IJETM

ISSN: 2394-6881

International Journal of Engineering Technology and Management (IJETM)
Available Online at www.ijetm.org
Volume 2, Issue 5; September-October- 2015; Page No. 21-29

Improvising Website Navigability by Analyzing the Usage Patterns and Link Structure

[1] **Harish Kumar B T**, Assistant Professor, Dept CSE, BIT, Bangalore

[2]**Jamuna B S**, M.Tech, Dept CSE, BIT, Bangalore

[3]**Vibha L**, Professor, Dept of CSE, BNMIT, Bangalore

[4]**Venugopal K R**, Principal, UVCE, Bangalore

ABSTRACT

Designing a well-structured Website is one of the biggest milestones for the Web Developers. Each organization adopts a unique hierarchy to organize their website contents. Thehierarchy varies from one firm to other making it unclear as to where a particular document is located in the website. Organizing the website structure to meet the requirements of users is acritical challenge for the Web developers. Web developer's view towards a website is totally different from the perspectives of users. In the recent years many papers have been proposed to improve the navigability in a website, they seem to be highly unpredictable with reorganized structure and increasing the cost of finding the links to the target pages. Looking forward to improve the navigability in a website, a framework is designed that facilitates easy navigation with minimal links and minimalmodifications to the structure of a website. To prove the effectiveness experiments have been conducted using real data sets from live website. Experimental results confirm an enhanced methodology or a framework to benefit the disoriented Users.

Keywords:Links, Personalization, Structure Website, Web Log File

I. INTRODUCTION

Internet undoubtedly has been a major invent where people gain knowledge in the domain of their interest. It is approximated that 1.75 billion users use internet every day [1]. This number is found increasing every year. A survey reports massive business openings to Entrepreneurs. To enhance the online reputation, organizations are investing prominently on marketing and promoting the business through website. Despite the firms spending large amounts for website creation and promotion it is observed that website owners are facing problems with decreased number of visitors to the website. Users are found to leave the website if they do not find the relevant information in few hits. Web page navigation patterns are complicated due to its versatility and the hierarchies followed by each organization differ substantially. It is difficult to find the well-defined rules that arrange navigation in a website. Most websites, with lot of content and functionality use navigation menus. Navigation menus become complex as the website content and functionality increases and cannot be accommodated in a single menu. Websites have at least two menus; primary and secondary menu. Primary menu is main menu that contains the top level hierarchical information links. Secondary menu carries the bottom level (sub menu) information links. Primary menu navigations are best suited to secondary menu navigation.

Motivation:

Surfing, navigation, and transactions are vital activities in web usage. Navigation is very essential after users find the required website using search engines and before they make any transactions. It has been observed that many users abandoned a website because of difficulties faced during navigation and switch to a more competitive website. Generally when a user uses many paths to identify his target page, indicated that he has experienced difficulty during navigation. Hence many of the website designers try to reorganize the structure of a website. Websites framed like this can lead to various issues.

- Complete restructuring could drastically change the position of familiar objects; the fresh website may confuse users.
- Restructured website arrangement is extremely arbitrary, and the cost of confusing users after the modifications remains unexamined.
- Lastly, as website restructuring methodologies could considerably modify the existing arrangement, reformation cannot be made repeatedly.

Personalised recommendations to individual users will improve the navigability.

Contribution:

In proposed work, personalized recommendations to the website users and suggestions to the website

Corresponding author: Harish Kumar B T

administrator are generated by analyzing the web log file to identify the sessions, mini sessions and source page, backtrack page in mini session and target page. Website structure is analyzed by mining the links in all the web pages of a website and minimum links between every pair of web pages are identified using the shortest path algorithm. This helps in faster retrieval of the target page and also improves the user navigation ability.

The remaining sections of the paper are structured as follows. Section II gives the overview of the related work in the specified research area. Section III presents the general architecture of the proposed work. In section IV problem statement, aims and objectives are discussed. Proposed methodology and working example is presented in section V. section VI contains the algorithms for the proposed methodology. Results obtained for the real live website are tabulated in section VII. Section VIII contains conclusions.

II. Related Work

Internet is one of the most attractive words for today's Generation. Researches have focused their studies on improving the navigation in a website by analyzing the browsing behavior by reading web log data and extracting useful fields. A brief survey of the related work in this is area is presented below.

T. Nakayama et al., in [2] has proposed a new technique that finds the gap between the website creators' anticipations and user behavior. The detection of pages that are almost related but infrequently co-occur in visits recommends areas where website design enhancement would be suitable.

M. Perkowitz and O. Etzioni in [3] havediscussed a novel methodology that produces index pages which contain links to pages concerning to specificareas to simplify user search which uses the co-occurrence rate of pages in user traversals'. Lazarin [4] has written a book titled user-centered web development guide which escorts the readers through the procedure of designing web centered resources based on the necessities of the users. This document guides the readers from the preliminary idea of developing a website through to defining the mission of the website.

M. Kilfoil et al., in [5] have dealt with research directions on adaptive websites challenges and approaches to the solutions which are amalgamation of data mining, machine learning, user modeling, Human Computer Interface (HCI), optimization theory and graph theory.B. Mobster et al., in [6] has presented two methods based on clustering of user transactions and clustering of page reviews in order to determine intersecting cumulative profiles that can be efficiently used by recommender system for real time web personalization.

B. Mobasher, Robert Cooley et al., in [7][8] has described a methodology to personalization based on web usage using web mining techniques. This method makes the personalization process both automatic and dynamic. Techniques were developed to preprocess web usage logs, clustering URL references into sets called user transactions'. Palmer in [9] has presented a report on website usability, design and performance metrics containing download interval, navigability, site content, interactivity and responsiveness.

V. McKinney et al., in [10] have proposed hypothetically reasonable concepts for determining web-customer satisfaction by splitting the website quality into information quality (IQ) and system quality (SQ). To fulfill the growing demands of online customers, corporations are heavily spending in the development and maintenance of their websites. Internet Retailer [11] reports that the overall website operations spending increased in 2007, with one third of site operators hiking spending by at least 11 percent compared to that in 2006. R Srikant et al., in [12] has suggested how to improve a website without making too many modifications to the website by analyzing the web log. In [13] R Gupta et al., has proposed a experimental scheme that uses collective user preferences data to relink the pages to improve the page navigability.

Chang-Chun Lin et al., in [14] proposed models to condense the information overload, search depth for user surfing the web and ant colony system to reorganize website structures. In [15] Y. Fu et al., proposed an approach for re-organizing websites based on user access patterns. Robert Cooley et al., in [16] presented several data preparation techniques in order to identify unique users and user sessions.

III. Architecture and Modeling:

Inputs to the proposed work are web log file and website www.sunninedeal.com as shown in Figure 1. Web log file is preprocessed to remove the entries with 400 status code, .gif, .jpg. Web users and sessions are identified, for every user session mini sessions are identified based on the backtrack points. Source and target page of user sessions are identified and stored in the database. URL of the corresponding web log file is taken and all links are extracted from it using jsoup parser. Duplicate links and missing links are eliminated from the list of all extracted

links. Connectivity between every unique link is tested and a link path matrix is built. Finally the source page and the target page are obtained from the web log file for specific user, and the link path matrix is computed by analyzing the website structure.This is fed as an input to the recommendation algorithm which produces the personalized recommendation in terms of minimum links to reach the target.It also suggests the website owner as to how they need to improve their website structure.

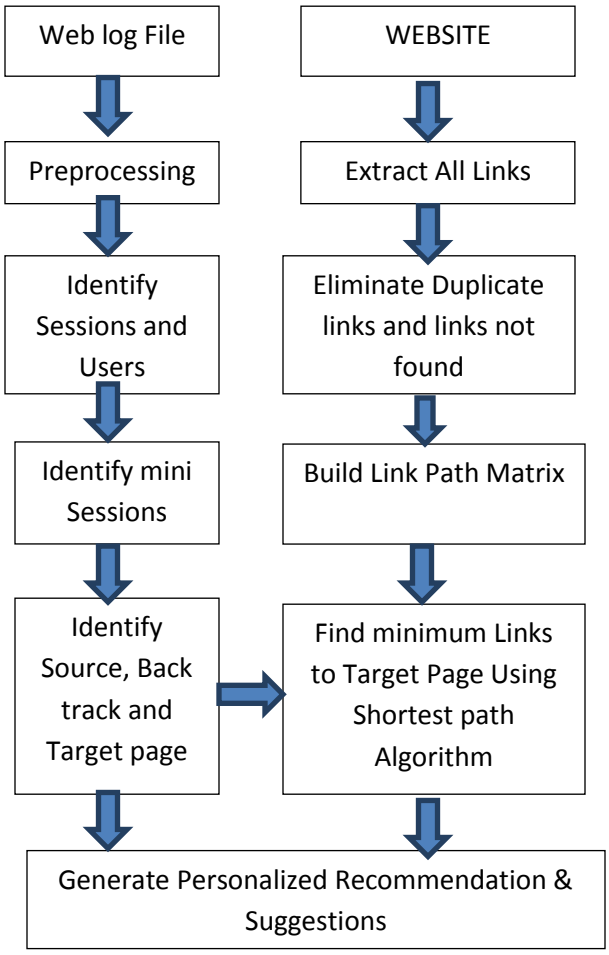

Figure 1: Architecture

IV. Problem Statement:

In the fast growing era of Internet, competition among the Online Business is exceeding the expectations. Holding back the users to their Website is becoming a difficult task. If users do not find the relevant information in less than few clicks, then theyare most likely to navigate away to another competitive Website.

Objectives:

- To develop an algorithm that helps to find the source and target page of a user.
- To design a model that improves the user navigation in a website.

Assumptions and Limitations

- Target page is identified with the page stay heuristics assumption.
- Can be applied only to art of state web log file and website structure.

V. Proposed Methodology:

To address the problems related to navigation website is modeled as a directed graph with nodes representing the web pages and links showing the connectivity in the website. In the proposed work web log file from E-commerce website www.sunninedeal.com is used. The log file is preprocessed to eliminate irrelevant fields (Time Zone, Method, Resource path, Bytes Accessed) and is stored in a database for further analysis. The user is then identified with his IP address and a session is created. Session contains set of web pages accessed by the user within a given time frame. The process of grouping the web pages that fall within a given time threshold is termed as Sessionaisation.

Once different sessions are identified, target page of the users are identified. Identifying the target page plays a key role in personalizing the navigation for each individual. The main idea behind providing a personalized navigation to the user is to make the user reach the target in lesser than few clicks.This can guarantee an increased usage rate of the website as well a good navigation experience to the users.

Target page is identified based on page - stay heuristics. According to page - stay heuristics users are more likely to spend more time in the page that interests them, so this is referred to as the target page. The next phase is determining the mini sessions. Mini sessions are set of pages or links or path that is taken to reach the target page in a single session. In mini session WebPages are grouped with backtrack point as shown inTable I.

Table I: Mini Sessions

Session Id	Mini Sessions
S1	{{3,1},{4},{2,6}}
S2	{{5,3},{4,1},{2,6}}
S3	{{1,4,2},{6,4}}
S4	{{2,3},{2,1},{5,4}}
S5	{{4,1},{2},{5},{2,6}}

In Table I S1, S2,S3, S4, S5 represent a mini session. Consider the mini session S1 here the user starts from page 3 and from the page stay heuristics webpage 6 is marked as the Target page. User navigates from page 3 to page 1 after failing to find the relevant page i.e. 6 he backtracks and visits page 4 and again backtracks and goes to page 2. A user finally navigates to the target page 6 via webpage 2. On the same basis in the mini Session S2 user begins from webpage 5 where the target page is found to be page 6. Backtracking at webpage 3, 1 and finally reaches the target via webpage 2. Mini session S3, S4, S5 also works in similar manner with target pages 4, 4, 6 respectively.

Backtrack point is simply a webpage where preceding and succeeding pages in the log files are identical. In some browsersthat do not support caching facility backtrack points have to be identified and set. The HTTP standard states that the browser should not request the page again when using the browser's history mechanism. Most browsers read to the cached pages when the visitor clicks on Back button.The fact isthat if there is no connectionamong pages P1 and P2, the usermighthave hit the "back" button in the browser to go from P1 to P2.

Website Structure Analysis:

Links from the website are extracted using the jsoup parser and every link is tested for the existence. If the link exists it is added to the hash table. Hash table is filtered to remove the duplicate links. Reachability from every link to every other link is checked and a link path matrix is constructed as shown in Table II.

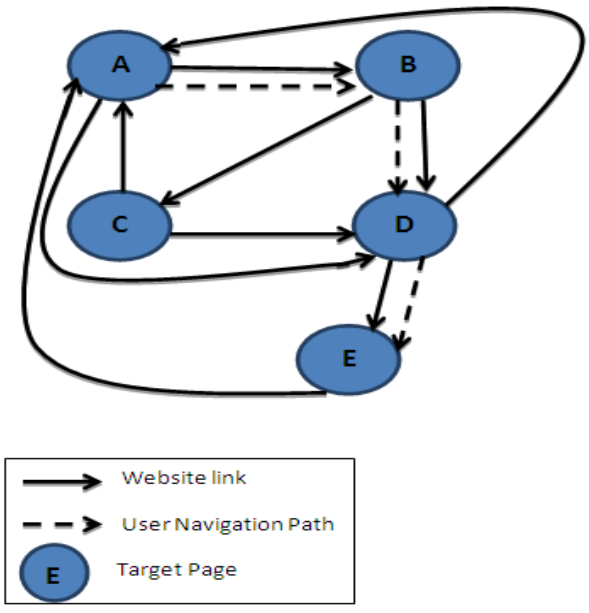

Figure 2: Website Structure and User Navigation Pattern

Table II: Link Path Matrix

	A	B	C	D	E
A	0	1	0	1	0
B	0	0	1	1	0
C	1	0	0	1	0
D	1	0	0	0	1
E	1	0	0	0	0

Working:

Source, target page and the link path matrix are given as the input to the shortest path algorithm to find the shortest path between every pair of source and the target page. The shortest path thus obtained is compared with the session path of the users and personalized recommendations are generated to every user. If there are pages that are not reachable from any of the pages, suggestions are given to the Web administrator to improve the website structure. Figure 2 shows the directed graph of the sample website with 5 web pages named A, B, C, D and E. Dotted directed arrows in Figure 2 shows the actual path taken by the user to reach the target page E (A->B->D->E) but the target page E can be reached by the user in the shortest path (A->D->E).

VI. Algorithms:

Identifying Source and Target Page Algorithm:

Input: Web Log File
Output: Source page and target pages for everyuser session.

```
/* Preprocessing */

Variable String Line;

While(Line=readLine())!=NULL)

{

//split line into tokens using space as delimiter
```

```
        String [] tokens=Split(Line, "   ");

//Entries with status code not equal to zero or

//last 3 characters of referred URL equal to JPG,
//PNG, GIF

If((token[http_status!=200) or
trim_left(token[referred_url],3)==jpg,png,gif))

        {

        Discard;

        }

        Else

        {

       Insert into log values(

        token[ip],

token[date],

token[time],

token[referred_url]);

        }

/* distinct user identification */

ArrayListDistinctusers[]; //arraylist

DistinctUser[]="Select distinct IP fromlog";

For(each Ui in DistinctUsers[])

{

//Initialize  first session ID as 1

Integer  SessionID=1;

For(each Pi ofUi)

{

If(Pi access time  is with in time frame)
```

```
insert Ui,SessionID,Pi to Session Table;

Else

SessionID=SessionID+1;

}

}

/* BackTrack Page Identification */

For(each Ui in DistinctUsers[])

        {

        For(each Si of Ui)

                {

                        For(each Pi in Si)

                        {

                        If(Pi+1==Pi)

                        {

Add Pi to BackTrackList[];

        Pi is added to the MiniSessionji(pagei);

                        }

                        Else

                        {

        Pi is added to the MiniSessionji(pagei);
                        }
}
        }
        }

/*Target Page Identification */

Targetpage[Ui,Si]=MaxTimeSpent{BackTrackList[]};
Sourcepage[Ui,Si]=Firstpage(Si);
```

Creation of Link Path Matrix Algorithm:

Input: Website Link, Source page and Target page
Output: Link Path Matrix, personalized Recommendation and Suggestions

```
/* connect to website using jsoup parser and extract all links */

Jsoup.connect(URL);

//Extract all links form the input website

Elements Links=Select("a[href]");
For(each Li in LINKs)
{
        If(Li exists) then
        {
        Add Li to the HashTable;
        }
        Else
        {
        Discard(Li);
        }
Remove(duplicate LiFromHashTable);
Integer LinkCount=Count_Entries(HashTable);
Boolean Matrix LinkPath[LinkCount][LinkCount];

For(inti=0 to Linkcount)
        {
        For(int j=0 to Linkcount)
                {
                LinkPath[i][j]=0;
                }
        }
For(each Li; in Hash table)
        {
        For(each Lj; in Hash table)
                {
                        If(Li!=Lj)
                        {
                        If(Lj is reachable from Li)
                        {
                        LinkPath[i][j]=1;
                        }
                        }
                }
        }
```

Recommendation Algorithm

Input: Source Page, Target Page, LinkPath[][]
Output: Personalized shortest path

```
String MinPath;

For(each source and target page)

   {

MinPath= diskjtra'salgm(source page, target page,

LinkPath[][]);


        If(MinPath<Sessionpath)

        {

        Recommend(MinPath to Ui)

        }

   }
```

VII. Results:

Proposed algorithms are tested on the www.sunninedeal.comanE-Commerce website structure and web log file. The web log file consisted 3days data from 14-05-2015 to 16-05-2015. Web log consisted of 200 entries from 11 different users. Raw log is preprocessed to remove the entries with 400 status code, .gif, .jpg, .pdf etc., and filtered to 72 entries.Table 3, Table 4 and Table 5 shows the sessions obtained for 11 different users for 10mins, 20mins and 30minssession threshold time respectively.

Table Header Description:

H1-users.

H2-Number of Sessions.

H3-Number of Recommendations.

H4-Number of links in original path to reach target.

H5-Number of links in the recommended path to reach the target.

Table 3: Recommendations Results (Threshold time=10min)

H1	H2	H3	H4	H5
106.51.226.113	16	8	3	1
			5	1
			3	1
			5	1
			3	1
			5	1
			3	1
			5	1
106.51.233.29	2	2	6	1
			3	1
112.79.35.48	1	0	4	--
117.251.67.71	1	0	1	--
122.172.75.247	1	1	7	1
14.98.149.118	1	0	1	0
160.62.13.190	1	1	5	1
168.235.195.41	1	0	4	0
27.63.32.165	1	0	2	0
49.205.125.130	1	0	4	0
61.135.190.103	1	0	4	0

Table 4: Recommendations Results (Threshold time=20min)

H1	H2	H3	H4	H5
106.51.226.113	16	8	3	1
			5	1
			3	1
			5	1
			3	1
			5	1
			3	1
			5	1
106.51.233.29	1	1	9	1
112.79.35.48	1	0	4	--
117.251.67.71	1	0	1	--
122.172.75.247	1	1	7	1
14.98.149.118	1	0	1	--
160.62.13.190	1	1	5	1
168.235.195.41	1	0	4	--

27.63.32.165	1	0	2	--
49.205.125.130	1	0	4	--
61.135.190.103	1	0	4	--

Table 5: Recommendations Results (Threshold time=30min)

H1	H2	H3	H4	H5
106.51.226.113	12	4	8	1
			8	1
			8	1
			8	1
106.51.233.29	1	1	8	1
112.79.35.48	1	0	4	--
117.251.67.71	1	0	1	--
122.172.75.247	1	1	7	1
14.98.149.118	1	0	1	--
160.62.13.190	1	1	5	1
168.235.195.41	1	0	4	--
27.63.32.165	1	0	2	--
49.205.125.130	1	0	4	--
61.135.190.103	1	0	4	--

VIII. Conclusion:

The proposed work aims to improve the user navigability within the website and provides suggestions to the web owner to improve the website structure. The resulting system analyses user access behavior based on the web log data and website structure by building the link path matrix. The proposed work can be used to ease thenavigability of the website users in large informational websites and e-commerce websites.

References

[1] Pingdom,"Internet 2009 in Numbers", [Online]Available: http://royal.pingdom.com/2010/01/22/internet-2009-innumbers/,2010.

[2] T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gap between Website Designers' Expectations and Users' Behavior," *Computer Networks, vol. 33, pp. 811-822, 2000*.

[3] M. Perkowitz and O. Etzioni, "Towards Adaptive Websites: Conceptual Framework and Case Study," *Artificial Intelligence, vol. 118, pp. 245-275, 2000.*

[4] J. Lazar, "User-Centered Web Development", *Jones and Bartlett Publishers, 2001.*

[5] M. Kilfoil et al., "Toward an Adaptive Web: The State of the Art and Science," *Proceedings of Communication Network and Services Research Conf., pp. 119-130, 2003.*

[6] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," *Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 61-82, 2002.*

[7] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," *Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.*

[8] B. Mobasher, R. Cooley, and J. Srivastava, "Creating Adaptive Web Sites through Usage-Based Clustering of URLs," *Proceedings of Workshop Knowledge and Data Engineering.Exchange, 1999.*

[9] J. Palmer, "Web Site Usability, Design, and Performance Metrics," *Information SystemsResearch, vol. 13, no. 2, pp. 151-167, 2002.*

[10] V. McKinney, K. Yoon, and F. Zahedi, "The Measurement of Web- Customer Satisfaction: An Expectation and Disconfirmation Approach," *Information Systems Research, vol. 13, no. 3, pp. 296- 315, 2002.*

[11] Internetretailer, "Web Tech Spending Static-But High-for the Busiest E-CommerceSites," *http://www.internetretailer.com/dailyNews.asp?id = 23440, 2007.*

[12] R. Srikant and Y. Yang, "Mining Web Logs to Improve Web Site Organization," *Proceedings of*

10th International Conference.World Wide Web, pp. 430-437, 2001.

[13] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," *INFORMS J.Computing, vol. 19, no. 1, pp. 127-136, 2007.*

[14] C.C. Lin and L. Tseng, "Website Reorganization Using an Ant Colony System," *Expert Systems with Applications, vol. 37, no. 12, pp.7598–7605, 2010*

[15] Y. Fu, M.Y. Shih, M. Creado, and C. Ju, "Reorganizing Web Sites Based on User Access Patterns," *Intelligent Systems in Accounting, Finance and Management, vol. 11, no. 1, pp. 39-53, 2002*

[16] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowledge and information Systems, vol. 1, pp. 1-27, 1999*