International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Secure and Efficient Web Usage Mining : A Novel Framework for Prediction of Consumer Behaviour

Sonia Sharma[1]*, Dalip[2]

[1]Hindu Girls College, Jagadhri, Haryana, India

[2]Department of MMICT &BM, Maharishi Markandeshwar Deemed to be University, Mullana (Ambala), Haryana, India

Article Info

Volume 9, Issue 1

Page Number : 545-551

Publication Issue

January-February-2022

Article History

Accepted : 01 Jan 2022

Published : 20 Jan 2022

ABSTRACT

To protect the weblog data from various threats of active attacks security of data is the major challenge for the organization. To obtain maximum profit keeping intact the interest of consumers is some other major aspect of an organization. Web usage mining techniques prove a benchmark in the field of web development consumer behaviour web intelligence. In the past to analyze consumers behaviour conventional Apriori algorithm was used and improved version i.e. T_Apriori algorithm was introduced by the researcher but problems of space complexity time complexity and no concept of security of the data is found in these algorithms. The present paper presents a Novel Secure and Efficient Web Usage Mining(SEWUM)which not only ensure the security of data but also accurately predicts consumer behaviour with the security of data by exhausting the Hash table Hashing Encryption Message-Digest (MD5) algorithm encoding and decoding and anti-monotonicity Apriori property has a distinct edge over the classical Apriori algorithm and T_Apriori algorithm. The performance evaluation of the proposed system is carried out in python on real data (weblogs). Experimental results show that the proposed technique is more effective for the organization and results can be obtained in less memory consumption and less execution time.

Keywords : Security, Consumer Behaviour, Web Usage Mining, Data Integrity, Active Attacks

I. Introduction

In this era of technology, everything has undergone transformation, instead of offline business online business has started gaining ground, because it is more convenient, time and energy-saving. Consequently, the daily consumers visiting the site started shooting up giving phenomenal rise to raw data (Weblogs)[17]. It became a huge challenge to control, monitor, and process the raw data which keeps flooding in and piling up enormously. In this regard, the most genuine method is web usage mining [1] [2] [3] which is competent enough to discover patterns hidden in raw data. Most of the organizations are using web analytics tool [9] to analyze traffic on the web site but these tools only provide traffic statistics but by applying web usage mining techniques, web intelligence, market analysis, exploration of consumer browsing behaviour can easily be obtained and in this way an organization can garner maximum profit. However, with the mining of weblogs (raw data), the security of the piled up data is another serious challenge in this field. The integrity of relevant data forms the foundation of dependable and justified results [8]. Any analysis emanating from corrupt data can lead to serious consequences. Even companies' future may be in the doldrums. Several methods are adopted by the organizations for data security such as passwords, firewall, etc. The proposed approach explores the browsing behaviour of consumer with data integrity by using web usage mining process. The proposed work is a marked deviation from the earlier researches which either focused on mining the weblogs by using classification, clustering, association rule mining or secure data communication but the authors have attempted to conduct their research work from a new angle i.e. exploration of consumer browsing behaviour using web usage mining technique (association rule mining) and side by side assuring the security of data by consuming less memory ,taking less execution time and still maintaining high accuracy rate in predicting behaviour. Section-II describes the study done by various researchers .Research motivation and author contribution is mentioned in section-III .Section-IV clearly depict the process carried out by the author for novel methodology SEWUM. Various results obtained are shown in section V.

II. Review of Literature

In recent years, a plethora of research has been conducted concerning web usage mining for user web browsing behaviour. The author [6] studied the Apriori algorithm which is based on association rule mining and suggests that various needs of business analytics, consumers, and service providers can easily be understood with the Apriori algorithm and the business sale can be increased and it is the best technique for web usage mining. The authors [5] felt motivated for research to discover the patterns generated on weblog server. They compared Apriori algorithm and FP growth algorithm for their research. [4] Authors offered enormous assistance for e-business and gives an overview of various security threats that can be occurred on e-commerce web site and also show the concept of web mining in a secure way. Authors [7] designed Lempel-Zbyiv-Welch algorithms by keeping in mind that for data, information and message transferred or stored online, security and confidentiality is the major aspect under consideration.Also, the authors emphasized upon the digital representation of data which requires compression of data. Apriori Algorithm is one of the oldest classical algorithms proposed by the authors R. Aggarwal and R. Srikant to find the association among data items [13]. This algorithm is mainly used for market-based analysis but presently the organizations are using this algorithm in web usage mining, analysis of medical data, and educational data [14], [15], [16]. T_Apriori is proposed by Xiuli Yuan which is an improved version of Apriori algorithm for mining association rules, It utilizes overlap strategy to generate association rule. Author checked its performance with other modified algorithms such as I-Aprioi, BITXOR and also with conventional Apriori algorithm and shows its performance better than existing ones.

III. Research motivation and Contribution

After studying literature it is found that a lot of improvement in the methodology is required for the accurate understanding of consumer's behaviour by actual implementation of methodology efficiently w.r.t memory usage and execution time. After thorough scrutiny, it came to the fore that the Apriori algorithm suffers from two problems. Firstly, it keeps filtering the database time and again and thereby generating, unnecessarily, an enormous number of rows. Secondly, there is no concept of the security of data. Similarly in T_Apriori algorithm , problem of space complexity and foolproof security of data is found. There is a need of novel mechanism of exploration of consumer behaviour with security of data. By considering security of data as main parameter our research focuses on designing novel methodology which not only secures data from active attacks but also predicting consumer behaviour accurately and provides quick results thereby helping organizations to increase their profit. Therefore, researcher has selected conventional Apriori algorithm as benchmark for the study. While undertaking this research there was a time when multiple ideas started cropping up in researcher's mind regarding use of Apriori algorithm as a new angle of research however, deep analysis and holistic study of all the angles in the given research area pushed the researcher to remain confined that researcher decided to undertake. Even the study of improved versions of Apriori algorithm strengthened my convictions of sticking to conventional Apriori algorithm for designing Novel Secure Framework but side by side not negating the advantages of T_Apriori. The researcher has decided to go to T_Apriori for its advantages.Actual implementation of the methodology is done by utilizing Hash table, Hashing Encryption (Message Digest algorithm MD5), association rule mining on the weblogs (31 October 2017 to 25 Dec 2017) of website www.viralsach.xyz**. Results are analyzed by using Answerminer tool. This novel framework is far ahead of the conventional Apriori algorithm as well as T_Apriori algorithm. Not only its effectiveness and security of information is unquestionable but its accuracy in predicting the consumer's future conduct is also found to be unparalleled.

IV. Novel Methodology (SEWUM)

By considering the security of data as the main objective and to reduce the problems that occurred in the Apriori algorithm as well as in T_Apriori algorithm . We have to propose a secure and efficient web usage mining technique (SEWUM). We use hash table [11] [12] and anti-monotonicity property and hashing encryption (Message Digest Algorithm MD5) together to get results securely and efficiently. All the process carried out for novel methodology (SEWUM) is shown in Figure-I. At Step-I For experimental purposes, the raw data is collected from the genuine site www.viralsach.xyz for the period 31 Oct 2017 to 25 December 2017 and it contains 33172 rows and it is converted into .csv format.

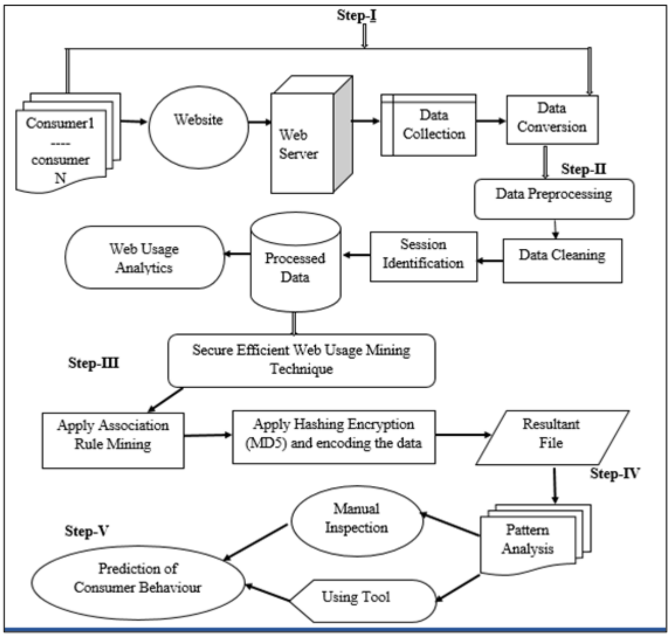


Figure-I : Process of SEWUM

At Step-II data preprocessing which is the most important step of Web Usage Mining is performed. Anonymous data is removed at this step. After data preprocessing data size is reduced to 18216 rows. This data is considered as processed data for further steps. At step-III,proposed methodology i.e. secure pattern discovery algorithm is applied by utilizing hash table, MD5 and Apriori property and setting of minimum support level. After getting secure resultant file as per the minimum support count, next step is to analyze the results. In our work we haveanalyzed the data manually and also by using Answerminer tool [10] .

V. Results

Performance of algorithms Apriori algorithm,T_Apriori algorithm and the proposed algorithm (SEWUM) is checked by executing all algorithm under the same environment(chosen Python preferably spyder 3.2) on same data set at the same minimum support by taking memory usage and execution time and security of data as the main parameter. Toensure the better performance of proposed algorithm.The raw data of18216 rows were reduced into different data sets while progressing in the system, Data set-I consists of 18216 rows, data set-II consist of 10576 rows, data set-III consist of 6656 rows and data set-IV consist of 2567 rows. Numbers of experiments were performed at various minimum support levels. Comparison in terms of memory usage & execution time for different data set is shown in figure-III & IV. To studythe performance of both algorithms in terms of accuracy of predicting behaviour,we analyze the results obtained from data set IV. From resultant file we arbitrarily select some IP address& referring URL and checked up the chances of occurrence of IP address and referring URL w.r.t row number. Table-II shows arbitrarily chosen IP address and name assigned andTable-I showsreferring URL and name assigned. Results obtained are shown in figure-V and figure-VI.Figure-II depicts the difference in memory usagefor allalgorithmsat different dataset. As the memory reduced at data preprocessing at 50% then after implementation of proposed algorithm(SEWUM) memory there is again reduction of memory usage & Figure-III depicts the difference in execution time for different data size at best minimum support level. Figure-IV depicts the difference in accuracy of predicting chances of IP address for all algorithm .The chances of occurrence of IP address U1 is .44% in case of SEWUMand by Apriori algorithm is .25% and by T_Apriori algorithm is 0.35%.. For U2 chances are 3.45 by proposed system and 2.22 by Apriori algorithm and 2.94% by T_Apriori algorithm. Similarly difference of predicting result is obtained for U3, U4 and U5. Occurrence of U4 is 3.98% more by proposed system as compared to Apriori algorithm and T_Apriori algorithm. Figure-V depicts the difference in accuracy of predicting Referring URL for both algorithms. A, B.C.D name is assigned to referring URL for work. Predicting of chances of referring URL by proposed algorithm is more in less rows where as in Apriori algorithm chances are less in more rows as well as in case of T_Apriori algorithm. Chances of A are 32.53% in SEWUM and 21.63% in Apriori algorithm and 28.23% in T_Apriori algorithm. The proposed system (SEWUM) gives a 10.9% more accurate result as compared to Apriori algorithm.

Table-I: Notation used for Referring URL

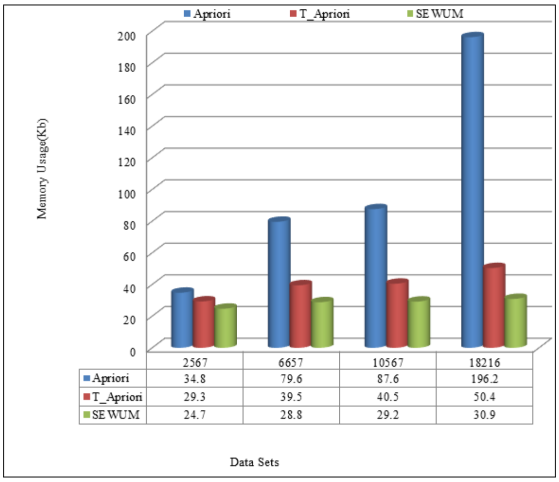
Randomly picked Referring URL	Name Assigned
http://viralsach.xyz/	A
http://viralsach.xyz/bulk-sms-india-2/	B
http://viralsach.xyz/tehacher-update-time-time/	C
http://viralsach.xyz/life-easy-due-research/	D

Table-II: Notation used for IP Address

Randomly picked IP address	Name Assigned
106.192.241.96	U1
101.226.33.218	U2
122.173.226.14	U3
111.125.230.213	U4
101.226.66.187	U5

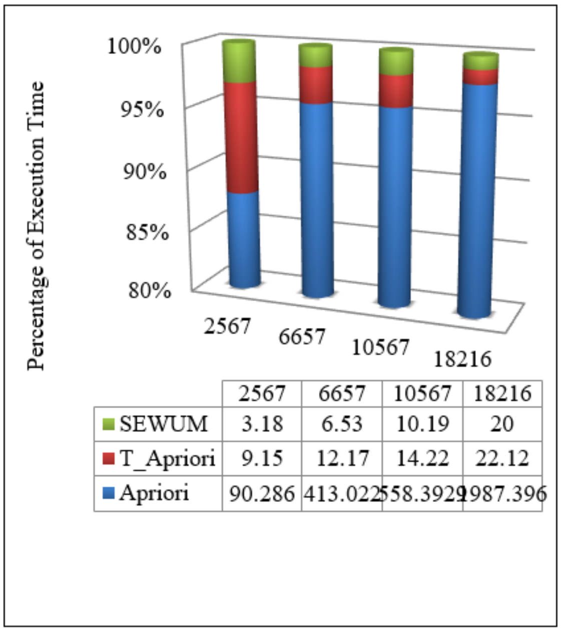
VI. Discussion

SEWUM shows results in terms of less memory usage for the different data sets. SEWUM takes care of unwanted data using Filters. After security filters data gets reduced in size for access and this is evident in the results shown in graph. Here the gain in terms of memory can be referred to new proposed scheme functioning at optimum level. Here not only filtering is done but data has been screened using security algorithm (like Hashing). This may lead to a bit of Offhead overload but the gain is more significant than other factors. The trend that the proposed scheme has firmly established keeps gaining strength when more and more data is loaded. Also improvement in percentage is more visible and effective. It may be worth mentioning that in many cases proposed scheme is able to reduce memory liability by more than 50 % in many cases. SEWUMshows results in terms of less execution time for different data set. Execution time has been taken as one metrics. in case of small data size as in dataset- I(2567) it is seen that execution time has come on to a scale of 23 % . This reduction can be seen in a context that when lesser data is there for accessing and carrying out operations, then obviously time of execution will be lesser. Though efforts were made to take execution time in terms of milliseconds, but due to speed and other hardware related issues it has been noted in seconds. Reduction in execution time reaches to a maxima of 80 to 90 percent. Here it can be seen effectively that SEWUM has taken more aspects here. As hashing and MD5 have collectively reduced mass data volume, time is saved in bulk which in turn will increase effective rating of scheme. As our scheme kept modifying using more and more parameters for security aspects then data has been taken in filter mode and more and more data gets reduced to give way to quality data and this trend is visible in all metrics. SEWUM is more accurate in terms of predicting behaviour. As in the analysis phase, it is shown that chances of occurring of IP address are more in small data. Due to the security of data, more accuracy, less execution time, and less memory usage, the efficiency of SEWUM is more as compared to the Apriori algorithm and T_Apriori algorithm .



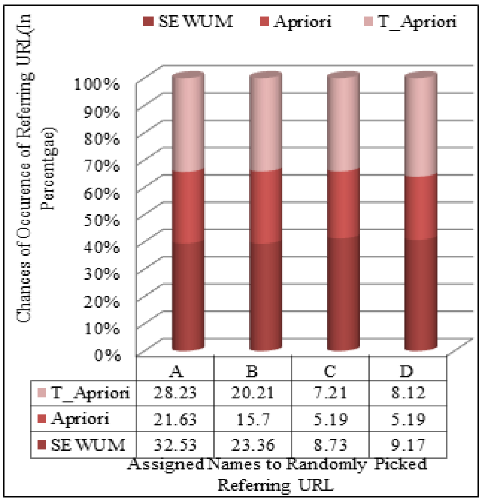
	2567	6657	10567	18216
Apriori	34.8	79.6	87.6	196.2
T_Apriori	29.3	39.5	40.5	50.4
SE WUM	24.7	28.8	29.2	30.9

Figure-II: Difference in Memory Usage



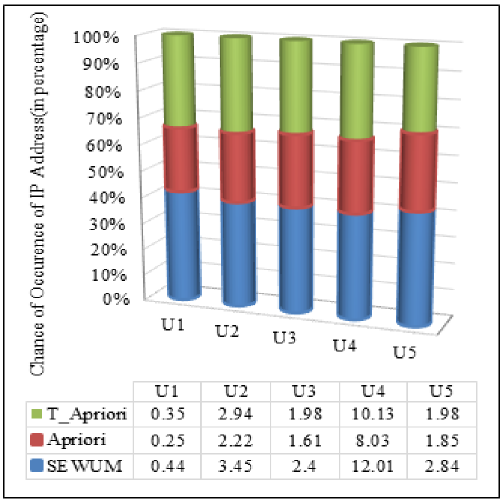
	2567	6657	10567	18216
SEWUM	3.18	6.53	10.19	20
T_Apriori	9.15	12.17	14.22	22.12
Apriori	90.286	413.022	558.392	1987.396

Figure-III: Difference in Execution Time



	A	B	C	D
T_Apriori	28.23	20.21	7.21	8.12
Apriori	21.63	15.7	5.19	5.19
SE WUM	32.53	23.36	8.73	9.17

Figure-V: Performance of algorithms predicting Occurrence of randomly picked URL



	U1	U2	U3	U4	U5
T_Apriori	0.35	2.94	1.98	10.13	1.98
Apriori	0.25	2.22	1.61	8.03	1.85
SE WUM	0.44	3.45	2.4	12.01	2.84

Figure-IV: Performance of algorithms predicting Occurrence of randomly picked IP address

VII. Conclusion

For e-business, keeping in view the concern of consumers, security of data is prerequisite to defend it from active attacks such as masquerading; message modification on the data .This paper presents the very effective, efficient, secure approach for an organization by choosing python as a platform which will be fruitful for the organization to explore the behaviour of the consumer securely. By taking the security of data as the foremost phases, SEWUM uses the message digest algorithm (MD5) hash function with encoding and decoding of data so it satisfies the main attributes of security such as data integrity,availability,non-repudiation, and data freshness. Experimental results show that SEWUM is taking less memory usage and less execution time for different data set as compared to Apriori algorithm as well as T_Apriori algorithm and accuracy of predicting results in the proposed algorithm is more as compared to Apriori algorithm and T_Apriori algorithm. Due to the amalgamation of security, hash table, Apriori property, encoding decoding of data, results are efficient in terms of less memory usage, less execution time, more accuracy of predicting results, less computation cost. Therefore proposed system is more efficient and secure.In future to improve the performance of the proposed system, meticulous tests should be conducted to maintain its accountability and more securemethods such as digital signature should be imposed. Prediction of consumers with security can be imposed with deep learning techniques.

References

[1]. Li Yong-hong; Liu Xiao-liang(2010), "Research of data mining based on e-commerce," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.4, no., pp.719-722.

[2]. Chaoyang Xiang; Shenghui He; Lei Chen (2009), "A Studying System Based on Web Mining," Intelligent Ubiquitous Computing and Education, 2009 International Symposium on, vol., no., pp.433, 435.

[3]. Li Mei; Feng Cheng(2010), "Overview of Web mining technology and its application in e-commerce," Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, vol.7, no., pp.V7-277,V7-280.

[4]. Prof. Dr. M. E. Mohammad pourzarandi, R. Tamimi(2013), " The Application of Web Usage Mining In E-commerce Security", International Journal of Information Science and Management, IJISM, Special Issue (ECDC 2013)

[5]. B. Uma Maheswari, Dr. P.Sumathi (2015), "A Comparative Study of Rule Mining Based Web Usage Mining Algorithms", International Journal of Science and Research (IJSR), ISSN: 2319-7064, Vol-4, Issue-11, pp: 2540-2543.

[6]. Ashish Vitthalrao Galphade and Dhiraj Bhise (2016), "Suggestion of an Apriory Algorithm for Web Recommendation System", Imperial Journal of Interdisciplinary Research (IJIR), ISSN: 2454-1362, Vol-2, Issue-8, pp: 997-1001.

[7]. Ida Bagus Ary Indra Iswara, Ketut Sudarsana(2018)," Application of Data Encryption Standard and Lempel-Ziv-Welch Algorithm for File Security," International Journal of Engineering & Technology, 7(3.2), pp. 783-785.

[8]. Sonia Sharma, Dalip (2020), "A Novel Secure Web Usage Mining Technique to Predict Consumer Behaviour" International Journal of Advanced Science and Technology. Vol. 29, No. 5, (2020), pp. 5633 – 5640.ISSN: 2005-4238 IJAST.

[9]. Sonia Sharma, Dalip (2019),"Comparative Analysis of various tools to Predict Consumer Behaviour" Journal of Computational and Theoretical Nano science Vol. 16, 3860–3866, 2019.

[10]. www.answerminer.com

[11]. K.Vanitha and R.Santhi (2011), "Using Hash Based Apriori Algorithm to Reduce the Candidate 2-Itemsets for mining Association Rule" .Journal of Global Research in Computer Science. Volume 2, No. 5, April 2011

[12]. Arwa Altameem and Mourad Ykhlef(2018). Hybrid Approach for Improving Efficiency of Apriori Algorithm on Frequent Itemset, IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.5, pp.151-156.

[13]. Z. Chen, S. Cai, Q. Song, and C. Zhu. Retracted Article (2011)," An improved Apriori algorithm based on pruning optimization and transaction reduction", 2nd Int. Conf. Artif. Intell. Manag.Sci. Electron. Commer. AIMSEC 2011 - Proc., pp. 1908–1911.

[14]. Q. Zhang(2013), "The Application of Apriori Algorithm in Analysis on Admitted Students of Colleges and Universities, Appl. Mech. Mater., vol. 321–324, pp. 2578–2582..

[15]. Y. Guo, M. Wang, and X. Li (2017), "Application of an Improved Apriori algorithm in a Mobile ecommerceRecommendation System", Ind.Manag. Data Syst., vol. 117, no. 2, pp. 287–303.

[16]. N. Duru. An Application of Apriori Algorithmon a Diabetic Database, Database, vol. 3681 LNAI, pp. 398–404, 2005.

[17]. Sonia Sharma, Dalip, "Web Logs - A Roadmap to Online Consumer", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 1, pp. 576-581, January February 2019.