International Journal of Advance Research in Science and Engineering
Vol. No.5, Special Issue No. (01), February 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

PERSONALIZED AUTOMATIC RECOMMENDATION WITH PAGE RANKING SYSTEM

Priyanka Panchal

Department of Information Technology,
Madhuben & Bhanubhai Patel Women's Institute of Engineering for Studies & Research in
Computer & Communication Technology (MBICT), New V. V. Nagar, (India)
Gujarat Technological University (GTU)

ABSTRACT

The user's accesses to Web sites are stored in Web server logs. Pre processing of the Web log data is an essential and pre-requisite phase. Web usage mining, a classification of Web mining, is the application of data mining techniques to discover usage patterns from click stream and associated data stored in one or more Web server. We are here developing a system that will recommend the user of the Web Site to the next page that he might be interested to refer according to the use of the previous users by using this web log file. Also a page ranking system that updates the page rank of the web page according to the use of the user.

Keywords: Page Ranking Algorithm, Web Log File, Web Mining, Web Recommendation, World Wide Web,

I. INTRODUCTION

The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet. The popularity of WWW is largely dependent on the search engines. Search engines are the gateways to the huge information repository at the internet. Now anyone can quickly search for helpful cleaning tips, music lyrics, recipes, pictures, celebrity websites and more with search engines. The database is a warehouse of the pages downloaded and processed and search engine results engine and digs search results out of the database. With the Internet usage gaining popularity and the steady growth of users, the World Wide Web has become a huge repository of data and serves as an important platform for the dissemination of information. Web Server Logs file contain user's navigation which is stored the user's access to web sites. However, the data stored in the log files do not present an accurate picture of the users' accesses to the Web site. Hence, pre processing of the Web log data is an essential and pre-requisite phase before it can be used for knowledge-discovery or mining task. The pre processed Web data can then be suitable for the discovery and analysis of useful information referred to as Web mining.

Web Mining classified into three categories – Web Content Mining, Web Usage Mining, and Web Structure Mining of Web data. Web usage mining collapse under on among the category of data mining which is used to discover user's browsing pattern from click stream over WWW and linked data stored in one or more Web servers.

The proposed system will recommend the user’s next navigation in Web Site that might be interested by his/her. That can be recommended according to the use of the previous history of users by using their web log file. Also a proposed page ranking system that updates the page rank of the web page according to the access of user’s.

II. RELATED WORK

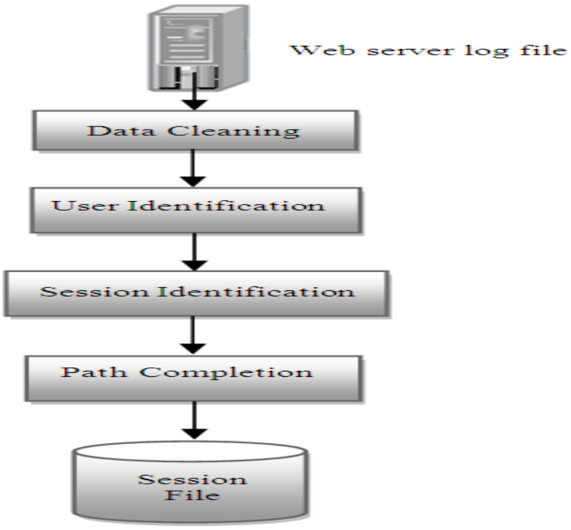
Most of users are heavily dependent on internet and other web technologies for reasons like online shopping, searching information on the web, taking surveys for giving feedback etc. Recommender systems came into existence as users were interested in retrieving either personalized queries or for fetching more elaborate and diversified results related to their searches from a very large collection of items or information. The basic function of recommender systems was to provide recommendations to the users based on items or information related to their interests and in some cases, to provide guesses or ratings to each item which the user may prefer Recommender Systems are classified into two categories based on web user’s navigation behavior and web users Ratings. Navigation behavior is the visiting of pages in a particular session of a user [5]. Page Rank, which was developed at Stanford University by Larry Page and Sergey Brin, is the most popular link analysis algorithm used to rank the results returned by a search engine after a user query. Page Rank models the behavior of a web surfer who browses the Web. The Web surfer starts from a random node on the graph, he/she clicks on hyperlinks forever and picks a link uniformly at random on each page to move on to the next Page. The number of times the surfer has visited each page is counted. Page Rank of a given page is this number divided by the total number of pages the surfer has browsed. Page Rank is a static ranking of web pages in the sense that a Page Rank value is computed for each page off-line and it does not depend on search queries [2]. Web personalization refers to any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users by using the knowledge procured from the navigational activities and individual interests of users recorded in the web usage logs, in conjunction with the content and the structure of the Web site [6].

III. EXISTING PROBLEM

Generally a user of the World Wide Web services does not go through 10 to 20 pages for collecting any information. This work is usually done by the search engines. In spite of having well sophisticated search engines yet the user ends up receiving plenty of results that may not be useful to him or that does not match his search criteria. The reason is that most of the time a user wants a particular type of page like an index page to get the links to good web pages or an article to know details about a topic. The existing search engine is a proper classification of the search pages and ranking according to that. The advanced search options of search engines take very raw input like link to or from a particular website, or mandatory portion of a query etc, which are easier to use for a search engine software but difficult to interpret and use for non computer professional person. The proposed page ranking and recommendation system of the web pages will help the end users to gather accurate result information.

IV. DATA PREPROCESSING

The purpose of data pre-processing is to change a web data mining into reliable data. The normal procedure of data pre-processing includes the following steps.



A. Data cleansing

The task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are two kinds of irrelevant or redundant data to be removed. They are:

•Additional Requests:

A user's request to view a particular page often results in several log entries. Graphics and scripts are downloaded in addition to the HTML file, because of the connectionless nature of the HTTP protocol. Since the main intention of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Suffix part of an URL is checked and eliminates suffixes like gif, jpg, GIF, JPEG, css, map etc.

•Entries with error:

Status code shows the success or failure of a request. Entries with status code less than 200 and greater than 299 are failure entries which are to be removed. Only necessary fields like date, time, IP address, User Agent, URL requested, URL referred, time taken are considered for further experiments to reduce the processing time so attribute subset selection is done.

B. User identification

User identification can be done by IP address, cookies or user registration. User identification is identifying each user from the weblog who access website. User identification group together the record for the same user from log records which are recorded in a sequential manner as they are coming from different user.

The aim of the user identification process is to find out the different users from the web access log file. Different users are being distinguished by using their Internet Protocol (IP) addresses. The method used for this process is a referrer-based method. User identification is complex due to the presence of resident caches, firewalls and proxy servers. To deal this problem, we can employ the WUM methods that rely on user cooperation. However, it's difficult because of high security and privacy. We have the following heuristics used in our testing methodology to identify the user:

1) Each IP address represents one user

2) If the IP address is same for more logs, but the agent log displays a change in browser or operating system, the IP address represents a different user

3) If there is a same IP address, browser and operating system, the referrer information can be considered. If a user requested page is not directly accessible by a link from any of these pages, hence with the same IP there is another user.

C. User session ID

Used to find all page references made by user. We differentiate the entries into different user sessions through a session timeout. If the time between page requests exceeds a certain limit, it is assumed that other user-session has started. We have used 30 minute timeout for session's timeout property value. The aim of the user session identification is to find out the different user sessions from the web access log file. A set of user clicks usually referred to as a click stream, across Web servers is defined as a user session. The user session identification involves - dividing the page accesses of every user into separate sessions. At present, we have the methods which will identify user session mainly include timeout mechanism and maximal forward reference. The following rules deployed to identify user session in our research

1) If there is a new user, and hence, there is a new session

2) If the refer page is null in one user session, there is a new session

3) It is presumed that, the user is starting a new session, if the time frame between page requests exceeds a limit (usually 25.5 or 20 minutes)

D. Path Completion

There are many important user accesses that are not being recorded in the access log due to the existence of proxy server and local cache. The aim of the path completion is to acquire complete user access path by filling up the missing page references. The incomplete access path is recognized based on user session identification. We can employ the same methods which used for user identification. For example, a user requested for a page, that is unlinked to the last page. We can use the referrer log to check what page the request came from? If the page is available in the users recent history, it is anticipated that the user has backtracked using the back button, bringing up the cached versions of the pages till a new page requested. The site topology can be used to if the referrer log in unclear to this effect. If in a start of the user session, Referrer as well URI has a data value, delete value of the referrer by adding a delimiter Web log pre processing helps in removing unwanted click-streams from the log file and also reduces the original file size by 50-55%.

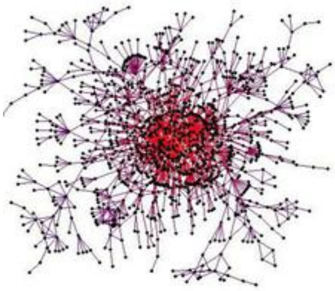
E. Personalized recommendation

Most of users are heavily dependent on internet and other web technologies for reasons like online shopping, searching information on the web, taking surveys for giving feedback etc. Recommender systems came into existence as users were interested in retrieving either personalized queries or for fetching more elaborate and diversified results related to their searches from a very large collection of items or information. The basic function of recommender systems was to provide recommendations to the users based on items or information related to their interests and in some cases, to provide guesses or ratings to each item which the user may prefer

Recommender Systems are classified into two categories based on web user's navigation behavior, and web users Ratings. Navigation behavior is the visiting of pages in a particular session of a user. Web users 'ratings are based on the ratings given by the users to the products in e-Commerce sites Implicit, explicit and both the ratings for web pages, products and users are possible in Recommender System. Generally, it provides information about the items which are recommended by the system.

V. PAGE RANKING ALGORITHM

Page rank is a link analysis algorithm for each numeric weighting to each web page, with the measure of relative importance. Page rankling usually based on hyperlink map. The Web is treated as a directed graph G = (V, E), where V is the set of vertices or nodes, i.e., the set of all pages, and E is the set of directed edges in the graph, i.e., hyperlinks. In page rank calculation, especially for larger systems, iterative calculation method is used. In this method, the calculation is implemented with cycles. In the first cycle all rank values may be assigned to a constant value such as 1, and with each iteration of calculation, the rank value become normalized within approximately 50 iterations under ε = 0.85 [2]



Hyperlinks into a page are called inlinks and point into nodes and outlink point out from nodes. When one page links to another page, it is effectively casting a vote for the other page. That more page defines more importance. The page rank equation defined as following.

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) where,

- PR(A) – Page Rank of page A
- PR(Ti) – Page Rank of pages Ti which link to page A
- C(Ti) - number of outbound links on page Ti
- d - damping factor which can be set between 0 and 1

A simple way of representing the formula is, (d=0.85) *Page Rank (PR) = 0.15 + 0.85*

The amount of Page Rank that a page has to vote will be its own value * 0.85. Accurate values are obtained through much iteration. Therefore more iteration is necessary while calculating Page Ranks

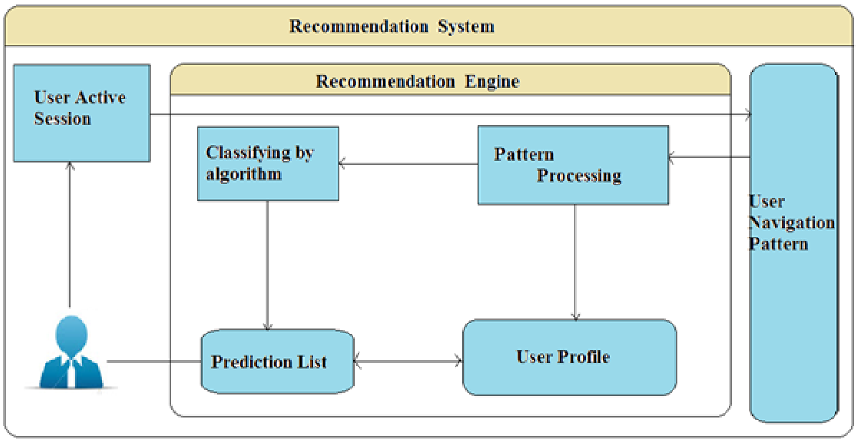
VI. PROPOSED SYSTEM

The system is mainly based on web usage mining. Web usage mining lets you to mine the details on the web application of the users and stores the navigation patterns and searched results into it. The Web Mining is done on the data that is known as the web server log files. These log files contains all the data regarding usage of the web site. The web log file consists of too many data that cannot be used simply hence data has to be pre processed for applying web mining. With the help of the web log files we can apply web mining step onto the system and create our required system.

Our system deals with the recommendation system that is personalized according to the user of the application. The recommendation engine will predict a list to the user according to his search patterns and that is compared to the end users patterns stored in the application based on web usage mining. It will predict the next web page based on the analysis pattern by known user to an anonymous user.

The system also deals with the page ranking mechanism that deals with giving ranks to each web page based on the user search queries. As the user searches for the page the ranker mechanism maintains the rank log of the web pages of the application. This is helpful while searching for a particular page the best result will be fetched first and end user will get the best output from the web application that also increases the usability of that web application This recommendation and page ranking system will be applicable for any web application by doing the analysis of the search pattern of that particular application.

A. Personalized Recommendation System

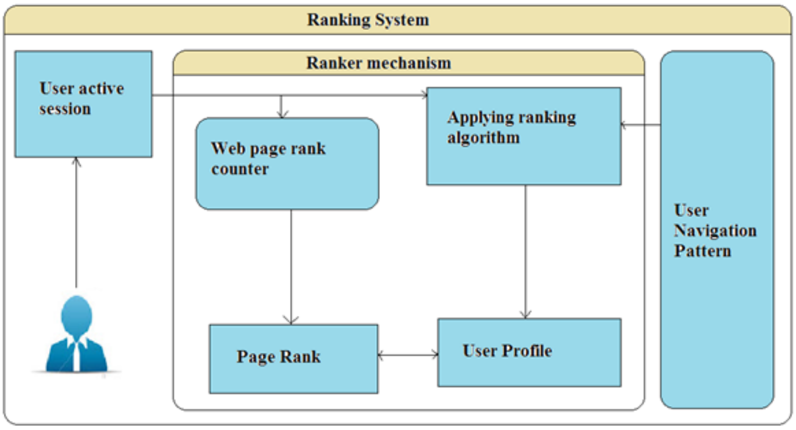


The Proposed Recommendation system model in which the process works as follows.

First the user logs in into a User Active Session that stores all the details regarding the user active mode. This session passes all the User Active Session information to the User Navigation pattern. This User Navigation patterns stores the history of search patterns of end users using the same web application. It also connects with the current user session and compares with the previous patterns.

This data is then forwarded to the recommendation engine which is responsible for generating the recommendation list to the current user. Herein the data enters into the pattern processing stage in which the active session and the previous patterns are compared and then classified by k-means clustering Algorithm. On the basis of this algorithm processing, the recommendation engine generates a prediction or recommendation list. This list is stored into the user profile and is directed to the user. Hence generating the user a recommendation list on the basis of his search patterns.

B. Page Ranking System



In the page ranking system model, in proposed system when the user logs in into a User Active Session, it is directed into the ranker machine that applies the Page Ranking Algorithm for that particular page of the web application. This is also done from the historic information of the User Navigation Patterns and the resultant of algorithm is applied on to the User Profile giving a page rank to the web page.

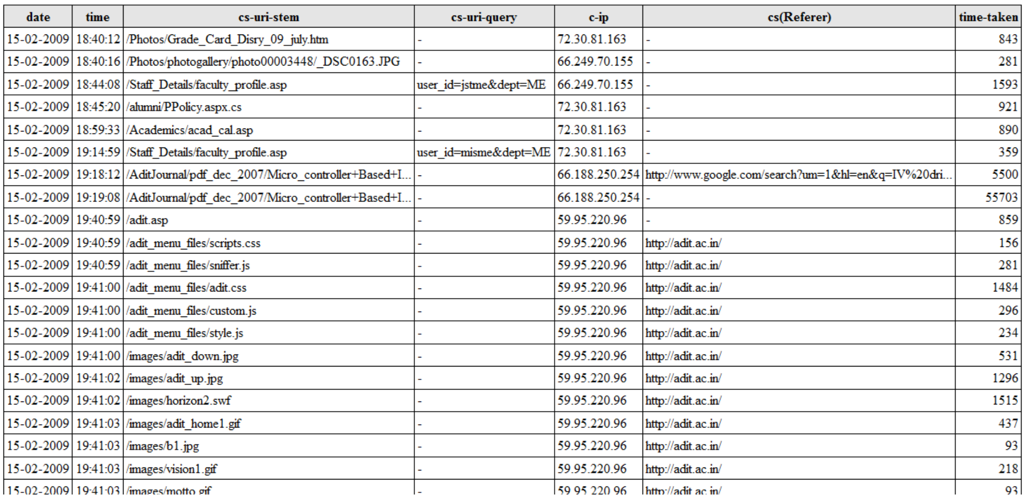
The page having existing rank also undergoes the same ranking mechanism in which the ranker system manages a web page rank counter that keeps on updating the rank according to the search of end user. Hence generating a page rank for each web page visited by users into the system.

VII. IMPLEMENTATION AND EXPERIMENTAL EVALUATION AND RESULT

Step-1 For the implementation and experimental result query one of the college web site log file has been taken as a input. Web log is an unprocessed text file which is recorded from the IIS Web Server. That web log file has number of different attribute which shows as record data in text file, that converted into sql file format. The attribute record as define below.

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

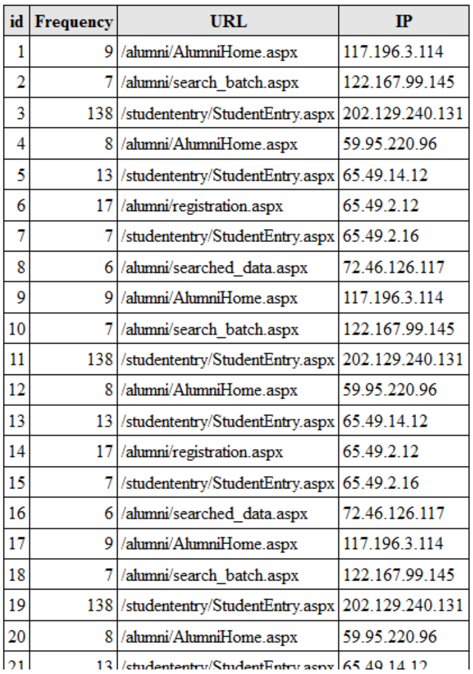
After converting data set text file input into sql file format the input Data set as shown in Table-1.



date	time	cs-uri-stem	cs-uri-query	c-ip	cs(Referer)	time-taken
15-02-2009	18:40:12	/Photos/Grade_Card_Disry_09_july.htm	-	72.30.81.163	-	843
15-02-2009	18:40:16	/Photos/photogallery/photo00003448/_DSC0163.JPG	-	66.249.70.155	-	281
15-02-2009	18:44:08	/Staff_Details/faculty_profile.asp	user_id=jstme&dept=ME	66.249.70.155	-	1593
15-02-2009	18:45:20	/alumni/PPolicy.aspx.cs	-	72.30.81.163	-	921
15-02-2009	18:59:33	/Academics/acad_cal.asp	-	72.30.81.163	-	890
15-02-2009	19:14:59	/Staff_Details/faculty_profile.asp	user_id=misme&dept=ME	72.30.81.163	-	359
15-02-2009	19:18:12	/AditJournal/pdf_dec_2007/Micro_controller+Based+I...	-	66.188.250.254	http://www.google.com/search?um=1&hl=en&q=IV%20dri...	5500
15-02-2009	19:19:08	/AditJournal/pdf_dec_2007/Micro_controller+Based+I...	-	66.188.250.254	-	55703
15-02-2009	19:40:59	/adit.asp	-	59.95.220.96	-	859
15-02-2009	19:40:59	/adit_menu_files/scripts.css	-	59.95.220.96	http://adit.ac.in/	156
15-02-2009	19:40:59	/adit_menu_files/sniffer.js	-	59.95.220.96	http://adit.ac.in/	281
15-02-2009	19:41:00	/adit_menu_files/adit.css	-	59.95.220.96	http://adit.ac.in/	1484
15-02-2009	19:41:00	/adit_menu_files/custom.js	-	59.95.220.96	http://adit.ac.in/	296
15-02-2009	19:41:00	/adit_menu_files/style.js	-	59.95.220.96	http://adit.ac.in/	234
15-02-2009	19:41:00	/images/adit_down.jpg	-	59.95.220.96	http://adit.ac.in/	531
15-02-2009	19:41:02	/images/adit_up.jpg	-	59.95.220.96	http://adit.ac.in/	1296
15-02-2009	19:41:02	/images/horizon2.swf	-	59.95.220.96	http://adit.ac.in/	1515
15-02-2009	19:41:03	/images/adit_home1.gif	-	59.95.220.96	http://adit.ac.in/	437
15-02-2009	19:41:03	/images/b1.jpg	-	59.95.220.96	http://adit.ac.in/	93
15-02-2009	19:41:03	/images/vision1.gif	-	59.95.220.96	http://adit.ac.in/	218
15-02-2009	19:41:03	/images/motto.gif	-	59.95.220.96	http://adit.ac.in/	93

Table – 1 INPUT DATA SET

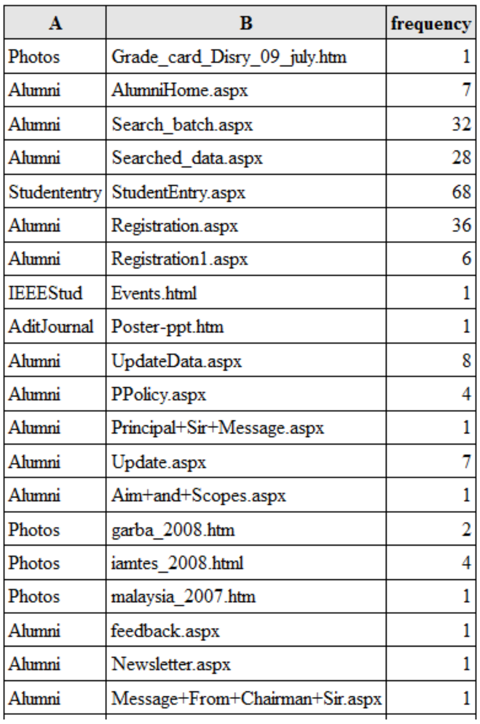
Step-2 Generally, Data Pre-Processing Task need to perform before performing any other web mining or page ranking algorithm on web server log files. Data Pre-Processing followed by several number of task such as Data Cleaning, User Identification, Session Identification, Path Completion. After that identify the frequency of each unique web page from the web log file of data pre-processing task. This is shown in Table – 2.



id	Frequency	URL	IP
1	9	/alumni/AlumniHome.aspx	117.196.3.114
2	7	/alumni/search_batch.aspx	122.167.99.145
3	138	/studententry/StudentEntry.aspx	202.129.240.131
4	8	/alumni/AlumniHome.aspx	59.95.220.96
5	13	/studententry/StudentEntry.aspx	65.49.14.12
6	17	/alumni/registration.aspx	65.49.2.12
7	7	/studententry/StudentEntry.aspx	65.49.2.16
8	6	/alumni/searched_data.aspx	72.46.126.117
9	9	/alumni/AlumniHome.aspx	117.196.3.114
10	7	/alumni/search_batch.aspx	122.167.99.145
11	138	/studententry/StudentEntry.aspx	202.129.240.131
12	8	/alumni/AlumniHome.aspx	59.95.220.96
13	13	/studententry/StudentEntry.aspx	65.49.14.12
14	17	/alumni/registration.aspx	65.49.2.12
15	7	/studententry/StudentEntry.aspx	65.49.2.16
16	6	/alumni/searched_data.aspx	72.46.126.117
17	9	/alumni/AlumniHome.aspx	117.196.3.114
18	7	/alumni/search_batch.aspx	122.167.99.145
19	138	/studententry/StudentEntry.aspx	202.129.240.131
20	8	/alumni/AlumniHome.aspx	59.95.220.96
21	13	/studententry/StudentEntry.aspx	65.49.14.12

Table-2 Frequency Count

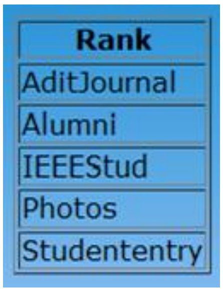
Step-3 After finding frequency of each web page, navigation between two web page frequencies is finding out. That's output as shown in below Table-3.



A	B	frequency
Photos	Grade_card_Disry_09_july.htm	1
Alumni	AlumniHome.aspx	7
Alumni	Search_batch.aspx	32
Alumni	Searched_data.aspx	28
Studententry	StudentEntry.aspx	68
Alumni	Registration.aspx	36
Alumni	Registration1.aspx	6
IEEEStud	Events.html	1
AditJournal	Poster-ppt.htm	1
Alumni	UpdateData.aspx	8
Alumni	PPolicy.aspx	4
Alumni	Principal+Sir+Message.aspx	1
Alumni	Update.aspx	7
Alumni	Aim+and+Scopes.aspx	1
Photos	garba_2008.htm	2
Photos	iamtes_2008.html	4
Photos	malaysia_2007.htm	1
Alumni	feedback.aspx	1
Alumni	Newsletter.aspx	1
Alumni	Message+From+Chairman+Sir.aspx	1

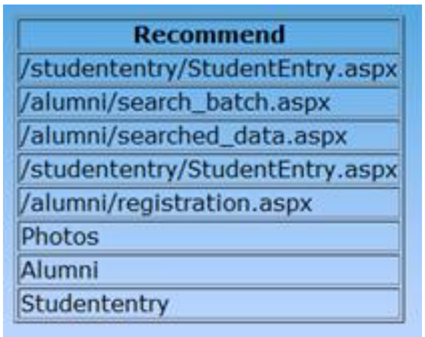
Table – 3 Navigation Frequency

Step-5 Using the proposed page ranking system approach generates the rank to the web site link page.



Rank
AditJournal
Alumni
IEEEStud
Photos
Studententry

Step-6 Recommendation is done by fetching the usage of the user. Proposed approach individually gave recommendation for each web page. Recommendation as well as page ranking is done in a single proposed system. Proposed approach deduces that recommendation can also be done on the basis of page ranking.



Recommend
/studententry/StudentEntry.aspx
/alumni/search_batch.aspx
/alumni/searched_data.aspx
/studententry/StudentEntry.aspx
/alumni/registration.aspx
Photos
Alumni
Studententry

VIII. CONCLUSION

In the traditional page ranking algorithm, multiple pages link to a single page. But proposed approach is that a single page is linked to multiple pages and ranking is done through the ranks of the multiple pages. Thus conclude research work after researching that using a-priori algorithm we get the same result every time for each web page that is not favorable. Hence proposed approach use page rank algorithm and recommend the user according to the ranking of the page and it will be automatically updated by the use of the web pages and its usage.

REFERENCES

[1] D. Ciobanu, C. E. Dinuca, "Predicting the next page that will be visited by a web surfer using Page Rank algorithm", International Journal Of Computers And Communications, Issue 1 - Volume 6, 2012.

[2] Phyu Thwe, "Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm", International Journal Of Scientific & Technology Research, Volume 2, Issue 3, March 2013

[3] Xxx

[4] Ms.Dipa Dixit, Mr Jayant Gadge, "Automatic Recommendation for Online Users Using Web Usage Mining", International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, August 2010

[5] K.Suneetha, P. Sunil Kumar Reddy, A SURVEY ON WEB RECOMMENDER SYSTEMS, IJPAPER-Vol 04, Special Issue01; 2013

[6] K. Suneetha and M. Usha Rani, “Performance Analysis of web page recommendation algorithm based on weighted sequential patterns and markov model”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 3, January 2013

[7] R. Suguna,D. Sharmila, “An Efficient Web Recommendation System using Collaborative Filtering and Pattern Discovery Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 70– No.3, May 2013

[8] Magdalini Eirinaki, Michalis Vazirgiannis, UPR: Usage-based Page Ranking for Web Personalization, 5th IEEE International Conference on Data Mining, (ICDM ’05)

[9] Priyanka Bauddha, Thirunavukkarasu, “Evolution of Web Usage Mining in Page Rank Algorithms”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[10] A.M. Sote, Dr. S. R. Pande, Application of Page Ranking Algorithm in Web Mining, IOSR Journal of Computer Science (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 47-51

[11] Anuradha, G.Lavanya Devi and M.S Prasad Babu, “Role of Web Mining Algorithms for Ranking Web Pages”, International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

[12] Rekha Jain, Dr. G. N. Purohit, “Page Ranking Algorithm for Web Mining”, International Journal of Computer Applications (0975 – 8887) Volume 13– No.5, January 2011