ISSN No.-2321-8711

International Journal of Software Engineering and Systems

Volume No. 11

Issue No. 2

May - August 2023

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

ISSN No.-2321-8711

International Journal of Software Engineering and Systems

Aims and Scope

Software Engineering has become very important with the ever-increasing demands of the software development to serve the millions of applications across various disciplines. For large software projects, innovative software development approaches are vital importance. In order to gain higher software standards and efficiency, software process adaptation must be derived from social behavior, planning, strategy, intelligent computing, etc., based on various factors. International journals of software engineering address the state of the art of all aspects of software engineering, highlighting the all tools and techniques for the software development process. The journals aims to facilitate and support research related to software engineering technology and the applications. International journals of software engineering welcomes the original research paper, review papers, experimental investigation , surveys and notes in all areas relating to software engineering and its applications. The following list of sample-topics its by no mean to be understood as restricting contributions to the topics mentioned:

Ø Aspect-oriented software development for secure software

Ø Dependable systems

Ø Experience related to secure software system

Ø Global security system

Ø Maintenance and evolution of security properties

Ø Metrics and measurement of security properties

Ø Process of building secure software

ISSN No.-2321-8711

ISSN No.-2321-8711

International Journal of Software Engineering and Systems

(Volume No. 11, Issue No. 2, May - August 2023)

Contents

ISSN No.-2321-8711

A Software Defined Wireless Network Based K-Way Spectral Clustering Algorithm (SDWN-KSCA)

[1]Nivine Samarji, [2]Muhammed Salamah
Computer Engineering Department, Eastern Mediterranean University, TRNC, Mersin 10, TURKEY
E-mail: [1]nivine.samarji@emu.edu.tr,[2]muhammed.salamah@emu.edu.tr

ABSTRACT

In WSNs too much energy consumption is considered to be critical as it causes network failure, and hence energy saving plays a major network performance metric which in case of absence, can lead to serious network problems resulting in the form of sensors failure, connectivity loss, and disconnected network. This paper introduces the use of a software defined network (SDN) controller in wireless sensor networks and shows how WSNs can benefit from the existence of SDN. Simulation has been conducted for the deployment of SDN controller at the base station, shows approximately 20% improvement in the overall network energy saving compared to the conventional energy saving approach based on network spectral clustering. In addition, our approach shows 25% improvement in the network lifetime.

Index Terms - Software Defined Network (SDN) controller, Network performance metrics, Last node to die (LND), K-way spectral clustering, Energy saving.

I. INTRODUCTION

Wireless sensor networks are typically composed of a number of spatially distributed power-limited sensor nodes with constrained resources that are frequently deployed in hostile environments for measuring or monitoring phenomena of interest. Being deployed in hostile environment, it becomes difficult to recharge or replace sensor nodes. Supplied with limited power, energy efficiency has become crucial for most WSNs, which depletes over time; precisely forwarding process consumes much more energy than sensing process. Gathering information from sensor nodes in WSN environment rely on one or more centralized BS or sink. Hence, Software Defined Network[1],[2] model with the controller being centralized at the sink can easily be adopted for WSNs [3]. Having SDN deployed in WSNs, sensor nodes are freed from routing and topology management, allowing them to become data plane forwarding elements whereas topology control, scheduling, routing, and network coverage and connectivity planning are the main responsibility of the centralized controller [1]. SDN is considered to be an emerging networking paradigm that eliminates the limitations of the traditional networking by providing consistent policy enforcement, scalability, and centralized management & network programmability[1]. Its main characteristic is decoupling the control plane and data plane. Control plane provides network topology, performance, and fault management. Data plane is responsible for packet forwarding based on policies provided by the control plane. Decoupling provides flexibility in updating policies without the need to add additional hardware cost [2], [4].

Forwarding information over long distances consumes more energy than sending the information over short distances, here arises the need for network partition [5] in order to minimize the energy consumption and thus leads to an extend in network lifetime. Our research paper is based on hierarchical clustering scheme namely K-Way Spectral Clustering Algorithm in Wireless Sensor Network (KSCA-WSN)[6], [5] with some essential modifications. This Algorithm takes into consideration the sensor nodes spectral positions in clustering and their residual energy for assigning cluster heads; however, our modifications to the mentioned algorithm will be discussed later. For improving the latter network lifetime and energy saving, we have introduced SDN controller to be deployed at the BS for this purpose. The remaining of this paper is organized as follows: Section II, we briefly highlight on related research work. Section III describes the proposed algorithm. Section IV presents the simulation results and interpretations.

II.RELATED RESEARCH WORK

Different clustering techniques have been found in literature. Previously, the most generally used clustering approach is the LEACH algorithm [7]. LEACH is an energy efficient communication protocol based on hierarchical clustering, where nodes are organized in clusters using distributed algorithm. CH election is based on probability for each period, where each node has an equal probability of becoming a CH, that is , this probability is N/K, where N is total number of nodes and K is the desired number of clusters.

To allocate nodes to CHs, each node will choose a CH based on the received signal strength, the highest is the preferred. CHs then create Time Division Multiple Access (TDMA) schedule for its associated nodes. Yong and Pei [8]presented a modified version of LEACH based on distance between sensor nodes and BS, and on the residual energy of nodes namely distance-energy cluster structure algorithm (DECSA).

This algorithm prevents direct contact of low energy CHs with BS by partitioning the network into three hierarchical levels for the purpose to reduce the energy consumption of Chs.

Elbihri et al. [9] proposed an algorithm to partition the network based on spectral bisection called Spectral Classification Based on Near Optimal Clustering in Wireless Sensor Networks. However their approach is restricted to specific number of clusters 2n in which desired number of clusters can't be achieved. Liu et al. [10] have proposed an energy efficiency communications approach for delay minimizing in internet of things. Their method is based on two aspects, in non-hotspot regions, nodes relay their information without considering their energy consumption since surplus energy exist in these regions, causing the delay to be minimized. On the other hand, nodes in hotspot regions take into account their energy consumption in relaying their information causing an increase in network lifetime. Hence, FFSC algorithm tends to reduce the delay while keeping network lifetime value unchanged.

III. PROPOSED ALGORITHM

Before describing in details our proposed approach, let's have a brief overview on the Ng-Jordan-Weiss(NJW) algorithm [6] being used. NJW algorithm is based on spectral clustering that cluster points using eigenvectors of the Laplacian matrix derived from the data, where clustering can be viewed as partitioning the similarity graph. First, a weighted graph is built where nodes represent data points and edges are the distance between the points. As spectral graph partitioning deals with graph bi-partitioning, one can recursively use this algorithm to have k-partitions graph. In other words, partitioning the graph into k clusters, where points belonging to same cluster are said to be similar and vice is valid too. This illustrates the similarity matrix characteristic of the graph. The main steps of the algorithm [5] is shown in Figure 1 and Figure 2 respectively:

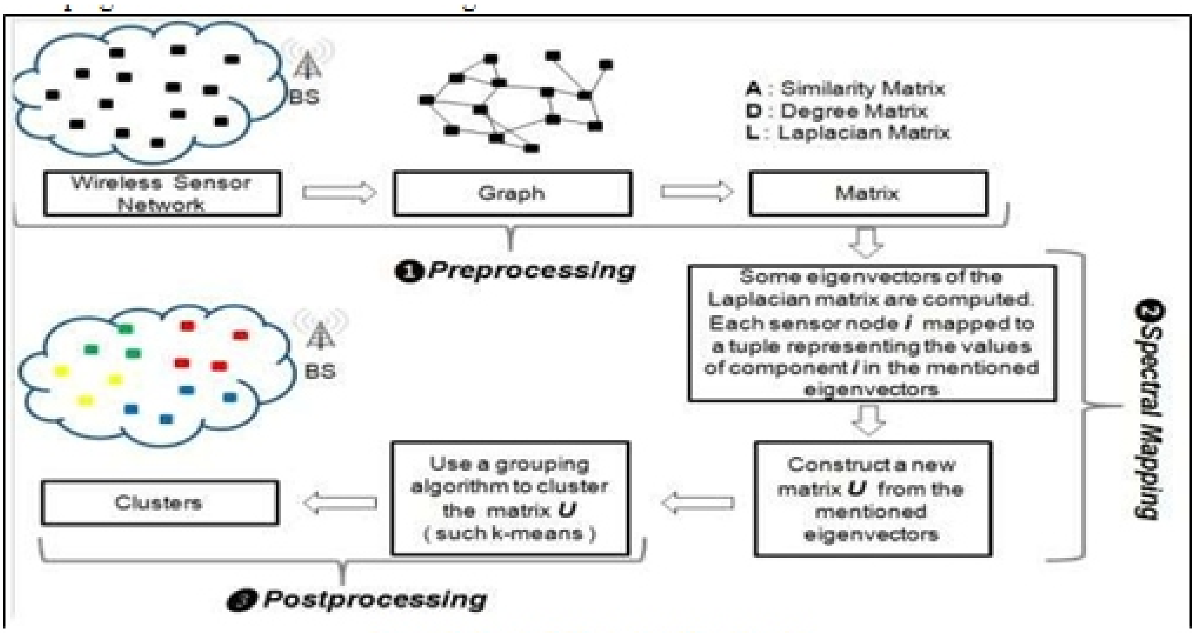

BS
A : Similarity Matrix
D : Degree Matrix
L : Laplacian Matrix
Wireless Sensor Network
Graph
Matrix
❶ Preprocessing
Some eigenvectors of the Laplacian matrix are computed. Each sensor node i mapped to a tuple representing the values of component i in the mentioned eigenvectors
❷ Spectral Mapping
BS
Construct a new matrix U from the mentioned eigenvectors
Use a grouping algorithm to cluster the matrix U (such k-means)
Clusters
❸ Postprocessing

Figure 1: Spectral Clustering Algorithm [5]

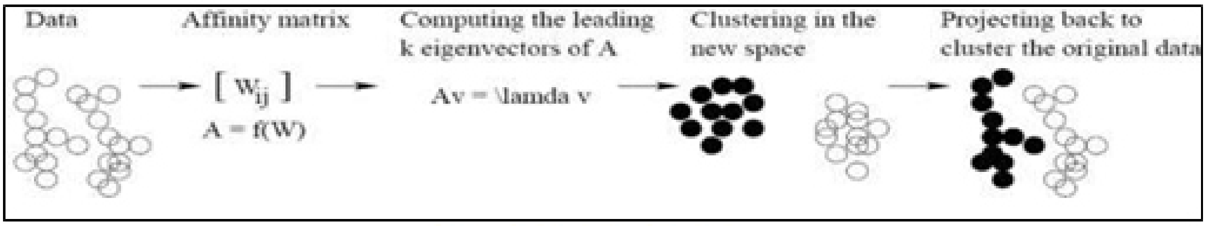

Data
Affinity matrix
[W_ij]
A = f(W)
Computing the leading k eigenvectors of A
Av = \lamda v
Clustering in the new space
Projecting back to cluster the original data

Figure 2: Clustering Illustration

The Pre-Processing step: Find the similarity matrix A which represents the network is given by :

$$\begin{array}{ll} \exp^{-1}{}_{2\delta^2} * d^2\, i, j \text{ if } i \neq j & (1) \\ 0 \quad\quad \text{otherwise} & \end{array}$$

d(Ij) is the Euclidian distance between nodes i and j. Matrix and the Laplacian one will be calculated. Then, the eigenvalues and the eigenvectors of the last matrix will be deduced accordingly.

2) Clustering step: KSCA-WSN Algorithm The optimal number of clusters K is given by[11]:

$$K = \sqrt{N} / \sqrt{2\Pi} * \sqrt{Efs / Ems} * \sqrt{\frac{M}{dtoBS}} \qquad (2)$$

Then constructing a new matrix U from the K eigenvectors related to the K largest eigenvalues of the Laplacian matrix, and iteratively applying K- means clustering algorithm until no changes take place.

3) Cluster head election step: Once the clusters are determined, the next step selects the CHs taking into consideration the node id in the cluster and its residual energy.For each round r of the simulation, the number Ck = (rmodCn) to elect a CH for the appropriate cluster; Cn is the total number of nodes in the cluster k. The node with id = Ck and residual energy Er greater than threshold Ermin will be the CH of the cluster k in the round r. A node is considered dead if its residual energy satisfies equation (3). The minimum residual energy is given by equation (4).

$$Er < \frac{Efs}{Ems} \qquad (3)$$

Efs is free space model's amplifier energy consumption, Ems is multiple attenuation model's amplifier energy consumption.

$$^{4}Ermin = L * (Cn + 1 * E_{elec} + Cn * E_{DA+\epsilon} * d_{ms}^{i} \qquad (4)$$

L is data size in bits to be transmitted to BS at a distance d, and Eelec is the energy consumption per bit in the transmitter, E is energy for data aggregation per bit.

4) Software Defined Wireless Network Based K-Way Spectral Clustering Algorithm (SDWN-KSCA)

An important modification that we did to KSCA- WSN[5] is we intentionally focus on the communication mode of any node to be through its associated cluster head and discard any communication to be taken between any node and the BS even if the distance between the node and base station is less than that to its associated cluster head otherwise this will violate the main functionality of the cluster head. Second, we deploy an SDN controller at the BS to analyze the network performance metrics such as network lifetime and energy saving, in comparison with the KSCA-WSN algorithm in [5]. By deploying SDN controller at the BS, all the routing policies, data flows, calculations will be done by the controller, eliminating the decision to be taken by each CH to choose the next CH according to the residual energy. By doing so, CHs energy will be saved. Third, SDN controller will create for each CH Time Division Multiple Access (TDMA) schedule needed for communication in its respective cluster, adding a benefit for CHs energy saving. Fourth, SDN controller will automatically determine for each node in each cluster its corresponding disjoint neighbor found in other clusters, so that in case of CH failure, the nodes will no longer broadcast their message, despite, each node will send the message to its nearest neighbor found in other clusters. This will ensure reliability by determining alternative paths.

IV. SIMULATION RESULTS & INTERPRETATION

In this section, simulation results were carried out using MATLAB for analysis. Our proposed SDWN-KSCA algorithm is based on Jorio et al. [5]to cluster the network based on spatial positions of nodes and residual energy. We have deployed SDN controller at the BS and executed the algorithm for 2500 rounds according to Table 1. We have analyzed two important network performance metrics: Network lifetime which is defined as the time needed for the first or last node to die in the network. We have carried out results based on first node to die (FND), half of nodes to die (HND), and last node to die (LND) and compared our results with KSCA algorithm. Another performance metric is Energy consumption which is important factor that cause death of nodes. The more energy is consumed, the less possibility for nodes to die.

Table 1: Experimental Parameters

Parameter	Value
E_{elec}	50nJ/bit
E_{fs}	10pJ/bit/m^2
Ems	0.0013pJ/bit/m^4
E_0	0.5J
M	200mx200m
BS coordinates	100m,250m
Message size	4000bytes
Number of Nodes	500

Every time a node sends, receives, or aggregates data, its initial energy decreases. Applying the given parameters to (2), we get 6 disjoint clusters as shown in Figure 3.

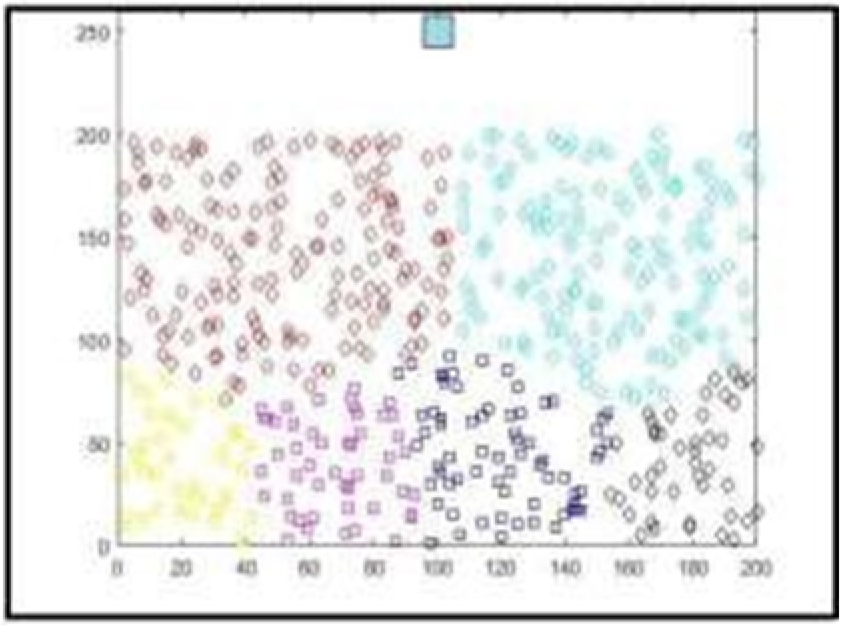


Figure 3: K-Way Spectral Clustering

Figure 4 shows the last node to die in the KSCA algorithm happens at 1302 rounds whereas, Figure 5 shows that LND happens at 1961 rounds for SDWN- KSCA.

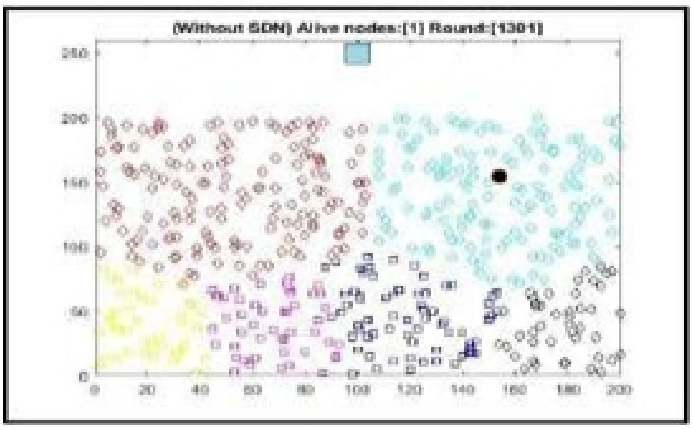


Figure 4: LND-KSCA

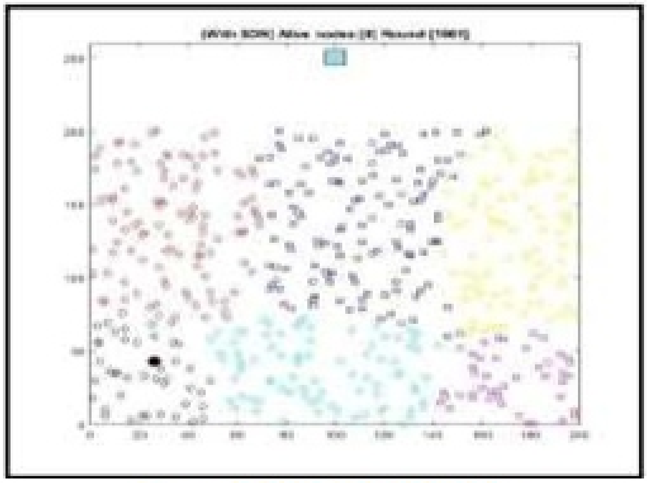


Figure 5: LND-SDWN-KSCA

Figure 6 shows number of alive nodes during each round for both KSCA and SDWN-KSCA.

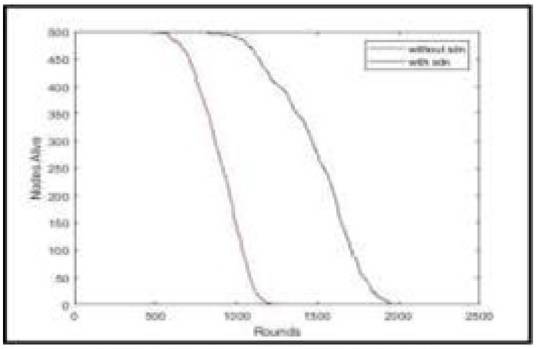


Figure 6: Alive Nodes Comparison

Figure 7 shows FND, HND, and LND respectively for both algorithms with initial energy0.5J.

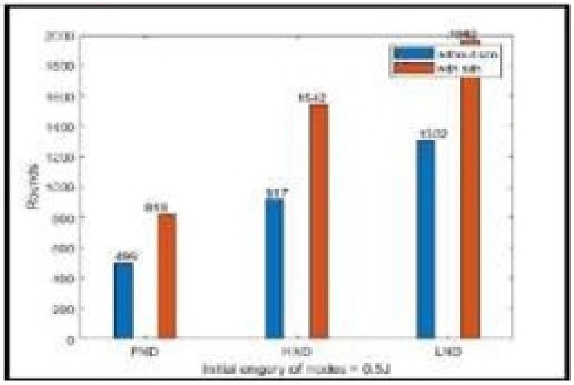


Figure 7: FND-HND-LND

Figure 4, 5, and 6 show how SDWN-KSCA outperforms KSCA in terms of number of alive nodes; approximately 25% improvement has been detected. Figure 8 shows the network remaining energy for every round for both algorithms. Again, SDWN-KSCA outperforms KSCA algorithm in terms of network remaining energy.

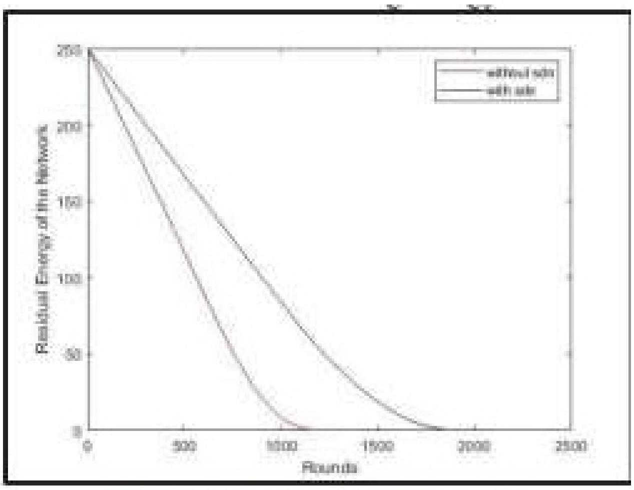


Figure 8: Network remaining energy

We conclude that SDWN-KSCA shows performance improvement in terms of energy and network lifetime.

CONCLUSION AND FUTURE WORK

Network lifetime and Energy saving are critical network performance factors in wireless sensor networks. SDN has proved to be promising solution for various challenges in WSNs. Many researchers have been conducted for WSN performance improvement based on this newly network paradigm. As our results show that using SDN controller at the BS improves the network performance. However, using single SDN controller is risky for being a single point of failure. To this extent, we are planning for future work to use three SDN controllers, to solve multi-objective combinatorial problem, for the same network topology used in this paper, however using heuristic algorithms for controllers' placement to ensure load balancing among controllers, network lifetime saving and minimizing total network delay and energy consumption.

REFERENCES

[1] I. Haque and N. Abu-Ghazaleh, "Wireless Software Defined Networking: A Survey and Taxonomy," vol. 18, no. 4, pp. 2713 - 2737, 19 May 2016.

[2] F. Hu, Q. Hao and K. Bao, "A Survey on Software-Defined Network and OpenFlow: From Concept to Implementation," IEEE Communication Surveys and Tutorials, vol. 16, no. 4, pp. 2181-2206, 2014.

[3] J. A. Puente Fernandez, L. J. Garcia Villalba and T.-H. Kim, "Software Defined Networks in Wireless Sensor Architectures: A Review," Entropy, 2018.

[4] W.-S. Kim and S.-H. Chung, "Proxy SDN Controller for Wireless Networks," Mobile Information Systems, vol. 2016, no. 4, pp. 1-14, 2016.

[5] A. Jorio, S. El Fkihi, B. Elbhiri and D. Aboutajdine, "A New Clustering Algorithm in WSN Based on Spectral Clustering and Residual Energy," in Seventh International Conference on Sensor Technologies and Applications, SENSORCOMM 2013, 2013.

[6] A. Y. Ng, M. I. Jordan and Y. Weiss, "On Spectral Clustering:Analysis and an algorithm," in Advances in Neural Information Processing Systems, 2002.

[7] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy Efficient Communication Protocol for Wireless Microsensor Networks," in The 33rd Annual Hawaii International Conference, 2000.

[8] Z. Yong and Q. Pei, "An Energy-Efficient Clustering Routing Algorithm Based on Distance and Residual Energy for Wireless Sensor Networks," Procedia Engineering, vol. 29, pp. 1882-1888, 2012.

[9] S. Elbhiri, S. EL Fkihi, R. Saadane and D. Aboutajdine, "Clustering in Wireless Sensor Networks Based on Near Optimal Bi-partitions," in 6th EURO-NGI Conference on Next Generation Internet, Paris, 2010.

[10] Y. Liu, A. Liu, Y. Hu, Z. Li, Y.-J. Choi, H. Sekiya and A. J. Li, "FFSC: An Energy Effieicency Communications Approach for Delay Minimizing in Internet of Things," Green Communications and Networking For 5G Wireless, vol. 4, pp. 3775-3793, 2016.

[11] W. B. Heinzelman, A. P. Chandrakasan and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks," IEEE Transactions on Wireless Communications, vol. 1, no. 4, pp. 660-670 , 2002.

[12] "https://towardsdatascience.com/spectral-clustering-for- beginners-d08b7d25b4d8," [Online].

[13] U. Luxburg, "A Tutorial on Spectral Clustering," Statistics and Computing, vol. 17, no. 4, pp. 395-416, 2007.

[14] B. Othmane, M. Ben Mamoun and R. Benaini, "An Overview on SDN Architecture with Multiple Controllers," Journal of Computer Networks and Communications, vol. 2016, no. 2, pp. 1-8, 2016.

[15] A. El-Mougy, M. Ibnkahla and L. Hegazy, "Sotware-Defined Wireless Network Architecture for the Internet-of-Things," in 40th Annual IEEE Conference on Local Computer Networks, Florida, 2015.

Review on Extreme Programming-XP

[1]Khusbhu Sahendrasingh Yadav, [2]Maleeha Arif Yasvi, [3]Shubhika

[1]Khusbhu Sahendrasingh Yadav, Shubhika

[2,3]Department of CSE , IIITD, Delhi, India

E-mail: [1]yadav18087@iiitd.ac.in, [2]maleeha18112@iiitd.ac.in, [3]shubhika18085@iiitd.ac.in

ABSTRACT

Extreme programming is an iterative software development methodology which aims to produce higher quality software and helps in providing an optimal solution. Extreme Programming differs from other software development methodologies as it focuses more on adaptability and responsiveness to the changing customer requirements. By using extreme programming as a software development methodology, better results have been obtained in software development.

Keywords - Extreme Programming, Testing, Pair Programming, Refactoring, Small Releases

I. INTRODUCTION

Extreme Programming is a software development model which primarily focuses on the software quality improvements and responsiveness to the changing customer requirements. It is a type of agile software development model. It advocates frequent releases in short development cycles. These releases are intended to improve the productivity and enhance the quality of the software by introducing certain measures that focus on initiating certain checkpoints at which the customer requirements can be adopted and met.

Extreme Programming focuses on light-weight processes. The main phases involved in the cycle of XP are- Planning, Design, Coding, Testing [2]. As it is an iterative model, the system is developed by dividing the overall project into small functions. The cycle of development from the design to the test phase is performed for one function. After executing for one function and debugging it properly, the developers then shift to the next [2].

XP is based on rapid release cycles and continuous communication between the developers and the stakeholders; that is, the customers [4]. It strongly relies on oral communication, frequent testing, code review and designing [5]. Communication is regarded as an important criteria in XP. The communication should frequently happen among the concerned parties like the developers, customers, and managers [6].

II.BACKGROUND

XP was created by Kent Beck during his work on Chrysler Comprehensive Compensation system (C3) payroll project. Kent Beck became C3 project leader in March 1996 and refined the development methodology used in the project and wrote a book on the Extreme Programming methodology in 1999.

XP is regarded to be better than the traditional waterfall model. It is increasingly adopted by companies worldwide for software development. Since in the waterfall model, the development process is completed on a single project and the requirements have to be known beforehand, so there is no scope of changing the requirements once the project development starts. So the waterfall model is not flexible to the changing requirement needs. On the other hand, XP focuses on iterations and allows for changing requirements even after the initial planning has been completed [2]. Also, the waterfall model does not require the participation of customers while as XP focuses on customer participation and satisfaction and regards it as a primary goal during software development process. So, XP is a preferred model in the software development these days.

XP is a framework of the agile software development model, and it primarily aims at producing higher quality software. Agile software development advocates adaptive planning, evolutionary development, continuous improvement and encourages rapid and flexible response to changes. It focuses on customer satisfaction, simplicity and continuous attention from the developers as well as the customers. XP is the framework of the agile software development and hence all these practices are implemented in XP.

III. TRADITIONALVS XP

The Waterfall model, often termed as a linear sequential development model has been used widely in the last few decades whereas extreme programming (XP) has shown its popularity in the recent past. There was a need to analyze both these methodologies to choose one that serves as the best fit for any software development process. The factors upon which decision lies are time constraint and risk mitigation and adaptability to the requirement changes during any phase of software development. The Waterfall methodology comprises of five phases including requirement gathering and specifications, design phase, implementation phase testing and deployment phase. As stated in [19] and [20], waterfall methodology spent most of its time in documentation i.e. during the first phase, and implementation phase also shows a significant amount of time consumption whereas extreme programming methodology devotes maximum time in the testing phase.

Extreme Programming (XP) being iterative in nature, proposes short releases periodically. It follows an automated testing strategy that prevents risk propagation in the later stages. To avoid any kind of risk, XP performs continuous checks in each phase and resolves it in that particular stage if found any. In the case of waterfall methodology, if the requirements are misunderstood in the initial phase then it may lead to the development of a faulty system as it lacks communication process throughout the development. As waterfall tries to find errors in the later phase, risk propagation in the upcoming phases increases which eventually end up costing high amount to fix the risks.

Waterfall model is sequential. Hence there is no way one could backtrack to the previous stage. Hence it causes a problem in making corrections. It is not possible to adapt to the changing business requirements in the later stages. In Extreme Programming, several checkpoints are being introduced which could be used for adapting the new customer requirements. However, adapting frequent requirement changes late in the development process may cause a delay in the completion of the project. It may also result in increasing the complexity of the project.

XP model helps to deliver fast with minimum risk exposure and is well suited for small sized software project whereas waterfall serves as the foundation for a large scale software project.

IV. KEYS TO SUCCESS IN EXTREME PROGRAMMING

- **Management Backing:** Management encourages the team to try different scenario related to the size of teams, duration of iterations.

- **Coach Role:** Technical lead and Project Lead on the project, guide and push the teams to follow the process to which team has committed.

- **Simple Coding:** XP works best when the simplest solution is implemented for the current iteration and refactoring is done after each iteration when stories require it.

- **Adaptation:** If the process is the same as one year before, then the process is not agile as it was a year ago. So, adaptation is an important step in the success of XP.

- **Team Coordination:** The coordination between the members of a team plays a vital role in the success of an XP project.

- **Tester Roles:** Testers are a critical part of the success of an XP project, while their role may look different on an XP team, make them part of the team.

- **Test First:** Testing saves developers from coding the same model again and again. Testing after almost every iteration shows the flaws in the system and hence provide a means to achieve good quality software.

A healthy XP project can run for a long time, but you need to change and adapt along the way. Change is good and necessary, including changing the process along way.14

V. PRACTICES

XP has 12 basic practices which are always implemented. These practices ensure better code readability and understandability and also help in enhancing communication between the developer and the customers so that a better product is made.

The 12 practices of XP are listed as:-

1. Planning Game
2. Small Release
3. Metaphor
4. Testing
5. Refactoring
6. Pair Programming
7. Collective Ownership
8. 40-hour week
9. On-site customer
10. Coding Standards

Planning Game: This involves the developmental strategy of releasing plans for the project and the meetings between the developers and the customers. The strategies to improve communication among the stakeholders are also focused upon in this practice of XP. The system releases plans and dates for the meetings and the project reviews [1]. General discussions to know about the progress of work are held through XP. Discussions among the stakeholders ensure that the queries are resolved and both the developer as well as the customer has a proper understanding of the system and the project flow [3].

Small Release:- The entire project is distributed into functions, and after completion of each function, the development team releases its version to the customers. The release cycles in XP are shortened to speed up the feedback from the customer. It is done to handle the risks and errors according to the release of each version [3]. It ensures better accountability and efficiency. After the release of each function, it is integrated with the previously released functions. So, continuous integration is done [5].

Metaphor:- It describes how the program looks. It is a document that describes the working of the system and expresses the evolving project vision that would define the system scope and purpose. The metaphors are directly derived from the principle and standards of the project architecture and requirements [4].

Testing:- XP focuses on frequent testing. Testing is done to ensure that the code is free from errors. Unit testing is performed in which every piece of code that is written of a particular function is tested before moving on to the next function. Apart from the testing involving rectification of the coding errors, acceptance testing is also done. Acceptance test verifies the requirements as understood by the developer and whether it meets the requirements of the user or not. Integration testing is also done which helps in testing the system when different functions of the system are integrated into one unit [1]. The testers and developers work together to find out faults. Therefore, in XP the tests are initially run on small codes and then the whole system [3]. In some of the software applications, J-Unit is used for testing. It allows testing localization classes, entire package or even entire project [8].

Refactoring:- It is a way by which the code is kept in an easy to understand form. It ensures the removal of duplicate code and makes the code easy to understand. The use of complex coding schemes is not appreciated in XP and stress is given on forming an easy code [3]. Long methods, use of unnecessary classes are avoided [5]. In this phase, restructuring is implemented. It is the change made to the internal structure of the software to make it easier to understand. Refactoring helps to modify the structure without changing its observable behavior [1].

Pair Programming:- It is the concept of having two programmers work together for a particular function code. One programmer writes the code while as the other is the observer who reviews the code written by the programmer [1]. It ensures better efficiency and helps in determining more alternatives[5].

As Extreme Programming focuses on pair programming for implementing the user stories, it gives rise to observe the personality traits which will impact the development process. It is being observed that the efficiency of pair programming depends on whether both the persons are working remotely or at the same location. It has been checked by a well- known personality test model namely Myer-Briggs Indicator which takes into account following traits: extraversion / introversion (how people get influenced by surrounding), sensing/intuition, thinking/feeling (how an individual takes a decision) and judging/perceiving (organizing plan of developers). The final results stated in [23] showed that overall efficiency reduces if two different personalities are working together on a specific module at a foreign location. On the other hand programmers with similar traits tend to understand and communicate easily regardless of the location thereby leading to enhance the efficiency.

Collective Ownership:- It is a practice in XP which ensures that all the team members should be familiar with the project code. This practice enables any programmer to change the code in the system at any time. It works like open-source programming [4].

40-Hour Work:- It is the practice in XP which ensures that the team members in a project should work the hours that they can sustain quality [1]. The important thing is to recognize the time agreeable by all the team members for the number of hours that have to put in for a week. After every week, the work done will be reviewed [3].

On-Site Customer:- To ensure that the developers have a proper understanding of the desired project and all the requirements are being met, a single customer from the customer team is always available to answer all the questions of the developers, resolve disputes and set small scale priorities all the time [2].

Coding Standards:- For ensuring collective ownership so that any programmer can change the code in the system at any time, the practice of collective ownership is implemented [6]. XP mandates the use of coding standards. The programmers must follow a common coding standard so that all the code appears to have been written by one person [4].

VI. LIFE-CYCLE OF XP

The iterative life-cycle followed in XP goes through the analysis, planning, designing, implementation and the delivery phase [1]. All these phases are executed following the practices of XP. Initially, stories are created that give the estimate, release plan, iteration duration, etc. In the analysis phase, the stories are analyzed by the developer to see if the implementation can be carried through them or not. The output of the analysis phase is the feasibility report. For forming the stories, brainstorming sessions are encouraged among the stakeholders [2]. Value graphs which help to answer questions such as "why", „how‟ are also followed. Value graph is a tool used for finding value and requirement functions [2]. Followed by the analysis phase is the planning phase in which the strategy is planned. It is done according to the planning game practice of XP. Failed stories will be analyzed in this phase and new stories will be created to overcome the failed stories.

After the analysis phase, designing and implementation phase is executed which ensures extensive testing. Pair programming is followed while coding.

The last stage of the life-cycle is the delivery stage. This phase includes the activities like the installation of the software, training of the customer to accustom with the enhancements or changes [1].

VII. APPLICATIONS OF XP

XP practices are widely used in the software development process.

XP practices are used in the development of web- based applications [7]. This is done to ensure software organization responsiveness while decreasing the developmental overhead [7].

By ensuring frequent meetings between the customer and the developer, XP has also been helpful in global software development [6]. When customers and developers are not co-related, then the feedback between them is not timely enough which may become an obstacle. So, XP practices when adopted result in a shortcut communication channel [6].

XP implements reverse engineering practices while developing the software model [5]. The developers comprehend the source code of the old system and deliver its features to the customer. The customer validates the specification set and the developers again revise the features and start building the system according to the XP practices [6]. In the designing of some software models, there are separate developers for reverse engineering phase and for implementation that is the forward engineering phase, separate developers are present. Reverse engineering group shares documents and specifications with the forward engineering group and consultation between the two groups happen [6].

XP practices have also been implemented in university projects among the students [8]. In some of the projects, JCVS (Java Concurrent Version System) tool is used for providing a standard for coding and providing collective ownership. The JCVS tool leads to a more consistent and relaxed method of coding. By containing all code in one server and accessing the server for the most recent version, all programmers can get to know the recent release.

For maintaining the documents of the project system and extracting stories, automatic generation tools are used which record the insights during the analysis phase and then the grouping of similar ideas is done to derive a proper description of the system [2].

The customers have ample knowledge regarding product requirements and background. The bidirectional communication between customers and the development team is essential as it allows developers to gain adequate insight regarding the requirements of the project so that they can develop a high-quality product. Hence XP communication is divided as internal communication among the project members, communication between the customers and communication between developers and customers been discussed in [22].

The XP bidirectional communication module involves formal (includes conferences) as well as informal communication (which includes team discussion, customer discussion). But sometimes customers may specify requirements in abrupt manner i.e, there could be a case when they specify the requirements in a fuzzy manner leading to requirement deficient condition or sometimes may over expect from the developers resulting in providing excessive requirements which in real sense would not be possible to satisfy in the given time frame of the project. This problem could be solved by introducing the two dimensional characteristic model in XP that analyzes the satisfactory situation of customers leading to the developments of a high- quality product with the best features. It can be achieved by comparing the customer‟s requirements and developer needs. The customer requirements could further be classified as Expectation, No Opinion, Bear and No Way, whereas the developer needs could be classified as Expectation, Possible, Reluctantly and Refuse. Both the sides then reach a consensus to fix the requirements that need develop and thereby eliminating the undesirable needs.

The demand modules of XP serves as a base to improve the product quality and increasing the customer awareness of the project along with reducing the risk factor involved in the project development. The demand module comprises of concept stage, prospect stage, requirement stage high- quality requirement stage and tool stage. The concept stage performs preliminary analysis to identify the main objective of the project eliminating the problem of deficient requirement. In the prospect stage simple planning is done using all the key records to get the characteristics of the product thereby designing the work process and managing the flood requirements. The requirement stage focuses on combining the use cases modeled by developers and stories being prepared by customers and analyze the data to avoid frequent changes. In the high requirement stage where both customers requirement and developer need to meet a consensus followed by tool stage, where defects are being tracked and version control is looked upon. This leads to enforce a better insight for the developer to understand the requirements and functionality clearly.

The efficiency of the various software methodologies could be enhanced by incorporating XP principles within their life cycle. One of them includes the Rational Unified Process (RUP) which is a process oriented development methodology that focuses more over on the process control and spending a significant amount of time to meet the process requirements. In contrast to XP, it does not pay much attention to the developing program and improving its agility. Hence RUP could be improvised by adopting agility feature of XP as mentioned in [24] and XP could incorporate certain RUP principles to strengthen the theoretical foundation.

The requirement changes often lead to a considerable amount of rework in the software development process. Extreme programming helps to reduce this rework by forecasting the changes based on business and technical perspective using enhancing story card and adjusting the planning game i.e. whenever the customer presents a story, the developers analyze the stories and break them up into various features and organize them into components. These components are then assigned risk, cost, schedule and likeliness of getting changed discussed in [25]. Thereby components with minimum chances of getting changed were implemented first, except few times where there exists dependency among the components and therefore those components are developed sequentially.

Security is a matter of concern while developing a web-based application as these applications are highly prone to vulnerabilities. The most popular threats include Cross-Site Scripting and Structured Query Language Injection as discussed in [21]. It becomes important to incorporate certain security measures while adopting any software development methodology. Extreme programming could be improved by tightening security controls all over the development stages.

The improved model thereby involves both the development team and the business representatives in the early phase that helps to identify security threats and to work over it at the early stages in the development lifecycle. This way they define security requirement and acceptance testing requirement within the company policies. Misuse cases along with use cases are introduced in the model that is being carried out from the requirement and design stage as it gives a proper insight regarding how will the system be exposed to threats. It also emphasizes on following secure coding standards that enforce secure naming convention, remarks, and security infrastructure. Pair programming also helps to review code by double checking the areas that are prone to vulnerabilities. There is an iterative risk assessment during the entire life cycle that keeps track of the security issues as few of the threats may go undetected in the early phases.

Various other applications of XP are as follows:

A. Extreme Programming Development through Dialog

The two primary roles in XP are customer and programmer. The customer is responsible for identifying the features (known as stories) that the programmer must implement, providing detailed acceptance tests for those stories and assigning priority to them. On the other hand, programmers are responsible for estimating how long it will take to implement those stories [10].

With this division of responsibility in place, a project plan represents a dialog between the customer and the programmers. The protocol of that dialog is also clear. The customer tells the programmers what stories he or she wants in the next iteration. The programmers add up the estimates for those stories and tell the customer whether they are possible [10]. The customer can remove or swap stories but cannot get more in the iteration than the programmers estimate is possible. The programmers can tell the customer how long something will take, but cannot select stories to be implemented.

B. Extreme Usability (XU) in Software Engineering Education

In software engineering, usability is a degree to which software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use. Extreme Usability (XU) is such that all best practices of Usability Engineering (UE) are kept in XP during planning games, with a restriction of the usability aspects in next iteration and the equal treatment of Usability and Functionality [11].

A practice of XP, which is often difficult to achieve in a realistic setting, is the customer-on-site because of the heavy time-restraints this poses on the customer [10]. In XU, this difficulty could be transformed into an advantage by allowing different customers to take part in different iterations, if not releases, thus solving two problems at once:

- From the point of view of standard XP, the requirement that the same customer has to be present all the times can be relaxed, thus possibly achieving a better overall coverage of customer time in the team.
- From the point of view of Usability Engineering, the usability of the real system can be tested all times on several different real end users, one at a time, but at any stage in greater depth with the possibility of redesigning the user interaction with the state of the system, for a cost that can be accurately specified.

The combination of Extreme Programming (XP) and Usability Engineering (UE) which leads to a new method: Extreme Usability (XU), is very promising, especially for Software Engineering education [11].

C. XP Labs Design in Universities

Planning Game, which is one of the 12 practices in XP help students to learn how to divide requirements into User Stories and how to prioritize and estimate the costs of those stories [6]. A preparation class conveys the principles of agile methods and XP practices to the students. Afterward, a multi-day block course is conducted, which is the main part of the lab. The students have to develop an application for a customer using the XP practices. The on-site customer is played by a tutor. Another tutor performs the coach role [12].

The best results were achieved when the project was for 9 block days and each iteration was of 2 days. Also, the team size played a vital role, so a team of 8 students was best.12 It is difficult to efficientlyen force test-first development when supervising a larger group of students. So, a small group shows a better result than the large one.

By tweaking the properties of the XP lab, we could improve the overall learning outcome over the years. The XP lab has been a very successful class in the computer science curriculum so far.

D. Industrial Extreme Programming implementation (IXP) in Rational Unified Process (RUP) on Agile Development theme

Rational Unified Process (RUP) is a case-driven and architecture-centric based IBM Theory. The main purpose of RUP is to ensure that the resulting software has high-quality as the user‟s need. It also ensures that software is produced within time and cost specified. RUP life cycle is divided into 4 different phases: Inception, Collaboration, Construction, and Transition. In each phase there are 9 processes, they are : Business Modeling, Requirements, Analysis and Design, Implementation, Test, Deployment, Configuration and Change Management, Project Management, Environment.

Industrial Extreme Programming (IXP) is a development of the failed Extreme Programming (XP) which is intended for large scale software development [23].

Implementation of Industrial Extreme Programming (IXP) in Rational Unified Process (RUP) is done for every process of RUP. Additions or modification are done on the basis of the workflow of RUP practices offered by IXP. IXP supported by RUP is a better scope for software development which can be used for large scale software development providing structured and applicable methods [12].

VIII. QUALITY CONTROL IN XP

The biggest challenge in XP is maintaining quality control over various iterations. Reworking on the code in different iterations can lead to an unmaintainable mess. But XP has different means to have control over quality. First, XP demands simplicity from the programmers- they must leave the code in the simplest possible state that passes all the acceptance test. Thus, when the code is reworked from iteration to iteration, it is also continually reduced to the simplest state programmer can find. Second, programmers are not allowed to work on code alone, the programmers team up in pairs. Finally, before adding code to the system, the programmers must write a failing unit test that the new code must make successful. This ensures that as the program grows, a copious suite of tests grows with it. These tests keep the quality of the software high and give the programmers the courage they need to continually rework the code into its simplest form—an operation known as refactoring.[10.]

IX. ADVANTAGES OF XP

- The greatest advantage of XP is that it allows software development companies to save their cost and time required for project realization. XP eliminates unproductive activities to reduce the cost and frustration of everyone involved. XP allows developers to focus on coding.
- The Simplicity of code. With different iterations over time, refactoring of code is done i.e, developers are restricted to write code in the simplest form and after every iteration, the code becomes more simple.
- XP reduces the risks related to programming and project failure. The customer gets what he or she wants at the end.
- Constant feedback after every iteration from customers helps programmers to proceed in the right direction.
- XP helps in increasing employee satisfaction and retention. The breakdown of the project into subcomponents and constant feedback helps employees in completing the project within deadline without overtime.
- This approach creates working software faster. Regular testing at the development stage ensures detection of all bugs, and the use of customer-approved validation tests to determine the successful completion of a coding block ensures implementation of only what the customer wants and nothing more.

X. XP AND DATA WAREHOUSE

Some of the practices of XP can be used in the data warehouse to enhance the efficiency of the data warehouse systems.

The data warehouse focuses on the subject of the decision to be taken. By subject, we mean that the information should be completely present for a particular problem. The XP practice of planning game and continuous interaction and feedback between the customer and developer can be a helpful practice to implement in the data warehouse. By interacting among the stakeholders and specifying the problem statement clearly, the data warehouse can be more helpful in providing clear information and hence, support in managing the decisions.

XI. CONCLUSION

Extreme Programming is hence categorized by short iterative cycles, incremental planning, evolutionary design and its ability to response to the changing business needs. It strongly practices oral communication among stakeholders. It encourages unit testing of each functional unit before the integration has to be done. It encourages the programmers to follow a coding standard and work in pair while coding. This is mainly done to achieve better results. XP is a software development model which emphasizes on better building and understanding of the system.

Extreme programming does not follow the traditional approach of maintaining voluminous documents for requirement specification rather it follows to maintain source code well documented. But in reality, the documentation plays a vital role as the complete product knowledge resides in the software requirement specification rather than merely on a development team"s intelligence.

To cope up with this problem, XP sets up an oral communication between the development and maintenance team immediately after the completion of the project as stated in [26]. It also requires skilled programmers to incorporate frequent changes in the project and to ensure that the simple design is maintained. XP is not well suited for mission-critical or safety-critical applications as large scale projects would not find it possible to arrange stand up meetings on a regular basis and having an oral handoff seems to be difficult.

Although extreme programming is a good software development practice, yet there are some of the problems that can be faced by the developers. The timing issues can be faced among the programmers who are involved in pair programming [8]. When standard data structures are used, then there is no such requirement of having two people involved for the same portion of code. Due to the practice of pair programming, the management might have to face the cost issues and expenses of two instead of one for a particular work. Even with some of the drawbacks that can be present in XP, XP still outperforms the traditional models of software development. It provides better efficiency and better system and software understanding.

REFERENCES

[1] J.Choudhari and Dr. U.Suman, Iterative Maintenance Life cycle using Extreme Programming, in International Conference on Advances in Recent Technologies in Communication and Computing in 2010.

[2] T. Goto, K.Tsuchida and T.Nishino, EPISODE:An iterative programming method for innovative software based on system designs in 3rd International Conference on Advanced Applied Informatics in 2014.

[3] B.Xu, Towards high quality software development with extreme programming methodolgy:Practices from real software projects.

[4] A.English, Extreme Programming: It"s worth a look.

[5] A.V Deurson, Program Comprehension risks and oppurtunities in extreme programming.

[6] Yang Xiohu, X.Bin,H.Zhijun, Extreme Programming in global software development.

[7] F.Maurer and S.Martel, Extreme Programming : A rapid development for web-based applications.

[8] J.Kivi, D.Haydon, J.Hayes, R.Schmeider, G.Succi, Extreme Programming: A university team design experience.

[9] S.Alsheri and L.Benedicenti, Prioritising CRC Cards as a simple tool in extreme programming in IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) in 2013.

[10]1Andreas Holzinger, 1Maximilian Errath, 1Gig Searle, Bettina Thurnher2, Wolfgang Slany,1Institute for Medical Informatics, Statistics & Documentation, Medical University Graz, 2 Institute of Software Technology and Interactive Systems, Vienna University of Technology, 3 Institute for Software Technology, Graz University of Technology, From Extreme Programming and Usability Engineering to Extreme Usability in Software Engineering Education (XP+UEoXU).

[11]Putu Edy Suardiyana Putra, Arlisa Yuliawati, Petrus Mursanto, Industrial Extreme Programming Practice"s Implementation in Rational Unified Process on Agile Development Theme, ICACSIS, 2012

[12]Yang Yong, Bosheng Zhou, Evaluating Extreme Programming Effect through System Dynamics Modelling, Beihang University, Beijing

[13] Brian Spears, The Bold New Extreme Programming-Now in its Ninth Year, Agile Conference, 2009

[14] Angela Martin, James Noble, Robert Biddle, Programmers are from Mars, Customers are from Venus: A practical guide for customers on XP projects.

[15]Sallyan Bryant, Double Trouble: Mixing qualitative and quantitative methods in the study of eXtreme Programming

[16]Kai Stapel, Daniel Lubke, Eric Knauss, Best Practices in extreme programming Course Design, Leibniz Universitat Hannover

[17]James Newkirk, Introduction to Agile Processes and Extreme Programming, Chicago, USA

[18]Pooja Sharma, Nitasha Hasteer, "Analysis of Linear Sequential and Extreme Programming Development Methodology for a Gaming Application", in international Conference on Communication and Signal Processing, 2016.

[19]Feng Ji, Todd Sedano, "Comparing Extreme Programming and Waterfall Project Results",

[20]Bala Musa S., Norita Md Norwawi,Mohd Hasan Selamat, Khaironi Yetim Sharif, "Improved Extreme Programming Methodology with Inbuilt Security", in IEEE Symposium on Computers & Informatics, 2011

[21]Zhai Li-li., Hong Lian-feng, Sun Qin-ying,"Research on Requirement for High-quality Model of Extreme Programming", in International Conference on Information Management, Innovation Management and Industrial Engineering, 2011

[22]Ramlall Poonam, Chuttur M. Yasser,"An Experimental Study to Investigate Personality Traits on Pair Programming Efficiency in Extreme Programming", in 5th International Conference on Industrial Engineering and Applications, 2018

[23]Xiaobo Wu, Chang Ge, "The Research on Necessity and Plan for Using Extreme Programming in Rational Unified Process"

[24]Xu Bin, Yang Xiaohu, He Zhijun, Srinivasa R. Maddineni, "Extreme Programming in reducing the rework of requirement change"

[25]Colin J. Neill, "The Extreme Programming Bandwagon: Revolution or Just Revolting? ", in IEEE Computer Society, 2003

[26]Sara Shahzad, "Learning From Experience: The Analysis of an Extreme Programming Process", in Sixth International Conference on Information Technology: New Generations, 2009

[27]Zahid Hussain, Martin Lechner, Harald Milchrahm, Sara Shahzad, Wolfgang Slany, Martin Umgeher, Thomas Vlk, "Optimizing Extreme Programming", in Proceedings of the International Conference on Computer and Communication Engineering, 2008

[28]H. Kiwan, Y. L. Morgan, Luigi Benedicenti,"Two mathematical modeling approaches for extreme programming",in 26th IEEE Canadian Conference Of Electrical And Computer Engineering (CCECE),2013

[29]Elmuntasir Abdullah, El-Tigani B. Abdelsatir,"Extreme Programming Applied in a Large-scale Distributed System", in International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), 2013

Comparative Analysis of Parameter Estimation Techniques Used in Software Reliability Growth Models

[1]Poonam Panwar, [2]Arvind Kumar Lal
[1]Department of Computer Science and Engineering,
Ambala College of Enginnering and Applied Research, Ambala, INDIA
[2]School of Mathematics, Thapar University, Patiala, INDIA
E-mail: 1rana.poonam1@gmail.com, 2aklal@thapar.edu

ABSTRACT

Software Reliability Growth Models (SRGMs) have been used by engineers and managers for tracking and managing the reliability of software. The main objective is to achieve the required standard of quality before the software is released to the customer. A number of SRGMs have been proposed in literature to estimate the reliability and quality of software. To apply a given reliability model, defect inflow data is fitted to the selected model equations. All these SRGMs model equations have some unknown coefficients to be estimated. The estimated value of these unknown coefficients can greatly affect the predictions made by these software reliability growth models. There exists two of the widely known and recommended techniques for parameter estimation, maximum likelihood estimation (MLE) and least squared estimation (LSE). Any of these techniques can be used by the software engineers to estimate the values of unknown coefficients. But the estimated values using these approaches differ in most of the cases. So, in this paper a comparative study is performed between these two estimation techniques for their usability and applicability in context of SRGMs. We have also validated the study by predicting the number of failures using different datasets.

Keywords - Software Reliability Growth Models (Srgms), Maximum Likelihood Estimation (MLE), Least Square Estimation (LSE).

I. INTRODUCTION

Software is playing an ever increasing role in our day today life. Most of the products and services we consume are now based on software or uses software in certain ways [1]. Over the years the complexity of software artifacts has been growing rapidly, while at the same time the demands for dependability of software systems have also increased. The link between complexity and software faults has been suggested for long, studies as early as 1980s. Ken et.al suggested that software complexity often affects its reliability. Thus while it is important to keep the complexity of software under check, it is also important to tack and monitor their reliability growth [2]. Software testing is still the main source of ensuring reliability and quality of software systems. Testing in the area of software products is highly resource intensive exercise some of the estimates put it around 50% of overall development cost [3]. But testing resource consumptions can be much more resource/cost efficient, if project managers are able to plan testing activities well [4].

Software reliability growth models have been used to estimate the reliability change in software products and use the reliability growth predictions for making testing resource allocation decisions. Since the software can rarely be made fully error free, project managers need to balance costs associated with software testing to cost of fixing bugs after release [5]. Software reliability can be modeled using reliability models which can be based on Non-Homogeneous Poisson Process (NHPP), Markov process or Bayesian models. One of the major difficulties faced when using Markov and NHPP models is with their parameter estimation [6]. A number of difficulties that may be encountered when applying SRGMs to defect data. In this paper we explore practical considerations when using two types of Techniques i.e. Least Square Estimation and Maximum Likelihood Estimation. We have compared between these two techniques and analyses how a method differs from other to predict the failures for future.

In the next section we described the various SRGMs and their related work, section 3 outlines the proposed research methodology for comparison of selected estimation techniques, in section 4 we have listed he results and the results are analyzed an conclusions are drawn in section 5.

II.SOFTWARE RELIABILITY GROWTH MODEL (SRGMS)

Software reliability engineering tends to focus on using engineering techniques for assessing and improving the reliability of software systems during development and post development [9]. Application of empirical reliability engineering techniques have led to two basic categories, the first class of models called software reliability models (SRMs) are static models that uses attributes of software source code to assess or predict its reliability, while the software reliability growth models (called SRGMs) or the dynamic models generally uses statistical distributions of the defect inflow patterns to estimate/predict the end-product reliability [10]. The SRMs and SRGMs could also be differentiated based on their access to source code which former being a white box models while the latter being black box modeling of software reliability.

Over the past 30 years, many SRGMs have been proposed for estimating reliability growth of products during the software development process. Each model could be shown to work well with a unique data set but no model appeared to do well on all data sets. Many researchers like Musa have shown that some families of models have, in general, certain characteristics that are considered better than others. For example, the geometric family of models tends to have better predictive quality than other models. These and other attempts to compare different models, by Schick & Wolverton and Sukert have led to an evolution from proposing a new model to proposing techniques for finding the best model for each individual application from among the existing models. Ideally, we would like to select, before starting, which model we should use. This ideal has proven to be a very difficult, almost impossible task. Brocklehurst et. al [11] suggests that it is the very nature of software failures that has made the model selection process in general a difficult task. They observed that software failures are caused by hidden design flaws, and not by the psychological sciences that will someday show us how to select the model beforehand.

Goel & Okumoto [12] published a paper describing a non-homogeneous Poisson process model (NHPP) from the finite exponential class of models. This model was one of the first non-homogeneous Poisson process models proposed. Goel & Okumoto validated this model by showing that it predicted well on a unique data set. Goel and others started describing processes for which each model would be tested to see how well the model fits the data, and predicts the future events. The assertion was that different models predict well only on certain data sets and that by comparing the predictive quality of different models, it is possible to select the best one for a given application. Khoshgoftaar & Woodcock [15] proposed a method to select a reliability model among various alternatives using the log-likelihood function. They apply the method to the failure logs of a project. The method selected an S-shaped model as the most appropriate. Lyu & Nikora [16] implemented Goodness-of-Fit (GOF) model selection criteria in their tool. Common statistical tests for GOF are Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests, as well as the simple (Rsq.) test. Rigdon presented detailed descriptions of these tests. The power of several of these statistical tests for GOF has been evaluated for a variety of reliability models including those based on a non- homogeneous Poisson process, Weibull distribution, and the Moranda model, by Gaudoin. Interestingly, the R-square value is, for most of these models, at least as powerful as the other GOF tests, and sometimes the most powerful.

Parameter estimation for SRGMs is generally achieved by applying MLE (maximum likelihood estimation) or LSE (Least squares estimation) technique. The maximum likelihood technique estimates parameters by solving a set of simultaneous equations. The method of least squares minimizes the sum of squares of the deviations between actual errors and expected errors based on the selected theoretical model.

III. PROPOSED APPROACH

This study is an attempt to compare the techniques MLE and LSE for the software reliability growth models. The proposed method works as under:

- Identify various NHPP SRGMs for proposed study.
- Collect experimental data from literature/industry.
- Unknown parameter estimation using Maximum Likelihood Estimation (MLE) and Least Square Estimation Method (LSE).

- Comparatively Analysis of parameter estimation technique by calculating the distance of predicted values from the original values as given below in equation 1 and equation 2.

$$distance_i = \sqrt{(observed\ value_i - predicted\ value_i)^2} \quad (1)$$

$$total\ distance = \sum_{i=0}^{i=n} distance_i \quad (2)$$

3.1. Unknown Parameter Estimation

We have used eight of the widely used software reliability growth models. These SRGMs and their mean value functions are listed below in Table 1. The main reason for their selection is their wide familiarity and availability of MLE simultaneous equations. The mean value functions have parameters a, which refers to total number of predicted defects and b, which is generally the shape parameter or growth rate parameter.

Model Name	Mean Value Function m(t)
Generalized Goel	$m(t) = a(1 - e^{-bt^c})$
Goel-Okumoto	$m(t) = a(1 - e^{-bt}$
Gompert	$m(t) = a(b^{k^t}$
Delayed s-shaped	$m(t) = a(1 - (1 + bt)e^{-bt}$
Logistic Growth	$m(t) = \frac{a}{1 + ke^{-bt}}$
Modified Duane	$m(t) = a[1 - \left(\frac{b}{b+t}\right)^c]$
Inflection s- Shaped	$m(t) = \frac{a(1 - e^{-bt})}{1 + \beta e^{-bt}}$
Musa-Okumoto	$m(t) = a\ln(1 + bt)$

Table 1.Summary of SRGM used in this study

After collecting data the testing data, next we have to evaluate how well the model fits the observed data (i.e. goodness of fit). The model is fitted to the test data by finding values of the unknown parameters used in its mean value function. But, the main problem in fitting SRGMs is to estimate their unknown parameters. There exist a number of parameter estimation techniques in literature like maximum likelihood estimation (MLE), least square estimation (LSE), GA etc. We have estimated the parameters in proposed approach using MLE and LSE technique i.e. One approach to estimate parameters is to input the data directly into equations for the parameters.

The most common method for this direct parameter estimation is the maximum likelihood technique (MLE) and the second approach is fitting the curve described by the function to the data and estimating the parameters from the best fit to the curve. The most common method for this indirect parameter estimation is the least squares technique (LSE). We have used these two techniques on two different datasets and then compared the results. The mathematical formulation of parameter estimation problem is used in MLE technique is depicted in equation 3. In this equation by maximizing the log likelihood function one can estimate the unknown parameters for any SRGM. Using the observed failure data (ti, yi) for =1, 2, ..., one can use the mean value function m(ti) as given in table 1 for any model to determine expected numbers of errors detected by time ti for i = n +1, n+2, etc.

$$\max(f) = \sum_{i=1}^{n} (yi - (yi - 1)).\log[m(ti) - m(ti - 1)] - m(tn)$$

where

yi = cumulative number of detected errors in a given time interval (ti)

i = 1,2,…, n is the failure index

ti = failure time index

(ti)= total number of failures observed at time ti according to the actual data.

IV. VALIDATING THE PROPOSED APPROACH

Case Study-1

The failure data set [7] come from three releases of a large medical record system, consisting of 188 software components. Each component contains a number of files. Initially, the software consisted of 173 software components. All three releases added functionality to the product. Over the three releases, 15 components were added. Between three and seven new components were added in each release. Many other components were modified in all three releases as a side effect of the added functionality. The proposed method has been applied to release 1 of this data set, given below in table 2.

Test Time (Weeks)	Failure Found	Test Time (Weeks)	Failure Found
1	28	10	125
2	29	11	139
3	29	12	152
4	29	13	164
5	29	14	164
6	37	15	165
7	63	16	168
8	92	17	170
9	116	18	176

Table 2. Failure data of large medical record system release

The parameters have been estimated using time 13 weeks because the weekly consumption of testing effort gradually decreased after the 13th week. The estimated values of the parameters have been represented in Table 3. We have used solver tool of MS Excel for MLE and Curve fitting tool of Matlab for LSE techniques. Using MLE and LSE method, the estimated values of the unknown parameters for each SRGMs used for DS1 using MLE are given in Table 3.

MODEL NAME	ESTIMATED VALUE OF UNKNOWN PARAMETERS USING MLE METHOD	ESTIMATED VALUE OF UNKNOWN PARAMETERS USING LSE METHOD
GENERALIZED GOEL	a= 233.0695,b=0.2,c=0.703	a= 1.029e+004,b=0.001356,c=0.9244
GOEL-OKUMOTO	a= 611.8981,b= 0.024	a= 1.068e+004,b=0.001104
GOMPERT	a= 26747.58,b=0.123,k=0.187	a= 412..1,b=0.01797,k=0.8896
DELAYED S- SHAPED	a= 303.2799,b= 0.1393	a= 451,b=0.1005
LOGISTIC GROWTH	a = 236.7015,b= 0.273,k= 15.4939	a=205.8,b=0.355,k=23.79
MODIFIED DUANE	a= 1.19e+10, b= 504.3466, c=5.4e-07	a= 3.762e+004,b=5081,c=1.589
INFLECTION S-SHAPED	a= 281.3323,b= 0.191, â=6.894	a=1.068e+004,b=0.001105,â=0.002329
MUSA-OKUMOTO	a= 171.705,b=0.123	a= 4.562e+004,b= 0.005941

Table 3. Parameter estimation using MLE-DS1

Next we have predicted the defects for the eight SRGMs using dataset-1 for the test time 13 to 18 weeks using MLE and LSE technique. Estimated defects is calculated using the mean value function equation by putting the values of the estimated parameters as shown in table 3 and the distance is calculated by taking the absolute value of the |actual defect – estimated defect| as given in equation 1.In figure 1 and 2 the comparison of predicted values of estimated defects with actual defects using MLE and LSE techniques for dataset-1 is done respectively. The estimated defects of the eight models are predicted for the test time 13 to 18 weeks for dataset- 1 using MLE and LSE technique. We observed that using MLE technique, the estimated values of defects are more close to the actual defects whereas using LSE technique, values deflect much from the actual defect.

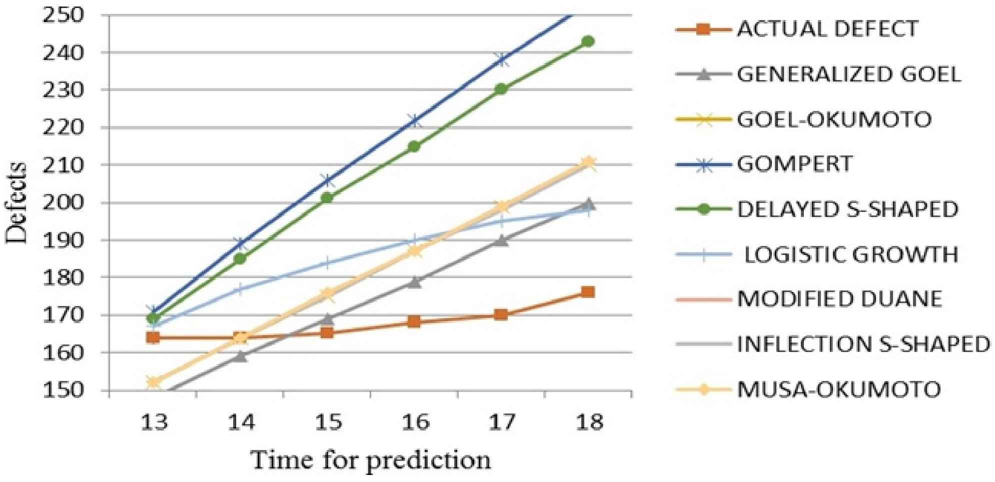


Figure 1. Prediction of defects using selected SRGMs by MLE method for dataset-1

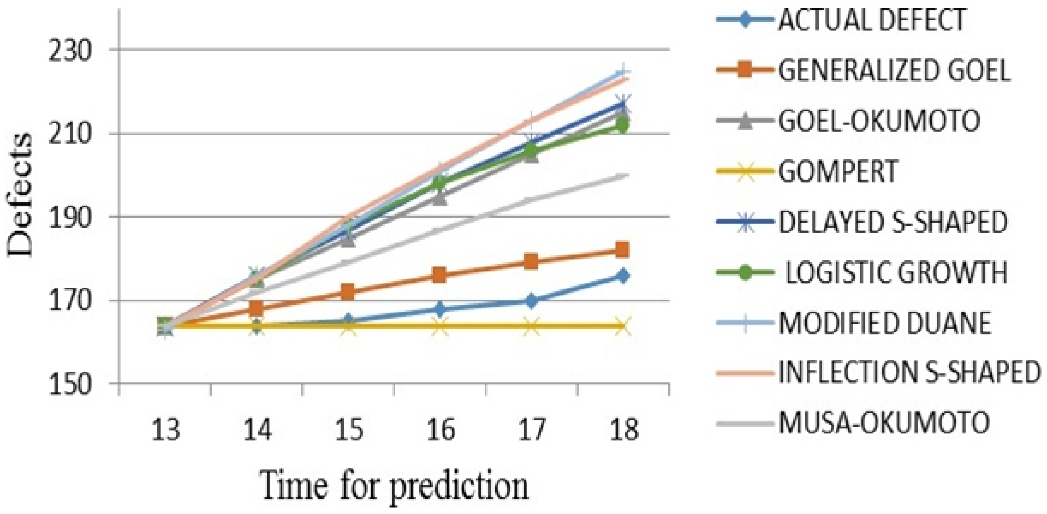


Figure 2. Prediction of defects using selected SRGMs by LSE method for dataset-1

In the figure 3, we have shown the Comparison of MLE and LSE using proposed approach for dataset-1 and we found that for dataset-1 LSE technique shows better results than MLE technique as out of the eight models, five models showed less distance when compared to the distance calculated from MLE.

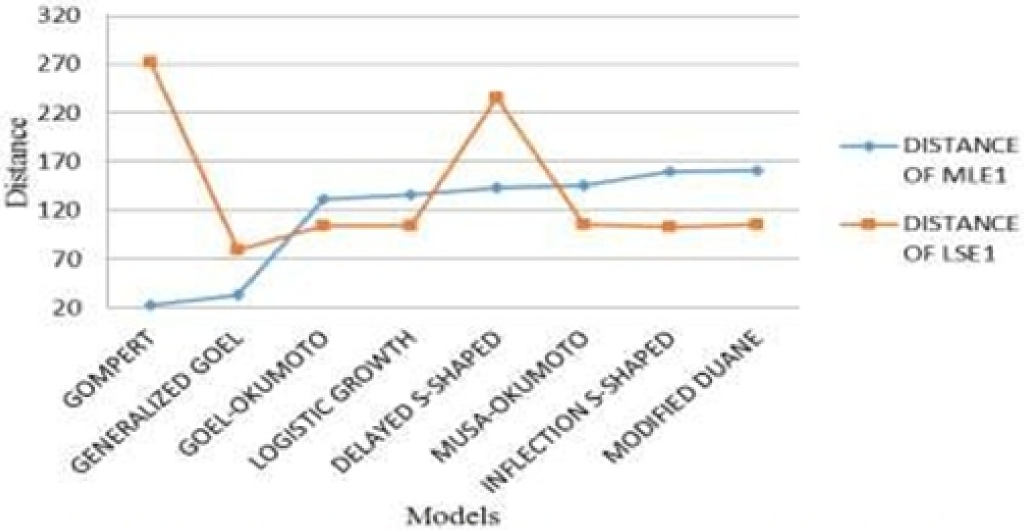


Figure 3.Comparison of MLE and LSE using Proposed approach for dataset-1

Case Study-2

The Dataset-2(DS2) having 100 reported defects have been taken from the open literature for evaluation, optimal selection. The dataset- 2 was collected from a subset of products for four separate software releases at Tandem Computers Company. To avoid confidentiality issues, the number of faults was normalized from 0 to 100. The dataset-2 is shown below in table 4.

Test Time	Failures Found	Test Time	Failures Found
1	16	11	81
2	24	12	86
3	27	13	90
4	33	14	93
5	41	15	96
6	49	16	98
7	54	17	99
8	58	18	100
9	69	19	100

Table 4. Tandem Computers Software Failure

The estimated values of the parameters have been provided in Table 5 and the same step-wise procedure has been followed for this data set.

MODEL NAME	ESTIMATED VALUE OF UNKNOWN PARAMETERS USING MLE METHOD	ESTIMATED VALUE OF UNKNOWN PARAMETERS USING LSE METHOD
GENERALIZED GOEL	a= 89.094,b=0.098,c=1.28	a= 3874,b=0.003053,c=0.7976
GOEL-OKUMOTO	a= 335.7978,b=0.024	a= 182.7,b=0.05215
GOMPERT	a= 125.92,b=0.091,k=0.859	a= 135.9,b=0.09073,k=0.8715
DELAYED S-SHAPED	a= 102.244,b=0.281	a= 96.09,b=0.293
LOGISTIC GROWTH	a=87.40,b=0.41,k=9.915	a=108,b=0.2691,k=6.546
MODIFIED DUANE	a= 8754.614,b=9.87,c=0.0123	a= 1988,b=14.54,c=0.0725
INFLECTION S-SHAPED	a= 96.38,b=0.345, â=5.22	a= 182.7,b=0.05215, â=1.874e-007
MUSA-OKUMOTO	a= 171.705,b= 0.123	a= 130.7,b= 0.07622

Table 5. Parameter estimation using MLE-DS2

We have calculated the estimated defects and distances for the eight SRGMs using dataset-2 for the test time 13 to 20 weeks using MLE and LSE technique in the same way as in the case1 and we found that for dataset-2 distance calculated using LSE technique gives better result than MLE Technique. In figure 4 and 5, we have shown the predictions of model estimated defects with the actual defects i.e. have taken the actual defect from the given dataset-2 for the consecutive 13 to 20 weeks. We have observed that using LSE technique, model's value are very close to the actual defect values whereas using MLE technique, values are farther from the actual ones .Thus we observe that for the dataset-1 MLE shows better results and for the dataset-2 LSE gives better performance.

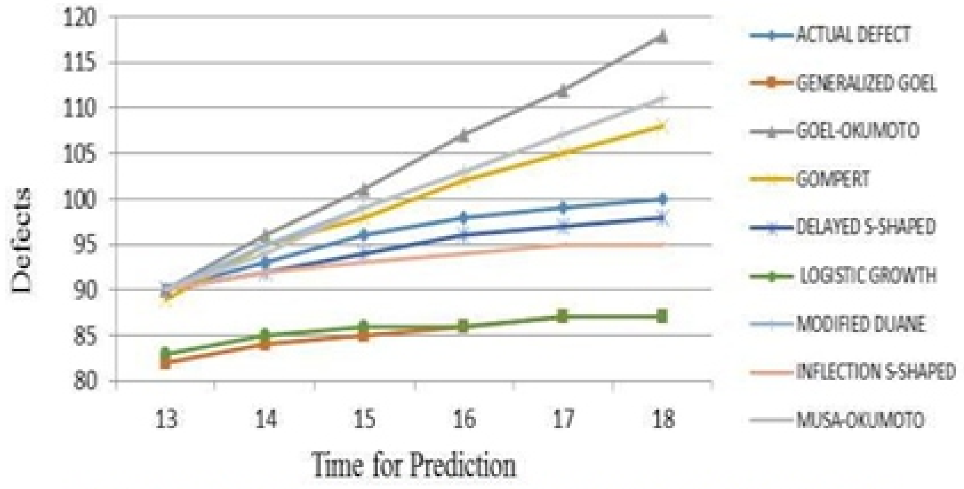


Figure 4. Prediction of actual defect with SRGMs defect using MLE for dataset-2

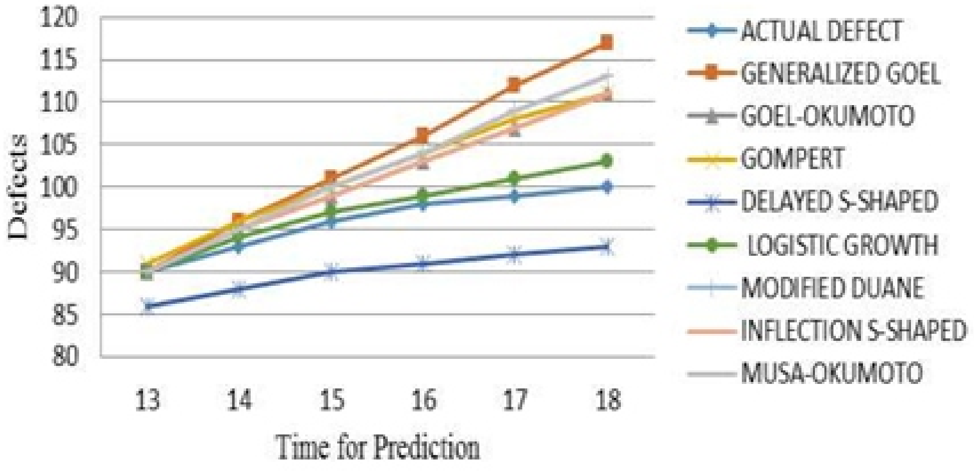


Figure 5. Prediction of actual defect with SRGM model defect using LSE for DS2

In the figure 6, we have shown the Comparison of MLE and LSE using DBA approach for dataset-2 and we find that for dataset-2, MLE technique shows better results than LSE technique.

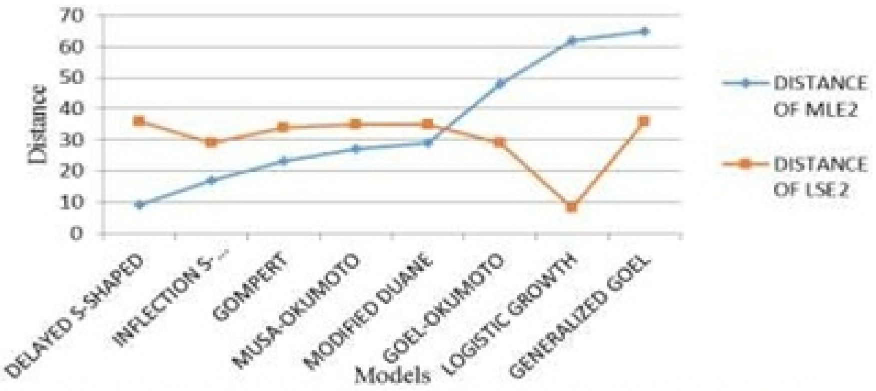


Figure 6: Comparison of MLE and LSE using proposed approach for dataset-2

V. CONCLUSIONS

In this study using data from literature two of the most widely recommended and used methodology for estimating parameters for the purpose of applying SRGMs to defect/failure data prediction is compared on the basis of distances in the actual and estimated values. It is noted in the study that while MLE is the recommended estimator with superior statistical properties, its usability and applicability in all situations is questionable. Further MLE is difficult to apply which limits its use in industry, especially due to lack of tools support. Although external validity of work presented here may be considered low due to use of only two datasets, the study provides important results that point towards different results obtained using different estimation procedures. The study provides useful and practical insights for industry practitioners and early researchers applying reliability modeling to defect/failure data. We further provide a metric for comparing the predictive accuracy which is symmetric for over and under prediction for distances. With results in this study suggesting that the fit, predict and predictive accuracy obtained using MLE and LSE estimators may be much different from one estimator to another, more research in this direction is needed to establish these differences in different contexts and thus helping to resolve the dilemma faced by reliability practitioners of which estimator to use and in which conditions a given estimator is better than other. Initial results presented here and properties of MLE and LSE estimators suggest that while LSE is good estimator for fitting the data to observed failure data. MLE is better estimator for making reliable predictions.

REFERENCES

[1] Kitchin, Rob, and Martin Dodge. Code/space: Software and everyday life. MIT Press, 2011.

[2] Lew, Ken S., Tharam S. Dillon, and Kevin E. Forward. "Software complexity and its impact on software reliability." Software Engineering, IEEE Transactions on14.11 (1988): 1645-1655.

[3]Rai, Arun, Haidong Song, and Marvin Troutt. "Software quality assurance: an analytical survey and research prioritization." Journal of Systems and Software40.1 (1998): 67-83.

[4]Lin, Chu-Ti, and Chin-Yu Huang. "Enhancing and measuring the predictive capabilities of testing-effort dependent software reliability models." Journal of Systems and Software 81.6 (2008): 1025-1038.

[5]Goradia, Tarak. "Dynamic impact analysis: A cost-effective technique to enforce error-propagation." ACM SIGSOFT Software Engineering Notes. Vol. 18. No. 3. ACM, 1993.

[6]Xie, M. "Software reliability models-past, present and future." Recent Advances in Reliability Theory. Birkhäuser Boston, 2000. 325-340.

[7]Stringfellow, Catherine, and A. Amschler Andrews. "An empirical method for selecting software reliability growth models." Empirical Software Engineering7.4 (2002): 319- 343.

[8]Sharma, Kapil, et al. "Selection of optimal software reliability growth models using a distance based approach." Reliability, IEEE Transactions on 59.2 (2010): 266-276.

[9]Lyu, Michael R. "Software reliability engineering: A roadmap." 2007 Future of Software Engineering. IEEE Computer Society, 2007.

[10]Kan, Stephen H. Metrics and models in software quality engineering. Addison-Wesley Longman Publishing Co., Inc., 2002.

[11]Brocklehurst, Sarah, et al. "Recalibrating software reliability models." Software Engineering, IEEE Transactions on 16.4 (1990): 458-470.

[12]Goel, Amrit L., and Kazu Okumoto. "Time-dependent error- detection rate model for software reliability and other performance measures." IEEE Transactions on Reliability, 28.3 (1979): 206-211.

[13] Knafl, George J., and Jerome Sacks. "Software reliability model selection. “Computer Software and Applications Conference, 1991. COMPSAC'91., Proceedings of the Fifteenth Annual International. IEEE, 1991.

[14] Ando, Takao, Hiroyuki Okamura, and Tadashi Dohi. "Estimating Markov modulated software reliability models via EM algorithm." Dependable, Autonomic and Secure Computing, 2nd IEEE International Symposium on. IEEE, 2006.

[15] Khoshgoftaar TM, Woodcock TG, “Software reliability model selection: a cast study. In 1991 IEEE International Symposium on Software Reliability Engineering 1991 May 17 (pp. 183-191).

[16] Lyu MR, Nikora A. Applying reliability models more effectively (software). IEEE software. 1992 Jul;9(4):43-52.

ISSN No.-2321-8711

A Software System Evolution In Human-centric Environment Driven By New User Intention Detection Using CRF

[1]Iqra Urooj, [2]Ming Yin, [3]Jijiao Jiang, [4]Wu Junsheng,
[5]Jahanzeb Jabbar, [6]Naqash Azeem

[1,2,4,5]School of Software and Microelectronics, Northwestern Polytechnical University Xi'an, P.R. China, [3]School of Management, Northwestern Polytechnical University Xi'an, P.R. China, [6]School of Mechanical Engineering, Northwestern Polytechnical University Xi'an, P.R. China

E-mail: [1]iqra.urooj@hotmail.com, [2]yming@nwpu.edu.cn, [3]jjj_leon@nwpu.edu.cn, [4]wujunsheng@nwpu.edu.cn, [5]jahanzeb.jabbar@hotmail.com, [6]iamnaqash@hotmail.com

ABSTRACT

The Software Service Evolution can easily determine through requests for changes, improvement, and enablement of knowledge development continuously from users', as compared to the other factors. It is unavoidable for almost all software and can be seen as the development of system-user interactions. The ability to precisely and effectively monitor users' volatile requirements is perilous that requires to make a timely improved system for adaptation of fast varying environments. In this research, a methodology applies Conditional Random Fields (CRF) as a mathematical foundation to discover the users' potential desires and requirements in order to deliver a quantitative exploration of system-user interactions. By examining users' run-time behavioral patterns, domain knowledge experts can predict how users' intentions shift. The results also show the effects of different regularization algorithms of CRF on the training model. Our supreme objective is to accelerate software service evolution by using machine learning techniques. To detect users' intentions using the CRF method, an experiment on an open-source software is performed.

Keywords - Conditional Random Fields, intention, requirement, software service evolution, target.

I. INTRODUCTION

The adaptation to ever-changing user requirements and the operating system is the main goal for the evolution of a software system [1]. Currently, most of the software systems required evolution that is considered an essential quality [2], [3] since the nature of human demands are ever changing and abrupt evolving [4] and real-world systems have volatile environmental changes [5].The advancement of system-user interactions can be viewed from the evolution of software services. The ability to precisely and effectively monitor users' volatile requirements is perilous that requires to make a timely improved system to adapt rapidly changing environments. Conventionally, manual analysis and based on business needs or delayed user feedbacks were used to elicit these new requirements [6], which made the base to continue the software evolution nowadays.

The inference process of Situ [7] framework is proposed that allows to make model and detect human intentions by taking the targets of an individual, including catching the behaviors through observations.

An interpretation procedure based on the Hidden Markov Model (HMM) considerably reduce service evolution cycle and creates instant individualized services definition at runtime possible. Situ [7] recommends that A human perceived situations according to its internal mental condition (target), i.e., actions, and the target achieved can be easily interpreted as exterior reflections. The externally observable of human behaviors are represented by some context value changes. If the user has a new approach for achieving a predetermined target or the user has a new target, both will result in a new intention detection. As these two cases reveal a user's new requirements, therefore, they are beneficial for the improvement of the system. Conditional Random Fields (CRF) is a statistical modeling method which is used in our research that detects users' targets by providing a quantifiable study of interactions between system and user. By examining users' run-time behavioral patterns, knowledge-baseddomain experts can have a great idea about changing in users' intentions. CRF model is constructed through the training data acquired by an experimentation on an open-source software was performed and the method of new user intention detection is demonstrated. CRF-L1 and L2 are regularizations to reduce the over fitting and improve the classification accuracy of model construction. Margin Infused Relaxed Algorithm (MIRA) is another algorithm to train the CRFs. An experimental comparison is also explained among regularizations algorithms to get the maximum accuracy that helps to provide accurate inference probability.

II.RELATED WORK

System Evolution and Requirements Elicitation

The software system grows continuously, as it increases the complexity of the system will also increase and will require more efficient methods. To make sure the flexibility and reliability, the target of software evolution is to perform significant modifications to the system. In general, software evolution defines as the study and management of the process of performing significant modifications to software in a timely manner. Throughout the whole system lifetime, the change identification continues and suggestions for change are the drivers of system evolution. A cooperative learning process of requirements engineer with the customer can be reflected in the traditional requirements elicitation procedure [8], as this procedure mainly depends on requirements engineer for performing the subjective analysis and prediction, it is typically inefficient and results in erroneous requirements. A cycle of modification, elicitation, development,and deployment takes too much time to finish. In order to turn timely and critical service individualization into reality, the researchers are facing many challenges to overcome unpleasantly long evolution cycles nowadays. New technologies should be introduced for fast growing software evolution.

The inference process of Situ Framework

The inference process of Situ [7] is proposed to detect human intention in context-aware service environments. It infers human target since most are hidden and captures subsequent context values by observing that provides people to detect human intentions and model. While detecting intention along with prediction mechanism, Hidden Markov Model (HMM), Situ is intended to build the instantaneous description of individualized services at runtime possible and greatly reduce service evolution cycle.

Situ describes a situation that includes the users' target as a time-stamped status together with users' behaviors (context value & action). Formally, the situation is expressed by 3-tuple {t, A, E) at time t, where t is a set of atomic target, A is set an of behavioral context and E is a set of environmental context. And intentionIis represented as I = seq (S1, S2, ...,Sk), where S1, S2, ...,Sk are temporal sequence of situations threaded through an atomic target t.

CRF (Conditional Random Fields)

CRF is a mathematical foundation that provides a method used to encode well-known relationships between observations and build reliable interpretations [9].CRF is mostly used for parsing or labeling the sequential data. A CRF model should be built that can encode users' standard behavior pattern so that it identifies new intention with low confidence is indicates users' divergent behaviors. The standard CRF model is developed from the training data that are collected by observing user behaviors who are expected to correspond to the system design. The CRF model will be used in the result of target inference to label with targets the divergent behaviors in low confidence, and in order to elicit user's potentially new intentions they can be easily singled out and examined. Apart from the divergent behaviors, new requirement or system drawbacks will also be detected by analyzing users' target transitions and erroneous behaviors.

To acquire the effective labeling for an observation sequence S, the CRF model marks every possible target-sequence T by summing the weighted feature functions f j over all data in the sequence:

$$\text{value}\, T\, S = \quad \overset{1}{\underset{j=1}{}} \quad \overset{V}{\underset{i=1}{}} \quad \omega_j f_j\, S, i, T_i, T_{i-1}$$

Where ωj is the weight correlated with feature function fj. The target-sequence T with the largest value (T | S) will be chosen as the labeling for the observation sequence S. The Viterbi Algorithm [10] is employed in value (T | S) for minimizing the computation complexity and Limited-memory BFGS algorithm is used for calculating the weights ωj of feature functions fj. To maximize the accuracy of inference results we did experiment on different CRF regularizations. Regularization is a machine learning technique to reduce learning model overfitting. Results shows in section 4.4.

CRF regularizations and MIRA algorithm are:

1. L2 regularization

It is also known as Ridge Regression.

$\max L Y X; w - \lambda wT w$

The parameter controls the degree of smoothing. Higher the smoothing, will be high value.

2. L1 regularization

It is also known as Lasso Regression.

$\max_w L Y X; w - \lambda_i wi$,

Again λ controls the degree of smoothing. Regularization with L1 make training model complex.

2. MIRA

Margin Infused Relaxed Algorithm is an online learning algorithm developed by Crammer and Singer (2003).In each round during training, the model algorithm updates the weight vector. MIRA processes one data set at a time in compared to others offline algorithms of CRFs.

III. NEW INTENTION DETECTION

Basic Concepts

The CRF model involves various feature function and each feature function specify the relations between target and observations. some basic idea used in the CRF model is described as follows:

Action: It is the user's operations on the system interface.

Context: an entity in the system domain that can be used to describe the status of any information.

Observed-status: A set of observations (action & context values) that are captured at a specific time.

Intention: an execution path for accomplishing specific target.

Situation: A situation that describes a time-stamped, including the user's behaviors (actions, context value),andtargets.

Target: the status or goal that user wants to accomplish.

Target Inference and New Intention Detection - As the prerequisite for detecting new intention, target inference is to infer users' target for each observation at a specific time point. To work out on this, an inference model should be built by encoding the interactions between targets and observations (actions & context values). We build inference model by using CRF in this research article. Target inference and new intention detection can be defined as follows:

Target Inference: Inference model acquires input asequence of observations and gives users' target for each observation at a specific time point as an output in target inference

New intention detection: In new intention detection users' new behavior sequence pattern accomplishing a target, which is not predefined in the knowledge base domain.

We detect new intention using the following three methods in this research article:

a) **Method I:** Comparison of divergent with identified behaviors;
   This method belongs to detecting users' divergent behaviors by comparing it with the identified behavior patterns.

b) **Method II:** Obtaining a new intention;
   In this method, a new intention based on target evolutions is obtained and it can be assumed to make the target changeovers more efficiently and competently when two targets appear successively.

c) **Method III:** Get erroneous behaviors;
   New targets from users' erroneous behaviors are found when users have new targets i.e. problem in uploading a file and system didn't mention the maximum and minimum size.

New Intention Detection using CRF

The mathematical method Conditional Random Field (CRF) is used in this paper to build the model for new intention detection and target inference. A linear chain CRF method is used. The CRF model acquires input asa sequence of observations and gives users' target inference with the highest confidence.

To find the output target inference in method 1 which indicate the detail process of users' low confidences is as follows:

1) **Stage 1:** Create a CRF model. The CRF model is built based on the training data which acquired through the users' observed-status.

2) **Stage 2:** After building the CRF model, record user's behaviors (actions & context values).

3) **Stage 3:** Inference probability is calculated from the CRF model of each output inference targets.

4) **Stage 4:** By Finding the output targets with low confidence, examine the corresponding users' divergent behaviors.

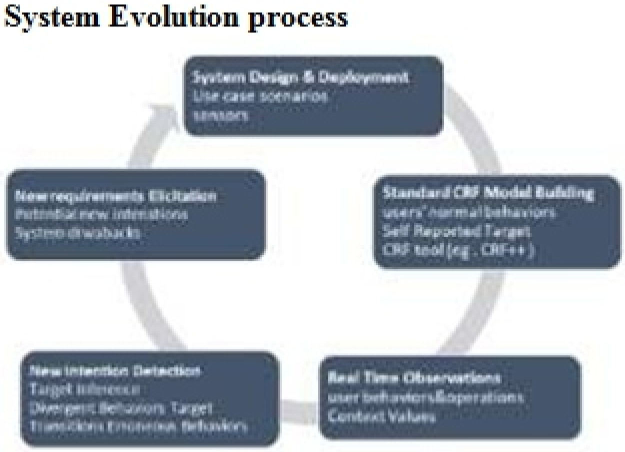


Figure 1. System Evolution Process Using CRF.

System drawbacks are eliminated by system engineers who differentiate between the new intentions and detected potentially system drawback that is mentioned in step 1 of flow chat. This step is purely performed by a domain expert, the system can't perform automatically. Identifying the system drawbacks is not a tough task but new intentions are implied and ambiguous. Knowledge-based domain experts have two options, first is to directly ask the users if possible and other is to estimates the new intention sand then verify them. The system can be redesigned based on users' new intention satisfaction or by providing a solution to the disclosed problems. Figure.1 shows the system evolution process on user new intention using a CRF. The complete system evolution process to detect new intentions by applying the CRF model is shown in Figure 1.

IV. EXPERIMENTATION ON A SYSTEM

Platform using for Experiment

To detect the new intentions using CRF method and to improve system evolution, an experiment is conducted on an open-source software. We took an online system to share views on research papers. To monitor users' behaviors an embedded program is deployed in software as an agent to record data. Users can submit papers/comments on papers, view/download papers and filter the papers. User's actions, as well as the contents on the selected web page

Process for Experiment

Users' actions, context values, and targets are recorded as raw data. The experiment is run in two rounds to demonstrate the evolution process. The procedure for the experiment shown below:

1) Firstly, Install the System version 1;
2) Run experiment of the first-round: call users to gather data of users' actions, targets, and context-values, after that evaluate the predicted users' new intentions;
3) Based on the predicted new users' requirements to revise the system and deploy version 2;
4) Run experiment of the second-round: call users again to gather data of users' actions, targets, and context-values, evaluate users' new requirements and estimate the system enhancement.

Data using for Experiment

Around 7,000 data records and 300 sessions taken from the raw data record, after cleaning the raw data that contains the users' actions, context-values, and self-reported targets at a specific time. A training data which is processed has the following format:

Table 1.Training data segment

Observed-status	Time	Target
click Menu My Profile	20s	View Profile
click Login & Login Good	5s	View Profile
click Menu My Profile	60s	View Profile
click Cancel My Profile	60s	View A Paper Info
click Menu All Papers	10s	View A Paper Info

Observed-status, time-interval,andtargetare three elements of data record:

Observed-status: the format of observed-status is action&context-values;

Time-interval: it is the time between two following observed-status. Intervals 5, 10, 20, 30, 60(seconds), (m)minutes and (h)hours that used from raw data to processed data;

Target: goals that reported by participants.

Creating the CRF Model

To detect users' divergent behaviors a CRF model is created and used.Around 5,000 data records are used as the first-round experiment. A standard CRF model iscreated on the training data records using CRF++. To build the standard CRF model feature template is required to define the specific formats of training data as shown in Table 2.

Table 2. Unigram and Bigram Feature templates.

Unigramfeatures	Bigram features
U01:%x [0,0]	B01:%x [0,1]
U02:%x [?1,0]/%x [0,0]	B02:%x [0,0]/%x
U03:%x [0,0] / %x [1,0]	[0,1]
U04:%x	B03:%x
[?2,0]/%x[?1,0]/%x [0,0]	[?1,0]/%x[0,1]
U05:%x	B04:%x [?1,0]/%x
[?1,0]/%x[0,0]/%x[1,0]	[0,0]
U06:%x	/%x [0,1]
[0,0]/%x[1,0]/%x[2,0]	

In this paper, it is also observed that the CRF algorithm with L2regularization gives the highest accuracy in compared with L1 regularization and MIRA (Margin Infused Relaxed Algorithm) as shown in Figure 2. L1 and L2regularizations are two related techniques used for reducing the training model over fitting. The term smoothing is also used for regularization. Single-best MIRA is an online- learning algorithm, makes a prediction on the training data that gathered from the model. 93.40% accuracy results from L2regularization helps in our experiment, that is used to create a CRF model for better inference probability output.

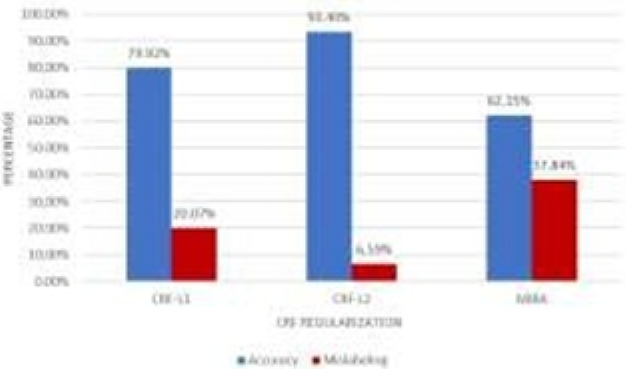
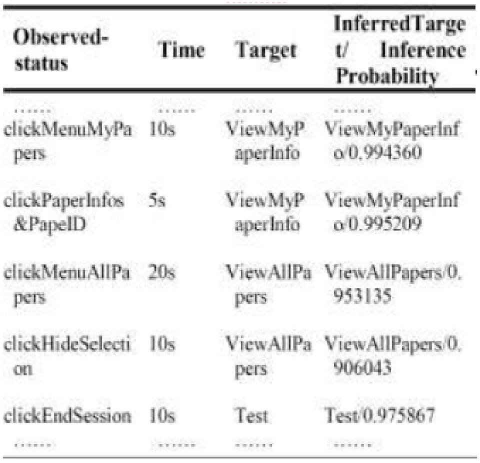


Figure 1. Comparison of CRF regularization algorithms

4.5 Experiment in the First-Round (Inference Result Analysis)

Around 2,000 data records and 106 experiment sessions are used for test data. For accurate results test data and train data should be the same. Inference results are shown:

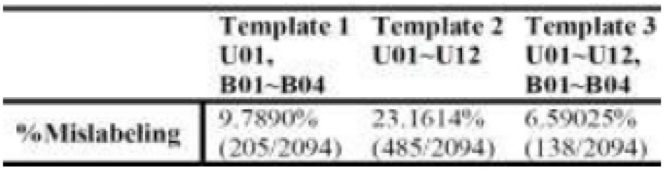
Table 3.The format oftarget inference output



Observed-status	Time	Target	InferredTarget/ Inference Probability
......	......	......	......
clickMenuMyPapers	10s	ViewMyPaperInfo	ViewMyPaperInfo/0.994360
clickPaperInfos &PapeID	5s	ViewMyPaperInfo	ViewMyPaperInfo/0.995209
clickMenuAllPapers	20s	ViewAllPapers	ViewAllPapers/0.953135
clickHideSelection	10s	ViewAllPapers	ViewAllPapers/0.906043
clickEndSession	10s	Test	Test/0.975867
......	......	......	......

Each target inference outputcovers the inferred target and inference probability, generated from the CRF model. We have observed that different feature templates give the different inference accuracy as result shown in Table 4. The complete feature template is shown in Table 2.

Table 4.Mislabeling % of different templates



	Template 1 U01, B01~B04	Template 2 U01~U12	Template 3 U01~U12, B01~B04
%Mislabeling	9.7890% (205/2094)	23.1614% (485/2094)	6.59025% (138/2094)

The CRF model cannot express divergent behaviors with users' real targets.For studying the targets with low inference probability, inference probability should relate to inference accuracy. As results are shown in figure 3 as inference probability range is increased, wrong inference will decrease. It is better to focus on learningobserved-statuses with low inference probability for improving the proficiency of new intention detection in method 1.

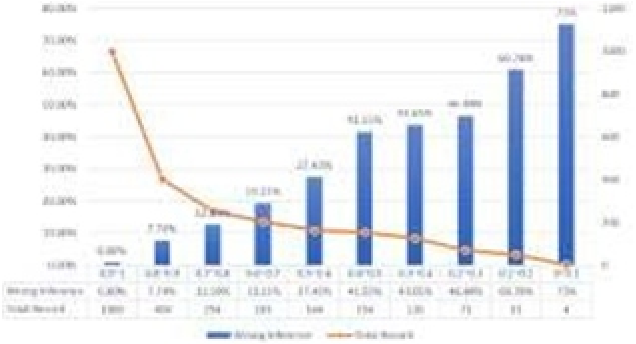


Figure 2. Comparison between Wrong Inference Percentage and Total Record

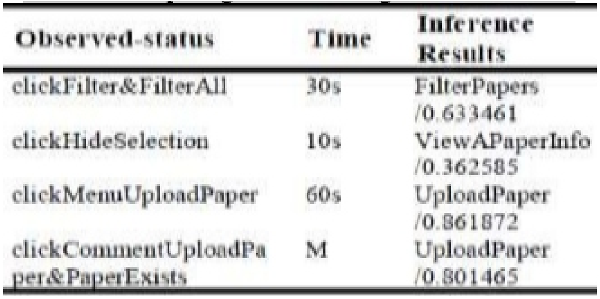
Experiment in the First-round (New Intention Detection)

The first-round experiment is shown some examples for new intention detection:

1) Method 1 to detect new intention;Study observed-statuses with low confidence values i.e. (inference probability <= 0.5). To validate users' divergent behaviors and evaluate them to improve the system, we give some examples here.

a) Example 1: “hide the above section form” is often clicked by users' to hide the selection form.

Table 5. Example segment in the target inference results



Observed-status	Time	Inference Results
clickFilter&FilterAll	30s	FilterPapers /0.633461
clickHideSelection	10s	ViewAPaperInfo /0.362585
clickMenuUploadPaper	60s	UploadPaper /0.861872
clickCommentUploadPaper&PaperExists	M	UploadPaper /0.801465

The observed-status “click Hide Selection” is labeled with inference target “View A Paper Info” with a probability lowest 0.362585, and its preceding observed-status is also labeled with low probability. However, its next observed-statusis labeled with a very high inference probability. This observed-statusof “Hide Selection” can be easily pulled out for analysis. The selection form delays users to see the information page of paper because it is too vast. We analyzed and applied a modification to the system: simply, a link “Expand it” is added to show the complete selection form.

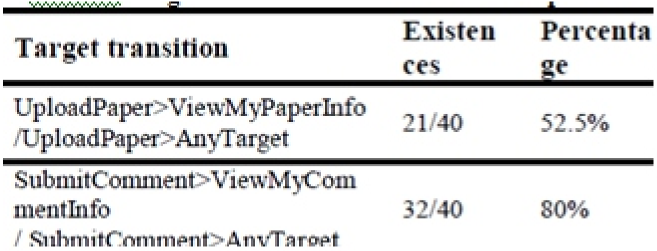
b) On the web-page of “submit/edit a comment” button is clicked twice: The reason of this mistake is that users are not clear about it, so the proposed method is to show a popup message after clicks the button “submit”.

c) The user may copy the password with an extra space, this becomes reason for the wrong password. A possible solution is to prompt the message to show that the password contains a space.

2) New intention detection method 2

To find users' potential new intentions is to study the target transitions is another way.

Table 6. Target transitions in the first-round experiment



Target transition	Existences	Percentage
UploadPaper>ViewMyPaperInfo /UploadPaper>AnyTarget	21/40	52.5%
SubmitComment>ViewMyCommentInfo / SubmitComment>AnyTarget	32/40	80%

To simplify users' behaviors and make efficient transitions the new intention can be expressed as the target transitions, so that system can be modified. E.g., for the target transition “Upload Papper>View My Paper Info” existences are 21/40, a new action link can be displayed of the submitted paper, right after the paper is submitted successfully.

1) Common users' erroneous behaviors

It is observed that a common error can be seen asif an erroneous behavior occurs in the observed-status more than 6 times. In the first-round the detected some common erroneous behaviors are:

a) When modifying the information of a paper, no file is uploaded. Users don't want to upload a file while editing the information that causes erroneous behavior. A possible modification is to allow the user to upload no file when modifying the information;

b) For submitting the comment, a number of words are not enough when users' is submitting it, so the error occurs because specific numbers of words are required. The system should count the words while the user is going to submit it.

Experiment in the Second-round (System Enhancement)

After the first-round experiment, modifications are applied, and the system is enhanced to satisfy users' new requirements. A standard CRF model of the second round is build using around 5,000 training data records and around 2,000 to test data to predict the target inference. The combined target inference accuracy is:

Table 7.comparison on target inference accuracy

Experiment	%Mislabeling	%Accuracy
First-Round	6.59025%	93.4097%
	(138/2094)	(1956/2094)
Second Round	1.9924%	98.0075%
	(42/2108)	(2066/2108)

System improvements in the second-round and estimation of these improvements are following:

1) In the new second-round system after deployment, a link "Expand it" is added for the selection form, to improve users' divergent behavior;

   We also studied the users' behaviors when users went to the page "All Papers", there are some kinds of behaviors are located: a) Filter papers by using the form; b) to filter the paper expand the selection form on all papers; c) others;

2) Target transitions from "Upload/Edit paper" or "Submit/Edit Comments" to "View My Comment/Paper"; In the new evaluated system, a link is added for newly submitted Comment or Paper. Another example of target transition is to add a link of "Download Paper", on the page of all paper's information.

As shown in table 7, system performance is improved in version II, as the rate of mislabeling significantly decreases, the success rate is increased.

CONCLUSION

A methodology of new intention detection and target inference by applying the CRF method is presented in this experiment. The CRF method is used to encode the connections of observable actions, existing targets and context values. We applied a technique to infer targets by applying the CRF model. We also demonstrate that CRF regularization L1 and L2 and MIRA algorithms have different accuracy result, thus L2 regularization gives best results. Furthermore, users' divergent behaviors are labeled with targets in low inference, they will be identified users' potential new intentions detection or system drawbacks on the basis of target inference. Thus, CRF method for detection new intentions and divergent behaviors and system errors make possible recommendations for system evolution.

ACKNOWLEDGMENTS

The author would like to thank all authors who participated in this research project. We also thank the researchers who are listed in the references. This research is supported by Shaanxi Social Science Fund in China (2018S28), Natural Science Basic Research Plan in Shaanxi Province of China (2018JM7008), Humanities and Social Sciences of Ministry of Education Fund Project (18YJA630043, 16YJA630068), Aeronautical Science Fund of China (2016ZG53071), and the Fundamental Research Funds for the Central Universities of Northwestern Poly technical University(G2017KY0101).

REFERENCES

[1] Rajlich, V. , 2014. Software evolution and maintenance. In: Proc. the on Future of Software Engineering, pp. 133–144 .

[2] Charrada, E.B., Koziolek, A. ,Glinz, M., 2015. Supporting requirements update during software evolution. J. Softw. 27 (3), 166–194 March .

[3] C. Sutton and A. McCallum, An Introduction to Conditional Random Fields for Relational Learning. MIT Press, 2006.

[4] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proceedings of 18th International Conference on Machine Learning, Morgan Kaufmann, pp. 282–289, 2001.

[5] Nehaniv, C.L. ,Wernick, P. , 2007. Introduction to software evolvability. In: Proc. Third Int'l IEEE Workshop on Software Evolvability, pp. 6–7 .

[6] Salinesi, C. ,Etien, A. , 2003. Compliance gaps: a requirements elicitation approach in the context o system evolution. In: Proc. 9th International Conference on Object- Oriented Information Systems (OOIS 2003), pp. 71–82 .

[7] Carl K. Chang, Hsin-yi Jiang, Hua Ming, and Katsunori Oyama, "Situ: A Situation-Theoretic Approach to Context-Aware Service Evolution" IEEE Transactions on Services Computing, VOL. 2, NO. 3, July-September, 2009.

[8] SotiriosLiaskos, Sheila McIlraithand Shirin Sohrabi, "Representing and reasoning with preference requirements using goals," Technical report, Dept. of Computer Science, University of Toronto, 2006.76

[9] Charles Sutton and Andrew McCallum,"An Introduction to Conditional Random Fields," Foundations and Trends in Machine Learning 4 (4), 2012.

[10] G. David Forney, "The Viterbi algorithm," Proceedings of the IEEE 61 (3), pp. 268–278, March 1973.

Survey on Web Usage Mining Based Intelligent Recommender System

[1]Eshetu Tesfaye, [2]Pooja Y

[1]M. Tech (Software Engineering), Computer science and Engineering, Sharda Univeristy, Greator Noida, India

[2]Associate Professor, Computer Science and Engineering, Sharda University, Greater Noida, India

E-mail: [1]eshe384@gmail.com, [2]pooja.1@sharda.ac.in

ABSTRACT

A recommendation system is software which supports customers by an illustrated way to recommend products and provide recommendations about the merchandise he or she interested in. In E-commerce, recommender system suggests products for the users intelligently by helping users to select their favorite items based on their past information. This paper examines the latest papers researched in the area of e-commerce and identifies the drawbacks and future work of the reviewed papers. In the end, the research gap of the study is prepared in the form of a table for output which will help the researchers and easily insights them when they are looking to do their research on the field of recommender system commonly for E-commerce domain.

Keywords - Recommendation System, Intelligent Techniques, Web Usage Mining, E-Commerce

I. INTRODUCTION

GENERAL OVERVIEW

The World Wide Web has been increasing at the highest rate which has changed how we handling our daily activities. It has become a serious supply of knowledge and it continues to extend in size and use.

[16] With the excellent advancement of network technologies and e-commerce sites, an increasing range of online utilities turning into well-liked like Yahoo, Bing, Google, Netflix and YouTube for observation videos, etc. One alternative well-liked usage of the WWW is for on-line looking, wherever the shopping for and commercialism of merchandise and services square measure conducted electronically [1, 3].

These online services cause a rise within the volume data of knowledge on the online that is named information overloads downside. Therefore, users have to be compelled to lose rather more time to seek out their fascinating things among an oversized range of alternatives. Recommender systems have even been a helpful approach to drive this downside which helps users to find their likes with exceedingly possible time [10]. Currently, recommender system incredibly abundant helpful in the E-commerce like electronic commerce, education like recommending education materials and books, and for travel and guidance by recommending historical places, Hotels, flights events and, etc.. In e-commerce, recommender systems advocate things like music, videos, news, and analysis articles, etc to their consumers. It first captures the past behavior of a client and depending upon the appreciation that the user has given for the items in the past, the engine will then recommend the item a user may well seem to shop for.

Nowadays, e-commerce applications are developed to sell merchandise and services on the net like Netflix. It provides a new business entrance and an oversized quantity of product info, thus customers pay a lot of and longer browsing information superhighway to seek out the correct info or product [33]. The main concept of the recommender system is to use customers past information to recommend the newly coming items to the users [16]. The satisfaction of the customers depends on the quality of the result returned by the recommending engine. Along these lines, building up an amazing method is an urgent issue to improve the presentation of a recommender framework.

OVERVIEW OF WEB USAGE MINING

We can say a recommender framework is intelligent if it has the behavior of intelligence to learn and reasoning capacity without the involvement of a human being through the action [8]. Intelligent techniques are helpful to investigate knowledge and predict data with a well-formed structure for the customers brilliantly. Some common intelligent techniques used for decision-making are fuzzy logic, data mining, a neural network, machine learning, and evolutionary computing [8, 24]. Fig.1 below describes some intelligent decision-making techniques.

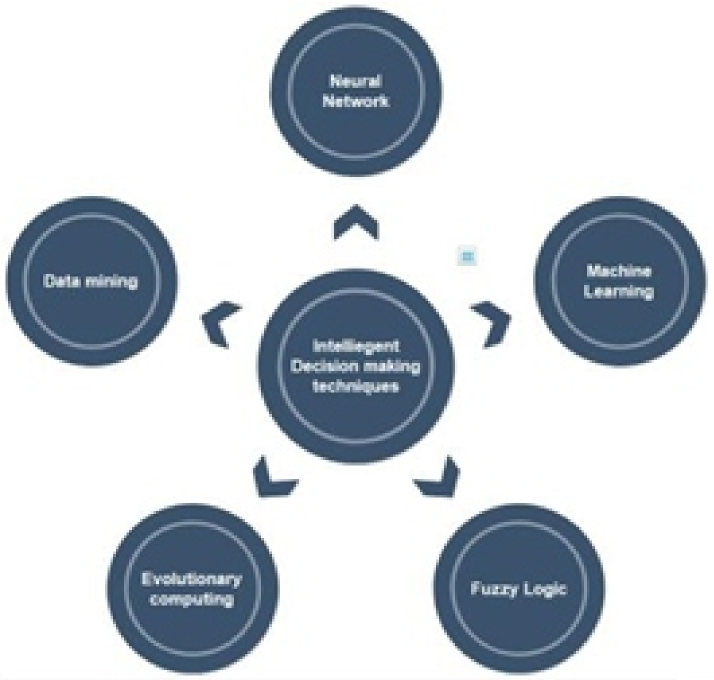


Figure 1: Intelligent decision making techniques

Web usage mining is one of the processes of the data mining technique to explore interesting figures from the net [33]. It helps us to know the information and activities of the visitors who have frequently visited and accessed the website. Some of the popular [2] web mining techniques are clustering, classification, frequent pattern analysis, and association rule. The process [3] of web usage mining can be split into four stages such as data collection, data preprocessing, pattern discovery, and pattern analysis. Fig. 2 below illustrates them as follow.

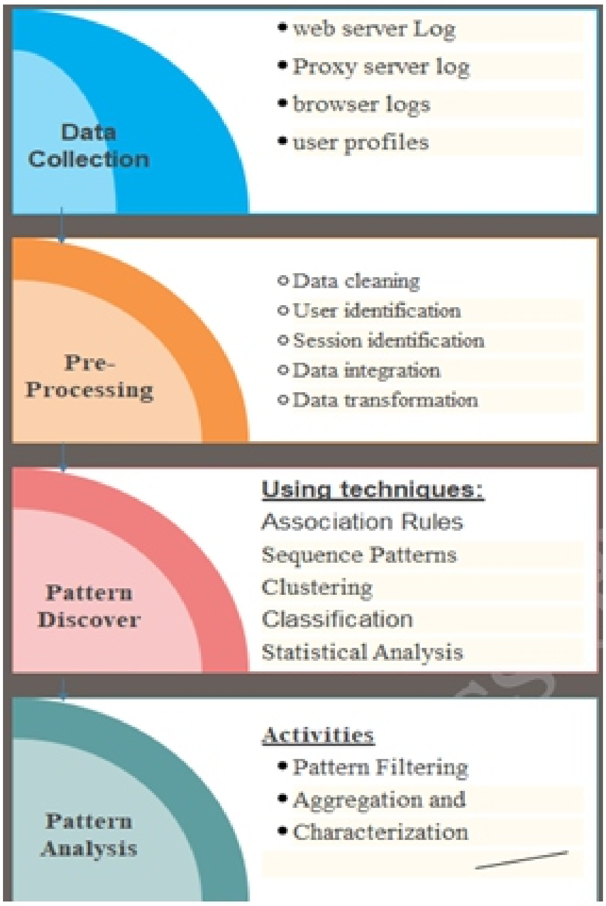


Figure 2: phases of a web usage mining

A. DATA COLLECTION

In web mining, most of the data are collected from browser logs, web server log, and proxy server [4, 33, and 34]. Among these data sources, the most important source of data is web server logs. These are the log records of the learners who are visiting the websites. In [6], the data collection step is also called a data acquisition process where the data is highly unstructured and inconsistent. In the recommender system, data might be gathered either directly by recording the information of the users called explicitly or gathering indirectly during the time the users are visiting the website known as implicitly. According to the studies [4, 3, 6, 25, 31], the customers' data is collected using session and cookies methods which implicitly records profile of visitors while visiting the web portal. Gathering the locations of visitors using the GPS method is another way to gather visitors' information [13, 18, and 21].

B. DATA PREPROCESSING

Data preprocessing is the next step after the data acquisition phase that aims to reform the web server log data. The log files are unstructured and unclean data which are not sufficient for exploring the patterns since it contains dirty data. According to [2, 3, 4, 6, 25], there are some common pre-treatment steps to make the data in a proper format such as data transformation, user identification, data integration, and session identification.

C. PATTERN DISCOVER

Pattern discovery is a key process after the preparing of data in data mining. It discovers arrangements of data from the huge volume of a data set. [13] It covers the relevant data mining algorithms such as sequential pattern analysis , classification, clustering, and association rules which are also helpful in various areas like machine learning, pattern recognition, image processing and etc. According to the studies [4, 6, 10], they used a combination of clustering and classification for clustering the pages frequently visited by users and to classify the learners respectively.

D. PATTERN ANALYSIS

Pattern analysis is the last step of the web usage mining activities. Its activity is to analyze and value the rules and models obtained from the pattern discovery process as patterns [9, 13].

OVERVIEW OF RECOMMENDER SYSTEM APPROACHES

There are three approaches of traditional RS, these are collaborative based, content-based, and hybrid filtering. Among them [1, 12], CF is the most widely used approach in the development of the recommendation system. Each recommendation system approach has its own downside and strengthens. Mostly, a collaborative filtering approach faces trust problem, cold start and sparsity obstacles, while content-based filtering has over specialization and privacy problem. When the users profile is empty and have not given any appreciation for the item which means new users who visits a website for the first time, this issue is called cold start problem [14, 16]. Sparsity problem occurs due to the lack of information which means when the user gave appreciation to view items; it's difficult to find his correct group to whom he has a similar taste [16, 18]. To overcome the problems facing in the traditional Recommender systems, a lot of research papers [6, 9, 10, 11, 15, 16, 23, 24, 25], has been done in CF by combining traditional recommendation approaches with that of present-day recommendation ways. Let us see below the categories of traditional recommendation approaches one by one.

A. CONTENT-BASED METHOD

Content-based filtering approach works based on the information of users recorded at the time of registration. Content-based filtering [14, 16] helps to avoid cold start problem by recording information of users or items in the time of recording profile. During the recommendation phase, the system finds the similarity of the items that the user positively rated, and he/she didn't rate. In this filtering approach [33], similar products that were rated by similar users in the past are generated as a recommendation.

B. COLABORATIVE-BASED METHOD

CF is a highly implemented technique of recommender system [12]. Its concept is finding similar users that share the same appreciations. In a collaborative filtering method, when two users having similar rating value interval, then the grouped into one cluster. Then users will get a recommendation of items that they haven't rated previously but already rated by a similar user in their group. So, in a collaborative filtering recommendation, items preferred by the people whose rating value is similar to that of user's rating value are generated as a recommendation. CF technique is the most prominent method to generate recommendations [20]. Collaborative filtering [9, 13] can be either user- based or item-based approaches.

a) USER-BASED APPROACH

In this approach, similarity is calculated between the users which mean users functions the major aspects. Then these users have the same taste will be grouped into one cluster. Nitin Pradeep and Zehen zhen Fan [5], stated that recommendations are given to the user when users from the same group whom he/she shares the same taste in the group. So, in the user-based, items which are positively appreciated by the users in the same group will be suggested to the user members in the group [13, 32].

b) ITEM-BASED APPROACH

The item-based approach is also a collaborative filtering approach which makes a recommendation on the concept of finding similarities among the items. Farida Karimova [1] stated that, in item-based collaborative filtering, users receives recommendations of items that are the same to that they positively appreciated in a past [5]. However, in a user-based, recommending the item is based on similar user where the similarity calculation is among the users, not items.

c) HYBRID APPROACH

As we see from the introductory part, both content- based and collaborative filtering have their own problems and limitations so to overcome those problems and get an effective result, researchers combine those two traditional approaches by hybridizing their techniques together [3, 16]. The concept of hybrid recommendation system [31] is aggregating of different techniques will obtain a more accurate and effective recommendation than a single traditional recommendation since one technique can be taken by another algorithm.

II. LITERATURE REVIEW

Until the present time, many papers have been presented to resolve the problems of traditional recommendation systems [1, 32]. Collaborative filtering [1, 5, and 12] was the most popularly used approach to build recommendation systems. However, there are some issues which down sides its performance like sparsity, cold-start, and scalability problems. So to overcome various approaches were proposed to discuss these problems. Nitin Pradeep and Zhenzhen [5] proposed a hybrid collaborative filtering method to discuss sparsity and scalability problems and meet a better-personalized item recommendation. For the sparsity issue, to fill the unfilled cells in the rating matrix they used a mean value method by the help of Euclidian similarity measure and complimented self-organizing map (SOM) with genetic algorithm (GA) for clustering users to solve scalabilty problem. On the other side, Vimala Vellacichamy and Vivekanandan Kalimuthu [10] have also proposed a model based collaborative RS to resolve scalibility and sparsity problems. They applied fuzzy c-mean (FCM) method for clustering the users and bat optimization to meet the first value to which they grouped the rest of the users. According to [10], fuzzy c-means method with bat algorithm is operated in two stages. In the first stage, FCM clusters similar users together according to similar tastes and then apply bat optimization to give most cluster center point. For measuring the accuracy of the proposed recommendation system, [5] they used MAE (Mean Absolute Error) as a statistical accuracy measure. According to their experimental result, MAE (Mean Absolute Error) of the classic Item-based approach is 0.22 and MAE for they proposed is 0.15 for each of 5-fold validation. Which indicates their proposed method provided a better suggestion quality than the classic item based. To overcome the problem of classic recommender systems, now most of the researchers apply heuristic techniques to get the best results [9]. Sambhav, Vikesh and Sushama [9] have applied the bat optimization algorithm with a collaborative RS to find a better section for the active users. Nitin Pradeep and Zhenzhen Fan [5] have also proposed a hybrid of Genetic optimization (GA) and Self-Organizing Map (SOM) to the collaborative filtering. A clustering algorithm is the most well-known algorithm used in recommendation systems commonly united with heuristic optimization methods to improve the effectiveness of the recommendation system. Clustering techniques are widely used in machine learning, image segmentation, data compression, pattern recognition, and statistical data analysis. In general [28, 34], the algorithms used in clustering methods categorized into two categories: hierarchical and partitioning. K- mean algorithm is a clustering technique that clusters data by selecting k-clusters randomly. It's easily implemented and processes a huge amount of data. However, its main problem is it selects the first k centric values randomly which results in a local ideal value. In the present time, researchers mostly combine the k-means algorithm with nature-inspired algorithms to get globally ideal solutions. For example, Rahul Kataria and Prakash verma [11] applied cuckoo search algorithm with k-means to the movie lens data set. They followed two steps; initially, they applied a K-means algorithm to movie lens dataset to have k number of clusters. Then, they grouped the rest of the users to the first k clusters depending on their difference distance from each centric. Next, to get the correct first centric value they applied a cuckoo search algorithm to find the ideal solution.

As indicated by the audit, there are such a significant number of RS papers executed utilizing web uses mining systems. In a RS, web usage mining techniques help to analyze weblogs and for suggesting users' interest. As we have discussed in the introductory part, users' information are captured and maintained in web log files [6, 19]. Sunil and Prof M.M Doja [3] developed an e-learning recommender framework utilizing web utilization mining. They developed a web portal and used web server logs to record learners who are visiting the website. In this e-learning recommendation engine, developers used a classification method to classify learners and clustering method for grouping pages often visited by the users. Also, Jinhyun Joao, Sanqulon Bangb, and Gueunduk Parka [4] developed a coupon RS using association rule and collaborative filtering. They used a K-NN algorithm to find customers shopping patterns and a priori algorithm for analyzing associaztion rule. Prajyot Lopes and Bidesha [6] proposed a RS on e-commerce area using web usage mining technique by analyzing lexical patterns. Through their work, they connected a clustering procedure for session grouping, which works explicitly for unregistered clients and client- based system for enlisted clients.

III. SYSTEMATIC REVIEW

In order to select the investigation papers for this paper, we have used various journal databases for our study such as research gate site mostly, IEEE Explore, Science direct and Springer link. This review has two main goals:

1. To identify which web mining algorithms are mostly used in an intelligent recommender system.
2. To identify which type of filtering methods for the recommendation they are used merely.

In the scope of this work, the authors are concerned with the review of finding publications, case studies and website links related to the field; therefore, this paper was investigated on:

- Related keywords of RS in E-commerce.
- The articles must address the current limitations.
-

Recently published articles and recommender systems that are improved by optimization techniques

IV. SYSTEMATIC REVIEW RESULT

Authors	Year	Methodology (Algorithm)	Aim
- Nitin Pradeep & Zhenzhen Fan	-2015	-Clustering with GA-SOM -Collaborative filtering	To solve the problem of scalability & sparsity issue.
- Vimala Vellaichamy & Vivekananda Kalimuthu	-2017	- Fuzzy C-Means -Bat Algorithm	- To diminish information sparsity and versatility issues..
Sambhav Yadav, Vikesh, Shreyam, Sushama Nagpal	-2018	-Clustering Bat Optimization	To improve drawbacks of collaborative & content-based filtering.
- Rahul Kataria & Om Prakash Verma	-2017	-k-mean algorithm and cuckoo search	To overcome the limitation of collaborative RS.
-Sajad Ahmadian, Mohsen & majid	-2019	-Clustering	To solve cold-start and data sparsity problems based on different reliability measures.
- Vibhar Kant, Tanisha Jhadani & Pragya Daivedi	-2017	- (FNBCF) fuzzy set theory and naïve Bayesian collaborative filtering	To handle ambiguous choices.
-Sunil & Prof M.M.Doja	-2017	-Clustering -Classification -Hybrid filtering	-To improve the structure of the E- learning website
- JinHyun Joao, SangWon Bangb, & Geun Duk Parka	-2016	-Association rule-based K- mean algorithm	Implementation of a mobile coupon recommendation system
- Prajyoti Lopes & Bidisha Roy	-2015	-Clustering	To reduce false positive errors.
-Ammar Abdusalam Neamah	-2018	-Classification methods (Decision Tree (DT), Naïve Bayesian & Nearest Neighbor	-To help students to find their desired content on E-learning
-T. Mombeini, A. Harounabadi, and J. Rezaeian Sheshdeh.	-2014	- Neural Network and fuzzy c-means algorithm	-To build a web RS using neural network.
- Prajyoti Lopes and Bidisha Roy	-2014	-Clustering	-To provide a real-time recommendation.
- Susi Maulidiah, Imas S. Sitanggang, and Heru Sukoco	-2018	-Map sequential pattern mining using Bitmap Representation (CM-SPAM) algorithm	-To explore hidden patterns from visitors profile.
- Hernando A, Bobadilla J, and Ortega F	-2016	-Matrix factorization and model-based RS.	Present a novel procedure for foreseeing the flavor of clients in RS dependent on the factorization of the rating lattice.
- Hamidreza Kooh, Kourosh Kiani,	-2016	Fuzzy C-means method	-To build CF for movie lens dataset using fuzzy c-mean clustering.
-Dimple Trivedi, and Swati Tahiliani	-2018	-K-nearest Neighbors (KNN) search algorithm -Correlation Coefficient	-To produce the most suitable web page prediction according to the past web navigation history using WUM and web content mining (WCM)
- Phongsavanh Phorasim and Lasheng Yu	-2017	-K-means Clustering with collaborative filtering	-To develop a movie RS using CF technique and K-means
- Boli, Yibin Liao, and Zheng Qin	-2014	-Clustering techniques (density-based clustering (DBSCAN), and hierarchical clustering)	-To implement movie RS using distance matrix based.

Table 1: Summary of the related work

V. RESEARCH GAPS

Pub. No	year	Used Dataset	Future Work
[5]	-2015	-Movie lens dataset	- More experiment will be done on other data set. - Apply different similarity measures like cosine similarity. N.B. They used Euclidian distance similarity measure
[9]	-2018	-Jester Data set	- Fuzzily the rating into good, bad, and average. - Use social networked data.
[11]	-2017	- Movie lens	- Use other nature-inspired methods instead of cuckoo search. - Its limitation is efficiency may decrease when initial partition will not work well.
[10]	-2017	-Movie lens	- Further work on other features such as demographic information, trust, and contextual. N.B. Their aim was to improve Data sparsity and scalability problems.
[12]	-2017	-CiaoDVD -Movielens -Extended Epinions	- Implement on socio-contextual approaches. N.B. They implemented with various recommendation approaches such as baseline, and trust. And also presented process of preparing data set.
[23]	2017	-Movie Dataset	- Extend the work on other domains like music, books, and jokes. - For further improvement, incorporate the notation of trust and risk. - Furthermore, investigate to handle data error entry by users.
[18]	-2017	-Sobazaar (fashion clothes data)	- Extend the work on linearly dependent auxiliary feedback. - -User actions such as purchase(target data), product wanted, product _clicked, product_detail_viewed) N.B. their aim was to resolve data sparsity issue using multi-type auxiliary feedback (click, bookmark, remove, and reply as target feedback)
[21]	-2018	-Web access log data August 2016	- The frequent sequence patterns are used to formulate the recommendations for the development of the LKP XYZ managerial unit. However, with the current website materials that were accessed by the visitor cannot be extracted from the access log detail. Further research using web structure mining and web content mining should be done to obtain comprehensive recommendations for the LKP XYZ website improvement.
[33]	-2018	-Web access log data	- The proposed work is not considering the cold start problem so in the near future for the cold start problem the work is extended. - The proposed work just considers the restricted variables of web use and web content mining so sooner rather than later the web application factors are additionally included for improving the suggestions.
[34]	- 2017	-Movie lens dataset	- Further use fuzzy c-means technique to provide a more effective segmentation.
[35]	- 2014	-Articles, News, & Music	- Further analyze thus clusters having both a lot of negative and positive votes.

Table 2: Research gab summary

VI. CONCLUSION

Generally, this paper researched diverse prediction techniques in recommendation systems to circle studies and practices in the area of recommendation systems. To find the direction and sight of research in e-commerce recommender system for researchers to the future, this paper specifically focused on recently published papers. The others have analyzed each of paper in detail and presented the future works in the table form which is summarized by table 4. The study results collaborative filtering and hybrid method plays a prominent role in e-commerce domain particularly for movie recommendation. According to the survey, CF is widely applied for filtering method and combined with natural-inspired optimization techniques. Currently, CF is the uttermost outstanding recommendation technique. Moreover, the study shows that, the researchers have been focused to overcome the limitations of collaborative recommendation algorithms such as solving sparsity and scalability problems, improving computational complexity, and improving recommendation accuracy, etc. As you see from Table 2, the research gap summary table shows that there are still major limitations to be solved. On that account, this paper emphasizes to continue researching on the recommendation system confirm several issues remain to be addressed. To the future, researchers have to apply nature-inspired algorithms in more to improve the presentation of the recommender system.

REFERENCE

[1] Farida Karimova, "A survey of e-commerce recommender systems." European Scientific Journal December 2016 vol.12.No. 34 ISSN:1857- 7881. DoI:10.19044/esj.2016.vol. 2, Issue34, p75

[2] Annupama prasanth, "Web personalization using web usage mining Techniques". International journal of current engineering scientific Research, ISSN:2393-8374, Issue-3, 2016.

[3] Suni, prof. M. N. Doja, "Recommender System Based on Web Usage Mining for Personalized E-learning platforms". Department of Computer Engineering, Faculty of Engineering & Technology, Jamia Millia Islam, New Delhi, India, International Journal of modern Computer Science (IJMCS) ISSN: 2320-7868 (online) vol. 5, Issue 3, June 2017.

[4] Jinhyun Jooa, SangWon Bangb, Geun Duk Parka, "Implementation of a Recommendation System using Association Rules and Collaborative Filtering " Procedia Computer Science 91 (2016) 944-952

[5] Nitin Pradeep Kumar, Zhenzhen Fan. "Hybrid User-Item Based Collaborative Filtering". Procedia Computer Science 60 (2015) 1453 – 1461, Institute of Systems Science, National University of Singapore25 Heng Mui Keng Terrace, Singapore- 119615

[6]Prajyoti Lopes, Bidisha Roy. "Dynamic Recommendation System using web usage mining for E-commerce users". International Conference on Advanced Computing Technology and Application (ICACTA-2015).

[7]Wikipedia. https://towardsdatascience.com

[8]Jose Aguilar, Priscila Valdiviezo-Dı´az, Guido Riofrio. "A general framework for intelligent recommender systems". Applied Computing and Informatics (2017) 13, 147–160

[9]Sambhav Yadav, Vikesh, Shreyam, Sushama Nagpal. "An Improved Collaborative Filtering Based Recommender System using Bat Algorithm". International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Procedia Computer Science 132 (2018) 1795–1803, Netaji Subhas Institute of Technology, Sector-3, Dwarka, Delhi 110078, India

[10]Vimala Vellaichamy, Vivekanandan Kalimuthu. "Hybrid collaborative Movie Recommender System Using Clustering and Bat optimization". International Journal of Intelligent Engineering and systems, vol.10, no.5.2017 DOI:10.22266/ijies2017.1031.05

[11]Rahul Katarya, Om Prakash Verma. "An effective collaborative movie recommender system with the cuckoo search". Egyptian Informatics Journal 18(2017) 105-112, Department of Computer Science & Engineering, Delhi Technological University, Delhi, India.

[12]Maryam Jalloulia, Sonia Lajimi, Ikram Amous, "Designing Recommender System: Conceptual Framework and Practical Implementation". International Conference on Knowledge- Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseile, France. Peer- review under responsibility of KES International 10.1016/j.procs.2017.08.195

[13]Daniar Asanov. "Algorithms and Methods in Recommender Systems". Berlin Institute of Technology Berlin, Germany.

[14]Santosh Kumar Uppada. "Centroid Based Clustering Algorithms- A Clarion Study". (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014, 7309-7313, ISSN: 0975-9646, PYDHA College of Engineering, JNTU-Kakinada Visakhapatnam, India.

[15]K Lakshmi, N Karthikeyan Visalakshi And S Shanthi. "Data clustering using K-Means based on Crow Search Algorithm". https://doi.org/10.1007/s12046-018-0962-3

[16]Sajad Ahmadian, Mohsen Afsharchi, And Majid Meghdad, "A novel approach based on multi-view reliability measures to alleviate data sparsity in recommender system." https://doi.org/10.1016/j.asoc.2019.01.026.

[17]"A novel approach based on multi-view reliability measures to alleviate data sparsity in recommender systems". https://doi.org/10.1007/s11042-018-7079-x

[18]Ammar Abdusalam Neamah (2018). "Developing a recommender system using web mining". DOI:10.13140/RG.2.2.24910.72001

[19]Guo G, Qiu H, Tan Z, Liu Y, Ma J, Wang X (2017) Resolving data sparsity by multi-type auxiliary implicit feedback for recommender systems. Knowl-Based Syst 138:202–207

[20]T. Mombeini, A. Harounabadi, and J. Rezaeian Sheshdeh. "Constructing a web recommender system using web usage mining and user's profiles". Doi: 10.5267/j.msl.2014.11.010

[21]Sonali B. Ghodake and Ratnamala S. Paswan. "Efficient recommender system using collaborative filtering technique and distributed framework". Vol. 03 Issue:09/sep-2016 e-ISSN:2395-0056

[22]Susi Maulidiah, Imas S. Sitanggang, and Heru Sukoco. "ICT recommendation using web usage mining". I.J. Information technology and computer sconce, 2018, 12, 21-26. DOI:10.5815/ijitcs.2018.12.03

[23]Hernando A, Bobadilla J, Ortega F (2016) "A non-negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model" Knowl- Based Syst 97:188–202

[24]Vibhor Kant, Tanisha Jhalani, Pragya Dwivedi. "Enhanced multi-criteria recommender system based on fuzzy Bayesian approach". DOI 10.1007/s11042-017-4924-2.

[25]T.K.Das. "Intelligent Techniques in Decision Making: A Survey". Indian Journal of Science and Technology, Vol 9(12), DOI: 10.17485/ijst/2016/v9i12/86063, March 2016 ISSN (Online) : 0974-5645. School of Information Technology and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India.

[26]Ms. Dipa Dixit, Mr. Jayat Gadge. "Automatic recommendation for online users using web usage mining". International Journal of Managing Information Technology (IJMIT) Vol.2, No.3, August 2010. DOI: 10.5121/ijmit.2010.2303

[27]Prajyoti Lopes and Bidisha Roy. "Recommendation system using web usage mining for users of E-commerce site" International Journal of Engineering Research & Technology (IJERT), ISSN:2278-018. Vol. 3 Issue 7, July-2014

[28]Zolghadr-Asli B., Bozorg-Haddad O., Chu X. (2018) "Crow Search Algorithm (CSA). In: Bozorg-Haddad O.(eds) Advanced Optimization by Nature-Inspired Algorithms". Studies in Computational Intelligence, vol 720. Springer, Singapore

[29]Alireza Balavand, Ali Husseinzadeh Kashan, and Abbas Saghaei, "Automatic clustering based on Crow Search Algorithm-Kmeans (CSA-Kmeans) and Data Envelopment Analysis (DEA)". International Journal of computational intelligence systems, Vol. 11 (2018) 1322-1337.

[30]Alireza Askarzadeh,"A novel metaheuristic method for serving constrained engineering optimization problems: Crow Search

[30] *Alireza Askarzadeh, "A novel metaheuristic method for serving constrained engineering optimization problems: Crow Search Algorithm”. http://dx.doi.org/10.1016/j.compstruc.2016.03.0 01 0045-7949/ 2016.*

[31] *F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, “Recommendation systems: Principles, methods, and evaluation.” Egyptian Informatics Journal (2015) 16, 261–273*

[32] *Hamidreza Kooh, Kourosh Kiani, “User-based collaborative filtering using fuzzy c-means.” http://dx.doi.org/10.1016*

[33] *Dimple Trivedi, Swati Tahiliani, “An implementation of web recommendation system using web usage mining techniques.” International Journal of Computer Applications (0975 – 8887) Vol. 182 – No. 19, October 2018*

[34] *Phongsavanh Phorasim and Lasheng Yu, “Movies recommendation system using collaborative filtering and k- means” International Journal of Advanced Computer Research, Vol 7(29) ISSN (Print): 2249-7277 ISSN (Online): 2277-7970.http://dx.doi.org/10.19101/ IJACR.2017.729004. School of Information Science and Engineering, Changsha, Hunan, China*

[35] *Bo Li, Yibin Liao, and Zheng Qin, “Precomputed Clustering for Movie Recommendation Systemin RealTime” Journal of Applied Mathematics. Vol. 2014, Article ID 742341, 9 pages. http://dx.doi.org/10.1155/2014/742341*

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.

The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005

International Journal of Software Engineering & Systems

SUBSCRIPTION FORM

SUBSCRIPTION PRICES

India	Institutional	Individual
Print	3000 INR □	2000 INR □
Print + Online	3000 INR □	3000 INR □
Online	1500 INR □	800 INR □

Rest of the world	Institutional	Individual
Print	300 USD □	200 USD □
Print + Online	400 USD □	300 USD □
Online	200 USD □	150 USD □

Contact Person : ______________________________

Designation: ______________________________

Institution Name: ______________________________

Address : ______________________________

____

City : ____________________ Pin : __________ State : __________

Tel : ____________________ Fax : __________ Email : __________

PAYMENT OPTIONS

□ Cheque / DD is enclosed in favour of **" Apex Subscription Pvt. Ltd."** Payable at Mumbai.

Amount : ____________________

Cheque / DD No : ____________________ Dated : ____________________ Drawn on Bank

____

□ **NEFT / RTGS :**

Beneficiary Name : Apex Subscription Pvt. Ltd.

Bank Name : HDFC Bank Ltd, BRANCH : Goregaon (East),

Bank A/C No : 02122320004035, IFSC Code: HDFC0000212

Note : For more subscription detail follow the link www.apexsubs.com

Editorial Office
Enriched Publications Pvt. Ltd
S-9, IInd Floor Above Federal Bank,
Plot No-5, Sector-5, Dwarka, New Delhi- 110075,
Phone: - + (91)-(11)-47026006
Email : info@enrichedpublications.com
Website : www.enrichedpub.com

Subscription Manager
APEX SUBSCRIPTION PVT LTD.
304, De-Elmas, Sonawala Cross Road 2,
Goregaon (East) Mumbai - 400063.
Tel : +91-022-26853114/5 Fax : 022-26853116
Email : enquiry@apexsubs.com
Website : www.apexsubs.com

Notes: