ISSN:0975-9646

Implementation of Navigation Pattern Mining in Dot Net Framework

Aditi Shrivastava , Nitin Shukla
Shriram Institute Of Engineering and Technology
JABALPUR(M.P), INDIA

Abstract-Web user navigation pattern is a heavily researched area in the field of web usage mining with wide range of applications. log mining is application of data mining techniques to discover usage patterns from web data, in order to better serve the needs of web based applications. The user access log files present very significant information about a web server. This paper is concerned with the in-depth analysis of Web Log Data to find information about a web site, top priority pages, navigation pattern of the users of particular website etc, which help system administrator and Web designer to improve their system by determining the patterns and the usage of web pages. The obtained results of the study will be used in the further development of the web site in order to increase its effectiveness.

Keywords:web usage mining, web Log. Navigation pattern mining

I. INTRODUCTION

With the explosive growth of knowledge available on World Wide Web, which lacks an integrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for web masters to organize the contents of websites to cater to the need of user‟s. Analyzing and modeling web navigation behavior is helpful in understanding demands of online users. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining website maps, and web based personalization system and improving searching accuracy when seeking information. Nevertheless, an online navigation behavior grows each passing day, thus extracting information intelligently from it is a difficult issue. Web Usage Mining (WUM) is process of extracting knowledge from Web user‟s access data, by exploiting Data Mining technologies. Typically, the Web usage mining prediction process is structured according to two components performed online and off-line with respect to Web server activity. Offline component builds the knowledge base by analyzing historical data, such as server access log file or web logs which are captured from the server, The access data of the users visiting a given web site is provided by the Server Log Files. They provide details about file requests to a web server and the server response to those requests. In the access log, which is the main log file, each line describes the source of a request, the file requested, the date and time of the request, the content type and length of the transferred file, and other data such as errors and the identity of referring pages.

In our paper, we present architecture for forming clusters of navigation pattern of users in the tabular form. This table consists of clusters of pattern of pages visited by user**.** The results represent that improved accuracy of clustering. The rest of this paper is organized as follows: In section 2, we review some previous approaches. Section 3 defines some useful definintions which is used in implementation. Section 4 explains the solution framework used for generation of clusters of navigation pattern of users with experimental results. Finally, section 5 summarizes the paper and introduces future work.

II. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [9].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests and Mobasher (2003) presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) [10] proposed a hybrid approach for analyzing the visitor click sequences. Jalali et al. (2008a [7] and 2008b [8]) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) [5] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. But all the graphical approaches are complex in nature and also very cost expensive.So here we introduced the new approach for finding pattern of path of users while browsing the site.

III BASIC DEFINITIONS USED

Before implementation of algorithm here are some important terms related to it.

(A) Regular Expression - A regular expression is a way for a computer user or programmer to express how a computer program should look for a specified pattern in text. It Search for occurrences of one of single patterns as well as multiple patterns in a text file.

A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules. Although the syntax is somewhat complex it is very powerful and allows much more useful pattern matching than say simple wildcards like ? and *.

[-a-z0-9]+(\.[-a-z0-9]+)*

Regular expression is a notation to stipulate set of strings.

Operation	Example	In Set	Not in set
Concatenation	aabaab	aabaab	Every other string
Wildcard	.u.u.u	cumulus jugulum	succubus tumultuous
Union	aa | baab	aa baab	Every other string
Closure	ab*a	aa abbba	ab ababa
	a(a |b)aab	aaaab abaab	Every other string
Parentheses	(ab)*a	a ababababa	aa abbba

(B) Hash Tables - A **hash table**, or a **hash map**, is a data structure that associates *keys* with *values*. The primary operation it supports efficiently is a *lookup*: given a key , find the corresponding value . It works by transforming the key using a hash function into a *hash*, a number that the hash table uses to locate the desired value. The concept of a hash table is a generalized idea of an array where key does not have to be an integer. We can have a name as a key, or for that matter any object as the key. The trick is to find a hash function to compute an index so that an object can be stored at a specific location in a table such that it can easily be found.

Technically we define hash table as

"A hash table is a data structure for storing a set of items, so that we can quickly determine whether an item is or is not in the set". The basic idea is to pick a hash function h that maps every possible item x to a small integer h(x). Then we store x in slot h(x) in an array. The array is the hash table.

Let's be a little more specific. We want to store a set of n items. Each item is an element of some

finite set U called the universe; we use u to denote the size of the universe, which is just the number of

items in U. A hash table is an array T[1.. m], where m is another positive integer, which we call the

table size. Typically, m is much smaller than u. A hash function is any function of the form

$$h: U \rightarrow \{0, 1, \ldots, m-1\},$$

mapping each possible item in U to a slot in the hash table. We say that an item x hashes to the slot

T [h(x)].

Hash tables are used to speed-up string searching in many implementations Therefore in our work we use concept of hash table for forming cluster of patterns

(C) Support - Support is the one of the measure of interesting Here we find the support of each pattern that we have discover. It tells us the measurement of webpages in the pattern that appear together.

Here support is calculated by

○ Input :-

Total Transaction in DB

No. of occurrences each item {x,y}

$$\text{support} = \frac{\text{No. of occurrences}\{x,y\}}{\text{Total transaction in DB}}$$

(D) Confidence – Confidence is used how confident we are about our patterns By finding confidence we get information regarding tendency to appear web pages after another one.

○ Input:-

Total occurrence for item X

Total occurrence for item X and Y

$$\text{Confidence} = \frac{\text{Total occurrences for item X and Y}}{\text{Total occurrence for item X}}$$

IV Proposed System

We would like to propose a system which would discover interesting patterns in these weblogs. Weblogs has information about accesses to various Web pages within the Web space associated with a particular server. In case of Web transactions, it capture relationships among page views based on navigation patterns of users.

A. Steps involved in the proposed system are-

1) The input is a set of Weblogs from which we will discover navigation pattern of users to.
2) The server logs contain entries that are redundant or irrelevant for data mining tasks.
3) The Data cleaning process will select a subset of fields that are relevant for the task.
4) These selected attributes are then stored into a database.

5) Now from these logs we will find the general and access statistics by implementing algorithm and find the patterns of the path that user had follow while browsing the site.

6) Then support and confidence of each pattern is calculated to know their interestingness.

7) As a result , interesting patterns can be discovered and client's web usage can be evaluated.

V Implementation With Experimental Results

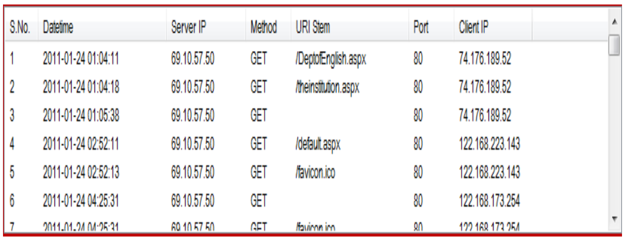
Implementation of project is done in dot net framework all programs are written in Microsoft Visual studio 2010 as front end and SQL server for database management. Here we take weblogs as input the sample weblog file are

```
#Software: Microsoft Internet Information Services 7.0
#Version: 1.0
#Date: 2011-01-24 01:04:12
#Fields: date time s-sitename s-ip cs-method cs-uri-stem cs-uri-
query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus
sc-win32-status sc-bytes cs-bytes time-taken
2011-01-24    01:04:11    W3SVC115    69.10.57.50    GET
/DeptofEnglish.aspx-80-74.176.189.52
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+GTB6.
6;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+.NET+CLR+3.
0.4506.2152;+.NET+CLR+3.5.30729) 404 0 0 1765 677 456
2011-01-24    01:04:18    W3SVC115    69.10.57.50    GET
/theinstitution.aspx-80-74.176.189.52
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+GTB6.
6;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+.NET+CLR+3.
0.4506.2152;+.NET+CLR+3.5.30729) 404 0 0 1766 678 97
2011-01-24 01:05:38 W3SVC115 69.10.57.50 GET / - 80 -
74.176.189.52Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT
+5.1;+GTB6.6;+.NET+CLR+1.1.4322;+.NET+CLR+2.0.50727;+.N
ET+CLR+3.0.4506.2152;+.NET+CLR+3.5.30729) 403 14 0 1417
546 286
```

These weblogs contain different attributes The raw log files consists of 19 attributes such as Date, Time, Client IP, AuthUser , ServerName, ServerIP, SetverPort, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, protocol Version, Host, User AGENT, Cookies, Referer. These all attributes are not necessary for our work so we select some attributes which are relevant to our work. By applying the concept of regular expression, regular expression are used for matching the strings which can be character or number here we match the particular format of attributes which is required to the fields for eg date field format should be dd\mm\yyyy if this type of pattern is present in web logs then that entries will be extracted and add to the list and consider as date

```
string pattern = @"^([1-9]|[1-9][0-9]|1[0-9][0-9]|2[0-4][0-9]|25[0-5])(\.([0-9]|[1-9][0-9]|1[0-9][0-9]|2[0-4][0-9]|25[0-5])){3}$";
string dtPattern = @"^([0-9]{4,4})-([0-9]{2,2})-([0-9]{2,2})$";
Regex ipReg = new Regex(pattern);
Regex dtReg = new Regex(dtPattern);
```

By this we get the ip address and date field in the same way we can extract other fields too. Now after getting these entries in list in a tabular form.



S.No.	Datetime	Server IP	Method	URI Stem	Port	Client IP
1	2011-01-24 01:04:11	69.10.57.50	GET	/DeptofEnglish.aspx	80	74.176.189.52
2	2011-01-24 01:04:18	69.10.57.50	GET	/theinstitution.aspx	80	74.176.189.52
3	2011-01-24 01:05:38	69.10.57.50	GET		80	74.176.189.52
4	2011-01-24 02:52:11	69.10.57.50	GET	/default.aspx	80	122.168.223.143
5	2011-01-24 02:52:13	69.10.57.50	GET	/favicon.ico	80	122.168.223.143
6	2011-01-24 04:25:31	69.10.57.50	GET		80	122.168.173.254
7	2011-01-24 04:25:31	69.10.57.50	GET	/favicon.ico	80	122.168.173.254

Fig 1 log files in tabular form

For further processing this data should be saved to database. For saving data we open new connection

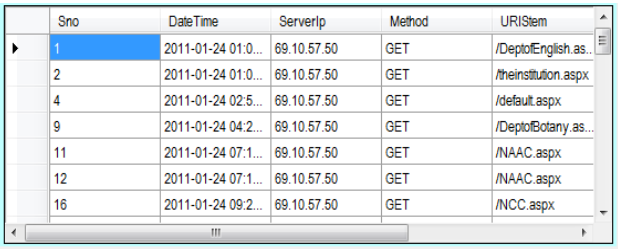
```
OleDbConnection conn = new
OleDbConnection("Provider=Microsoft.ACE.OLEDB.12.0
;Data
Source="+Environment.CurrentDirectory+"\\LogMiningData.accdb");
```

and save all the necessary entries which we have extracted from the log file.

This database contain all the entries including text file ,images , multimedia files etc which is irrelevant so cleaning of server logs will be done In this phase all the files which have extension other than .aspx are removed, and the entries having .aspx as the extension will be now in the database.

```
OleDbCommand cmd = new OleDbCommand("delete from
ProcessedData where URIStem not like '%.aspx%'", conn);
```

After removing all the irrelevant entries database will look like this



	Sno	DateTime	ServerIp	Method	URIStem
▶	1	2011-01-24 01:0...	69.10.57.50	GET	/DeptofEnglish.as...
	2	2011-01-24 01:0...	69.10.57.50	GET	/theinstitution.aspx
	4	2011-01-24 02:5...	69.10.57.50	GET	/default.aspx
	9	2011-01-24 04:2...	69.10.57.50	GET	/DeptofBotany.as...
	11	2011-01-24 07:1...	69.10.57.50	GET	/NAAC.aspx
	12	2011-01-24 07:1...	69.10.57.50	GET	/NAAC.aspx
	16	2011-01-24 09:2...	69.10.57.50	GET	/NCC.aspx

Fig2 relevant attributes of log file

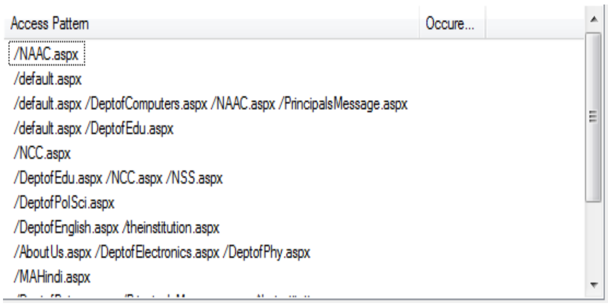
Now process of clustering is applied and clusters of all possible patterns of users are formed by applying hash table.

For implementing this two hash table are used first we named as cip which means it contain the client ip address and other one is uri list which includes url's

```
        Hashtable CipList = new Hashtable();
        Hashtable UriList = new Hashtable();
```

Now select all client ip address from the database and also select url for specific client ip which had a .aspx extension. Here the key of hash table will be the web pages. If the web page is already accessed then we move to the next entry In this way the pattern of pages of particular users will be formed. Here concept of hash table is advantageous because the particular fact is stored in hash table and system does not require to search in whole database again and again.

```
        OleDbCommand cmd = new OleDbCommand("select
distinct ClientIp from ProcessedData", conn);
        cmd.Connection.Open();
        OleDbDataReader
reader=
cmd.ExecuteReader(CommandBehavior.CloseConnection);
        while (reader.Read())
        {
          CipList.Add(reader.GetString(0), "");
        }
        cmd.Connection.Close();
OleDbCommand cmd1 = new OleDbCommand("select ClientIp,
URIStem from ProcessedData where URIStem like '%.aspx%'
group by ClientIp, URIStem", conn);
```



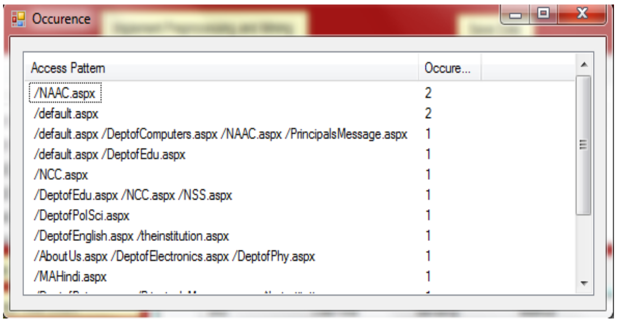
Access Pattern	Occure...
/NAAC.aspx	
/default.aspx	
/default.aspx /DeptofComputers.aspx /NAAC.aspx /PrincipalsMessage.aspx	
/default.aspx /DeptofEdu.aspx	
/NCC.aspx	
/DeptofEdu.aspx /NCC.aspx /NSS.aspx	
/DeptofPolSci.aspx	
/DeptofEnglish.aspx /theinstitution.aspx	
/AboutUs.aspx /DeptofElectronics.aspx /DeptofPhy.aspx	
/MAHindi.aspx	

Fig 3 patterns formed from logs

The occurrence of each pattern is also calculated by the function dictionary entry which is used for reading hash tables.

```
foreach (DictionaryEntry de in CipList)
{
UriList[de.Value.ToString()]                    =
Convert.ToInt16(UriList[de.Value.ToString()]) + 1;
}
```

The url are read from the hash table uri lst and search for same patterns if same pattern occurred then the counting of that pattern will be increased.



Access Pattern	Occure...
/NAAC.aspx	2
/default.aspx	2
/default.aspx /DeptofComputers.aspx /NAAC.aspx /PrincipalsMessage.aspx	1
/default.aspx /DeptofEdu.aspx	1
/NCC.aspx	1
/DeptofEdu.aspx /NCC.aspx /NSS.aspx	1
/DeptofPolSci.aspx	1
/DeptofEnglish.aspx /theinstitution.aspx	1
/AboutUs.aspx /DeptofElectronics.aspx /DeptofPhy.aspx	1
/MAHindi.aspx	1

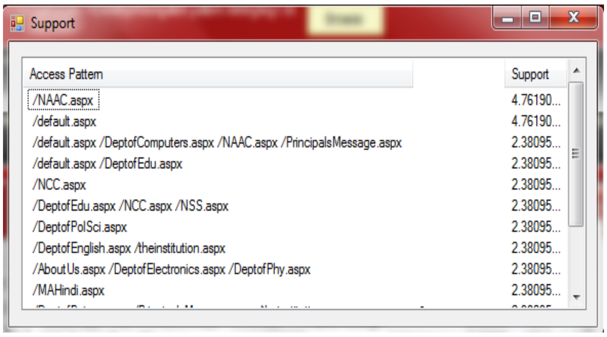
Fig 4 occurence of pattern

Support and confidence of each pattern is calculated support is the measure of interestingness which gives information about web pages that appear together

```
for (int i = 0; i < listView4.Items.Count; i++)
        {

            listView4.Items[i].SubItems[2].Text=Convert.ToS
            tring(Convert.ToDecimal(listView4.Items[i].SubIt
            ems[1].Text)
            / cnt5 * 100);
        }
```

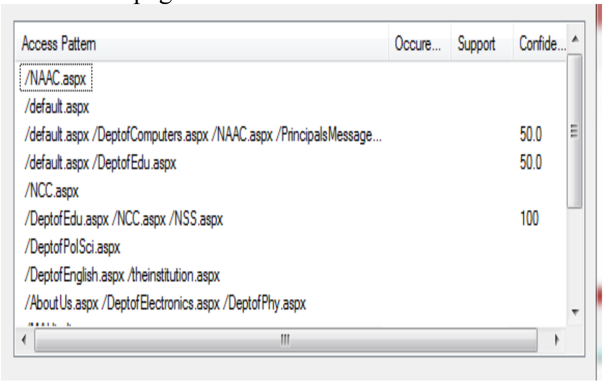
List view 4 contains the navigation pattern of users and cnt 5 contain the total transaction in database



Access Pattern	Support
/NAAC.aspx	4.76190...
/default.aspx	4.76190...
/default.aspx /DeptofComputers.aspx /NAAC.aspx /PrincipalsMessage.aspx	2.38095...
/default.aspx /DeptofEdu.aspx	2.38095...
/NCC.aspx	2.38095...
/DeptofEdu.aspx /NCC.aspx /NSS.aspx	2.38095...
/DeptofPolSci.aspx	2.38095...
/DeptofEnglish.aspx /theinstitution.aspx	2.38095...
/AboutUs.aspx /DeptofElectronics.aspx /DeptofPhy.aspx	2.38095...
/MAHindi.aspx	2.38095...

Fig 5 support count

Confidence is the measure by which we conclude that occurrence of pages one after the another.



Access Pattern	Occure...	Support	Confide...
/NAAC.aspx			
/default.aspx			
/default.aspx /DeptofComputers.aspx /NAAC.aspx /PrincipalsMessage...			50.0
/default.aspx /DeptofEdu.aspx			50.0
/NCC.aspx			
/DeptofEdu.aspx /NCC.aspx /NSS.aspx			100
/DeptofPolSci.aspx			
/DeptofEnglish.aspx /theinstitution.aspx			
/AboutUs.aspx /DeptofElectronics.aspx /DeptofPhy.aspx			

Fig 6 confidence

CONCLUSION

The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These features have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for detecting pattern. In this paper we tried to give a clear understanding of the web server logs, and interesting patterns extracted from the web logs.also The support and the confidence values of extracted patterns are considered for obtaining the interest of the web visitors. Collecting the interesting patterns using the required interestingness measures, which help us in discovering the sophisticated patterns that are ultimately used for developing the site. This will also help in evaluating address campaigns, restructuring and redesigning of website The experimental results represents that our approach can improve the quality of clustering for user navigation pattern in web usage mining systems by covering all page views in the clusters. For the future, these clusters can be used for classification methods for classifying user request. In addition, more log files of longer periods of time (such as months) are required to fabricate more reliable and more useful cluster of patterns , which will improve further the performance of the Web Servers.Since this is a huge area, and there a lot of work to do, this is the starting point

.

REFERENCES

[1] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, vol. 43, pp. 142-151, 2000.

[2] F. Masseglia, P. Poncelet, and R. Cicchetti, An Efficient Algorithm for Web Usage Mining, *Networking and Information Systems Journal (NIS)*, 2(5-6), pp. 571-603, 1999.

[3] R. Cooley, Web Usage Mining: Discovery and Application of Interesting patterns from *Web* Data, Ph. D. Thesis, University of Minnesota, Department of Computer Science, 2000.

[4] P. Pirolli, J. Pitkow, and R. Rao, Silk From a Sow‘s Ear: Extracting Usable Structures from the Web, *Proceeding on Human Factors in Computing* Systems *(CHI'96)*, ACM Press, pp. 118-125, 1996.

[5] M. Spiliopoulou, and L.C. Faulstich, WUM: A Web Utilization Miner,*proceeding of EDBT Workshop on the Web and Data Bases (WebDB'98)*, Springer Verlag, pp. 109-115, 1999.

[6] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 1(2), pp. 12-23, 2000.

[7] Jalali, M., Mustapha, N., Sulaiman, M. N. B. And Mamat, A. (2008a) OPWUMP: An Architecture for Online Predicting in WUM-Based Personalization System, Communications in Computer and Information Science, Advances in Computer Science and Engineering, Springer Berlin Heidelberg, Vol. 6, Pp. 838–841.

[8] Jalali, M., Mustapha, N., Sulaiman, N. B. and Mamat, A. (2008b) A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems," 12th International on Information Visualisation,IV'08, London, UK, Pp. 302-307.

[9] F. Bonchi , F. Giannotti , C. *Gozzi* , G. Manco, M. Nanni , D. Pedreschi, C. Renso , S. Ruggieri, Web log data warehousing and mining for intelligent web caching , *Data Knowl Eng*, 39(2), pp. 165– 189, 2001.

[10] B. Hay , G. Wets, K. Vanhoof , Mining navigation patterns using a Sequence alignment method , *Knowl Inf Syst*, 6(2), pp.150–163, 2004.