International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 5, Sept-Oct-2023, ISSN (Online): 2395-566X

An Efficient Approach to Predict the User's Interest Based upon the Current Data

M.Tech. Research Scholar Sarabjit Kaur , Professor Gurpreet Singh
Deptt. of Computer Science Engineering,
St. Soldier Institute of Engineering & Technology, Jalandhar, India

Abstract- In the field of Information Technology, finding the right information can be tough. The internet is filled with a massive amount of data, and keeping it organized is crucial but challenging. As the internet keeps growing, we have a chance to understand how people use it by looking at web access logs. To make sense of all this data, we use data mining techniques. However, web data is messy and not easy to work with using standard data mining methods. So, we came up with a way to study how users behave based on weblogs. Our method helps us mine useful information from these logs. We use what we learn about user behaviour to come up with new strategies for similar groups of users. This approach can be useful in areas like social media, online shopping, and figuring out what users are interested in from their web activity. Our experiments have shown that our method works better than older ones. The key to our success is the way we organize the data, making it easy to extract different types of information.

Keywords - Web Usage Mining, Social Network Services (SNS), Weblog Data, Information Technology, Online Shopping.

I. INTRODUCTION

Web mining is a field that brings together computer science and data mining. Its main goal is to dig out useful knowledge and understanding from the enormous and varied data found on the World Wide Web. It involves various methods and strategies for examining web content, how websites are built, and how users interact with them. Data mining is a technology that extracts valuable insights from vast and massive datasets, transforming it into a format that is easily comprehensible to humans.[1] The subject of identifying user interests has generated substantial research interest due to its intriguing and multifaceted implications, as well as its strategic relevance in the business world.

These studies have explored various domains, including social networks [2], and have also taken into account users' geographical locations [3]. Simultaneously, the methods and techniques employed for analyzing user behavior have undergone significant evolution. This strategy also involves the generation of patterns and the monitoring of user actions. The proposed algorithm can provide recommendations by analyzing these user behaviors[4].

Data mining can be described as an advanced data search technique that employs statistical algorithms to uncover patterns and associations within datasets [5]. It essentially reveals concealed patterns and relationships that exist within the data [6]. By employing data mining techniques like clustering, decision trees, or association analysis, it becomes feasible to unearth crucial traits and patterns from customer details. These extracted characteristics can then be potentially harnessed for making predictions in the future [7]. Social networking service (SNS) is one of the most encouraging directions of web applications [8]. In this context, the data is gathered in the Internet Information Services (IIS) file format, primarily for uncovering concealed insights about visitors. NASA-HTTP collects this log data, and you can find the log file at the Internet Traffic Archive (ITA) [9].

II. METHODOLOGY

Web mining is the practice of extracting valuable knowledge and insights from the immense volume of data accessible on the World Wide Web. It entails applying data mining techniques to web-related information, covering aspects such as web content, structure, and user behavior. The primary objective of web mining is to uncover concealed patterns, emerging trends, and valuable information that can be put to use in numerous ways.

These applications include website enhancements, better user experiences, and informed decision-making based on datadriven insights. Preparing weblogs, or log data, for analysis involves a critical step called preprocessing. This process is vital for cleaning, structuring, and converting the raw data into a usable format for analysis Shown Figure 1 Following the preprocessing of weblog data, which involves the removal of unsuitable attributes and structuring the data appropriately, it became evident that different online users visited the site at various times. To streamline the data for analysis, a segregation process was implemented, resulting in the arrangement of weblog data using five parameters

International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 5, Sept-Oct-2023, ISSN (Online): 2395-566X

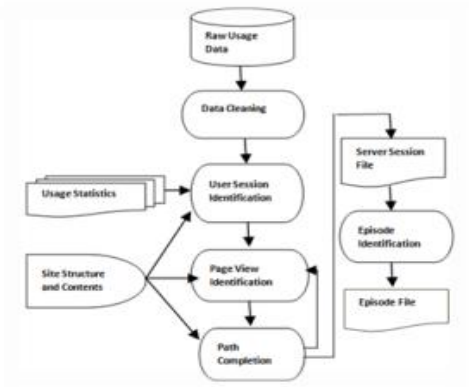

Figure 1 Pre-Processing Model for Weblog Data

1. Uid (User Identifications): This parameter identifies individual users.
2. Iid (Image Identifications): It helps identify specific images or content.
3. Tid (Time Slot): Time values from the weblogs were converted into timestamps and organized into time slots. These time slots divide the day into 24 segments, such as T1, T2, and so on upto T24 with each slot representing a maximum duration of 60 minutes.
4. Tweight (Number of Logins): This parameter tracks the number of times a user logs in during a specific time slot. 5. Iweight (Number of Visits): It records the number of times a user visits specific content or images.

Table 1 Time slots and Time Intervals

Time Slot	T_1	T_2	-	-	T_{22}	T_{23}	T_{24}
Time Interval	00:01-01:00	01:01-2:00	-	-	21:01-22:00	22:01-23:00	23:01-24:00

Five parameters such as Uid , Tid , Tweight , Pid and Iweight. There are following description of those parameters. Uid is a user ID, this id can be repeated in the table because same user can be interacted in same or different time slots at Social Network Services (SNS). From weblog data the Value of those five parameters is shown in Table 2.

Table 2: Parameter for Identifications

Uid	Tid	Tweight	Pid	Iweight
1	21	5	P_{101}	3
1	22	4	P_{102}	5
2	23	3	P_{104}	7
2	21	2	P_{101}	5
2	22	7	P_{103}	11
3	22	6	P_{103}	7
3	24	5	P_{105}	18
4	21	6	P_{102}	4
4	22	8	P_{103}	6
5	23	3	P_{105}	3

In the weblog data, a focused analysis was conducted on five specific users, denoted as U1 to U5, who exhibited the highest activity on a particular website during the time slots T21 to T24. Figure 2 illustrates the weightage or intensity of these users' behaviors during this time period. Through the visualization of these user behaviors, a robust and discernible pattern has emerged, providing valuable insights into how these users interacted with the website during these specific time slots. This pattern can offer significant information for further analysis and decision-making. shown in Figure 3.

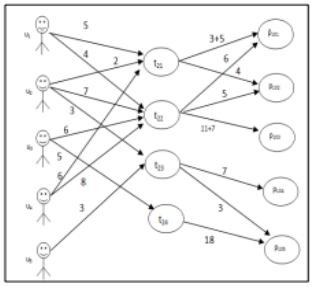

Figure 2 User's Behavior in Weblog

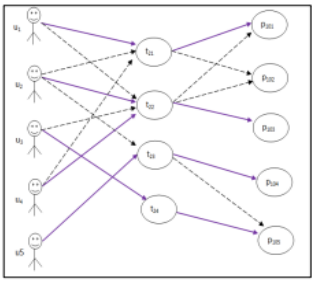

Figure 3: Strong Pattern generations

Using this graph, useful relation is generated corresponding to the weight value. These relationships are calculated using the concept of Related Intensity (RI), and they focus on pairs of parameters such as (Uid, Tid) and (Tid, Pid). The calculation of RI involves the use of predefined values for α1 and α2, which must satisfy the relationship α1 + α2 = 1. In this particular case, α1 is set to 0.6, and α2 is set to 0.4. Additionally, a minimum support threshold of 0.3 is applied. The specific formula and parameters chosen are aimed at achieving meaningful results in discovering strong patterns within the data. $\{weight\ of\ U \bowtie T$ = [x1 $Highest\ (x)$, x2 $Highest\ (x)$ … … … … … … xn $Highest\ (x)$] }……..eqn 1. $\{weight\ of\ T \bowtie P$ = [y1 $z1$, y2 $z2$, y3 $z3$, y4 $z4$ … … … … … … … yn zn] }……..eqn 2. Results Using the formula and parameters mentioned earlier, the relationships for user U1 between Uid and Tid, as well as between Tid and Pid, have been established as follows. Weight of U⋈ T (U1, T21) = 5/5 = 1 (U1, T22) = 4/5 =0.8 Weight of T⋈ P ((T21, P101) = 8/18 = 0.44 (T21, P102) = 4/18 = 0.22 < 0.3(REJECT) ((T22, P101) = 6/18 = 0.33 ((T22, P102) = 5/18 = 0.27 < 0.3(REJECT) ((T22, P103) = 11/18 = 0.61 Relationship Intensity RI = [E1(ui, ta) × α1] + [E2(ta, ib) × α2] U1 → T21 → P101 = [1 ×0.6] + [0.4 × 0.4] = 0.76 U1 → T21 → P102 = [1 ×0.6] + [0.4 × 0.11] = 0.64 Here RI of different relation is arranged in decreasing order as RI (U1 → T21 → P101) > RI (U1 → T21 → P101) The relationship intensity method, as described, is applied uniformly to all users to calculate the RI (Related Intensity) values.

These RI values are then utilized in MATLAB to conduct a more in-depth analysis of the patterns within the data. The results for all five users are obtained and summarized

in a dataset, as shown in the table.The MATLAB analysis likely provides a comprehensive view of the patterns and relationships between various parameters for each user. User1.Table suggests the information that User 1 was actively engaging with P103 from 8:01 PM to 9:00 PM with weightage of 18 and then switched to interacting with P101 from 09:01 PM to 10:00 PM with 8 weightages. It might be valuable to analyze the context or content of these images or sites to understand the user's interests better is shown in Table 3.

Table 3 Activity of User-1

U_{id}	T_{id}	P_{id}	T_{weight}
1	21	P_{103}	18
1	22	P_{101}	8

User 2 Based on the data in Table 4, it can be inferred that User 2 was actively involved with P104 from 10:01 PM to 11:00 PM, and during this time, it had a weightage score of 7. Subsequently, within the same time slot, User 2 transitioned to interacting with image ID P105, which had weightages of 3.

Table 4 Activity of User-2

U_{id}	T_{id}	P_{id}	T_{weight}
2	23	P_{104}	7
2	23	P_{105}	3

User 3 Analyzing the information from Table 5, it is apparent that User 3 was actively engaged with P101, P102, and P103 between 09:01 PM to 10:00 PM. During this period, he exhibited weightage scores of 6, 5, and 18 respectively,

Table5 Activity of User-3

U_{id}	T_{id}	P_{id}	T_{weight}
3	22	P_{101}	6
3	22	P_{102}	5
3	22	P_{103}	18

User 4 Based on the details provided in Table 6, we can conclude that User 4 actively interacted with P101 from 08:01 PM to 09:00 PM, during which the weightage score was 8. Following that, during the same time window, User 4 shifted their focus to engaging with image ID P102, which had a weightage of 4.

Table 6: Activity of User-4

U_{id}	T_{id}	P_{id}	T_{weight}
4	21	P_{101}	8
4	21	P_{102}	4

User 5 Based on the data found in Table 7, it can be concluded that User 5 was actively interacting with P104 from 10:01 PM to 11:00 PM, and during this period, it had a weightage score of 7. Subsequently, within the same time frame, User 5 transitioned to engaging with image ID P105, which carried weightages of 3.

Table 7: Activity of User-5

U_{id}	T_{id}	P_{id}	T_{weight}
5	23	P_{104}	7
5	23	P_{105}	3

In the combined activity chart, it is found that on Image ID P101 three users U1, U3 and U4 were active in which user U1 and U4 visited on-site at the same time slot T21, whereas the other user U3 was active on the same site but in a different time slot T22. on the other hand, it is found that U3 also visited on the image id P102 and P103 at same time slot with having a weightage of 5 and 18 respectively which shows the behaviour of U3 that he has many times visited on P103.It also shows the nature of user U1 which he had also visited many more times on image ID P103 (at T22 timeslot) as compared to P101. furthermore, U2 and U5 showed their interests in the Image ID P104 and P105 in between the T21 timeslot as shown in Figure 4.

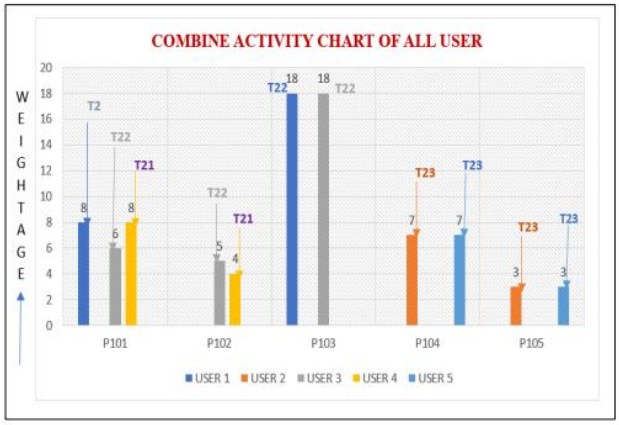

Figure 4: Combine Activity Chart.

In our research, we conducted a performance comparison of our proposed algorithms against an Apriori-like algorithm. This evaluation involved assessing the sensitivity of these algorithms by varying parameters related to different data characteristics. Additionally, we examined how the strength of relationships among the nodes (Uid, Tid, Pid) changed. as the log data for the Social Networking Service (SNS) is typically retained by the server for a fixed period.

Through this comparison, we found that our proposed approach outperformed the Apriori-like approach. Specifically, our approach required less time to generate strong frequent patterns. Shown in Figure 5. Interestingly, as we increased the minimum support value, we observed that the performance of our proposed algorithm only slightly decreased. In contrast, the Apriori-like algorithm exhibited a more significant decrease in performance as the minimum support value increased. This suggests that our proposed algorithm is more robust and efficient in handling changes in the minimum support threshold, making it a better choice for pattern generation in this context.

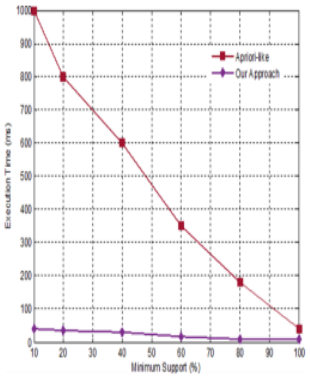

Figure 5: Comparison with Apriorism-like

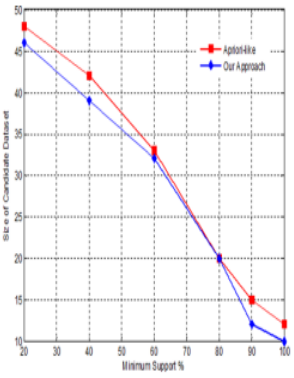

Figure 6 : Size of candidate dataset

The experimental results indicate that the proposed approach can obtain better overall performance for various pattern generations as compared with traditional approach. Because Apriori-like approach takes too much time to generate candidate itemset with different minimum support value, but our proposed approach calculates RI value. Based on those RI values, it can eliminate weak patterns. Those weak patterns are not computed next time. It only considers the strong pattern for those reasons our approach gives better results. user accessed "P101, P103 image in SNS between 10:00 to 10:30 and that user again enter the SNS between 10:31PM to 11:00PM then, according to our approaches, he might have accessed P102 and P103 images. Experimental analysis applied on user U2; in this we have found a few relations of the user (U2).

U2 → T22 → P103 U2 → T23 → P104 In this if the user interacts with the SNS between 10:00 to 11:00 and he accesses the image P103. On the other hand, if the user interacts with SNS between 11:01 to 12:00 then he might access P101 image. In a similar way, we have found different patterns for different users". Conclusion The user behavior analysis approach proves highly effective in uncovering strong patterns among interconnected parameters. These patterns hold substantial potential for businesses, enabling them to introduce new services and improve existing ones. Experimental findings consistently affirm the superiority of this approach in terms of overall pattern generation compared to traditional methods.

Furthermore, this approach extends its utility beyond pattern generation. It finds practical applications in diverse domains such as social network service advertising, e-commerce optimization, and strategic business planning. By leveraging insights gleaned from user behavior analysis, businesses can make informed decisions that lead to more successful and targeted strategies in these areas. The Related Intensity based approach establishes frequent patterns among different access parameters, including user, time, and image, through matching joins operations. These patterns are generated based on predefined support values, making this approach valuable for recommending and predicting services for various users, as confirmed by experimental results with simulated data.

REFERENCES

1. Singh, Gurpreet, and Er Rajwinderkaur. "Review Paper On Decision Tree Data Mining Algorithms To Improve Accuracy In Identifying Classified Instances Using Large Dataset." International Journal Of Computer Science Engineering And Information Technology Research (Ijcseitr) 7 (2017): 67-70.
2. Antelmi, A. (2019). Towards an exhaustive framework for online social networks user behaviour modelling. In Proceedings of the 27th ACM conference on user modeling, adaptation and personalization, UMAP '19 (pp. 349–352). New York, NY, USA: Association for Computing Machinery.
3. Preoţiuc-Pietro, Daniel, and Trevor Cohn. "Mining user behaviours: a study of checkin patterns in location based social networks." In Proceedings of the 5th annual ACM web science conference, pp. 306-315. 2013.
4. Bakariya, Brijesh. "Efficient approach of analyzing and generating intrinsic information from weblog." National Academy Science Letters 44, no. 6 (2021): 525-527.
5. Newton's Telecom Dictionary, Harry Newton, CMP Books, Edelstein H. Data mining: exploiting the hidden trends in your data. DB2 Online Magazine. http://www..htm
6. Arora K. Rakesh, Badal Dharmendra," Admission Management using Data Mining using WEKA", IJARCSSE Vol. 3, Issue 10, October 2013
7. Bakariya, Brijesh, and G. S. Thakur. "Pattern mining approach for social network services." National Academy Science Letters 40 (2017): 183-187.
8. Internet Traffic Archive Repository. http://ita.ee.lbl.gov/, 2013.