Indian Journal of Science and Technology, *Vol 11(3), DOI: 10.17485/ijst/2018/v11i3/118440,* January *2018*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Deriving user Interest by Mining User Navigation Patterns

Nirali Honest*

Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charusat – 388421, Gujarat, India; niralihonest.mca@charusat.ac.in

Abstract

Objectives: The paper focuses on mining the usage patterns to help the user find the pages easily while visiting the website. In order to carry out discovery patterns first the patterns are created and then analyzed for the performance issues. **Methods/Statistical Analysis:** The quality of a website can be measured by more than one ways, one can consider the aesthetic view of the pages, others can view how the pages are ordered to make it convenient for the user to navigate. We focus on the second perspective of the quality, as it aims at helping the user what it want to find while visiting the website.In order to know what the user wants when he visits the website, it is important to know their browsing patterns. **Findings:** A lot of good information can be derived from mining log files generated as part of user access to the website. Log files can be mined for basically two reasons 1) To predict the pages that user may prefer to visit and 2) To know the core reason for visiting the site. In order to know the motive of the user visiting the website, it is important to build the browsing patterns of the user. **Application/Improvements:** In this paper, we present a new way of building patterns to derive user interest by applying predefined classes and assigning weights to the pages accessed during the visit of the user. The algorithm shows better accuracy in predicting the web pages that can be referred by the user in future while navigating the site.

Keywords: Classification, Web Log Analysis, Web Usage Mining, User Interest, User Navigaiton Pattern Identification

1.Introduction

Web mining research can be classified into various categories[1,2] as, Web Content Mining (WCM), Web Structure Mining (WSM), Web Usage Mining (WUM). WUM is the process of deriving significant behavior approaches of users who visit the site[3], which can be of great help to website's administrator by analyzing the generated behavior. To analyze the behavior of users it is first necessary to build the data required to form significant patterns. The process of building the data is depicted in Figure 1.

The first step is to collect the raw log file, in our work we have W3C Extended Log File Format, which is converted in the CSV file and then cleaning and preprocessing is applied to prepare the data for future processing. The details of the preprocessing phase is covered in[4-6] Authors in[7] presents a method to predict the future clicked pages of a user, after the user has input a query based on the page url, the future pages can be predicted. For building the user navigation patterns we read the user sessions and work on mainly three parameters to form the navigation patterns, like page order, number of times the page occur, knowing the adjacent pages for a given length and checking these parameters in regular access as well as based on some special event like the admission process, recruitment process, and many more events related to a university website, the details of building the pages is covered in[8]. Authors[9] performs page ranking by identifying URLs that are alike user's past queries and clicked pages. Forming user navigation pattern is very important as the user interest is derived based on this pattern. Authors[10] and[11] applied association rules for capturing related pages. Based on previous page visits future page accesses can be formed[12].

*Author for correspondence

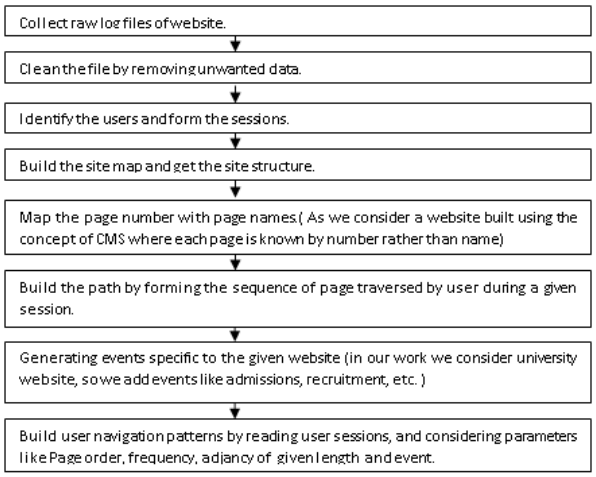

Figure 1. Steps Involved in Building user Navigation Patterns.

After forming the user navigation we can derive the user interest in navigating the websites, our approach for forming the user interest is described in the next section. Previous work is carried out for knowing the user interest and it is mainly concerned with the semantic context of pages. Authors[13] focus on the documents from which the words are extracted for considering the recent pages of interest by the user, they don't consider the access pattern of users. Authors[14] consider the web usage logs and semantics for web personalization. Authors[15] decide the type of pages as auxiliary and content pages based on the time spent by the user on a given page. They use the Maximum Forward Length approach to identify the number of pages accessed by a user before a backward reference is made. They use time based heurastics to decide the type of page, we use the name of page to classify the page in a in a class based on the website hierarchy, we don't focus on the content of the page, we analyze the user pages accessed and classify them for a given class. The pages identified in the class decide the final user interest in accessing the website. All the authors have considered different websites having static name of pages,we have considered website which is designed using the concept of Content Management System (CMS) where pages don't have the name they are accessed using a unique page id. Page name plays a significant role in such sites. We first map the page id with name and then use the name of pages in determining the user interest.

2. Building Patterns for User Interest

the user's intuition in visiting the website is one of the important patterns in web usage mining. This pattern can help the administrators to know what kind of pages are accessed by a given user interest. The general process of building pattern is depicted in the Figure 2.

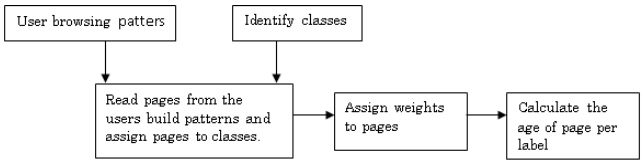

Figure 2. Assigning Classes and Weights of Page and Calculating Age.

2.1 Define the Class Labels

For carrying out this work we define class labels and assign pages to classes. While assigning pages to class label, we also assign weights to each page based on the level of the page from the root. Classification consists of assigning a class label to a set of unclassified cases. In Supervised Classification the set of possible classes is known in advance. We define classes based on the type of pages from the given website. Class assignment is shown in Table 1. We define classes based on the type of pages from the given website.

Table 1. Assigning Pages under the Predefined Classes

Class Label Ci	Pages Falling under this label Pi(w) (where i is the ith page and w is the weight of ith Page.
C1	P1(1),P2(2) , ------,Pm(n)
C2 ‘ ‘ ‘	P1(1),P2(2) , ------,Pm(n) ‘ ‘ ‘
Cn	P1(1),P2(2) , ------,Pm(n)

Algorithm 1. Define Classes and weights to pages of the website

Input: Pages of a given website

Output: The Class Label (C_i) and Weight (w_i) assigned to the page.

Define the main class C

Define the classes under C by C1, C2, C3,……,Cn. where n is the total number of sub-classes identified under the main class C.

Read the set of pages P1, P2,….., Pn of a given website W, where W={P1,P2,P3,…..,Pn} where n is the total number of pages for a given website W.

Read the page Pi where i is the ith page in the set W={P1,P2,…Pi,……,Pn}

parse the page name in m parts where m is the length of the path until leaf node

Pi(length) = (p1/p2/…/pm)

Read the p2 name to classify the page in the Class Ci where i is the ith name of the class

in the set C={C1,C2,..Ci,…,Cn}

Find the length of the page in the order of occurance using Maximum Forward length and create weight w for the page by calculating the order of page in the hierarchy of site pages.

Assign the page Pi with weight wi, to Class Ci.

Repeat the steps for all the pages of a Pi of W.

Store the details

We consider the University website and based on that we have identified the below classes.

Defining the class labels help to derive the user interest based on the knowledge gained from the maximum pages accessed from a given label and the order in which the pages are accessed form the given class label. The steps for defining classes and labels are shown as below,

After the execution of steps we can form the class labels with pages as follows,We consider the University website and based on that we have identified the below classes , and the pages are assigned to the classes based on the page name, as shown in Table 2.

Table 2. Assigning Pages under the Predefined Classes for a given Website

Class Label	Pages Falling under this label
University Information	University at glance, About University, About Turst, Campus, Donars, etc.
Institute Information	Visit sub institute pages
Faculty Information	Visit faculty information pages
Admissions	UG and PG pages showing information for admission
Downloads	Syllabus, advertisements, etc.

After forming the patterns for page accesses and page classification based on the page name, we can analyze the patterns in the next phase.

2.2 Derive the user Interest

Author in[16] has proposed the concept of classification for finding the index and the content page. They work on static html pages where classification is based on the html page. We use the classification concept in determining the general category of a page which we are predetermining.

Classification consists of assigning a class label to a set of unclassified cases.

- Supervised Classification: The set of possible classes is known in advance.
- Unsupervised Classification: Set of possible classes is not known. After the classification, we can try to assign a name to that class. Unsupervised classification is called clustering.

Supervised Classification includes methods like Bayesian classifier and Robust Bayesian Classifier are used. For Prediction of User Interest Stochastic dominance and Weak Dominance methods can be used.In Stochastic dominance we first assign a class label as cj and if pmin(cj/ek) is greater than pmax(ch/ek) for all h≠j.Each page is assigned a predefined class and weight, which can be further used in calculation of user interest. Considering assign class label as cj if pmin(cj/ek) is greater than pmax(ch/ek) for all h≠j

The process for deriving the user interest consists of two steps, 1. Build the user interest pattern from the accessed pages and 2. Calculate the interest.

For building the user interest, we read the already built navigation patterns for each user which consist of all the pages for a given session as shown in below formulas.

Ui={S1,S2,……,Sn} where U indicates users and S indicates session.

Sj={P1,P2,….,Pm} where P indicates pages accessed during each session.

Based on the weights and classes assigned to the pages of the website, we read them and assign the them to the pages read during the given user session. We read each page from the given user session and allocate the class and weight as predefined earlier. In case of reading a page from user session if the page is repeated , we reassign the weight α by adding the old weight w with the new level determined by β , the calculation is shown as below,

$$\alpha = Pi(w+\beta) \quad (1)$$

where α is the new value of the weight, which is derived by calculating β

$$\beta \propto L \quad (2)$$

where L is the level of the page and is defined by the sequence of page in the path of the user session.

Equation (1) and (2), are used in determining new weight α to a page, only if the page is repeated in the

sequence of user sessions, otherwise the original weight w, is assigned to the page.

The general process of calculating the user interest is depicted in below algorithm.

Algorithm 2. Build the pattern for pages of user interests

Input: Pages of a given user session

Output: Pages with calculated weights and class label .

Read the pages from the navigation pattern build for a given user. U1={S1,S2,…,Sn}

Read the page Pi of session Si assign class to page, assign weight to the page P1.

Read the next page in sequence page Pj, assign class Ci to page,

If the page is unique

Assign the weight w to page in sequence.

If the page is repeated

Change the weight by calculating α and assign the new calculated weight to the page.

$$c(i) = \sum_{n=1}^{m} (Cn) \quad (3)$$

Repeat the above steps until all pages of all the sessions are read for the given user.

After reading the page from user sessions and assigning the class labels as well as weights, now the pattern for pages is ready for calculating the user interest, which is derived as below,

In equation 3 we calculate the row wise weights of each page assigned to the class Ci, after calculating the row wise weight of each class we find the maximum weighted class Ci as shown in equation 4 by calculating the user interest UI

$$UI = MAX(WCi) \quad (4)$$

Algorithm 3. Calculate the interest of the user

Input: Pages of a given user session with the pattern build for class label and weight.

Output: User Interest class label based on the user navigation.

Read the pages from the build weigths and assigned classes of a given user.

Read the first page P1 note the label of class, read the page weight.

Read the next page in sequence page P2, note the label of the class

If the label is stored earlier

Add the weight of the page under the same label.

If the label is new

Add the label and store the weight of page under it.

Repeat the above steps for all the pages of a given user.

From the above algorithm we generate the following tables. In Table 3 we assign the weights to the pages based on the order of pages as per the page hierarchy of a given website. Based on the page name, the pages are assigned to the predefined classes as shown in Table 4. The presence of page for a given class is mentioned by 1, other cells are zero padded to mark the absence, to increase readability, only the present pages for a given class are stored.

After finding the class for each page the predefined weights are read for each page. In Table 5 the pages from a given user session and navigation discovered for a given user are shown and they are assigned weights based on the class hierarchy. After the pages are assigned weights the class wise summation is performed for each class label as shown in Table 6. The class label having the maximum weight is derived as the user interested class label.

According to the pages accessed from a given class the maximum weights is for the class University Information, so it is derived that the user accessing pages for a given session is interested in university information. In the table, no page is repeated so weight w is used to assign the weight , if the page is repeated in the access path, then new weight α is calculated and assigned to the page.

Table 3. Constructing Weights of the Pages Pi for a given Site

Page Class	P1	P2	P3	P4	P5
University Information	1	2	3	4	5
Institute Information	1	2	3	4	5
Faculty Information	1	2	3	4	5
Admissions	1	2	3	4	5
Downloads	1	2	3	4	5

Table 4. Assigning Classes to Pages Accessed from a user Session

Page Class	P1	P2	P3	P4
University Information	1	0	1	1
Institute Information	0	0	0	0
Faculty Information	0	0	0	0
Admissions	0	0	0	0
Downloads	0	1	0	0

Table 5. Assigning Weights to the Pages Accessed from a user Session

Page Class	P1	P2	P3	P4
University Information	1	0	3	4
Institute Information	0	0	0	0
Faculty Information	0	0	0	0
Admissions	0	0	0	0
Downloads	0	2	0	0

Table 6. Calculating Weights of Pages for the given Class

Page Class	P1	P2	P3	P4	Total weight
University Information	1	0	3	4	08
Institute Information	0	0	0	0	00
Faculty Information	0	0	0	0	00
Admissions	0	0	0	0	00
Downloads	0	2	0	0	02

3. Conclusion

Pattern discovery and analysis are one of the most important steps in web usage mining to find out meaningful information like, which pages are accessed the most, which browsers and operating systems are used the most, predicting user accesses and knowing the user interest. The user navigation patterns can be used to predict the future page accessed by a given user as well as we can derive the motive and interest of user in visiting the website. The paper describes a novel approach to calculate the page weight in determining the interest of users. This approach is well suited for the websites which consist of many pages that can be grouped among the similar classes. Single page cannot help in identifying the interest area of the user in visiting the website, but the number of pages belonging to a single class can provide better accuracy in determining the user interest in visiting the website.

4. Acknowledgement

The authors would like to thank Charotar University of Science and Technology (CHARUSAT) for providing the necessary resources for accomplishing the work.

5. References

1. Kosala R, Blockeel H. Web Mining Research: A Survey. Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter. 2000 June; 2(1):1-15. Crossref
2. Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: discovery and applications of usage patterns from Web data. Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining. 2000 January; 1(2):12-23. Crossref
3. Marquardt CG, Becker K, Ruiz DD. A pre-processing tool for Web Usage Mining in the Distance Education Domain. Proceedings of the International Database Engineering and Applications Symposium. 2004; p.78-87. Crossref
4. Honest N, Patel B, Patel A. Applying Web Usage Mining to a University Website Access Domain. International Journal of Applied Information Systems. 2012; 2(9):7-14. Crossref
5. Honest N, Patel B, Patel A. Preprocessing phase for University Website Access Domain. International Journal of Scientific Engineering and Research. 2013 June; 4(6):3071-75.
6. Honest N, Patel B, Patel A. Sessionization Process for the Pages Designed with the Concept of CMS. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 September; 3(9):362-68.
7. Wang X, Tan B, Shakery A, Zhai C. Beyond hyperlinks: organizing information footprints in search logs to support effective browsing. In Proceeding of the 18th Association for Computing Machinery Conference on Information and Knowledge Management. 2009; p. 1237-46. Crossref
8. Honest N, Patel B, Patel A. A Study of User Navigation Patterns for Web Usage Mining. International Journal of Advent Research in Computer and Electronics. 2015 January; 2(1):5-7.
9. Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback. In Proceedings of the 28th annual international Association for Computing Machinery's Special Interest Group on Information Retrieval conference on Research and development in information retrieval. 2005 August; p. 43-50. Crossref
10. Ban Z, Gu Z, Jin Y. An online PPM prediction model for web prefetching. New York. NY USA: Association for Computing Machinery. 2007 November; p. 89-96.
11. Hipp J, Guntzer U, Nakhaeizadeh G. Algorithms for association rule mining—A general survey and comparison. New York. NY USA: Association for Computing Machinery Press. 2000 June; 2(1):58-64.
12. Lin K, Wang C, Chen H. Predicting next search actions with search engine query logs. Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intel-ligence and Intelligent Agent Technology. 2011; 1:227-34. Crossref

13. Bauer T, Leake D. Word Sieve: A method for real-time context extraction. In Modeling and Using Context, International and Interdisciplinary Conference. 2001 July; 2116:30-44. . Crossref
14. Eirinaki M, Vazirgiannis M, Varlamis I. SEWeP: Using site semantics and a taxonomy to enhance the web personalization process. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003; p. 99-108. Crossref
15. Godoy D, Amandi A. Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance. New Generation Computing. 2008 May; 26(3):259-75. Crossref
16. Fu Y, Creado M, Ju C. Reorganizing web sites based on user access patterns. In Proceedings of the tenth international conference on Information and knowledge management. 2002 August; 11(1):583-85. Crossref