IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A Web Usage Mined and Web Structured Enriched Web-Page Recommendation Model

Anura Khede[1], J S Raikwal[2]

M.Tech Student, Department of Information Technology, IET DAVV, Indore, India [1]

Assistant Professor, Department of Information Technology, IET DAVV, Indore, India [2]

Abstract: Web-Page recommendation is of outstanding significance in today's dynamic world of internet. Intelligent web systems discover useful data using web mining techniques so as to do effective web-page recommendation. The proposed model is based on domain knowledge of web site and integrates web usage and structure mining for enriched connectivity-based recommendations. It is basically a semantic enhanced web page recommendation whose foundation is semantic network (Knowledge map) creation of a website. This network represents domain terms, Web-pages and the relations between them. Each web page has associated PageID which helps to calculate PageHits and the analysis of links between the pages helps to calculate PageRank.The entire topology of the website can be restructured after analysing user's behaviour through PageHit counts and PageRank and a well-structured website attracts more number of user and this in turn generates a better website ranking in search engine results. A well maintained site architecture can entertain multiple users simultaneously and promotes optimum bandwidth utilization.

Keywords: Recommendation systems, Web usage mining, Web structure mining, Semantic network.

I. INTRODUCTION

The rapid growth of World Wide Web resulted in tremendous web page generation and only a small portion of the web's pages contain truly relevant information [7]. Web page recommender system helps user to find pages of their interest and which provide suggestions to them based on user's navigation pattern on a website [1]. Web logs are very helpful to find out the browsing pattern and to find which pages are visited by the user the most. The recommendation provides link to mostly viewed pages of the website. Web usage mining uses historical data to capture and analyse user's browsing activities from web logs like time spent on pages, pages downloaded and pages with user feedback and deduces conclusions like which pages in a website have maximum hits and minimum hits and web structure mining uses graph theory to understand the link structure which connects web pages thus predicting trends in data.

The main objective of web recommender system is to effectively predict pages that will be visited from a given web-page of a website [2]. Good web page recommendation can improve website usage. Web page ranking is an optimisation technique used by the search engines for ranking hundreds and thousands of webpages in relative order of their relevance. To rank aweb page various types of criteria are usedby ranking algorithms. For example, some algorithms consider the link structure of the web page while others look for the content of the web pages to rank. Broadly, Page Ranking algorithms can be classified into two groups Content-based Page Ranking and Connectivity-based Page Ranking [11]. Connectivity-based ranking is major motivation behind the research work.

This paper is organized as follows. Section I is the Introduction. Section II highlights the related work. Section III provides a deep insight about the proposed work. Section IV provides experimental results. Section V concludes.

II. LITERATURE REVIEW

A. Recommendation Systems

Mining interesting knowledge from web logs is the foundation for recommendation systems. The purpose of the recommender systems is to predict meaningful suggestions to the user. These systems have changed the life of people in the way that it provides suggestions and help people find products, information, places etc. These are a Sub-class of Information Filtering Systems [10] that provides informative items (web pages, movies, songs, books, news, images, holiday destinations etc.) that might interest the user. The aim of Information filtering is to expose users to only information that is relevant to them. When the delivered information comes in the form of suggestions, an information filtering system is called a "Recommender System". These systems provide suggestions not only to registered users but also to unregistered users or random net surfers.

Recommendation systems (Content-based recommendations and Collaborative recommendations) were developed to gain insight to web user experience in order to model the interaction between users and items described on web pages and to recommend interesting items to the users. As a point of conception, the recommendation problem is reduced to the problem of estimating ratings for the items that have not been seen by

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

a user. Intuitively, this estimation is usually based on the ratings given by this user to other items and on some other information [1].

Market-Basket analysisis the backbone for recommendations based on E-commerce [15]. Consider an example where user A and user B gives similar rating to item I or have similar behaviour like watching movie, online shopping, etc. Then they may have same area of interest thus a system can suggest items to user A which are previously referred by user B or vice versa. Famous E-commerce sites like Flipkart, eBay recommend users about what they might like to buy, based on their past history of purchases or item searches. IMDb offers its customers a wide range of movie suggestions based on their choices, ratings. These systems are based on the fact that people who agreed on an item in past will agree on the same item in future also. Similar point in a recent publication by Nguyen, Yan Lu &Jie Lu[2] also mentions that web logs helps to understand the transition links between the web-pages. For the current visited web page (referred to as state) P and K-previously visited pages the web-pages that will be visited in next navigation can be predicted. The size of training data set decides prediction accuracy, bigger the data set the higher the prediction accuracy is. The success of the web page recommendation systems is entirely dependent on the web access sequence discovered from the web logs[15].

B. Web Mining

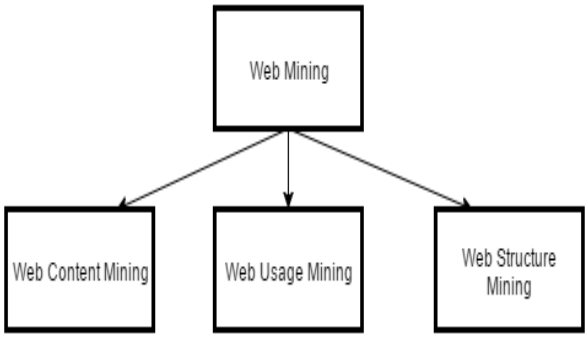

Fig. 1 Categories of Web Mining

Web Mining [6] is the application of data mining techniques to extract information from unstructured raw data. It is a knowledge discovery process. Web mining aims at finding and extracting relevant information from the perpetual changing World Wide Web. Web mining when performed combining usage and structure data, helps to improve web-site recommendation thereby providing an effective ranking in search engine results [4] also it helps a user to personalize their navigation.

Web Mining can be categorized into three types as shown in fig.1

1. Web Content Mining- It is the process of extracting relevant information from the web documents present inside the HTML or XML tags. It focuses on the content that is present in the form of text, images, audio, video, or any kind of structured records. The collocations and co-occurrences of terms in a user query is matched against a document's content for generating results. With the advancement in Search engine mechanism, Natural Language Processing based engines emphasize on the semantics of the keywords [7].

2. Web Usage Mining- It uses Web server logs, Application server logs to find out most accessed pages and analyses the user's browsing pattern from the Web access sequence.The predicted pages are often limited within the discovered web access sequence. Usage data also provides an efficient support for web designing.

Meaningful patterns are discovered from the web logs generated due to client-server transactions. Information is stored in the form of 'Sessions'. Sessionization [8] is important in order to store the activities of a client.

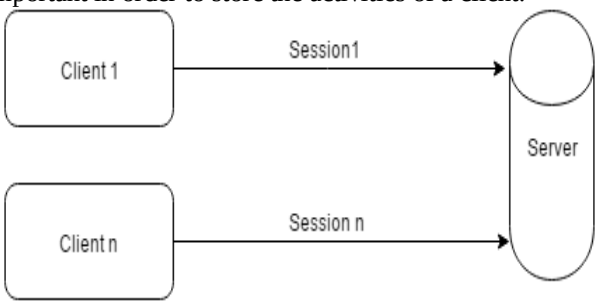

Fig. 2Sessionization

3. Web Structure Mining- "Web is a Graph" where Pages are nodes and Hyperlinks are edges [6]. Web Structure Mining is the discovery of the link structure of the web. Hyperlinks are the sources of pure navigation. It helps to understand which web pages are linked to which next set of web pages. Link Mining is an on-going area of research where link analysis is done to find out the importance of a web page. Famous *PageRank* algorithm proposed by Larry Page and Sergey Brin is based on the link structure of WWW.

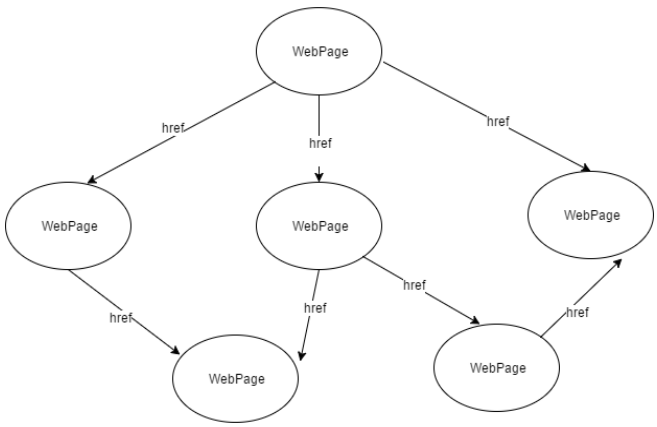

Fig. 3 Web as a Graph

C. Usage and Structure Mining as an Integrated Approach

The usage data and the structure of web pages, and in particular those of one website, reflect the intentions of authors and designers of the web pages and the underlying information architecture. The actual behaviour of the users of these resources reveals additional structure [3]. In the World Wide Web, there are two sources of information

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

- Web Structure*:* Reflects the author's viewpoint of browsing behavior.
- Web Usage*:* Reflects the user's browsing behavior.

Any conflicting evidence from these sources of information would be termed "interesting".[6]

It is useful to combine WUM with WSM in order to make sense of the observed frequent paths and pages on these paths. A publication by Miller and Remington [16] pointed out that the structure of linked pages has a decisive impact on the usability.Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites [14]. Since web is an ocean of information, web site attracts users of all age groups and the same topic carry different weightage to different people hence developers viewpoint and user's behavior both play an important role in way the website can be structured and the way it has been accessed. For example, sequence mining shows that many of the users who visited page A later went to page C, along paths that indicate a prolonged search. This can be interpreted to mean that visitors wish to reach C from A, did not find any direct path that connects C to A . This information structure can be modified to create a hyperlink from C to A which improves web site structure [3]

Web usage mining aims to discover useful browsing patterns from web usage data such as click streams [4], user transactions which are stored in web logs. The conception of data sets can be accomplished by selecting a random or for pre specific criteria like choose only those clickstreams that have accessed the website in last 3 hours or 7 hours according to the need. A web server log stores user activities in session of visiting different set of web pages. After successful processing of logs (data set) a Directed graph is generated which helps in analysis of user browsing patterns [15].

Knowledge Discovery[6] process results in a lot of conclusions which helps to take decisions on factors like the web pages with maximum hit counts will be most popular, the sequence in which web pages were accessed in Session S1= (P1-P3-P1) S2= (P2-P3-P4) etc, time spent on a specific web page decides it interestingness measure. If recommendations are based only on usage data then solutions to log entry that repeatedly states "redirect" will not be notified. In such cases the involvement of developer helps to reframe the structure of website.

Thus avoids unsolicited pages from getting recommended [9]. Also when a visitor from a session is lost the navigation data stored in web log can deceive future recommendations. Hybridization of algorithms based on structure mining and usage mining improves quality of web page recommendations. Analysis of links along with web page access sequence gives higher precision results.

III. PROPOSED WORK

The proposed system for web page recommendation is based on Semantic Domain Term Generation Model also known as TermNetWP [2]. It's a graph that explains the domain terms, web pages and relations including the collocations of domain terms and associations between domain terms and webpages.Domain terms are nothing but keywords present in TITLE tag. Most efficient way to deal with all types of data sources is to model them in the form of graph. Fig-4 depicts the framework of the proposed recommendation model. TITLE tag in HTML plays an important role in understanding the semantics of a web page as page titles are usually given higher weights by the search engine. Initially the proposed system will take input from user browsing history and a graph TermNetWP is constructed using this data. Based on this graph we can query following things [2]-

- Domain terms of a given web page.
- Web pages mapped to a given domain term.
- Occurrences of PageID decidesPageHit.
- Association between web pages sharing same set of domain terms.

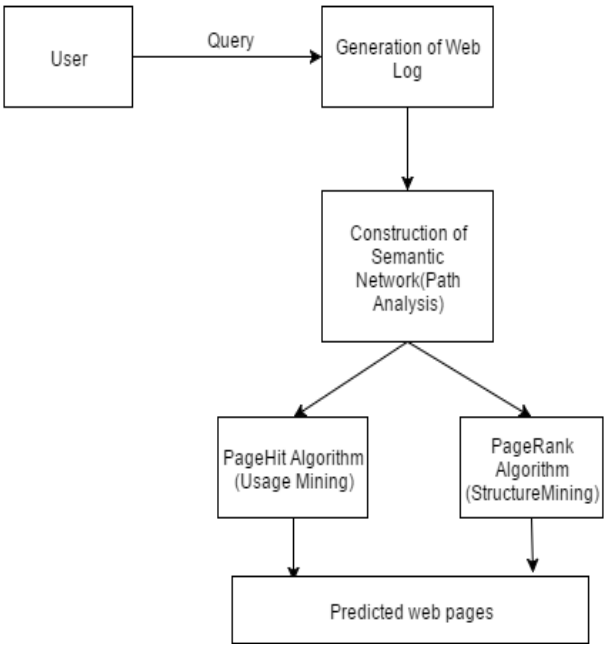

Fig. 4 Proposed Model

A. Formulated Strategy

The Directed graph TermNetWPformed helps in the analysis of the web pages associations with each of the other web pages and the associations of web pages with the domain terms. Every web page is given a PageID with which a page can be uniquely identified. PageHit can be calculated by occurances of pages by comparing the access between two web pages. PageRank associates every existing web page with just one number, its PageRank,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

which expresses the "importance", or rank, of that Web page in the whole Web, here TermNetWP. PageRank is calculated for a web page with maximum incoming links pointing to that web page. Rank of each web page depends on the total number of available webpages.

B. Algorithm to construct TermNetWP

Input: TSC (Term Sequence Collection)
Output: G (TermNetWP)

Process:

Let TSC = { PageID , X=t1,t2….t_m, URL }
Initialize G
Let R= root or start node of G
Let E= the end node of G

For eachpageID and each sequence at X in TSC{
Initilaize a WPage object identified as PageID

For each term $t_i \in$ X {

If node $t_i$is not found in G, then
Initialize an Instance object I as a node of G
Set I Name = t_i

Else
Set I = the Instance object named t_i in G

Increase I .iOccur by 1

If (i==0) then
-Initialize an OutLink R- t_i if not found
-Increase R- t_i .iWeight by 1
-Set R- t_i .fromInstance = R
-Set R- t_i .toInstance = I

If (i>0 & i<m) then
-Get preI = the Instance object with name t_i-1
-Initialize an OutLinkt_i-1 - $t_i$if not found
-Increase t_i-1 - t_i.iWeight by 1
-Set t_i-1 - t_i.toInstance =I
-Set t_i-1 - t_i.fromInstance= preI

If (i==m) then
-Initialize an OutLinkt_i-Eif not found
-Increase t_i-E.iWeight by 1
-Set t_i-E.toInstance = E
-Set t_i-E.fromInstance= I
-Set I.hasWPage=PageID

Add term $t_i$into PageID.Keywords
}

}
Algorithm for Enhanced TermNetWP

Input: TermNetWp, Pair Of user and List of pages visited
Output: TermNetWP with added hitCounts per page

For each user and pageIdpid {
- Get Page Object P
- increment P.numAccess by 1
}

Algorithm for PageHit

Input: TermNetWp, pageIdpid of page, user is on
Output: List of ranked pages based on hit counts

Ni be the set of neighbouring pages for page pid
Arrange *Ni* based on decreasing order of numAccess counts.

PageHit is based on the fact that each webpage has a PageID and by incrementing the occurance of PageID we can calculate Hits. Each web page"P" shares common set of domain terms and forms a *Web* of webpages linked to another set of webpages.

C. PageRank Algorithm

This algorithm computes the score for pages at the time of indexing of the pages [12]

PageRank algorithm as defined by Larry Page and Sergey Brin [13]-

We assume page A has pages T1...Tn which point to it. The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

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

Where:
PR(A) = PageRank of page A
T1….Tn= Set of all pages that are linked to page A
PR(Ti) = Page rank of page Ti
Q(Ti) = the number of pages to which Ti links to
d = Damping factor which is set between 0 and 1
"Damping factor decides for how much time old web pages should be included in iterations."
PR(Ti)/Q(Ti) = PageRank of Ti distributing to all pages that Ti links to.
(1-d) = To make up for some pages that do not have any out-links to avoid losing some page ranks. Generally kept 0.15

Step 1: Initialize the rank value of each page by 1/n. Where n is total no. of pages to be ranked. Suppose we represent these n pages by an Array of n elements. Then,

A[i] = 1/n where $0 \leq i < n$

Step 2: Take some value of damping factor such that 0<d<1.e.g 0.15, 0.85 etc.

DOI 10.17148/IJARCCE.2016.5591

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Step 3: Repeat for each node i such that 0≤ i< n. Let PR be an Array of n element which represent PageRank for each web page. PR[i] <- 1-d

For all pages Q such that Q Links to PR[i] do
PR[i] <- PR[i] + d * A[Q]/Qn*where Qn = no. of outgoing edges of Q*

Step 4: Update the values of A
A[i]= PR[i] for 0≤ i< n

Repeat from step 3 until the rank value converges i.e. values of two consecutive iterations match.

IV. EXPERIMENTAL RESULTS

PageHit and PageRank algorithm will take input from the graph TermNetWP and will generate web pages according to the relevancy (maximum hits and ranks) that a web page received by user access. On the basis of results achieved, web page sequences can be changed for better website usage [5]. Web pages in a web site can then be classified into three categories as

- Excellent- Highest hits and rank
- Average- Average hits and rank
- Poor- Lowest hits and rank

Fig.5 (a) & (b) shows the output of proposed webpage recommendation model. In TermNetWP a PageID 101 is connected with other pageID's sharing common set of domain terms and are linked together. After multiple iterations and calculations we can see that value convergence takes place and these web pages can be then arranged as excellent, average and poor. Hence it becomes important to not only focus on navigation pattern but also to emphasize on the connectivity information. The anatomy of the website can thus be restructured in such a way that the excellent web pages will be moved very near to the home page, at next level average ranked web pages be moved and so on…
Website designing and restructuring should be done in such a way that customers can easily navigate to find information or products. Achieved results conclude that-

- If a particular page is of utmost importance – implementing a hierarchical site structure with the excellent pages at the "top" or near to Home page.
- Where a group of pages may contain outgoing links – increase the number of internal links to retain as much PageRank as possible.
- Where a group of pages do not contain outgoing links – the number of internal links in the site has **no** effect on the site's average PageRank.
- Building a Site Map is very useful in retaining PageHits and PageRanks.

After incorporating these changes a website provides fast response to user. This hybrid approach of usage and structure mining helps in filtering unwanted browsing patterns and web pages and focuses on the relevant information to be introduced while designing website, since a poorly formed website will entertain users visiting spontaneous pages and rating and recommending unsolicited web pages.

```
Enter a valid webpage user is on:
101
Enter the page rank algo to use (1 for pageHit, 2 for Larry/Sergey pagerank: )
2
PageId: 108 PageRank: 0.9775  Hits: 2
PageId: 102 PageRank: 0.9233125  Hits: 1
PageId: 104 PageRank: 0.9233125  Hits: 1
PageId: 103 PageRank: 0.85  Hits: 1
PageId: 110 PageRank: 0.85  Hits: 1
```

Fig. 5(a) Results

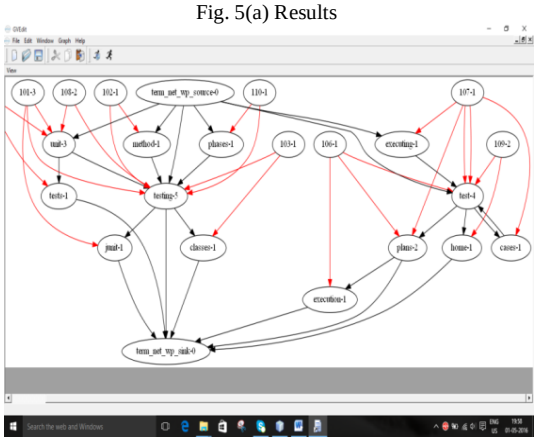

Fig.5 (b) showing Illustration of TermNetWP

V. CONCLUSION

The WWW has two different perspectives in which data is available, one is web structure view that reflects the author's viewpoint of browsing behaviour and second is web usage view that reflects the user's browsing behaviour. By using the effective web structuring methods and web usage mining we will have a website information architecture that will ensure altogether a completely different web navigation experience. The web service providers want to find the way to predict the users' behaviour and personalize information to reduce the traffic load and design the web site suited for the different group of users.

ACKNOWLEDGMENT

My heartfelt thanks to my Guide, Mr. Jagdish Singh Raikwal for the guidance he provided and for his constructive and positive feedback during the preparation of this paper.

REFERENCES

[1] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *Knowledge and Data Engineering, IEEE Transactions on* 17.6 (2005): 734-749.

[2] Nguyen, Yan Lu &JieLu,Web page recommendation based on web usage and domain knowledge,in Web Mining vol 26,2014. DOI: 10.1109/TKDE.2013.78

[3] Stumme, Gerd, Andreas Hotho, and Bettina Berendt. "Usage Mining for and on the Semantic Web." *National Science Foundation Workshop on Next Generation Data Mining*. Vol. 143. 2002.

[4] SaloniAggarwal, VeenuMangat, "Application Areas of Web Usage Mining", *ACCT*, 2015, 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT) 2015, pp. 208-211, doi:10.1109/ACCT.2015.115

[5] Mrs Geeta R.B., Prof. ShashikumarG.Totad, Dr. Prasad Reddy PVGD "Amalgamation of Web Usage Mining and Web Structure Mining" , International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[6] J.srivastava et al., "Web Usage Mining": Discovery and Applications of Usage patterns from Web data," SIGKDD explorations, vol .1, no. 2, 2000, pp12-23.

[7] Han, Jiawei, and Chen-Chuan Chang. "Data mining for web intelligence."*Computer* 35.11 (2002): 64-70.

[8] Mobasher, Bamshad. "Web usage mining." Web data mining: Exploring hyperlinks, contents and usage data 12 (2006).

[9] Li, Jia, and Osmar R. Zaïane. "Combining usage, content, and structure data to improve web site recommendation." *E-Commerce and Web Technologies*. Springer Berlin Heidelberg, 2004. 305-315.

[10] Hanani, U., Shapira, B., &Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, *11*(3), 203-259

[11] Gupta, A., Dixit, A., & Devi, P. (2015, March). A novel user preference and feedback based Page Ranking technique. In *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on* (pp. 1335-1340). IEEE.

[12] Prabha, S., K. Duraiswamy, and J. Indhumathi. "Comparative Analysis of Different Page Ranking Algorithms." *World Academy of Science, Engineering*and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering 8.8: 1486-1494.

[13] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." *Computer networks* 56.18 (2012): 3825-3833.

[14] Wang, Yan. "Web mining and knowledge discovery of usage patterns." *Cs 748T Project* (2000): 1-25.

[15] Khede, Ms Anura, and Mr JS Raikwal. "Applying Web Usage and Structural Mining for Web-Page Recommendations: A Survey." (2015).

[16] Miller, Craig S., and Roger W. Remington. "Modeling information navigation: Implications for information architecture." *Human-computer interaction* 19.3 (2004): 225-271.

BIOGRAPHIES

MsAnuraKhede, received B.E.(Hons) degree in Information Technology from RGPV university in 2011. She has published a survey paper entitled "Applying Web Usage and Structural Mining for Web-Page Recommendations: A Survey" in IRJET Volume 2, Issue 9 December 2015 She is currently pursuing her Master's Degree in Information Technology from IET DAVV, Indore

.

MrJagdish SinghRaikwal received his B.E. and M.Tech from UIT RGPV and SATI Vidisha in 2005 and 2008 respectively. His field of specialization is Data Mining. He is currently working as Assistant Professor in IET DAVV Indore.