Improving User Navigation through Web Mining

Priyanka Dhas

Abstract - The growing availability of information on the internet has raised users utility of the webbing environment. Websites are good source of information but still they fail to provide desired information to user. They need to traverse many web pages in search of desired information which make user dissatisfied. Websites are developed by web masters, those work on business agenda and imagination of developer. User’s involvement is negligible also behaviour of users on web is not analysed. Due to this gap user face problem in navigation because expected location of information is different than existing location. Thus it creates poor navigation efficiency of website. This paper has explained a system which provides solution to improve user’s navigation by applying Web usage mining and Web structure mining.

Keywords — Navigation Efficiency, Mathematical model, TimeSpent metric, PageVisit metric, Web Usage Mining.

I. INTRODUCTION

An immense data is available related to internet in various forms such as large brings amount of web pages or in the form of usage data which get stored at server side. Web Mining [1] is the area which brings out interesting knowledge from Internet. Using web mining technique effectively [2] is an important task. It can be used efficiently if information on the internet is available in unstructured way. Numbers of websites are having loads of information on the desired topics but many times user did not get the targeted data. Low navigation efficiency [3] is the reason behind it which is arises due to lack of user involvement in web design. Making user navigation easy [4] is the key goal of today’s website developers which can be positively achieved if developers consider users usage pattern, how they have interest with website, their psychology behind traversal of pages in website. Web Transformation [5] is a process of modifying website structure with respect to all users navigation pattern or which will improve user’s navigation in common manner. This process requires common web usage data i.e. working on weblogs can provide us adequate data to solve problem of user navigation. Therefore improving user navigation through web mining can be achieved by focussing on web usage and web structure mining as aim is to enhance the link structure of website instead of reorganizing website fully. Using this process generalized restructuring of algorithm will be achieved in minimum time [6].

II. RELATED WORK

Website structures are developed with high investment and huge information is deployed [4] over it but does not work intelligently mostly. User starts browsing pages and may face difficulty in fetching required data reasons are extracted by various researchers through analysing web structure and user’s data. The best way to ensure that a Web site meets the users’ needs is to involve the users throughout the design or redesign process. The answer for a question how user can get involved is given in [7] also provided a process-oriented approach to designing user-centered Web sites. Information Foraging Theory is based on human information-seeking and sense-making behaviour [8][9] which is relevant to behaviour of animals for foraging food, also based on concept information scent. Same concept is applied in information scent where user’s behaviour and pattern for searching on websites is consider for content position on Website.

Identifying Users Traversal Path [10] is also important task in the defined problem. It works for
identifying number of pages user has traversed to locate expected location literature as provided term traversal path i.e. taken as measure to the number of times the user has attempted to locate one target.

**Extracting useful usage patterns from web logs** deals with mining weblogs to get information providing traversal pattern are explored by [11]. Research has derived algorithm MF to convert the original sequence of log data into a set of maximal forward references. Through this procedure impact of backward references get filtered and focused on meaningful user access sequences. Existing system in [3] has provided efficient solution for facilitating user navigation through website structure improvement. This system has defined problem as a special graph optimization problem where website is considered as directed graph. Main objective is to provide goal for navigation to web masters so they can concentrate on the improvement of web structure with respect to user’s pattern of access. This system has used mathematical model to achieve improvement in navigation.

### III. SYSTEM ARCHITECTURE & WORKFLOW

The proposed system is working for improving user navigation on the web site by applying various web mining techniques. This is specifically working in web transformation approach. Outcome of the system is in terms of statistical data which will help web masters for suggesting how many links can be improved and what kind of links can be considered for refinement of website structure. The system has processes server web logs and step by step it works to get links from usage data which can be modified for improving navigation efficiency. Proposed system has worked on three metrics independently. Following figure express general architectural block diagram of proposed system. First block represent input for the system which can be any server log file with similar format of data. Next block is data pre-processing, logs used in this system is already cleaned in general manner so this module pre-process it as per system requirement. In usage data mining module pre-processed logs are mined to get knowledge base for further processing. Mathematical module

![Architecture of Proposed System](image-url)

Mathematical module is implemented on the basis of mathematical formulation provided in existing system. Elaborating given formulation by considering additional metrics in the study.

The workflow is based on the architecture of the system which consists of the following steps:

- A. Server web log collection.
- B. Pre processing on available web logs.
- C. Extraction of metric as per constraints.
- D. Applying mathematical model individually for each case.
- E. Output in the form of matrix.
- F. Final links which can be improved.

#### A. Server web log collection

Logs are generally records of various activities done in particular span. Here in this system required data is collected from authenticated web server of DePaul University.

#### B. Pre processing on available web log

It shows pre processing level wise. First level initially separates all session by using stringtokenizer method with the help of tokens used in session file for separation of individual sequence. Extracted session is then converted into session number by...
analyzing page sequences traversed by user in each session and then replacing each sequence with page id. Finally outcome of first level is stored in form of array in textual format. Second level uses outcome of first level and further process for calculating page visits of each page of the website. This page visit record is stored in the form of matrix.

C. Extraction of metric as per constraints

This module comes under usage data mining where system is trying to demarcate user’s usage data in the form of metrics by applying some theories provided through web usage mining literature survey.

Page Visit as a metric: In this how many times page is visited by a user is considered as a prime factor to demarcate target pages from the session. This is achieved by calculating click threshold values on the basis of methods provided in literature. Relevant links which can be forwarded for improvement are extracted through process.

Time Spent as metric: This metric is used in existing system as a prime factor for extraction of relevant links which can be improved. Survey of existing system indicates time spent on any page can be considered for identification of target pages. The dataset used in system provides timestamp of each page visited by user in every session.

Time Spent and Page Visit as a metric: Working under this parameter will extract relevant links for improvement by considering both above mentioned parameter. It is influenced with Page Stay Algorithm and Time Oriented Algorithm. It considers both the parameter to achieve improved links hence provides optimal solution as compared to individual case.

D. Applying mathematical model

Mathematical model is applied individually for described metrics. It is implemented in two ways one as per existing parameters applied in formulation. Second is modified with respect to additional metrics and applying threshold on the basis of dataset requirement. This phase is applying algorithm individually for every metric. For example considering TimeSpent as a metric it extracts time spent on every link by users in each session. When both the parameters are considered it shows more precised target page and extracts links which can be improved.

E. Output in the form of matrix

This step generates individual files which stores the values of page on the basis of metric applied. These values are time and number of page visit of every page. All the above illustrated tables are shown with few values from 683 pages.

F. Final links which can be improved

Final links can be identified by providing constraints on the parameters which are considered in target pages. These are taken in the form of threshold values. Output of each case is shown in further session.

IV. EXPERIMENTAL WORK

As per workflow explained above experiment on the web logs are set with the help of java platform. Output on every step for TimeSpent and PageVisit is illustrated below.

A. Output on the basis of TimeSpent

TimeSpent is a metric used in first module where algorithm extracts time spent on every page by all session. It is calculated by reading all URL in an every session and summing up time if same URL is visited repeatedly. Following formula is used to calculate time.

Let’s consider

\[ S = \text{Set of all mini session from } 1 \text{ to } n. \]

Where \( n = 20950 \).

\[ Ts = \text{Time spent on each page}. \]

\[ P(t) = \text{Time spent on each page occurring } n \text{ times}. \]

\[ Ts = \sum_{S=1}^{n} P(t)_{i=0 \text{ to } n} \]
Result is described in below given table. Calculated time for every page in a dataset is shown with few examples in a table.

Table I Time Spent on Page Id 0 to 2

<table>
<thead>
<tr>
<th>Page Id</th>
<th>Time in Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4761.609</td>
</tr>
<tr>
<td>1</td>
<td>22.783</td>
</tr>
<tr>
<td>2</td>
<td>5.479</td>
</tr>
</tbody>
</table>

B. Output on the basis of Page Visit

Page Visit considers number of time any page visited by users in a session. Applying algorithm on web logs first page sequences in every session is extracted and replaced with page Id of each page link. On every appearance of a link in a session page visit is stored in the form of matrix. Also a file is generated which stores output. Following tables shows some output values. Following formula is used to calculate number of page visit on each page. Let’s consider

\[ S = \text{Set of all mini session from 1 to n}, \]
\[ \text{Where n}=20950. \]
\[ P_i = \text{Page id of a page, where i is page id from 1 to k}; \]
\[ \text{Where k}=683 \]
\[ \sum_{s=1}^{n} P(\text{occurrence})_{i=1 \text{ to } k} \]

Table II Page Visit count of Page Id 0 to 4

<table>
<thead>
<tr>
<th>PagelId</th>
<th>Count of Page Visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1120</td>
</tr>
<tr>
<td>1</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>208</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
</tr>
</tbody>
</table>

On the basis of these counts target pages are extracted which can be used further for improvement of Web structure.

V. RESULTS

Following table shows result for Page Visit metric. This table is providing values got through a result of page visit applying click threshold 1000 and path required to reach target page. The path in the table is described as how many pages user needs to traverse for getting desired information. Path 1 illustrates within a single click target is while Path 2 represents users traversal path to get target in two clicks.

Table III Count of links for improvement

<table>
<thead>
<tr>
<th>Page Id</th>
<th>Path 1</th>
<th>Path 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homelink 0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Homelink 59</td>
<td>136</td>
<td>135</td>
</tr>
<tr>
<td>Homelink 415</td>
<td>155</td>
<td>0</td>
</tr>
<tr>
<td>Homelink 574</td>
<td>79</td>
<td>58</td>
</tr>
<tr>
<td>Total</td>
<td>376</td>
<td>193</td>
</tr>
</tbody>
</table>

This illustrates how links to be improved are changing if we take Path 1 it improves 376 links out of 683. Similarly Path 2 which need to traverse at least for 2 or three pages to get desired links improve 193 links out of 683. All the data representation is shown only for the HomeLinks. Time Spent and Page Visit are applied together shows more precise extraction of target pages. It is entered count on the basis of page visit threshold 500 which is approx average value for all 683 pages and threshold value for time spent is calculated through median of all the time spent values for all pages gives total 9 links for improvement purpose. This links are having approximate values for time spent and page visit.
VI. CONCLUSION

Identifying user’s interested links is a heuristic. It cannot be extracted with cent percent accuracy. On the basis of combined metric in a system these targeted pages can be identified as compared to the existing system. These time and visit parameters have approximately similar proportion of values. It consider the pages can be target page for user if those are visited multiple time by various users as well as time spent on particular page is also more. This combination extracts page more precisely as compared to individual metric. This study provides wide vision to extract links from heavy weblog so target page extraction will be done more precisely. When combined metric are used accuracy for extracting links for improvement is arise. This will ease web developer to adopt the system for improvisation of their website structure in terms of navigation efficiency. This system can be enhanced in a future by applying personalization and content mining to enhance efficiency. It will increase accuracy to identify target pages. A tool can be developed which will compatible any structure of web logs.

Acknowledgment

The author would like to take this opportunity to express profound gratitude and deep regard to my Project Guide Prof. Sarika M. Chavan, for her guidance, valuable feedback and constant encouragement throughout the duration of the project work. I express deep gratitude to Prof. Sanjay B. Kalyankar, Head of Dept. of Computer Science and Engineering. I am thankful to DePaul University for making availability of their Server web log for the purpose of project work.

REFERENCES