

Applying Web Usage and Structural Mining for Web-Page Recommendations: A Survey

Ms Anura Khede¹, Mr J. S. Raikwal²

¹ Post Graduate Student , Department of Information Technology , Institute of Engineering & Technology DAVV Indore, Madhya Pradesh, India

² Assistant Professor, Department of Information Technology , Institute of Engineering & Technology DAVV Indore, Madhya Pradesh, India

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 and integrates web usage and structure mining for enriched connectivity-based recommendations. It is basically a semantic enhanced web page recommendation which is based on 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 hybridization of two algorithms leads to efficient results. The entire topology of the website can be restructured after analysing user's behaviour(hit counts) through web logs which provides fast response to user, saves memory dimensions of servers and hence reduces HTTP requests and provides optimum utilization of bandwidth.*

Key Words: Recommendation systems, web usage mining, web structure mining, Domain knowledge, Semantic network.

1. 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 find make conclusions and Web structure mining uses graph theory to links which connects web pages.

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.

This paper is organized as follows. Section 1 defines recommendation systems. Section 2 highlights the related work. Section 3 gives an overview about Web Mining and its categories. Section 4 presents the proposed technique of we page recommendation. Section 5 discusses benefits of hybridization of algorithms. Section 6 concludes.

1.1 Recommendation Systems

The main 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 that provide informative items (web pages, movies, songs, books, news, images, holiday destinations etc.) that might interest the user. These systems provide suggestions not only to registered users but also to unregistered users or random net surfers.

Market-Basket analysis is the backbone for recommendations based on E-commerce. Consider an example where user A and user B gives similar rating to item I or have similar behavior 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. *Goodreads.com*, *WhatShouldIReadNext.com*, are sites which provides suggestion to user based on their taste like fiction, non-fiction etc. IMDb offers its customers a wide range of movie suggestions based on their choices, ratings.

2. LITERATURE REVIEW

This section provides the study on the recent contribution placed in the domain of Web mining and Web page recommendation using Web Usage and Structure Mining. 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 [1]. The popularity of recommender systems has increased since mid 90's.

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 a recent publication [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.

Web mining is an invaluable help in transformation from human understandable content to machine understandable semantics. In the late 2000's semantic web changed the WWW. Semantic web mining [3] has emerged as an advanced and effective methodology over traditional method of recommendation that used tree based algorithms. 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. 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. Amalgamation of Web usage mining and Web Structure mining [5] pointed out that the structure of linked pages has decisive impact on usability. 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 behaviour both play an important role in way the website can be structured and the way it has been accessed.

3. WEB MINING

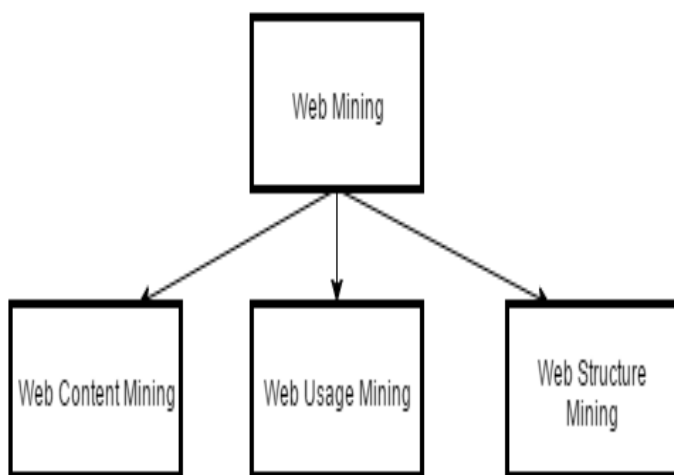


Fig -1: Taxonomy of Web Mining

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

3.1 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].

3.2 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.

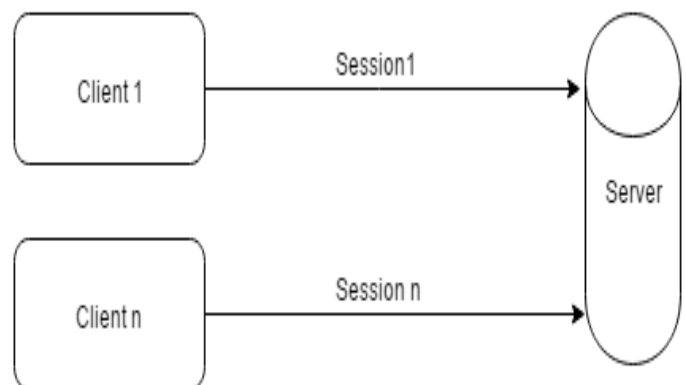


Fig- 2: Sessionization Process

3.3 Web Structure Mining- "Web is a Graph" where Pages are nodes and Hyperlink 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 ongoing 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.

4. 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 web pages. Domain terms are nothing but keywords present in title. Most efficient way to deal with all types of data sources is to model them in the form of graph. Fig-3 depicts the work structure of 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 decides PageHit
- Association between pages sharing same set of domain terms.

PageHit and PageRank algorithm will take input from this graph and will produce 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 and to gain high ranking in search engine results.

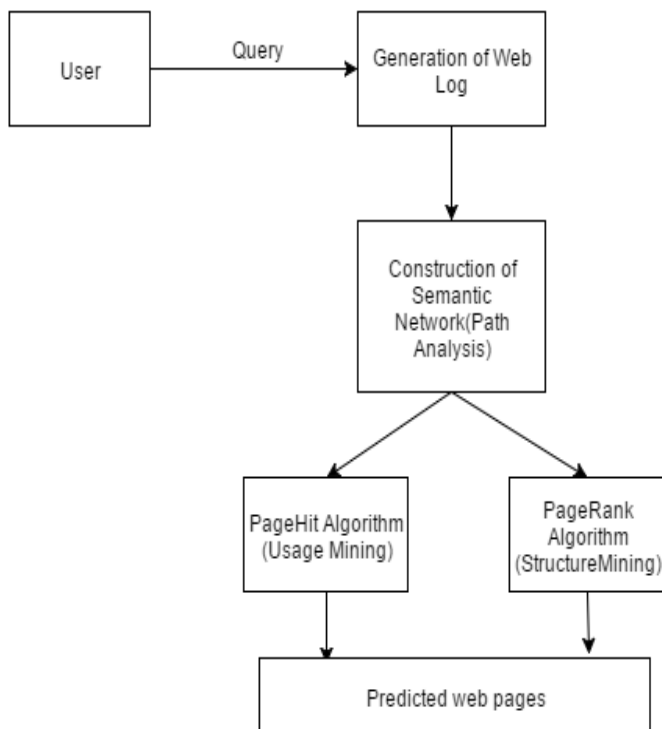


Fig-3: Webpage recommendation model

Table below illustrates the Page accessed in terms of PageID , Title tag and the associated URL.

Table -1: Processed Sample Web Log

Page	Title	URL
P1	International AutoRoute	/autoroute
P2	Corporate Advertising Content	/ads
P3	Internet News	/internet
P4	Internet Explorer	/ie
P5	For Developers Only	/devonly

5. BENEFITS OF HYBRIDIZATION

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 its 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.

6. CONCLUSION

Recommendations solely performed using web usage data addresses challenges like when a user is lost the clickstream data stored in the server by this user can mislead further recommendations, since a poorly formed website will entertain users visiting spontaneous pages and rating and recommending unsolicited web pages, 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 after analysing user’s browsing behaviour through web logs which provides fast response to user. This hybrid approach of usage and structure mining helps in filtering unwanted browsing patterns and web pages.

ACKNOWLEDGEMENT

I am immensely grateful to my Guide, Mr. Jagdish Singh Raikwal 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 & Jie Lu, 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] Saloni Aggarwal, Veenu Mangat, "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. Shashikumar G.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.

BIOGRAPHIES



Ms Anura Khede, received B.E.(Hons) degree in Information Technology from RGPV university in 2011. She is currently pursuing her Master's Degree in Information Technology from IET DAVV, Indore.



Mr Jagdish Singh Raikwal 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.