Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 02, February -2018

Survey of Web Usage Mining Techniques for Web-based Recommendations

Dave Mansi B. [1], Prof. Pratik B. Chauhan [2]

[1]*Computer Engineering, Atmiya Institute of Science and Technology*
[2] *Computer Engineering, Atmiya Institute of Science and Technology*

Abstract — *Data mining is simply mining of the huge data sets into useful information. The Web Mining is the set of data mining techniques which automatically extract and discover some useful knowledge from the web. Web usage mining is the process of extracting and discovering interesting usage patterns from Web data which are required for many web applications. Web based Recommendation is one of the applications of Web Usage Mining. Web based recommendation systems are developed to provide suggestions for set of items or relevant information for the users.*

Keywords*- Data Mining, Web Mining, Web Usage Mining, Recommendation, Web-page Recommendation*

I. INTRODUCTION

Data mining is simply mining of the huge data sets into useful information. Among very large amount of information available on the web, there are very few amounts of data actually useful or relevant to particular user. To find this relevant information from web there is a requirement of mining the web [1].

The Web Mining is the set of data mining techniques which automatically extract and discover some useful knowledge in the form of web documents, images, audios, videos, etc. different kind of data present on the web [2].

Web mining can be categorized into three types based on extracting knowledge:

a) Web Content Mining
b) Web Structure Mining
c) Web Usage Mining

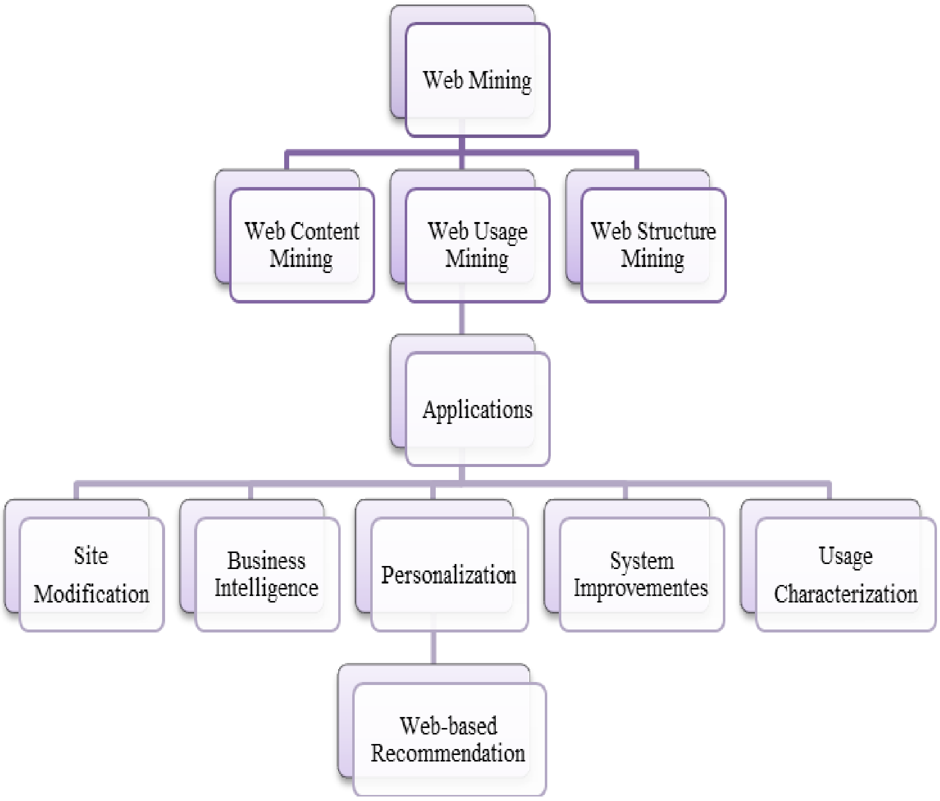

Fig.-1: Classification of Web Mining[6]

a) Web Content Mining:

Web content mining can be referred as the mining, extraction and integration of useful data, information and knowledge from heterogeneous web page content or web resources such as HTML - XML documents, digital libraries, database queries responses and these web resources are related to traditional information retrieval techniques [3].

b) Web Structure Mining:

Web structure mining or structured data mining is a process of discovering structural information from the topology of the link structure among web documents or hyperlinks. The structural mining can be performed on either document level or hyperlink level. Graph theory is used in this type of mining to analyse the node and connection structure of a web site [5].

Based on the type of web structural data, web structure mining can be divided into two types:

i. Extracting patterns from hyperlinks in the web:

A hyperlink is a structural element that connects the web page to a different location.

ii. Mining the text structure:

Tree-like structure of page structures is analyzed to describe HTML or XML tag usage.

C) Web Usage Mining:

Web Usage Mining is the application of data mining techniques. It is the process of extracting and discovering interesting usage patterns from Web data like web logs in order to understand and better serve the needs of web-based applications [5]. The usage patterns include server data i.e. IP address, Application server data i.e. web logic, and Application level data i.e. events [2].

Web usage mining can be categorized in three phases:

i. Data Pre-Processing
ii. Pattern Discovery
iii. Pattern Analysis

i. Data Pre-Processing

Data pre-processing aims to reformat the original logs to identify user's sessions. Data pre-processing is applied to the web server access log to split the log based on weblog format and to improve the quality of data, the efficiency and ease of the mining process [3]. There are four different tasks of data pre-processing namely, data cleaning, user identification, session identification, and path completion.

ii. Pattern Discovery

In this phase pre-processed information is further analyzed to extract valuable patterns. Patterns are mined using statistical methods and machine learning methods [2].Pattern discovery techniques include: clustering, sequential pattern, path analysis, classification, association rules [1].

iii. Pattern Analysis

Pattern analysis is used for filtering out the extracted pattern from discovery phase by eliminating monotonous patterns. This phase discovers knowledge of exciting trends out of the entire available trends.

❖ <u>Web Mining Applications</u>

There are many applications of web usage mining briefly explained as below:

- **Site Modification**

Web usage mining provides detailed feedback of user behavior. It provides the Web site designer information on which websites can be redesigned; web pages structures can be changed.

- **Business Intelligence**

Web usage mining is used here by defining Web log data hypercube that will associate Web usage data along with marketing data for various e-commerce applications.

- **System Improvements**

Web usage mining provides the key to understand web traffic behavior, which can be used further for developing policies for web caching, network transmission , load balancing and data distribution. Web usage are also useful for intrusion detection, fraud, attempted break-ins, etc.

- **Usage Characterization**

In this area of web mining application user's interaction with the browser interface as well as the navigational strategy used for browsing a particular site and client side events occurrence statistics are characterized.

- **Personalization**

Personalization is the application of web usage mining which is used to tailor the pages according to individual user preferences or characteristics. Web usage mining is used to model interest of users and personalize web applications based on that.

- **Web Based Recommendation**

Web based Recommendation is one of the applications of Web Usage Mining. Web based recommendation systems are developed to provide suggestions for set of items or relevant information that the users are most likely to interact with, but perhaps would not find in the huge number of available items.

II. TECHNIQUES FOR WEB PAGE RECOMMENDATION

There are many techniques used for web page recommendation. Some of them are mentioned and briefly explained as below:

A. *K means clustering & Boyer Moore Pattern Matching technique*
B. *Differential Semantic Technique*
C. *Web log mining techniques*
D. *User Rating and Synonyms Based modified Ranking Technique*
E. *domain knowledge and web usage knowledge based technique*
F. *Time and Semantic Relatedness based technique*

A. *K means clustering & Boyer Moore Pattern Matching*[7]

In this hybrid approach product information with user's access log data is integrated in the form of architecture and then a set of recommendations for that particular user are generated.

The architecture of an online web recommendation system based on web usage mining basically consists of three phases:

i. *Data Preprocessing* – It is an offline phase involves transforming the web access logs and user profiles into appropriate format of the system.

ii. *Pattern detection* – It is an offline phase involves using of clustering, sequential pattern mining or association rule mining like data mining techniques.

iii. *Generating recommendations* – It is an online phase which generates recommendations and provide customized links or data to the user based on detected patterns.

Existing architecture of recommender systems are simple as they use only one data mining algorithm, a sequential pattern mining algorithm which is applied to whole of the user logs to discover frequent navigation patterns of the user so that the next page request can be figure out and predicted by the system.

Disadvantage: The new user obtains the recommendations only on the basis of his current navigation.

Where the advanced system involves integrating additional information about the users like profile of users and uses two data mining algorithms like clustering and pattern matching algorithms. In this system, one of the clusters first classify the new users and then, the patterns of the corresponding clusters are used to customize the recommendations based on user's current navigation and other analogous users in the cluster.

The advanced architecture is divided in two phases; **offline phase** and **online phase**.

Offline Phase of the Architecture:

Data pre-processing:

In this phase the original web logs are pre-processed to discover the web access sessions. Clustering is also carried out in this stage. Here k-means clustering algorithm is applied.

- *K-Means Clustering Algorithm:*
  It is a partitioning method which aims to minimize the distance of the objects with respect to the centroid of each cluster. This algorithm moves objects between clusters until it cannot be diminished further.

Knowledge Base:
It follows the data pre-processing step, where various products' features are combined with the extracted user session data from the logs.

Online Phase of the Architecture:

Generating Recommendations

Here in this phase recommendations are generated by utilizing some refining parameters like value, brand, rating and so forth. For generating recommendations Boyer-Moore Pattern Matching Algorithm is used.

- *Boyer Moore Pattern Matching Algorithm:*

  An effective pattern matching and string searching technique which scans the characters of the pattern from right to left beginning with the rightmost one. In case of a mismatch occurs, Two precomputed functions are used to shift the window to the right. These two shift functions are called the good-suffix shift and the bad-character shift.

B. A Differential Semantic Algorithm for Query Relevant Web Page Recommendation[8]
A methodology in which the semantic heterogeneity is computed between the keywords, content words and query words for web page recommendation is incorporated.

The architecture of the proposed system can be explained as the input query from the user is pre-processed at first, which involves parsing and tokenization of the multi word query and elimination of redundant words in the query. After, that query keyword set is formulated which comprises of unique query words.

The proposed system architecture incorporates a very large repository called URL Base that consist a large volume of URLs collected from several web sources. The procedure for query pre-processing is repeated for the URLs and at the end of process parsed and normalized URLs are obtained. URL structure is elicited to extract the keywords which are extracted from the HTML title tag and content words which are obtained from the HTML body tags. The Semantic Heterogeneity between the keywords and the content words are computed using the proposed Adaptive Pointwise Mutual Information strategy, based on those values a feasible word set is formulated.

The computation of Semantic Similarity is carried out in order to obtain the URLs with a higher relevance to the query. Then, URLs are ranked based on the scores of the final semantic similarity. The finally ranked URLs which are the Web Page Links with high relevance to the input query are recommended to the user. The semantic similarity is computed twice using the Adaptive PMI strategy with heterogeneous thresholds, the proposed algorithm is termed as the Differential APMI algorithm.

When the adaptive coefficient coupled with the PMI value, it enhances the overall performance of the system. APMI increases the confidence in computing the semantic heterogeneity and hence it increases the overall relevance of the pages that are returned to the user.

C. Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques[9]
In this paper, a web recommendation approach is proposed which is based on learning from web logs and recommends user a list of pages which are relevant to the user by comparing with user's historic pattern. Finally, search result list is optimized using re-ranking the result pages. It is an efficient system which shows the pages desired by the user, on the top in the result list and thus reducing the search time.

The proposed architecture can be divided into two main phases named as Back end and Front end. In the back end phase, there are main two modules: Data pre-processing and sequential pattern mining.

i. Data pre-processing:

Data pre-processing is performed on web log to capture user navigation session. The pre-processing of web logs is complex and time consuming. It is done using following steps:

a) *Data Cleaning :* It is used to eliminate irrelevant items from the web logs.
b) *User Identification :* It is used to identify website is accessed by whom and which pages are accessed.
c) *Session Identification and reconstruction:* It is used to identify the different user session from very poor information available in log files and then to reconstruct the user navigation path within the identified session.
d) *Path Completion :* It is used for acquiring the complete user access path.

ii. Sequential pattern mining:

The next step in backend architecture is to determine the sequential patterns in each cluster. In this step only frequent patterns of events from the current access sequence are considered for generating the next candidate sequence. For all patterns in the candidate set, all URL sequences are processed once and the count is then incremented for each detected pattern. At each iteration, the module eliminates those candidate sequences whose support is less than support threshold.

In the front end phase, URL request of the user is processed by search engine then the recommended list of web pages relevant to user query is captured and after that rank updating algorithm is applied on them. As a result the popular and relevant pages are shown at the top position in the recommendation list.

D. User Rating and Synonyms Based modified Ranking Technique for Recommender Systems[10]
This paper proposes a recommender system for user rating and synonyms based ranking of the websites. In recommender system, when a keyword is searched by the user, result will be shown in the form of that keyword and also the related synonyms. Based on this search, the websites are displayed and ranked. Based on user rating the page rank of website is upgraded or degraded in the database. Here, the map reduce algorithm has been used to differentiate and reduce the data retrieved.

The working of proposed algorithm can be explained as following:

First, the two datasets are created which contains the page rank, website name and 5 keywords that relate to the website. The ranking in the given datasets is inserted on the basis of the algorithm designed by "google.com". Second, the input keyword is inserted by the user. The keyword is then matched with the data in the dataset.

Then map-reduce algorithm is applied on the data obtained. MapReduce groups everything by a key, in this case the word. So the reducer gets the word and sums up the counts of the word. The sites that match the keywords and the synonyms from both the dataset are obtained and then the data is filtered based on the keywords and synonyms. Then the top 5 websites are chosen and displayed. After this the user is asked to give the ranking of the website, then it updates page rank in the dataset according to the weightage given by user. At the end, the comparison between the base rank and the updated rank by the user is displayed and the message of updating the page rank is also displayed to the user.

E. A novel approach to provide Web page recommendation using domain knowledge and web usage knowledge[11]
To obtain the relevant data from the large World Wide Web, proposed system gives a novel approach to provide a Web page recommendation which consists of three knowledge based models. To improve the performance key information extraction algorithm is proposed. Then the results obtained from applying three knowledge based models are compared with models along with key information extraction algorithm. At the end, a page is recommended from weblog records.

The working of this approach can be described as below:

The input datasets are weblog records which contain irrelevant data. To convert irrelevant data into relevant data the pre-processing technique means stemming is used. After the pre-processing of data three models are constructed.

a) Ontological model:

In the construction of this model, titles of web pages are extracted then these titles can be split into domain terms by using term extraction algorithm.

b) Semantic network analysis model:-

In this model firstly domain terms obtained from ontological model are relisted in descending order of occurrence from each webpage and then each domain term belongs to which web page is found out.

c) The conceptual prediction model (CPM):-

This model uses Conceptual Prediction for recommending web pages correctly. This model gives visitor count of each web page, user's previously and next visited web page. So it gives web usage knowledge by visitors count and in link and out link of each user. Finally , now web page is recommended using these models.

Now, In order to improve web page recommendation performance the Key information extraction algorithm is used and Recommendation, execution time and accuracy calculated for proposed system and compared with the existing system.

F. Effective Web Personalization System Based on Time and Semantic Relatedness[12]

A novel web personalization system is proposed in this paper, that accepts the timing information, semantic information along with the navigational pattern, and then classifies the users according to their interest and behaviour on the site. The Web personalization model is constructed using the real and synthetic data and then used to validate the proposed model.

The model generates the user profiles based on browsing characteristics of the user, area of interest and recommendable areas with relevant information. This is a hybrid model which is constructed by combing the ideas from web usage mining and web content mining.

Working of the system:

Firstly , two datasets are generated the first is from the web page content and the second one is from the server log files. The text content of the first dataset is processed by using keyword extraction algorithm named as RAKE algorithm - Rapid Automatic Keyword Extraction. The output of the RAKE algorithm is set of keywords from webpage content.

Second, the log dataset contains log information of user. These input data is passed to the data preprocessing stage which contains four steps: cleaning, user identification, session identification and attributes identification. After generating sessions page navigation pattern of each user for a particular session is obtained, which is a significant data for the web personalization.

Next , the clustering is applied on time related attributes from the sever log files to identify browsing characteristics of user. After generating keywords the area of interest of user is found using TFIDF algorithm - Term Frequency-Inverse Document Frequency which is applied on user's session and group keyword list based on users session. Finally at the end of algorithm the web pages are selected based on TFIDF score which is higher than a threshold value as the area of interest of user.

After identifying the area of interest, the recommendable pages are found for a user by using semantic relatedness measure. The semantic relatedness between two pages can be found by selecting two pages one from interest list and other from remaining list. Then highest scored pages from remaining list are selected as the recommendable pages to the user.

After completing these necessary steps user's profile is generated which contains user id, number of time the site is visited, browsing characteristics, area of interest and recommendable areas.

III. COMPARISON OF TECHNIQUES

TABLE 1

Technique Proposed by	Algorithm/ Technique used	Datasets	Improvements
Hiral Y. Modi Meera Narvekar	K-Means Clustering & Boyer Moore Pattern Matching	User Logs	Reducing time consumption, Enhanced precision and recall
Gerard Deepak, J Sheeba Priyadarshini, M S Hareesh Babu	Differential adaptive pointwise mutual information algorithm	URL Base	Improved Accuracy, precision and recall
Anamika Rajput, Sushil Kumar Chaturvedi	Map-Reduce Algorithm	Website URLs	Faster speed
Priyaka Kolekar, Suchita Wakhade	Key Information Extraction Algorithm	Weblogs	Less execution time and more accuracy
G P Sajeev∗, Ramya P T	RAKE, TFIDF algorithm	Web page content ,server log files	Good Accuracy
Ravi Bhushan Rajender Nath	Matching query algorithm, Rank updating algorithm	Web logs	Improved relevancy with reduced time

IV. CONCLUSION

This paper studies various web usage mining techniques used for generating the recommendations. This paper shows the comparison between various techniques with the improvements obtained using these techniques. By using various algorithm the recommendations can be generated with better efficiency and accuracy which is very necessary to provide proper recommendations to the users.

REFERENCES

[1] Kaur, Kamaljit. "Web usage mining-current trends and future challenges." *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*. IEEE, 2016.

[2] Chavda, Sahaj, et al. "Recent Trends and Novel Approaches in Web Usage Mining." (2017).

[3] Sukumar, P., L. Robert, and S. Yuvaraj. "Review on modern Data Preprocessing techniques in Web usage mining (WUM)." *Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on*. IEEE, 2016.

[4] Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data." *Acm Sigkdd Explorations Newsletter* 1.2 (2000): 12-23.

[5] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[6] Wikipedia contributors. "Web mining." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 23 Oct. 2017. Web. 24 Nov. 2017.

[7] Modi, Hiral Y., and Meera Narvekar. "Enhancement of online web recommendation system using a hybrid clustering and pattern matching approach." Nascent Technologies in the Engineering Field (ICNTE), 2015 International Conference on. IEEE, 2015.

[8] Kolekar, Priyaka, and Suchita Wakhade. "A novel approach to provide Web page recommendation using domain knowledge and web usage knowledge." Communication and Electronics Systems (ICCES), International Conference on. IEEE, 2016.

[9] Bhushan, Ravi, and Rajender Nath. "Recommendation of optimized web pages to users using Web Log mining techniques." *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. IEEE, 2013.

[10] Sajeev, G. P., and P. T. Ramya. "Effective web personalization system based on time and semantic relatedness." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016.

[11] Deepak, Gerard, J. Sheeba Priyadarshini, and MS Hareesh Babu. "A differential semantic algorithm for query relevant web page recommendation." Advances in Computer Applications (ICACA), IEEE International Conference on. IEEE, 2016.

[12] Rajput, Anamika, and Sushil Kumar Chaturvedi. "User Rating and Synonyms Based Modified Ranking Technique for Recommender Systems." Computational Intelligence and Communication Networks (CICN), 2015 International Conference on. IEEE, 2015