A study of Path Completion Techniques in Web Usage Mining

Nirali Honest and Dr. Atul Patel
Smt. Chandaben Mohanbhai Patel Insitute of Computer Applications
Charotar University of Science and Technology, (CHARUSAT),Changa, India
niralihonest.mca@charusat.ac.in

Dr. Bankim Patel
Shrimad Rajchandra Institute of Management and Computer Application
Uka Tarsadia University, Bardoli, India
bankim_patel@srimca.edu.in

***Abstract* - Path completion is a critical and difficult task in the preprocessing phase of web usage mining. We mold the data preprocessing phase to accomplish our goal to mine websites designed using a content management system (cms). The data preprocessing phase includes data cleaning, user identification, session identification, site structure and link details formation, path completion and event generation. The paper includes work on path completion by considering different types of path generated in accessing the website designed using cms and gives a novel algorithm to form the path.**

Index Terms – Web usage, Preprocessing phase, Path completion.

I. INTRODUCTION

Web Usage Mining (WUM) carries out interesting and decisive information for an assortment of people based on different work domains. We focus to generate patterns for the administrator of a website that is designed using cms. We consider a university website and try to identify the useful and meaningful information which can help the website administrator to manage the website. We decide to prepare a new reactive approach which uses the web log data, site structure and academic calendar of the university in order to produce more specific behavior patterns for the University website access domain (UWAD). G. Castellano, A. M. Fanelli, M. A. Torsello [6] have designed LODAP (Log Data Preprocessor) that takes as input log files related to a Web site and outputs a database containing some statistics about pages visited by users and the identified user sessions but they don't form sessions for dynamic pages considered for our work.

The motivation behind the generation of this concept is because of three reasons, 1) WUM can be molded according to the specific goal w.r.t mining 2) There is no support for generation of reports for particular events, you need to remember the interval of the event for generating the report of the event 3) The websites designed using the concept of cms have master page and content page concept, so each content page may be dynamically loaded in the master page. This page may not have unique page names, instead they are stored by page id, the report generation for per page frequency is not supported by certain tools, and if supported the page name cannot be known if it is generated by the ID number.

The paper presents the path completion techniques adopted by different authors and suggest the new approach for path formation. The paper is organized into six sections, in the first section Introduction of the concept is given, in the second section Literature survey for path completion is discussed, in the third section an overview of Data preprocessing phase is discussed, in the fourth section path completion technique is discussed, in the fifth section pattern discovery and analysis is discussed and in the last section experimental results are shown followed by conclusion and future work.

II. LITERATURE SURVEY

Chungsheng Zhang and Liyan Zhuang [1] suggested that reconstruction of accurate user sessions from logs is a challenging task as the HTTP protocol is stateless and connectionless

Path completion is an important activity in preprocessing phase, as many patterns can be discovered and analyzed after forming the complete and accurate path. Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah [3] uses the link sequence information for prediction user links, D. W. Chueng, B. kao, and J. W. Lee [4] analyses the web pages visited by users and performs topic spotting.

In user session identification and path completion methods, the most common methods are timeout, maximal forward reference and reference length methods. Many authors have implemented path completion phase with different parameters, The method proposed by Cooley. R., Mobasher, B. & Srivastava, J.[2], assumes that the amount of time a user spends on a page depends whether the page is an auxiliary or content page. Z. Chen, A. Fu, J. Tang and F. Tung [10][11], they defined each session as the set of pages from the first page in a request sequence to the final page before a backward reference is made. Yan LI, Boqin FENG, Qinjiao MAO [9] they have implemented path completion algorithm using three steps. 1) The incomplete access path is identified, and path combination is conducted. 2) The content and auxiliary pages are identified by using the Maximal Forward References (MFR) and Reference Length (RL) algorithms. 3) The complete path for each user session is acquired by using referrer information and the reference length of some pages of this complete path is modified by using Average Reference Length Auxiliary Pages. G. Arumugam, S. Suguna [5] proposed User Session Identification Algorithm (USIDALG) containing two modules for the activities related to User Identification, and Session Identification.

Cooley, R., Mobasher, B. & Srivastava, J. [2] and Z. Chen, A. Fu, J. Tang and F. Tung [10] suggest that to process l

978-1-4799-6023-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CICT.2015.64

backward references among L logs the time complexity is O (N/2 * l) where N is the number of pages on the server. The complete path generation process fails to generate a correct path when the pages are referred from some other servers. So, level of accuracy is reduced. Cooley, R., Mobasher, B. & Srivastava, J. [2], there is no algorithm for generating a complete set of the USS. Yan LI, Boqin FENG, Qinjiao MAO [9], using SbSfxminer and Absfxminer the complete set of the USS is generated with the time complexity of $\bullet_{i=1}^{MFRS}|MFRS_i|$. G. Arumugam, S. Suguna [5] the performances are analyzed on parameters like 1. Generating a complete path 2. Time complexity to generate the complete set of User Session Sequence (USS) 3. Accuracy in generating a complete set of the USS. They find the Time complexity for generating the complete USS by applying the formula O (l* log (n/2)) /(No. of search / sec). In all the approaches the pages designed using cms are not considered , so this area requires focus and we contrive to work on this.

III. An overview of Data Preprocessing for UWAD

This phase is used to clean and process the data for making it available for analysis. Our preprocessing phase comprises various steps like, Data Cleaning, User Identification, Session Identification, Path Completion, Generating Site structure (Site Map, Mapping Page Number and Name) and Generating Academic Events. A detail description of steps is given by Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel [7] [8]. Figure 1 shows the architecture of the data preprocessing phase for UWAD.

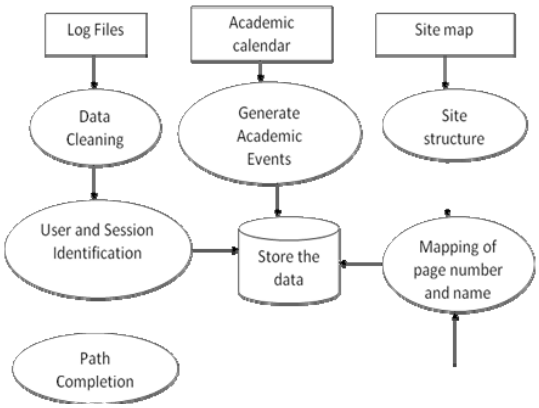


Fig. 1 Architecture of Preprocessing phase UWAD

Website Structure considered in our work is based on a website designed using Content Management System, which shares two characteristics, 1) The pages are generated with unique identifiers and 2) The pages may have logical names apart from actual name. These characteristics are important to understand before carrying out path completion. The website structure we consider is shown in Figure 2.

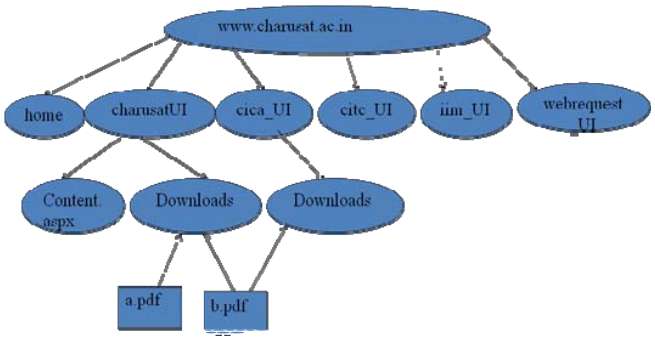


Fig. 2 Website structure

IV. Path Completion

In the literature survey, we have found that authors have worked for the static pages, in our work we consider the websites designed with dynamic pages. Web sites designed using the concept of CMS don't have a unique page name for every page, instead the pages have id by which the content of the pages can be retrieved. So performing path completion becomes difficult and complex. In the preprocessing phase, after identifying sessions, while attempting for path completion, we build the page name by reading the page id and page name from the xml files like RSS feed and site map, and during path completion, we Add missing pages in the session, Remove duplicate pages in the consecutive access within a given session and Map the name of pages with the page number. Apart from this we add the concept of the event as per the academic calendar in context with University environment. Adding events allow to mine the web logs based on temporal concept, as the accesses to web by the users are not same all the time. Based on event lots of patterns can be discovered and analyzed.

A. Definitions

While accessing the website, based on what user accesses, different path are formed. Before we attempt for path completion, it is necessary to understand the types of path. Below is the list of definitions for various path.

Definition of Path: A path p= {p1, p2, --, pn} where n is the number of pages traversed in a single session.

Definition Simple path: A path p containing the value as domainname/page.aspx?id=pagenumber

Definition ID and Name path: A path p containing the value as domainname/page.aspx?id=pagenumber&;name=title this path is formed if the pages of a tab is selected for the first time.

Definition UI and ID path: A path p containing the value as domainname/UIname/page.aspx?id=pagenumber the website we consider has more than one user interface (UI), which is included in the path.

Definition UI and key path: A path p containing the value as domainname/UIname/page.aspx?key=number this is used for right tab to and key indicates the tab number.

Definition UI and pOpen path: A path p containing the value as domainname/page.aspx?id=pagenumber&;pOpen=0 when a left tab has more than subtabs it is referred by pOpen, the value of pOpen referes to inner tab open on the left tab.

Definition resource path: A path p containing the value as domainname/UIname /download/advertisements/b.pdf a file ending in a pdf or an image is a resource file.

Examples of above definitions include the following path,

B. Site Structure and Link Details

For mapping the page id and page name it is We need to consider the site structure and store the link details of every link of the website. This we can perform by reading the RSS feed and a site map of the website. The file to have the following structure as shown in Figure 3.,

```
Site map
        <item>
          <link>http://www.charusat.ac.in/Content.aspx?ID=67 &amp;name=About_University</link>
          <ror:updatePeriod>week</ror:updatePeriod>
          <ror:sortOrder>0</ror:sortOrder>
          <ror:resourceOf>sitemap</ror:resourceOf>
        </item>
RSS feed
        <item>
        <guid>
              http://www.charusat.ac.in/charusatui/content.aspx?id=40&amp;name=about_trust
        </guid>
            <title>CHARUSAT Foundation </title>
            <link>
                http://www.charusat.ac.in/charusatui/content.aspx?id=40&amp;name=about_trust
            </link>
        </item>
```

Fig. 3 Site map and Rss feed files snapshot

After the files are read and parsed the details of link and site structure are stored as shown in below figure 4.

link_details

link_id	link	name
1	http://www.charusat.ac.in/CharUSATUI/MainWebsitePage2.aspx	
2	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=3&name=	CHARUSAT | About University
3	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=2	CHARUSAT | About University | CHARUSAT at a Glance | Vision
4	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=72	CHARUSAT | About University | CHARUSAT at a Glance | Mission
5	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=73	CHARUSAT | About University | CHARUSAT at a Glance | Quick Facts
6	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=74	CHARUSAT | About University | CHARUSAT at a Glance | Infrastructure
7	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=75	CHARUSAT | About University | CHARUSAT at a Glance | Campus Map
8	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=76	CHARUSAT | About University | CHARUSAT at a Glance | Getting To CHARUS
9	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=102	CHARUSAT | About University | CHARUSAT at a Glance | Cells at CHARUSAT
10	http://www.charusat.ac.in/CharUSATUI/Content.aspx?ID=64	CHARUSAT | About University | Presidents Welcome

Figure 4 : a)Link details

	domain_name	ui_name	page_name	name	id	k	pO	actual_name	lev
1	http://www.chan	CharUSATUI	MainWebsitePag						2
2	http://www.chan	CharUSATUI	/Content.aspx	About_Univer	ID=3			CHARUSAT | About University	4
3	http://www.chan	CharUSATUI	/Content.aspx		ID=2			CHARUSAT | About University | CHARUSAT at a Glance | Vision	3
4	http://www.chan	CharUSATUI	/Content.aspx		ID=72			CHARUSAT | About University | CHARUSAT at a Glance | Mission	3
5	http://www.chan	CharUSATUI	/Content.aspx		ID=73			CHARUSAT | About University | CHARUSAT at a Glance | Quick Fa	3
6	http://www.chan	CharUSATUI	/Content.aspx		ID=74			CHARUSAT | About University | CHARUSAT at a Glance | Infrastru	3
7	http://www.chan	CharUSATUI	/Content.aspx		ID=75			CHARUSAT | About University | CHARUSAT at a Glance | Campus	3
8	http://www.chan	CharUSATUI	/Content.aspx		ID=76			CHARUSAT | About University | CHARUSAT at a Glance | Getting T	3
9	http://www.chan	CharUSATUI	/Content.aspx		ID=102			CHARUSAT | About University | CHARUSAT at a Glance | Cells at C	3
10	http://www.chan	CharUSATUI	/Content.aspx		ID=64			CHARUSAT | About University | Presidents Welcome	3

Figure 4 : b)Site structure

C. Construction of path

Construction of path consists of the following steps,

Read user session Ui, Ui= {U1, U2, ----, UN} where n is the total number of sessions.

Divide first url from the session, into number of pages accessed, pi= {p1, p2, ---pn}, where n is the number of pages traversed in a single session.

Read pi and calculate the length of the page, checking it with the link name formed in the link details. Record the link name, uiname and level in the path.

Find the type of path in the existing links if the path then other than simple and path replaces it with the actual name of the link.

Read the second url in the session

IF url is same as first, don't add it in the path, read next url.

IF url is different then Record the link name, uiname, level and distance.

Compare the uiname, if same, enter the page name in the path, otherwise calculate the nodes to be traversed and list the new pages in between first and the second url.

Repeat the above steps for all urls in a given session.

Append the url for a given session into a single path.

Repeat the above steps for all sessions in a given file.

We consider the parameter like time to build the path and accuracy. Time to build a single path includes

$P(T_i) = \bullet_{i=1 to n}(T1*T2*T3*T4)/n$

T1= Time to read each session

T2= Time to calculate level and distance

T3= Time to add or remove pages

T2= Time to search and map page id and name

n=*Number of pages in a given session*

Total time to build paths for the given user sessions includes,

$P(TT_i)=\bullet_{i=1 to n} P(T_i)/n$

$P(T_i)$=*Time to build a single path*

n=Number of path

After the algorithm is applied to the resultant path will carry all the missing pages added, all the ID replaced by page names, so they are more legible while pages are used for discovering patterns.

- Example of path completion after mapping of name
  - /CharUSATUI/MainWebsitePage2.aspx/

- /CharUSATUI/Content.aspx?ID=3&name=About_University
- /CharUSATUI/Content.aspx/ID=6&name=Academics
- /CITC_UI/Content.aspx/ID=37
- /CharUSATUI/Content.aspx/ID=6&name=Academics
- /CharUSATUI/NewsAnnouncementDetail.aspx

- Example of pages after path completion
  - CharUSATUI/MainWebsitePage2.aspx|CharUSATUI/About_University|CharUSATUI/Academics/Syllabi|/CITC_UI/Syllabus_of_mechanical_Engineering|/CharUSATUI/Academics/Syllabi|/CharUSATUI/NewsAnnouncementDetail.aspx

D. *Event data generation*

Events can be anything based on the type of website. In case of online shopping events can be festival sale, end of season sale, brand wise sale, etc. The notion of event is to emphasize that web users may visit and access the website in a different way during different periods. In our work we try to capture the events and mining the web logs particular to those events in accordance with the regular access to the website.

In any University, there are lot many events that may occur during the particular academic year. An event is a special occurrence of an operation that occurs for a finite time. In the university academic events can be considered as Recruitment, Admission, Display of results, Announcement of workshop, etc. During these academic events the access of the website is different than the regular access. So in this paper, we try to show the insertion of academic events to be specified, so that the patterns of access during these events can be analyzed and compared to the normal access of the website. A variety of details are stored for generating academic events, like Academic year, Name of Institute, Name of event, Intervals of the event, i.e. start and end date, etc. This is the last phase of data preprocessing.

V. Pattern Discovery and Pattern Analysis

A. *Pattern Discovery*

After the completion of Data preprocessing phase, the next phase is pattern discovery.The purpose of Pattern Discovery phase is to produce meaningful patterns from the data stored after cleaning and reforming the data. In the context of pattern discovery following access patterns can be formed,Access Patterns that can be discovered for UWAD are as below and Figure 5 shows the process of pattern discovery,

- Operating systems and browsers used by the user while accessing the website.
- Access to website
  - Hourly, daily, monthly, yearly, (Regular, event based: admission, recruitment, etc.)
- Time spent by the user on the website.
- User Navigation
  - First page accessed by the user, last page accessed by the user, all the pages accessed, frequency of pages , order of pages accessed.

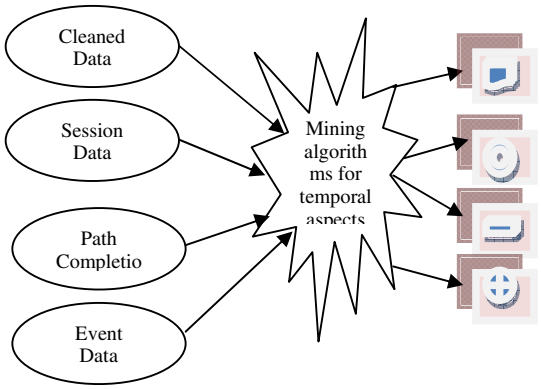


Fig 5. Pattern Discovery phase of UWAD

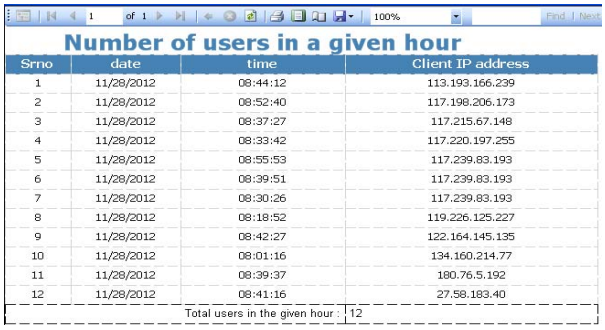
B. *Pattern Analysis*

Later in the patterns discovered the analysis is formed to find out, pages that are accessed the most, Browsers and Operating systems used the most, based on user navigation predicting user accesses and deriving the user interest in accessing the pages.

VI . Experimental Results

Experimental results derived for the patterns discovery are presented in below figures. Figure 6 a) shows the user interface for discovering pattern of user activities like number of users, sessions, pages accessed based on weekly daily and hourly dimension. Figure 6 b) shows the number of users in on a given date(s) for the selected hour. Figure 6 c) shows the number of files accessed per hour.

Fig 6. a) Pattern Discovery Interface



Number of users in a given hour

Srno	date	time	Client IP address
1	11/28/2012	08:44:12	113.193.166.239
2	11/28/2012	08:52:40	117.198.206.173
3	11/28/2012	08:37:27	117.215.67.148
4	11/28/2012	08:33:42	117.220.197.255
5	11/28/2012	08:55:53	117.239.83.193
6	11/28/2012	08:39:51	117.239.83.193
7	11/28/2012	08:30:26	117.239.83.193
8	11/28/2012	08:18:52	119.226.125.227
9	11/28/2012	08:42:27	122.164.145.135
10	11/28/2012	08:01:16	134.160.214.77
11	11/28/2012	08:39:37	180.76.5.192
12	11/28/2012	08:41:16	27.58.183.40
	Total users in the given hour :		12

Fig 6. b) No. of users on a given date & hour

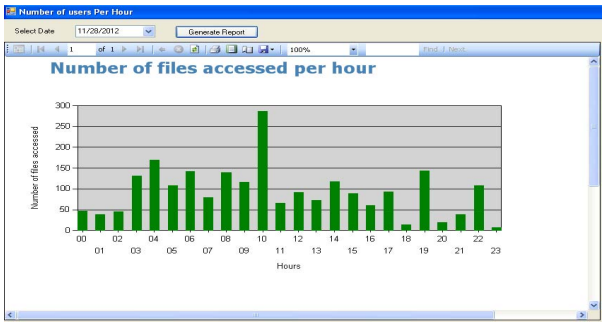


Fig 6. c) No. of files accessed per hour.

Figure 7 a) shows the user interface for daily based reports like browsers used, OS used and error types for a selected date. Figure 7 b) shows the browsers used on the given date. Figure 7 c) shows the OS used on the given date. Figure 7 d) shows the types of errors occurred on the given date.

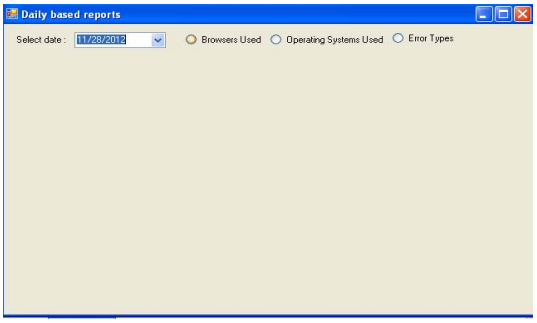


Fig 7. a) Daily based activity interface

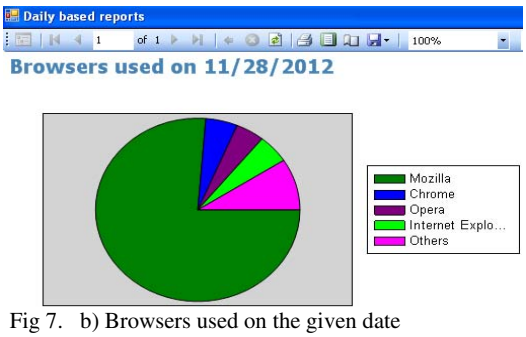


Fig 7. b) Browsers used on the given date

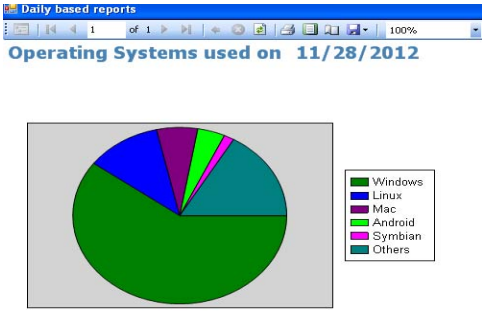


Fig 7. c) OS used on the given date

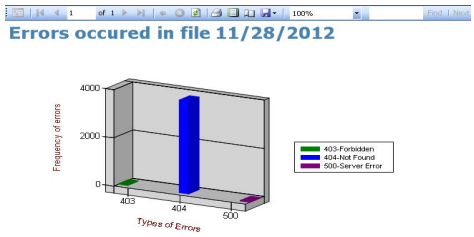


Fig 7. d) Errors occurred on the given date

Figure 8 a) shows the user interface for per page frequency based on individual files or based on events. Figure 8 b) shows the pages accessed on the given date with page id replaced with the meaningful page name.

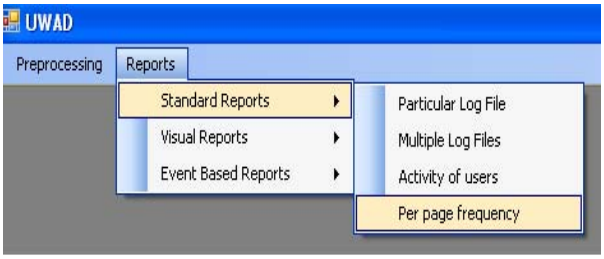
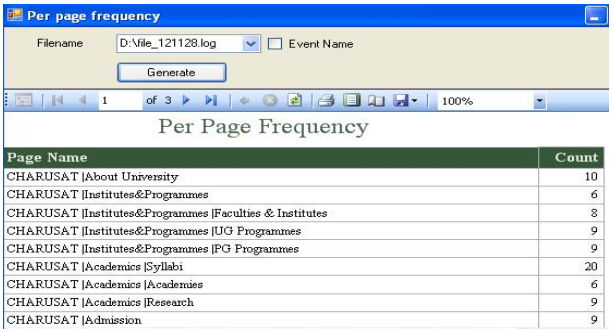


Fig 8. a) Reports interface



Page Name	Count
CHARUSAT |About University	10
CHARUSAT |Institutes&Programmes	6
CHARUSAT |Institutes&Programmes |Faculties & Institutes	8
CHARUSAT |Institutes&Programmes |UG Programmes	9
CHARUSAT |Institutes&Programmes |PG Programmes	9
CHARUSAT |Academics |Syllabi	20
CHARUSAT |Academics |Academics	6
CHARUSAT |Academics |Research	9
CHARUSAT |Admission	9

Fig 8. b) Per page frequency on the given date

CONCLUSION

The analysis suggested that the existing technologies need to focus on dynamic pages designed using a cms, as it adds complexity and requires further study and calls for further work to contribute to the existing algorithms.

ACKNOWLEDGMENT

The authors thank the Charotar University of Science and Technology (CHARUSAT) for providing the necessary resources to carry out this work.

REFERENCES

[1] Chungsheng Zhang and Liyan Zhuang , “New Path Filling Method on Data Preprocessing in Web Mining ,“, Computer and Information Science Journal , August 2008.

[2] Cooley, R., Mobasher, B. & Srivastava J., ,”Data preparation for mining World Wide Web browsing patterns”, Journal of Knowledge and Information Systems, I , Page(s): 5-32, 1999.

[3] Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah, "Kowledge Discovery from Users Web-Page navigation", , IEEE RIDE 1997.

[4] D. W. Chueng, B. kao, and J. W. Lee, "Discovering user Access patterns on the World-Wide Web", Proc. First Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-97).

[5] G. Arumugam, S. Suguna,” Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs”, Network and Service Security, 2009.

[6] G. Castellano, A. M. Fanelli, M. A. Torsello,” Log Data Preparation For Mining Web Usage Patterns”, IADIS International Conference Applied Computing, ISBN: 978-972-8924-30-0, 2007.

[7] Nirali Honest, Dr. Bankim Patel and Dr. Atul Patel. Article ”Preprocessing phase for University Website Access Domain”, International Journal of Scientific & Engineering Research, (IJSER) – ISSN : 2229-5518, 4, No.6, June 2013.

[8] Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel,” Sessionization Process for the Pages Designed with the Concept of CMS”, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X,Volume 3, Issue 9, September 2013 .

[9] Yan LI, Boqin FENG, Qinjiao MAO,”Research on Path Completion Technique in Web Usage Mining”, International Symposium on Computer Science and Computational Technology,2008.

[10] Z. Chen, A. Fu, J. Tang and F. Tung, “Optimal algorithms for finding user web access sessions”, Journal of World Wide Web: Internet and Information Systems, Vol. 6, Page(s): 259-279, 2003. Springer.

[11] Z. Chen, A.Fu, R.H. Fowler & C. Wang, “Efficient Web Mining of Frequent Traversal Patterns”, in Anthony Acime, Web Mining: Applications and Techniques, Page(s): 322-338, Idea Group Publishing, August 2004.