IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

A Novel Web Usage Mining Technique Analyzing Users Behaviour Using Dynamic Web log

Priyanka Sharma[1*], R.K. Gupta[2]

[1*] Dept. of Computer Science and Engineering, MITS, Gwalior, INDIA
[2] Dept. of Computer Science & Engineering, MITS, RGPV, Gwalior, INDIA

**Corresponding Author: 27shpriyanka@gmail.com*

Available online at: www.ijcseonline.org

Received: 17/May/2017, Revised: 28/May/2017, Accepted:14/Jun/2017, Published: 30/Jun/2017

***Abstract*—** WWW-world wide web is one of the most required resource for getting information and knowledge. Several organizations rely on websites to get new customers and to hold the existing one. The customer's surfing pattern can be obtained by web log record. This is a kind of work in the arena of web usage mining. The paper proposed the novel methodology by mining the usage of patters from web log records of real time. Ontology of the web content including user profiles and external data can be developed by Web usage mining. In this paper a method is proposed for discovering and tracking the growing user profiles along with domain specific information facets. To judge the quality of the obtained profiles after mining an objective validation plan is also used. Through this research organizations can take better decision by getting better recommendation

***Keywords*—** Data-mining, web-mining, web usage mining,etc

I. INTRODUCTION

With the expansive number of organizations utilizing the Internet to appropriate and gather data, learning disclosure on the web has turned into an imperative research range [1]. With the unstable development of data sources accessible on the World Wide Web, it has turned out to be fundamental for associations to find the utilization designs and investigate the found examples to pick up an edge over contenders. Jespersen et al [2] proposed a mixture approach for dissecting the guest click stream successions. A blend of hypertext probabilistic language structure and snap reality table approach is utilized to mine Web logs, which could be additionally utilized for general succession mining undertakings. Mobasher et al [3] proposed the web personalization framework, which comprises of disconnected undertakings identified with the mining if use information and online procedure of programmed Web page customization in light of the learning found. (LOGSOM, a framework that uses Kohonen's self-sorting out guide (SOM) to arrange website pages into a two-dimensional guide) proposed by Smith et al [4], uses a self-sorting out guide construct exclusively in light of the clients' route conduct, instead of the substance of the site pages. Logger proposed by Chi et al [5] develops client profiles by consolidating both grouping of client sessions and customary measurable activity examination utilizing k–means calculation. Joshi et al [6] utilized social online investigative preparing approach for making a Web log distribution centre utilizing access logs and mined logs. An exhaustive outline of web use mining examination is found in [1][3][7]. The web mining is an after effect of hybridization of the two regions i.e. information mining and second one is World Wide Web (WWW).

II. RELATED WORK

Chakarbarti [5] gives a study of information digging for hypertext. His paper principle accentuation on factual strategies like NPL for web content crosswise over administered, semi directed and unsupervised adapting additionally on interpersonal organization investigation procedures for web structure mining.

Garofalakis et al [6] survey a few information maining procedures and the calculations for web mining that particularly considers the hyperlink data.

NeeruMago[8] centres to give an up and coming review of the quickly developing territory of Web mining. With the development of Web-based applications, particularly electronic trade, there is noteworthy enthusiasm for dissecting Web substance, its structure and use of information to better comprehend and apply the learning to better serve clients.

Mohinder Singh and NavjotKaur [9] concentrate on portrayal issue on the procedure and learning calculation which depends on page positioning strategy.

Monika Yadav and Mr.Pradeep Mittal [10] manages a preparatory discourse of WEB mining, few key software engineering commitments in the field of web mining, the

noticeable fruitful applications and diagrams some encouraging regions of future research.

Ketul B. Patel, Jignesh A. Chauhan , Jigar D. Patel [11] talks about web mining in online business, the classes of web mining, design revelation strategies to discover intriguing examples, issues of web mining in web based business and utilization of web mining in internet business.

III. METHODOLOGY

Web mining can be categorized into three main areas. (i) *Web Content Mining*. (ii) *Web Structure Mining* and (iii) *Web Usage Mining.* We concluding the Web usage mining in table below

Web Mining				
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	DB View		
View of Data	Unstructured and Semi-structured	Semi-structured and website as DB	Link structure	Interactivity
Main Data	Text Document and Hypertext as document	Hypertext document	Link structure	Server log and Browser log
Techniques	Statistical (NLP) and Machine Learning	ILP and Modified Association on rule	Proprietary Algorithm	Machine Learning and Statistical

Table 1: Summary of Web usage mining

The profile based information extraction and assessment is as yet having the issue in the web use mining. The objective or target of web use mining is to catch, display and break down the behavioural examples and profiles of clients communicating with a site. The found examples are typically spoken to as accumulations of pages, items or assets that are every now and again gotten to by gathering of clients with basic needs or intrigue. The objective is acquiring is to prescribe the clients in light of their dynamic inquiries and web profile shaped in past sought. We suggest the clients in view of their scope and longitude unless indicate some area based question. We channel every one of the information which appears to be futile to clients contingent upon their dynamic web log and profile era.

Proposed Methodology has been described in the following

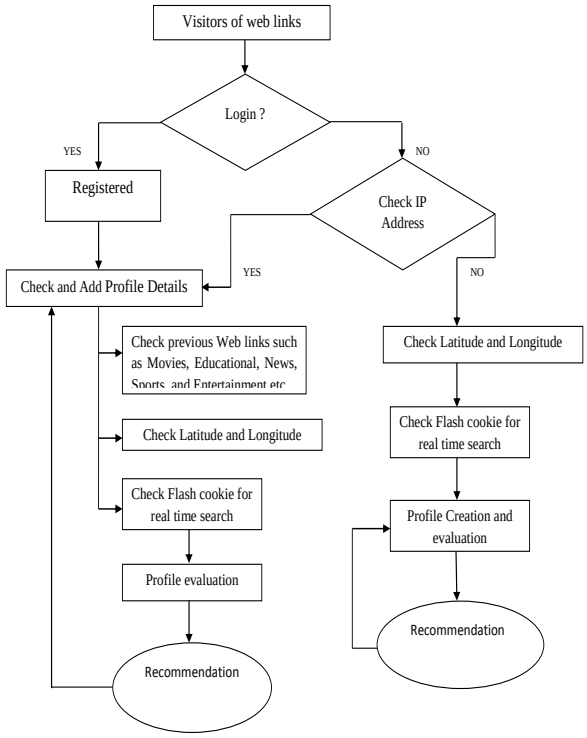

Figure 1: Proposed Enhanced Technique for Better Recommendation

We apply our model to two situations: in the main the likelihood of a client ending the route session is autonomous of the quantity of connections he has taken after up until now, and in the second the likelihood of a client ending the route session increments by a consistent each time the client takes after a connection. We break down these situations utilizing two arrangements of trial informational collections demonstrating that, in spite of the fact that the main situation is just a harsh estimate of surfers' conduct, the information is steady with the second situation and can in this manner give a clarification of surfers' conduct. In this review we introduce a total system for mining Web utilization designs with certifiable difficulties, for example, advancing access designs, dynamic pages, and outer information portraying philosophy of the Web substance and how it identifies with the business performing artists. The Web website in this review is an entrance that gives access to news, occasions, assets, organization data and a library. The Web website in our review is overseen by a not-for-profit association that does not offer anything besides rather just gives free data. Here we perform bunching of the client sessions extricated from the Web logs to parcel the clients into a few homogeneous gatherings with comparative exercises and after that concentrate client profiles from each group as an arrangement of significant URLs. Information mining

systems have been connected to concentrate utilization designs from Web log information; this procedure is known as Web use mining.

The proposed framework for the most part incorporates the accompanying a few strategies: information gathering, information pre-handling, session based bunching and multi parameter figuring and division.

4.1 Data collection

Information accumulation is the initial step of web utilization mining, the information legitimacy and integrality will specifically influence the accompanying works easily going ahead and the last suggestion of trademark administration's quality. Consequently it must utilize logical, sensible and propelled innovation to accumulate different information. At present, towards web use mining innovation, the principle information cause has three sorts: server information, customer information and center information (operator server information and bundle recognizing). The initial phase in the Web use mining process comprises of gathering the significant Web information, which will be examined to give valuable data about the clients' conduct[17][20]. There are two primary wellsprings of information for Web use mining, relating to the two programming frameworks connecting amid a Web session: information on the Web server side and information on the customer side. Likewise, when go-betweens are presented in the client–server correspondence, they can likewise move toward becoming hotspots for utilization information, similar to intermediary servers and bundle sniffers. We will consider each of these sources in the paper. Additionally we are attempting to relate the information gathering strategies with the necessities forced by various classes of personalization capacities.

4.2 Data Pre-processing

The motivation behind information pre-processing is to scrub the grimy/commotion information, separate and consolidation the information from various sources, and afterward change and change over the information into a legitimate arrangement. From the specialized perspective, Web use mining is the utilization of information mining procedures to use logs of substantial information stores. The motivation behind it is to create result that can be utilized to enhance and improve the substance of a site. In this stage, the beginning stage and basic point for effective log mining is information extraction. The following errand after information extractions is information cleaning and information sifting. Since the inception web logs information sources are mixed with unessential data[15][21], information preparing goes about as critical strides to channel and arrange just suitable data before exhibiting to any web mining calculation. A section of Web server log contains the time stamp of a traversal from a source to an objective page, the IP address of the starting host, the sort of demand (GET and POST) and other information. Numerous sections that are viewed as uninteresting for mining were expelled from the information documents. The sifting is an application subordinate. While by and large gets to inserted substance, for example, picture and scripts are sifted through. In any case, before applying information mining calculation, information pre-processing must be performed to change over the crude information into information reflection important for the further preparing.

4.3 Session Based Clustering

The motivation behind information pre-processing is to scrub the grimy/commotion information, separate and consolidation the information from various sources, and afterward change and change over the information into a legitimate arrangement. From the specialized perspective, Web use mining is the utilization of information mining procedures to use logs of substantial information stores [13][14][16]. The motivation behind it is to create result that can be utilized to enhance and improve the substance of a site. In this stage, the beginning stage and basic point for effective log mining is information extraction. The following errands after information extractions are information cleaning and information sifting. Since the inception web logs information sources are mixed with unessential data, information pre preparing goes about as critical strides to channel and arrange just suitable data before exhibiting to any web mining calculation. A section of Web server log contains the time stamp of a traversal from a source to an objective page, the IP address of the starting host, the sort of demand (GET and POST)[19][20] and other information. Numerous sections that are viewed as uninteresting for mining were expelled from the information documents. The sifting is an application subordinate. While by and large gets to inserted substance, for example, picture and scripts are sifted through. In any case, before applying information mining calculation, information pre-processing must be performed to change over the crude information into information reflection important for the further preparing.

4.4 Segmentation

Segmentation algorithms separate information into gatherings, or bunch, of things that have comparable properties. Client division is the act of separating a client base into gatherings of people that are comparative in particular courses significant to promoting, for example, age, sexual orientation, interests, ways of managing money, et cetera. Here in this framework the division has connected for the client perusing behavioral modular. The intrigue and utilization of web information were assembled by the number clients in the region. Cutting and dicing guest information gives more prominent perceivability into their conduct designs. Personalization can be a capable division strategy [23][24]. Many Web website content administration frameworks progressively show content in view of an

approaching guest's character. A guest signs into a Web webpage, for instance, and sees a customized welcoming. This review will be executed and assessed in .Net structure by making an individual site on the system.

All information that should be accessible to the application crosswise over various demands inside a similar session is called session state or session state information. Cases for such information are shopping wicker bin content, delegate consequences of database questions, and additionally approval state data. For putting away such information between solicitations, there are when all is said in done two conceivable outcomes:

•Sending the state data back to the customer. With the following solicitation the present state is transmitted to the server once more.

•Keeping the vital information structures on the server. No session information (with the exception of the referencing identifier) is transmitted to the customer.

Sending the session state information back to the customer is by and large not suggested. From a security viewpoint, the fundamental issue is that the session state can without much of a stretch be controlled on the customer side on the off chance that it is not ensured suitably[15] [16]. The reason for session distinguishing proof is to enable a web application to recognize related approaching solicitations all things considered.

The accompanying instruments for session distinguishing proof are normal:

- Unique identifier in treat.
- Unique identifier as URL parameter.
- Unique identifier in way segment of URL.

Client sessions (distinguished by methods for a treat based convention) are utilized to fabricate "Session Clusters" in the long run prompting a rundown of recommendations. It discovers gatherings of firmly related pages by dividing the chart as per its associated segments. Every segment thus speaks to an alternate class, or group, of clients. The associated segments are gotten in an incremental route by utilizing an induction of the notable Breadth-First Search (BFS) visit restricted to the hubs required in the demand. Fundamentally, we begin from the present page identifier and we investigate the segment to which it has a place. On the off chance that there are any hubs not considered in the visit a formerly associated segment has been part and should be recognized. We just apply the BFS once more, beginning from one of the hubs not went to. Besides, so as to farthest point the quantity of edges of the diagram we connected a threshold. Theerrand of client and session distinguishing proof is discovered the distinctive client sessions from the first web get to log. Client's recognizable proof is, to distinguish who get to site and which pages are gotten to. The objective of session distinguishing proof is to partition the page gets to of every client at once into individual sessions. A session is a progression of pages client peruse in a solitary get to. All the session data's will be assembled by utilizing the session based bunch strategy. Sessionization is the way toward portioning the client action record of every client into sessions, each speaking to a solitary visit to the site. The objective of a sessionization heuristic is to remake, from the snap stream information, the real grouping of activities performed by one client amid one visit to the site.

IV. RESULTS AND DISCUSSION

Recommendation is viewed as information retrieval task. In this Retrieve (recommend) all items which are predicted to be "good". Precision is a measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved [13][14] E.g. the proportion of recommended movies that are actually good.

$$Precision = \frac{tp}{tp + fp} = \frac{|good\ movies\ recommended|}{|\text{all recommendations}|}$$

Recall is a measure of completeness, determines the fraction of relevant items retrieved out of all relevant items E.g. the proportion of all good movies recommended

$$Recall = \frac{tp}{tp + fn} = \frac{|good\ movies\ recommended|}{|all\ good\ movies|}$$

Precision Vs Recall: When recommender system turn to increase Precision, Recall will decrease or vice versa.

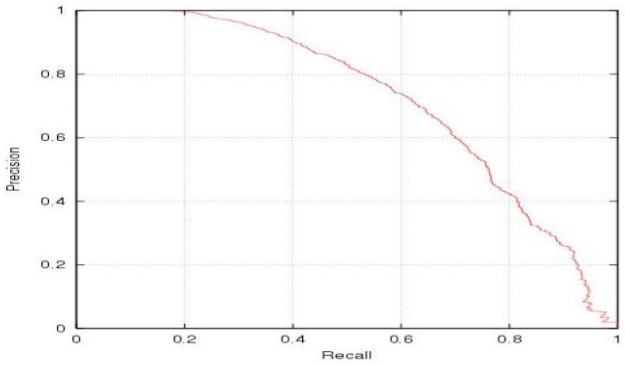

Figure.2

Table 2: classify the possible result of recommendation of an item to a user.

	Recommended	Not Recommended
Preferred	True Positive(tp)	False Negative(fn)
Not Preferred	False positive (fp)	True Negative(tn)

In numerous reasonable proposal applications the creator of the framework wishes to impact the conduct of clients. We are hence keen on measuring the adjustment in client conduct when cooperating with various suggestion frameworks. For instance, if clients of one framework take after the proposals all the more regularly, or if some utility assembled from clients of one framework surpasses utility accumulated from clients of the other framework, then we can reason that one framework is better than the other, all else being equivalent. The genuine impact of the proposal framework relies on upon an assortment of elements, for example, the client's expectation (e.g. how particular their data needs are, how much curiosity versus how much hazard they are looking for), the client's unique situation (e.g. what things they are as of now acquainted with, the amount they confide in the framework)??, and the interface through which the suggestions are displayed. Along these lines, the examination that gives the most grounded proof with regards to the genuine estimation of the framework is an online assessment, where the framework is utilized by genuine clients that perform genuine undertakings. It is most reliable to analyze a couple of frameworks web based, acquiring a positioning of options, as opposed to outright numbers that are harder to decipher. Consequently, numerous genuine frameworks utilize an internet testing framework [9], where different calculations can be analyzed. Normally, such frameworks divert a little rate of the movement to various option suggestion motor, and record the clients associations with the distinctive frameworks. There are a couple of contemplations that must be made when running such tests. For instance, it is imperative to test (divert) clients haphazardly, so that the examinations between choices are reasonable. It is likewise essential to single out the diverse parts of the recommenders. For instance, on the off chance that we think about algorithmic precision, it is imperative to keep the UI settled [8]. Then again, in the event that we wish to concentrate on a superior UI, it is best to keep the basic calculation settled. Now and again, such trials are unsafe. For instance, a test framework that gives superfluous suggestions may demoralize the test clients from utilizing the genuine framework until kingdom come. In this way, the test can negatively affect the framework, which might be inadmissible in business applications. Consequently, it is best to run an online assessment last, after a broad disconnected review gives prove that the competitor methodologies are sensible, and maybe after a client study that measures the client's state of mind towards the framework. This steady procedure diminishes the hazard in bringing about critical client disappointment. Online assessments are extraordinary in that they permit coordinate estimation of general framework objectives, for example, long haul benefit or client maintenance. Accordingly, they can be utilized to see how these general objectives are influenced by framework properties, for example, proposal precision and assorted qualities of suggestions, and to comprehend the tradeoffs between these properties.

TABLE 3 SHOWS TOTAL NO OF INTERESTING WEB PAGE, BOOK MARK AND UNINTERESTING WEB PAGE EVALUATED BY EXAMINEE.

Examinee	Interesting	Bookmark	Uninteresting	All
A	108	62	89	197
B	74	11	125	199
C	41	18	40	81
D	79	75	71	150
E	93	4	65	158
F	42	9	106	148

Table 3: Total No. of the each Web Pages

TABLE 4 SHOWS THE AVERAGE PRECISION IS 82% AND AVERAGE RECALL 63%.

Examinee	Browsing Time		Bookmark		Both	
	Precision	Recall	Precision	Recall	Precision	Recall
A	74%	30%	98%	56%	87%	77%
B	83%	51%	100%	15%	84%	57%
C	80%	29%	100%	44%	90%	63%
D	70%	33%	88%	84%	79%	86%
E	91%	33%	75%	3%	89%	35%
F	92%	57%	100%	21%	92%	57%
AVE	82%	39%	94%	37%	87%	63%

V. CONCLUSION

In any type of experiment it is important that we can be confident that the candidate recommender that we choose will also be a good choice for the yet unseen data the system will be faced with in the future. Several organizations rely on websites to get new customers and to hold the existing one. The customer's surfing pattern can be obtained by web log record. This is a kind off work in the arena of web usage mining. The paper proposed the novel methodology by mining the usage of patters from web log records of real time. Ontology of the web content including user profiles and external data can be developed by Web usage mining. In this paper a method is proposed for discovering and tracking the growing user profiles along with domain specific information facets. As we explain above, we should exercise caution in choosing the data in an offline experiments, and

the subjects in a user study, to best resemble the online application. Still, there is a possibility that the algorithm that performed best on this test set did so because the experiment was fortuitously suitable for that algorithm. To reduce the possibility of such statistical mishaps, we must perform significance testing on the results.

REFERENCES

[1] Soren E. Jespersen, JesperThorhauge, TorbenBachPederson, A Hybrid Approach to Web Usage Mining,Technical Report 02-5002, Department of Computer ScienceAalborg University, July 2002.

[2] Jespersean S.E., Throhauge J., and Bach T., A hybridapproach to Web Usage Mining, Data Warehousing andKnowledge Discovery, (DaWaK'02), LNCS 2454, SpringerVerlag Germany, pp73-82, 2002.

[3] BamshadMobasher, Robert Cooley, Jaideep Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, in Proceedings of the 1999 IEEEKnowledge and Data Engineering Exchange Workshop(KDEX'99), November 1999.

[4] Smith K.A. and Ng A., Web page clustering using a self-organizing mapof user navigation patterns, Decision Support Systems, Volume 35 , Issue 2 (May 2003) Special issue: Web data mining, Pages: 245 – 256.

[5] Chi E.H., Rosien A. and Heer J., Lumber Jack: Intelligent Discovey and Analysis of Web User Traffic Composition. In Proceedings of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM press, 2002.

[6] Joshi K. P., Joshi A., Yesha Y., Krishnapuram, R.,Warehousing and Mining We Logs, Proceedings of the 2nd ACM CIKM Workshop on Web Information and DataManagement, pp. 63-68, 1999.

[7] Robert Cooley, BamshadMobasher, and Jaideep Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web (A Survey Paper) (1997), in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

[8] NeeruMago, "Web Mining: Intelligent way of mining Web based data", Apeejay Journal of Computer Science And Applications, Vol. (3), January, 2015.

[9] Mohinder Singh and NavjotKaur, "A Review on Various Web mining Techniques with Purposed Algorithm of K-means Web Ranking", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013.

[10] Monika Yadav and Mr. Pradeep Mittal, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[11] Ketul B. Patel ,Jignesh A. Chauhan , Jigar D. Patel "Web Mining in E-Commerce: Pattern Discovery, Issues and Applications" International Journal of P2P Network Trends and Technology-Volume1Issue3- 2011.

[12] Dr.S.Vijiyarani and Ms. E. Suganya, "RESEARCH ISSUES IN WEB MINING", International Journal of Computer-Aided Technologies (IJCAx) Vol.2, No.3, July 2015.

[13] Kun Chang Lee and Sangjae Lee, " Interpreting the web-mining results by cognitive map and association rule approach", Information Processing and Management 47 (2011) .

[14] WenlongRen and Jianzhuo Yan, " An Improved CMAC Neural Network Model for Web Mining", 2015 8th International Symposium on Computational Intelligence and Design.

[15] NyomanKarna, IpingSupriana, NurMaulidevi, "Social CRM using Web Mining for Indonesian Academic Institution", 2015 International Conference on Information Technology Systems and Innovation (ICITSI) Bandung – Bali, November 16 – 19, 2015.

[16] RoyaHassanian-esfahani and Mohammad-javadKargar, "A Survey on Web News Retrieval and Mining", 2016 Second International Conference on Web Research (ICWR).

[17] Dr. Sanjay Kumar Dwivedi and BhupeshRawat, "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process", 2015 International Conference on Green Computing and Internet of Things (ICGCIoT).

[18]Khushbu Patel, AnuragPunde, KavitaNamdev, Rudra Gupta and MohitVyas," Detailed study of Web Mining approach- A survey", International Journal of Engineering Science & Research Technology, Patel 4(2) : February 2015

[19] S.chakarbarti ." Data Mining for Hypertext- A tutorial survey", ACM SIGKDD Explorations , 1(2): 1-11-2000.

[20] M.N Garofalakis, R. Rastogi, S. Seshadri and K Shim, " Data Mining and the web: Past, Present and Future", In Workshop of Web Information and Data Management, 1999 pp 43-47, 1999.

[21] Broder, A., R. Kumar, F. Maghoul, P. RaghavanandS. Rajagopalan et al., 2000.Graph structure in the web Computing.

[22] Xing, W. and A. Ghorbani, 2004. Weighted PageRank algorithm. Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May 19-21, IEEE Computer Society, Washington DC., USA., pp: 305-314. DOI: 10.1109/DNSR.2004.1344743.

[23] Web usage mining by B.Mobasher. Page No 449-483.

[24] A.G. Büchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J. G. Hughes, Navigation Pattern Discovery fromInternet Data, in WEBKDD, San Diego, CA 1999.

Author Profile

Mr A M Lee pursed Bachelor of Science and Master of Science from School of Computer Science, University of China, China in year 2009. He is currently pursuing Ph.D. and currently working as Assistant Professor in School of Computer Science, University of China, China since 2012. He is a member of IEEE & IEEE computer society since 2013, a life member of the ISROSET since 2013, ACM since 2011. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 5 years of teaching experience and 4 years of Research Experience.

