ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Customer Behaviour Prediction Using Web Usage Mining

[1]Himasree Karthakula,[2]Nandam Gayatri

[1]Department of Computer Science and Engineering,KITSW(Affiliated to Kakatiya University) Warangal, India – 506015,email:himasreekarthakula@gmail.com
[2]Department of Computer Science and Engineering,KITSW(Affiliated to Kakatiya University) Warangal,India506015,email:ng.cse@kitsw.ac.in

Abstract: Web usage mining is the use of data mining technologies to detect and service the needs of web-based applications by identifying and mining interesting usage patterns from web data. It entails first capturing client behaviour and flow on a website, then mining this data for behavioural patterns. It is a critical component of the ecommerce business that allows websites to review previously collected web traffic statistics. By finding the patterns in this data, e-commerce companies may improve their performance and recommend better items and services to customers. The system is set up to track various analytics data and record web shopping/buying behaviours in order to generate future prediction statistics. The system looks for user budget tracking, comparing it to prior years, user bounce rates (the amount of people who leave the payment page and come back), and other site usage indicators. In this paper, We present a Heuristic-based Distributed Miner (HDM) design to obtain consumer common behaviour patterns in real time, thereby solving the Web Usage Mining (WUM) problem. Also provided is an improved k-means clustering (IKM) algorithm for clustering web sites based on similarity function. Two assessment measures are used to evaluate the proposed IKM method to the standard k-means algorithm for cluster web sites: Sum of squared error (J) and Execution time (in mille seconds). The experimental results shows that the proposed algorithms shows high performance compared with previous algorithm.

Keywords: Web usage mining, customer behaviours, pattern mining, heuristic based Distributed Miner, Improved k-means.

I. INTRODUCTION

The web mining system [1] emphasizes the scale in which data mining techniques are contracted to obtain extensive data from the web. Web mining specializes in methodologies for extracting valuable knowledge from internet records. Data mining, technological knowledge to analyse records from unusual perspectives, and creating new and exciting patterns to discover significant correlations and trends from big data are followed in network statistics to realize patterns of expertise tools. This thesis presents research panels made to discover the use of data mining techniques and blueprint scanning techniques to discover exciting web patterns, in publicly available reference statistics, in the form of a webpage browsing sequence.

Web use mining is a new field of study gaining more and more interest in today's times. Learn about web mining techniques to gain perspective on Internet user behaviour [2]. It can be related to forecasting customer behaviour within a website on the internet, and predicting the next page can be a great diversion. For the web consumer, report searches for web clients and offer a more exciting and personalized webpage during their online journey etc. [3]. Therefore, prospecting for web usage

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

discovers its applicability in web cache design, network allocation, website restructuring and development, website traffic control, and business intelligence.

The automatic detection of a person's right to access web server patterns is known as web usage mining [4]. In their day-to-day operations, organisations generate large amounts of logs, which are frequently made utilising web servers gathered from the internet to access log files. It contains a wealth of information about how online users accessed the behaviour. Analysing those who gain access to stats can provide valuable data for general server performance updates, website restructuring, and direct marketing in e-commerce. Several algorithms have been integrated into the block chain exploration area over the past decade. Most have also been modified to help concise representations such as closed, maximal, additive, or hierarchical sequences. All the current research on serial mining patterns has dealt most simply with a list of one or dimensions with unprocessed information previously. But this does not provide many records. It will take the weblog from the server upon login and extract a frequently occurring sequential pattern using a pre-processed multi-dimensional internet log file. The reason for choosing the weblog document in this thesis is its convenience for producing long sequences.

The epidemic rise of Internet users is attracting many researchers' attention to delve into the field of the network in a position of exploring interesting and non-trivial methods. Discovering attractive and functional patterns can be a valuable resource for providing a better network environment for users, thus attracting more new users and keeping regular customers without losing their enthusiasm. Mining for interesting data related to various factors on the web is called net mining. Web mining finds its first application on electronic operator-based websites. Some e-services include e-commerce, finance, advertising, banking, learning, governance, etc. Web mining is classified into three categories: web structure mining, web usage mining, and web content mining.

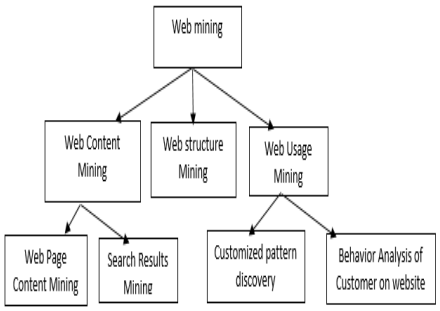

Fig.1 Web mining

Part of the data mining component is the problem of applying data mining techniques to web usage mining. It reveals realistic web usage patterns to understand and satisfy the desires of users browsing the web. Like every information mining challenge, web mining technology also consists of three main steps: (1) pre-processing, (2) pattern discovery (3) pattern analysis. The pre-processing step includes three separate levels (data cleansing, user identification, and session defining). In this work, pattern discovery applies common sequential sampling detection strategies introduced to the record's information. For this purpose, statistics must be transformed within the pre-

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

processing section to be used as input for the algorithms. The method for evaluating a pattern defines the results obtained with algorithms and concludes.

One of the number one mining packages for web use includes optimizing your internet cache. The web cache is a temporary space nearby to store linked files on the network, including HTML web pages and images, to reduce the delay caused by retrieving those pages from the server. We recently accessed pages are diagnosed with redundant opportunities; these pages can be cached to reduce users' waiting time. Other than that, if interesting categories of webpages are finally accessed from a previously recognized internet website, that may store the interesting pages etc. Therefore, identifying potential next pages or categories of interesting pages that were eventually reached allows for optimizing the web caching mechanism.

II. REVIEW OF LITERATURE

Web mining is significant because the size of content files is spread across the network, and a suitable pattern is selected. Manually correcting data is very difficult and time-consuming. This article's motivation is to explain internet mining and its three cool areas, equipment, and tactics. The primary focus of web mining is to find valuable records and extract them from the network. Information analysis of information on the Internet is sufficiently comprehensive. All kinds of data can be linked or unrelated to them on the Internet and are constantly negotiated. Since these exchanges are linked without delay to a specific industry type, it is necessary to adapt to these exchanges.

Rajesh et al [2020] The web is one of the most appropriate technologies for data collection, and its percentage and network utilisation are growing by the day. Computers are a data repository with a significant amount of information; Internet mining is an information mining procedure used to find facts and experience on the internet. Web mining collects logs from a server, client, proxy, or database. There are three types of web mining technologies: web content mining, internet architecture mining, and web mining. This work introduces the well-known concept of this type of net mining. Our assessment document primarily defines the mining of Internet usage and the technologies used in it. It also evaluates the mining application using the web.

Jeyalatha et al. [2019] Web usage mining includes the survey of facts about the weblog record. It can detect browsing patterns and be used to model the browsing behaviour of network clients. Log history analysis illustrates user options. The mining method cannot use statistics from the log file as miles. Recording record content must be pre-processed and anchored with a specific design. This article specializes in designing and implementing an algorithm for examining web usage history, specifically educational research programs. It accepts weblog records and consumer inquiries as inputs and outputs containing network log details and exact data. It is considered to extract the metadata to read the weblog report. The interest measures t-weight and d-weight are measured. User browsing, network architecture, and consumption habits would all benefit from the template described in this paper.

Kumar et al. [2017] there is a lot of data kept on the web. When a consumer searches for accurate data using search engines like Google, Bing and many more, it isn't easy because the web pages' complexity increases every day. Web mining plays a vital role

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

in addressing this problem. In prospecting for web use, we cultivate a suitable sample according to a person's roaming behaviour. The purpose of this document is to implement an Internet logging tool that specializes in Internet Server Log Report (Academic Institution Web Logs) to determine the behaviour pattern as well as profiles of users who connect with a website. The pattern of net-record mining use of an art institution. The logs related to the web are grouped into three parts, namely internet log, log access, error log and proxy log data, logs collection on internet server and application by web logging professional. Our experimental results help determine the number of tourists to the site and improve the site's usability. Web-related log information is in three types: proxy logs, weblog information, and error logs. We explore hobby statistics with a daily report based on weekly and monthly internet usage patterns. Internet usage exploration plays a vital website in improving your website data provision.

Dhanalakshmi et al. [2016] the increased online applications are causing a massive boom in web content. Most of the business enterprises participate in learning about the behaviour of network users to decorate their business. In this context, customer mobility in static and dynamic network software plays a vital role in cognitive research. Static mining strategies will not be suitable as they are far from dynamic Internet registry files and selection. Traditional weblog pre-processing approaches and blogging usage patterns face barriers to analysing the relationship between content and browsing records. This document specializes in various static internet log processing, mining strategies and related barriers to dynamic web mining.

Geng et al. [2015] Introduced new technology to find the navigation Based on a comparison of actual and predictable usage patterns, online usability concerns can be identified. By using log processing to recognise users, consumer sessions, and consumer-oriented transactions, and then applying a set of usage mining criteria to uncover patterns between actual usage paths, actual usage patterns for operational websites can be recovered from mechanically generated web server logs. Our best-in-class interactive consumer training courses, which are created by professionals totally based on their knowledge of consumer behaviour, capture planned use, which contains information about the course as well as the amount of time required for different tasks. To check data and detect user mobility issues, Oracle Look at the technique is used to calculate variance. This assessment's deviation records can aid us in identifying usability issues and recommending corrective steps to improve usability. A software tool has been created to automate a substantial portion of the tasks. We diagnose usability concerns, which industry experts test, and discover usability improvement through better job completion through a test on a small operator-oriented website fees and less effort. And the time of the specific obligations after the proposed corrections are made. This case note provides a preliminary verification of the applicability and effectiveness of our approach.

Yadav et al. [2012] with the exciting increase in records resources available on WWW, it has become a crucial tool for clients to identify, extract, clean, and evaluate preferred records and assets. This document's main reason is to monitor buyer behavior and use web mining techniques and software in e-commerce to undermine consumer behavior. This web mining idea describes the web statistics mining system in detail: gathering supply statistics, pre-processing information, pattern detection, sample evaluation, and batch evaluation using advanced data technology; servers are

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

capable of collecting and storing vast amounts of data, describing their various contributions and profiles. Unique Buyers, through which they seek records about consumer needs. Traditional methods are not suitable for these working conditions for finding customer behaviour. The principle of information mining is to aggregate consumer segments using a set of K-Means. The entry records come from the web history of various e-commerce sites. Therefore, define the relationship between web history mining and e-commerce and also implement the era of web mining in e-commerce.

Shekhar et al. [2011] this paper's main reason is to look at how web mining strategies, features and software (e-commerce and e-business) and their functional areas. Web mining is becoming more popular and widely used in various application areas (including commercial smart devices, e-commerce, and e-commerce). The impacts of e-commerce or e-business are enhanced by the benefit of mining strategies such as data mining and textual content extraction. Of all the mining strategies, web mining is fun.

Zhiwu et al. [2010] this paper mainly introduces the concept and type of web mining, discusses the method and methods for prospecting on an e-commerce network, and finally discusses web mining software in e-commerce. Some problems have been mentioned that need further study to take advantage of the advantages of the Internet and web mining technology, which have distinct advantages and play an increasingly vital role within the benefit of e-commerce.

Chaoyang et al. [2009] to provide the system's service over the Internet, a device-specific framework that relies entirely on web mining is proposed. Individual suggestions according to the user's interests and personalization can improve the first-class service.

Magdalini et al. [2008] Web personalization is a way to customize a website to specific customers' desires, leveraging knowledge gained from analysing a person's browsing behaviour (usage logs) about other accumulated facts in the web context. Specifically, the structure and materials of the content and the person's profile information. Due to the web's explosive growth, web customization received a tremendous boost in both studies and industry. In this newsletter, we provide a survey on using web mining to customize the web. We offer modules which incorporate the web customization system, which emphasizes the module of web mining. An evaluation of the frequent strategies which can be utilised is provided and technical issues which arise, including description of the famous equipment and packages.

Wang et al. [2006] Web Usage Mining (WUM) integrates strategies for two popular study fields: data mining and the Internet. By reading lists of hidden capacity in network logs, WUM helps personalize Internet content material delivery, improve network design, user satisfaction, and people browsing through search. Preview and buffering. This document introduces the classic data extraction algorithms: FPgrowth and PrefixSpan in WUM and are implemented in the real business case. The maximum forward-path (MFP) is also used to issue net mining use at some point in serial sample mining and PrefixSpan to reduce the interference of "false transition to" due to the browser cache and increase the common mining pointer. Across paths. Detailed analysis and vigilance on the corresponding effects are mentioned.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

III. PROPOSED METHODOLOGY

The main objective of the work is to perform monitoring on the websites of the various unique servers of the portal to determine a sample of expressions of user behaviour and characteristics. Predict the customer behaviours on web portals using web usage mining algorithms. To avoid aging of the end result, in this approach, We recommend the Heuristic-based Distributed Miner (HDM), which is based on heuristics for obtaining recurring patterns of behaviour from users and responding to the problem of real-time web usage mining.

We suggest that our method solves a number of issues that traditional Web Usage Mining approaches have:

- If a common behaviour pattern is concealed amid the enormous number of facts saved in the log file by means of the obtain entry, and if this pattern is the most common and simple for a range of specified time rather than complete file, then the common sample will not be displayed using a traditional methodology.
- Multiple unique regulations can be read at the same time where they fit with user behaviour. Remember that this sort of organisation has specified the method of movement of users near the relevant consumer, and that this organisation must then replace the other. Related users, in our opinion, can be deemed near to one another.
- For utmost web usage mining methods that rely solely on the log file for input analysis, The data mining process will be simpler if there are less clients who log in to that file. HDM is built to give real-time information to a wide range of clients.
- Finally, our approach is based on the fact that when a user connects to a website, the computational power available on the user's computer is not employed until the proposed architecture is employed in this work.

SYSTEM ARCHITECTURE

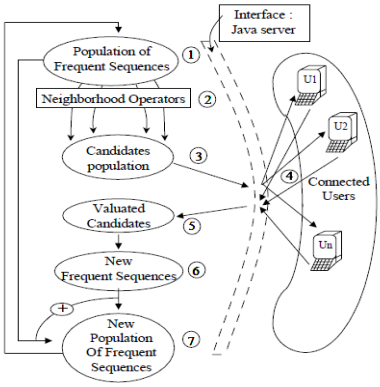

Fig.2 HDM approach

There are three main methods are included in web usage mining such as association rules, sequential patterns, and clustering. In this proposed work, after extracting the websites using HDM method clustering techniques implementing to cluster the web users based on their similarity interest on the web.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

Clustering strategies diagnose collection of comparable items among excessive amounts of information. This is done based on distance functions that estimate the degree of similarity between similarity items (websites). In web usage mining, clustering is used to group together similar meetings. The main thing in this type of research is the comparison between the consumer organization and the people organization. In this clustering, we can find two different types of interesting clustering 1) user clustering, 2) Web page clustering.

Using the similarity score and the amount of time consumed viewing a page to evaluation the similarity of assemblies. In the clustering algorithms improved k-means clustering is widely used data division method. The k-means clustering algorithm aims to partition m data points into k clusters (k m) given a set of m data points X=x I |i=1,2,…m where each data point is an n-dimensional vector. C = c1,c2,…,ck so as to minimise the within-cluster sum of squares, which is an objective function (or a cost function) of dissimilarity J(V,X). In most circumstances, the Euclidean distance is used as the dissimilarity metric. The objective function is a measure of how far each of the n data points is from its cluster centre. The below algorithms shows the basic k-means algorithms steps

K-Means clustering algorithm

Algorithm: *k*-Means Clustering
Input: No. of clusters k and Set of *m* data points $X=\{x_1, \ldots, x_m\}$
Output: Set of *k* centroids, $V=\{v_1, \ldots, v_k\}$, corresponding to the clusters $C=\{c_1, \ldots, c_k\}$, and membership matrix $U=[u_{ij}]$.
Steps:
1) Initialize the *k* centroids $V=\{v_1, \ldots, v_k\}$, by randomly selecting *k* data points from *X*.
2) **repeat**
   i) Determine the membership matrix *U* using (4), by assigning each data point x_i to the closest cluster c_j.
   ii) Compute the objective function *J(X,V)* using (2). Stop if it below a certain threshold ε.
   iii) Recompute the centroid of each cluster using (5).
3) **until** Centroids do not change

The proposed improved k-means clustering algorithm as given below

Algorithm: **Improved k − means clustering**

Input: The number of clusters is k, while the number of data points is m. $= \{x_1, \ldots . , x_m\}$

Ouput: Set of K − centroids, $V = \{v_1, \ldots . , v_k\}$, corresponding to clusters $C = \{c_1, . , c_k\}$

1. **for** j = 1 to k
2. **begin**
3. **if** j = 1 **then**
4. $v_1(1)$ = centroid of dataset X
5. $\mathbf{V_1} = \{v_1(1)\}$
6. **else**
7. **for** i = 1 to m

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

8. begin

9. Execute $k-$ means using $\{v_1(j-1), \ldots\ldots, v_{j-1}(j-1), x_i\}$ as
10. Initial centers
11. $v_j^i = j^{th}$ cluster center obtained using(5)
12. $j_i =$ objective function value using(2)

13. end

14. end if

15. $v_j(j) = v_j^k$, where, $1 \leq k \leq m$ and $j_k = \min_{i-1}^{m} j_i$
16. $V_j = \{v_i(j-1), \ldots, v_{j-1}(j-1), v_j(j)\}$

17. end

The improved k-means clustering technique provides a deterministic global optimization that is independent of cluster centre beginning placements. As a local search process, it employs the k- Means algorithm. Instead of selecting initial values for all cluster centres at random, the algorithm works in steps, adding one new cluster centre at a time.

IV. RESULTS AND DISCUSSIONS

In this part, we are taking experimental results on both standard k-means algorithms and proposed improved k-means algorithm in the two scenarios of execution time for cluster formation and sum of squared error (J) for number of clusters.

Table.1 Clustering results

Evaluation Measure	Clusters	K-means	Improved K-means
Sum of squared Error (J)	10	583.54	504.80
	20	443.06	358.95
	30	357.24	275.72
	40	284.09	208.01
	50	279.29	162.52
	60	260.64	127.17
Execution time (Milli seconds)	10	49	38369
	20	110	113623
	30	142	213973
	40	164	335883
	50	278	478573
	60	188	644976

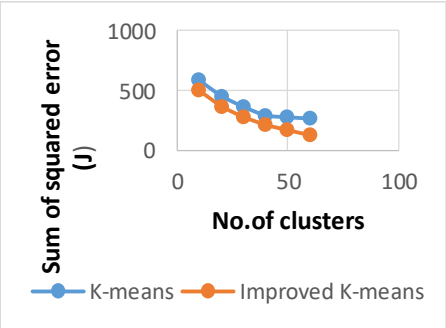

Fig.3 Clustering Error (J) Vs Number of clusters

As shown in figure 3, comparative taken between traditional k-means and improved k-means algorithms here, the clustering error calculated for cluster groups. The number of clusters is shown on the X-axis, while the sum of squared errors is shown

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

on the Y-axis (J). Graph indicates that the proposed algorithm have less cluster error when compared with previous algorithm.

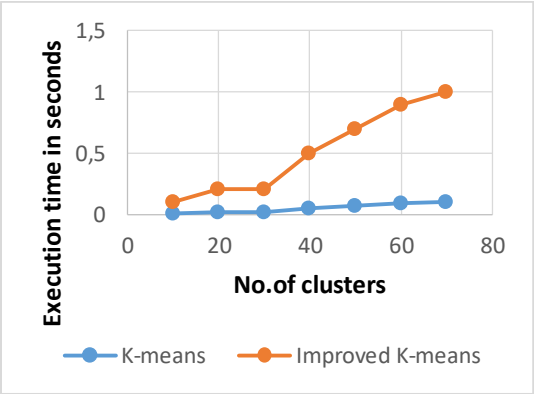

Fig.4 Execution time Vs No. of clusters

Figure4 indicates the Execution time comparison between traditional k-means algorithm and proposed improved k-means algorithm. The X-axis represents the number of clusters, while the Y-axis represents the execution time in seconds. Graph shows that the proposed algorithm provides the solutions much faster than previous k-means algorithm

V. CONCLUSION

Where there may be an increase in the widespread usage of online shopping sites, where customers can acquire anything they want with a single mouse click. Web use mining is one of the most important fields of research since it allows for easy prediction of browsing and data browsing. Various researchers have proposed a variety of technologies and methods for investigating site use mining. We studied the discovery of web user patterns in the education area, as well as the pattern analysis of employing similar web pages and similarity score. We have provided an algorithm that can automatically save the original value of k and select the correct prime factors according to the tools in the data sets. The algorithm proposed to classify similarity of clusters called improved k-means clustering. The proposed algorithm compared with previous k-means algorithm in the two measures of sum of squared error (J) for cluster error and execution time for cluster formation. The experimental results show that the proposed algorithm better, when compared with previous algorithm.

REFERENCES

1. R, Mobashar and Cooley, 1997, “Web Mining: Information and pattern discovery on the World Wide Web”, pp. 558-567.
2. Bari, P & Chawan, PM 2013, ‘Web Usage Mining’, Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, no. 6, pp. 34-38.
3. Zalane, OR & Luo, J 2001, Web usage analysis for a more effective web-based learning environment: Proceedings of Conference on Advanced Technology for Education, pp. 60-64.
4. Tug, E., Sakiroglu, and Arslan, A.M. “Automatic discovery of the sequential accesses from web log data IJCSI International Journal of Computer Science Issues, 20 ISSN (Online): 1694 0784 ISSN (Print) Vol. 7, No.4, pp.180-186, 2010.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

5. Dr. Rajesh and K Shukla, 2020, “WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining”, pp.1-6.
6. S. Jeyalatha and B. Vijayakumar, 2019, “Web Usage Mining Algorithm for an Academic Search Application”, pp.674-679.
7. Kumar M., Meenu. Analysis of visitor’s behaviour form Web Log using Web Log Expert Tool. International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017.
8. Dhanalakshmi P., Ramani K., Reddy E. The Research of Pre-processing and Pattern Discovery Techniques on Web Log files. IEEE 6th International Conference on Advanced Computing, 2016.
9. Geng R., Tian J. Improving Web Navigation Usability by Comparing Actual and Anticipated Usage. IEEE Transactions of Human-Machine Systems. Vol. 45, No. 1, February 2015.
10. Yadav, M.P.; Feeroz, M.; Yadav, V.K., 2012, "Mining the customer behaviour using web usage mining in e-commerce,” pp.1-5.
11. J.Shekhar and K.Ahmad, 2011,”Analysis of Web Mining Applications and Beneficial Areas”.
12. Zhiwu Liu; Li Wang, "Study of Data Mining Technology Used for E-commerce," Intelligent Networks and Intelligent Systems (ICINIS), 2010 3rd International Conference on, vol., no., pp.509-512, 1-3 Nov. 2010.
13. Chaoyang Xiang; Shenghui He; Lei Chen, 2009, "A Studying System Based on Web Mining," pp.433-435.
14. Magdalini Eirinaki, Michalis Vazirgiannis, 2008, “Web mining for Web personalization”.
15. Wang H., Yang C., Zang H. “Design and Implementation of a Web Usage Mining Model Based on Fpgrowth and Prefixspa”.