from the log file and also reduces the size of original file by 40-50%.

3.3Web Page Classification Algorithm: Web Page Classification algorithm, classifies the pages into two categories, index pages and content pages [3].

The page classification algorithm uses the following four heuristics.

3.3.1 File type

An index page must be an HTML file, while a content page may or may not be. If a page is not an HTML file, it must be a content page. Otherwise its category has to be decided by other heuristics.

3.3.2 Number of links

Generally, an index page has more links than a content page. A threshold is set such that the number of links in a page is compared with the threshold. A page with more links than the threshold is probably an index page. Otherwise, it is probably a content page.

3.3.3 End-of-session count

The end-of-session count of a page is the ratio of the number of time it is the last page of a session to the total number of sessions. Most Web users browse a Web site to look for information and leave when they find it. It can be assumed that users are interested in content pages. The last page of a session is usually the content page that the user is interested in [3]. If a page is the last page in a lot of sessions, it is probably a content page; otherwise, it is probably an index page. It is possible that a specific index page is commonly used as the exit point of a Web site. This should not cause many errors at large.

3.3.4 Reference length

The reference length of a page is the average amount of time the users spent on the page. It is expected that the reference length of an index page will be small while the reference length of a content page will be large. Based on this assumption, the reference length of a page can hint whether the page should be categorized as an index or content page. The reference length method for page classification is based on the assumption that the amount of time a user spends on a page is a function of the page category. The basic idea is to approximate the distribution of reference lengths of all pages by an exponential distribution [3].

Algorithm steps:

- Two thresholds set. Count threshold and link threshold.
- Set χ =1/(mean reference length of all pages).
- $t= -\ln(1-\text{Ỵ})/\chi$
- For each page p on the web site
- If P's file type is not HTML or P's end of session count > count _threshold
- Mark P as a content page else
- P's number of links > link _threshold
  Mark p as an index page else
- If P's reference length < t
  Mark P as an index page else P as a content page

3.4 Task Classification Algorithm**:** A user main task is classified into two main groups [2]. First task is careful searching in which user wants to find the precise and credible information, Second task is casual searching in which the credibility and accuracy of result are not important.

Algorithm steps:

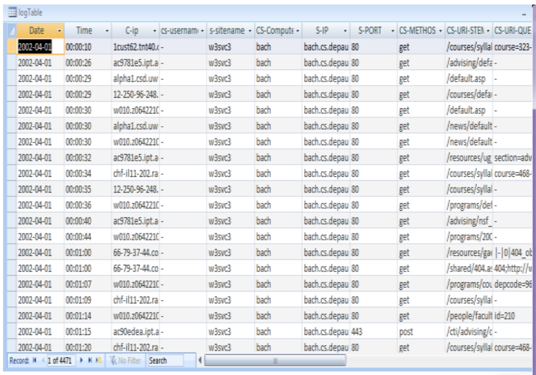
- User's task can be identified by the top level URL.
- Frequently visited URLs as indicators for the task type classification (cs-uri-stem) field.
- Web task is supposed to be kept some period of time .
- Sort all the element of frequently visited URLs.
- Checking how many times the Frequently Visited URLs visits.
- If frequently visited URLs are more than or equals to 5 then setting the user task is careful user otherwise user task is casual user.
- If frequently visited URL have query (cs-uri-query) and that query will be same then setting the user task is casual. Otherwise the user task is careful user.
- Total no. of the URL in casual searching was higher than total no. of URL in careful searching.

3.5 Discovery of contextual factors: In this paper we are implemented the Credibility is one the important factors in the web usage mining [1]. The credibility factor is mainly applicable in designing new web sites and sometimes in the E-business to implement the business value. The DePaul university log file gives the 76% of credibility.

IV. EXPERIMENTAL RESULTS

4.1 Collection of web logs which are in raw or unprocessed form.17 attributes are shown below:

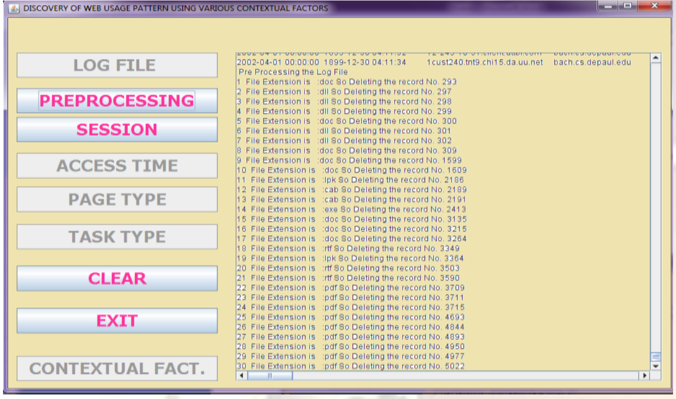
#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer



Date	Time	C-ip	cs-usernam	s-sitename	CS-Comput	S-IP	S-PORT	CS-METHOS	CS-URI-STEM	CS-URI-QUE
2002-04-01	00:00:10	1cust62.tnt40.	-	w3svc3	bach	bach.cs.depau	80	get	/courses/syllal	course=323-
2002-04-01	00:00:26	ac9781e5.ipt.a	-	w3svc3	bach	bach.cs.depau	80	get	/advising/defa	-
2002-04-01	00:00:29	alpha1.csd.uw	-	w3svc3	bach	bach.cs.depau	80	get	/default.asp	-
2002-04-01	00:00:29	12-250-96-248.	-	w3svc3	bach	bach.cs.depau	80	get	/courses/defa	-
2002-04-01	00:00:30	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/default.asp	-
2002-04-01	00:00:30	alpha1.csd.uw	-	w3svc3	bach	bach.cs.depau	80	get	/news/default	-
2002-04-01	00:00:30	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/news/default	-
2002-04-01	00:00:32	ac9781e5.ipt.a	-	w3svc3	bach	bach.cs.depau	80	get	/resources/ug_	section=adv
2002-04-01	00:00:34	chf-il11-202.ra	-	w3svc3	bach	bach.cs.depau	80	get	/courses/syllal	course=468-
2002-04-01	00:00:35	12-250-96-248.	-	w3svc3	bach	bach.cs.depau	80	get	/courses/syllal	-
2002-04-01	00:00:36	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/programs/del	-
2002-04-01	00:00:40	ac9781e5.ipt.a	-	w3svc3	bach	bach.cs.depau	80	get	/advising/nsf_	-
2002-04-01	00:00:44	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/programs/20C	-
2002-04-01	00:01:00	66-79-37-44.co	-	w3svc3	bach	bach.cs.depau	80	get	/resources/ga	|-|0|404_ob
2002-04-01	00:01:00	66-79-37-44.co	-	w3svc3	bach	bach.cs.depau	80	get	/shared/404.a	404;http://w
2002-04-01	00:01:07	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/programs/cou	depcode=96
2002-04-01	00:01:09	chf-il11-202.ra	-	w3svc3	bach	bach.cs.depau	80	get	/courses/syllal	-
2002-04-01	00:01:14	w010.z064221C	-	w3svc3	bach	bach.cs.depau	80	get	/people/facult	id=210
2002-04-01	00:01:15	ac90edea.ipt.a	-	w3svc3	bach	bach.cs.depau	443	post	/cti/advising/c	-
2002-04-01	00:01:20	chf-il11-202.ra	-	w3svc3	bach	bach.cs.depau	80	get	/courses/syllal	course=468-

Fig.2 Logfile To Databse Conversion

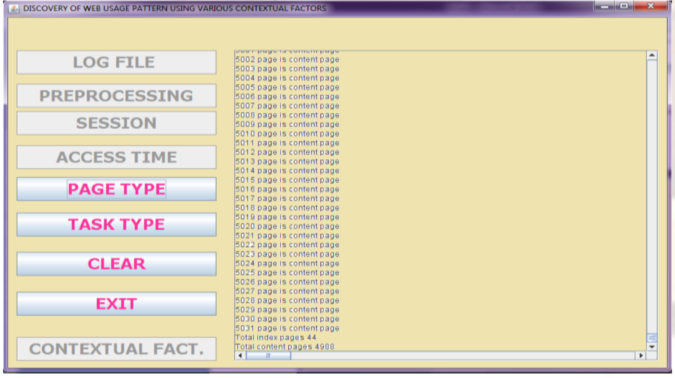
4.2 Preprocessing is done of 5048 web log records from CTI dataset. 34 records are removed from dataset. Page classification technique classifies pages into index and content page.



```
2002-04-01 00:00:00  1899-12-30 04:11:34     1cust240.tnt9.chi15.da.uu.net    bach.cs.depaul.edu
 Pre Processing the Log File
1  File Extension is   :doc So Deleting the record No. 293
2  File Extension is   :dll So Deleting the record No. 297
3  File Extension is   :dll So Deleting the record No. 298
4  File Extension is   :dll So Deleting the record No. 299
5  File Extension is   :doc So Deleting the record No. 300
6  File Extension is   :dll So Deleting the record No. 301
7  File Extension is   :dll So Deleting the record No. 302
8  File Extension is   :doc So Deleting the record No. 309
9  File Extension is   :doc So Deleting the record No. 1599
10  File Extension is   :doc So Deleting the record No. 1609
11  File Extension is   :lpk So Deleting the record No. 2186
12  File Extension is   :cab So Deleting the record No. 2189
13  File Extension is   :cab So Deleting the record No. 2191
14  File Extension is   :exe So Deleting the record No. 2413
15  File Extension is   :doc So Deleting the record No. 3135
16  File Extension is   :doc So Deleting the record No. 3215
17  File Extension is   :doc So Deleting the record No. 3264
18  File Extension is   :rtf So Deleting the record No. 3349
19  File Extension is   :lpk So Deleting the record No. 3364
20  File Extension is   :rtf So Deleting the record No. 3503
21  File Extension is   :rtf So Deleting the record No. 3590
22  File Extension is   :pdf So Deleting the record No. 3709
23  File Extension is   :pdf So Deleting the record No. 3711
24  File Extension is   :pdf So Deleting the record No. 3715
25  File Extension is   :pdf So Deleting the record No. 4693
26  File Extension is   :pdf So Deleting the record No. 4844
27  File Extension is   :pdf So Deleting the record No. 4893
28  File Extension is   :pdf So Deleting the record No. 4950
29  File Extension is   :pdf So Deleting the record No. 4977
30  File Extension is   :pdf So Deleting the record No. 5022
```

Fig.3 Reslut of Data Preprocessing

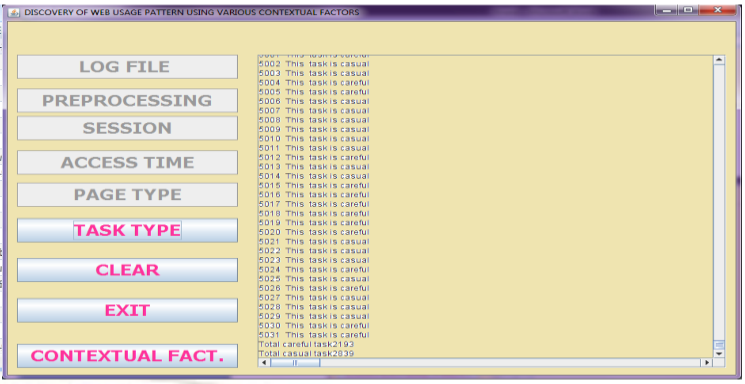
4.3 The page classification algorithm is used to classify the web pages. The result of page classification algorithm is 44 index pages and 5018 content pages [3].



```
5002 page is content page
5003 page is content page
5004 page is content page
5005 page is content page
5006 page is content page
5007 page is content page
5008 page is content page
5009 page is content page
5010 page is content page
5011 page is content page
5012 page is content page
5013 page is content page
5014 page is content page
5015 page is content page
5016 page is content page
5017 page is content page
5018 page is content page
5019 page is content page
5020 page is content page
5021 page is content page
5022 page is content page
5023 page is content page
5024 page is content page
5025 page is content page
5026 page is content page
5027 page is content page
5028 page is content page
5029 page is content page
5030 page is content page
5031 page is content page
Total index pages 44
Total content pages 4988
```

Fig.4 Reslut of Page Classification

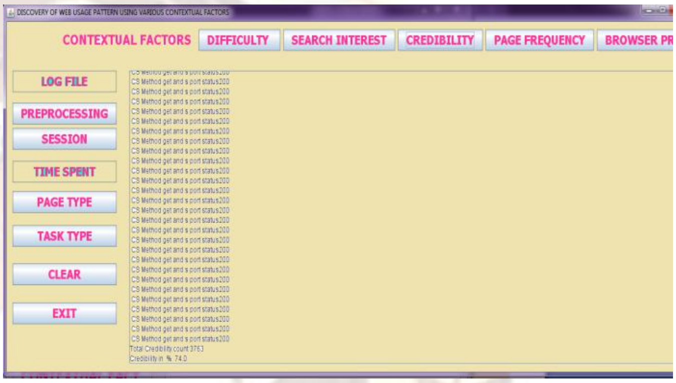
4.4 The task classification algorithm is used to classify the user's current task (careful user and casual user) on web pages. The result of careful user is 2193 and casual user is 2869.



```
5002  This  task is casual
5003  This  task is casual
5004  This  task is careful
5005  This  task is careful
5006  This  task is casual
5007  This  task is casual
5008  This  task is casual
5009  This  task is casual
5010  This  task is casual
5011  This  task is casual
5012  This  task is careful
5013  This  task is casual
5014  This  task is casual
5015  This  task is careful
5016  This  task is careful
5017  This  task is careful
5018  This  task is careful
5019  This  task is careful
5020  This  task is careful
5021  This  task is casual
5022  This  task is casual
5023  This  task is casual
5024  This  task is careful
5025  This  task is casual
5026  This  task is careful
5027  This  task is casual
5028  This  task is casual
5029  This  task is casual
5030  This  task is careful
5031  This  task is careful
Total careful task2193
Total casual task2839
```

Fig.5 Result of Task Classification

4.5 The credibility is one of the most important factor . Credibility or believeability gives whether the access page is crediblr or not . or site is credible or not [1]. The total cerdibility count is 3783 and percentage is 74 %.



```
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
CS Method get and s port status200
Total Credibility count 3763
Credibility in % 74.0
```

Fig.6 Reslut of Contextual factor credibility

Graphical Representation

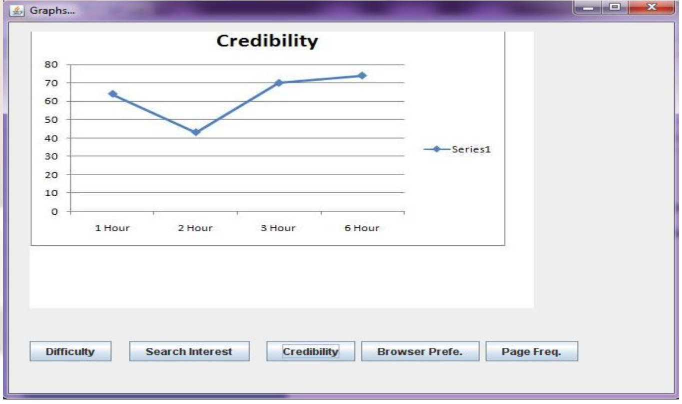


Fig.7 Graphical Result of Credibility Contextual Factor

V. APPLICATION

5.1 Application of Discovered Patterns: According to the application of web usage mining is classified into two main categories. Those dedicated to the discovery of web usage pattern (behavior) and hence customizing the web sites. Other dedicated to the architectural topology (related to the site topology) and improvement web sites effectively. These two applications are discussed below.

5.2 Web site Improvement: The logs are used to infer user interaction and hence provide implicit user rating .other browser provide sophisticated indicator such as bookmark , print , number of visits, save and more effective time calculation .

5.3 Performance Improvement: Another important application of web usage mining is to improve the web performance .here too we need intelligent system technique to identify, characterize, and understand user behavior [3]. In fact user navigation activity determines the network traffic and as a consequence, influences web server performance.

VI. CONCLUSION

Pattern Discovery system using various contextual factors helps webmasters of the website to pre process the web access log also finds the users problem which is caused by searching and accessing the site. System also captures the different types of pages which will helpful finding the credibility or trustworthiness finding. The proposed system helps in reducing the searching time of pages by the user on the web site. Thus, increase the website usability and provide better services to webmaster and users. In this paper a little attempt is made to provide an up-to-date survey of the rapidly growing area of Web Usage mining and how the various pattern discovery techniques help in developing business plans especially in the area of e-business. This article has aimed at describing different pattern discovery techniques and addressing the different problem phase user while accessing the web page. The WWW will keep growing, even in a somewhat different form than how we know it today.

REFERENCES

[1] J.Choi J. Seo,G. Lee, “ Analysis of web Usage pattern in consideration of Various Contextual Factors” , In Proceeding of 7th Workshop on Intelligent Techniques of Web Personalization & Recommender Systems at IJCAT ’09, July. pp. 1-12, 2009

[2] Jinhyuk Choi, Geehyuk Lee. “New Techniques for Data Pre-processing Based on Usage Logs for Efficient Web User Profiling at Client Side”, In Proceeding of International Joint Conferences on Web Intelligence and Intelligent Agent Technologies pp. 54-57, 2009

[3] Yongjian Fu, Ming – Yi shih , Mario Creado , Chunhua Ju , “Reorganizing Web sites based on user Access Patterns “ In Proceeding of 10th international conference on Information and knowledge management Atlanta , Georgia USAACM 2001 pp. 1-15 , 2001

[4] Peter I. Hofegang , De boelelaan ,” Methodology for preprocessing and Evaluating the time spent on web pages , In Proceeding of the IEEE / ACM International Conferences on web Intelligence pp. 1-8 , 2006 .

[5] Julia Schwarz, Meredith Ringel Morris,” Augmenting Web Pages and search Results to Support Credibility Assessment”, In Proceeding of ACM International Conferences, pp .1-10, 2011

[6] Kobra Etminani, Amin Rezaeian Delui, Noorali Raeeji Yaneshari, Modjtabani Dept of computer engg;” Web usage mining: Discovery of the user’s navigational Pattern using SOM” pp. 244-248, 2009

[7] Rajni pamnani, Pramila Chavan,” Web usage mining: Research area in web mining”, VJTI, Mumbai pp.1-5, 2010

[8] Byström K. and Järvelin K. 1995. Task complexity affects information seeking and use. *Information Processing and Management*, 31(2):191-213.

[9] Z. Baoyao, “Intelligent Web Usage Mining” Nanyang Technological University, 2004.

[10] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, third Edition,2011

[11] R. Kosala, H. Blockeel, “Web Mining Research: Survey: SIGKDD Explorations, vol. 2, Issue. 1, pp. 1-15, 2000.