

2017

**Proceedings of International
Conference on:
Information, Communication and
Computing Technology
(ICICCT 2017)**

13th May 2017



Organized by:



JAGAN INSTITUTE OF MANAGEMENT STUDIES

3, Institutional Area, Sector-5, Rohini,

New Delhi, India

Copyright © 2017 International Conference on Information, Communication and Computing Technology (ICICCT 2017)

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law and such quoting of the text or content of the book should be done in due manner of referencing to this book. For permission requests, write to the publisher, addressed “attention: permissions coordinator,” at the email below.


contact@edupediapublications.org

ORDERING INFORMATION:

Quantity sales. Special discounts are available on quantity purchases by corporations, associations, and others. For details, contact the publisher at the address above.

[Https://edupediapublications.org](https://edupediapublications.org)

BOOK SPECIFICATIONS

Name of Conference	International Conference on Information, Communication and Computing Technology (ICICCT 2017)
Vol. Editors	Dr.Latika Kharb & Dr.Deepak Chahal
ISBN NO. for Print Proceedings of Conference	978-93-86647-85-6
Publisher of Print Proceedings of Conference Papers	EduPedia Publications Pvt Ltd, New Delhi
Date of Conference	May 13, 2017
Venue of Conference	India International Centre (IIC) 40, Max Mueller Marg, New Delhi - 110003, India.
Conference Organizers Name and Details	Dr.Latika Kharb & Dr.Deepak Chahal Associate Professor (IT), JIMS,Sector-05,Rohini, Delhi,India.
Publisher & Printed By  P E N 2 P R I N T [®]	EduPedia Publications (P) Ltd D/351, Prem Nanar-2, Kirari, PIN-Code 110086, New Delhi, India Contact : +919557022047 or +919958037887 Email : contact@edupediapublications.org Website: https://edupediapublications.org Or http://edupediapublication.com

Preface

The International Conference on Information, Communication and Computing Technology (ICICCT 2017) was held on May 13, 2017 in New Delhi, India. ICICCT 2017 was organized by Department of Information Technology, Jagan Institute of Management Studies (JIMS) Rohini, New Delhi, India. The conference received 219 submissions and after rigorous reviews, 34 papers were included in Springer CCIS volume 750 and 18 papers are selected for this volume. The acceptance rate paper was around 15.5%. The contributions came from diverse areas of Information Technology that has been categorized into three tracks, namely:(i)Network systems &Communication security (ii)Software Engineering (iii)Algorithm & High Performance Computing.

The aim of ICICCT 2017 is to provide a global platform to the researchers, scientists and practitioners from both academia as well as industry to present their research and development activities in all the aspects of :Network systems & communication security, Software Engineering and Algorithm & High Performance Computing.

We thank all the members of the Organizing Committee and the Program Committee for their hard work. We are very grateful to **Dr. Vasudha Bhatnagar** from Department of Computer Science, University of Delhi as keynote speaker, **Dr.Maya Ingle** from SCSIT, Devi Ahilya Vishwavidyalaya, Indore as Guest of Honor and **Dr. Saroj Kaushik** from Department of CSE, Indian Institute of Technology (IIT) Delhi, **Dr. Sonajharia Minz** from Department of SCSS, Jawaharlal Nehru University (JNU), **Dr. Kapil Sharma** from Department of CSE, Delhi Technological University (DTU) as session chairs. We thank all the technical program committee members and referees for their constructive and enlightening reviews on the manuscripts.

We are pleased to present the second version of proceeding of the conference as its published record. We want to thank all of you, who have contributed to ICICCT-2017 and hope to see you in 2018 in our next international conference but with the same amplitude, focus and determination.

September, 2017

Patron

Dr.V.B Aggarwal

Conveners

Dr.Latika Kharb

Dr.Deepak Chahal

CONTENTS

S.no	Title	Page No.
1	Comparison of Group Deregistration Scheme and Explicit Deregistration Scheme in PCS Network Rajeev Ranjan Kumar Tripathi	1
2	ICT implementation- GIS and Smartphone application for disaster evacuation Dr.Bhoomi Gupta, Dr.Sachin Gupta	5
3	Implementing Cloud Based Approach to Maintain Household Routine Transactions Bansi Khimani, Prof. Kuntal Patel	17
4	Performance Analysis of K-Nearest Neighbor and K-means clustering to predict the diagnostic accuracy Kavita Mittal, Prerna Mahajan	26
5	Perception And Utilization Of Social Media Network Among Adolescents In Samaru Community, Sabongari Local Government Area, Kaduna State, Nigeria Hussaini Suleiman, Dr. Rajeev Vashistha	37
6	Error Detection by Checksum Dr. Anil Kumar Singh	48
7	Desirable Features for an Effective Sentiment Analysis System Sujata Rani, Parteek Kumar	54
8	An Efficient Algorithm for Data Field Extraction and Data Cleaning to improve performance of Web Usage Mining Preeti Rathi, Dr (Mrs). Nipur Singh	62
9	Comparative Approach Of Various Image Denoising Techniques Using Filters	73

	Amanpreet kaur	
10	Comparative study on various retinal vessel segmentation techniques Naina Singh, Aarti	91
11	Blood Mate – An Android Application for blood donors and receptors Kavita Pabreja, Akanksha Bhasin	102
12	Survey on big data analytics for cleaner manufacturing and maintenance processes Pawandeep kaur, Dr. Pankaj Deep Kaur	113
13	A Comparative Study of Classification Approaches for Entity Linking in Semantic Web Amit Singh, Aditi Sharan	126
14	New Paradigm For Software Design: ADML Namrata Sharma, Prerna Tyagi, Vaibhav Vyas, Rajeev G Vishvakarma	136
15	Analysis of Stability and Convergence on Perceptron Convergence Algorithm Vaibhav Kant Singh	149
16	A Review of Cyberbullying Detection in Social Networking Prankit Namdeo, R.K Pateriya, Sonika Shrivastava	162
17	Study on threats and improvements in LTE Authentication and Key Agreement Protocol Ritu Rani, Sandeep Kumar, Hitesh Sharma, Dr. Munish Mehta, Ms. Poonam Saini	171

18	A study on Self Healing Functionalities of Self Organizing Networks Anshita Singh, Srishti Shukla, Poonam	177
----	---	-----

Comparison of Group Deregistration Scheme and explicit Deregistration Scheme in PCS Network

Rajeev Ranjan Kumar Tripathi

Buddha Institute of Technology CL-1, Sector-7 , GIDA, Badgahan,
Gorakhpur, Uttar Pradesh(273209), India
rajeevranjankumartripathi@gmail.com

Abstract. Whenever an MT leaves its registration area and enters into a new registration area but the serving VLR remains same, location update of the MT takes place at the VLR end only and the HLR is kept unchanged. Due to move when MT enters into a new registration area which is being served by a new VLR this change in location is updated at both end current VLR and the HLR and the HLR also informs the previous VLR to delete the profile of this moved MT from its database. In this way deregistration of an MT takes place in PCS network and this scheme is referred as explicit deregistration scheme. Explicit deregistration scheme allows the removal of stale entries in real-time when an MT changes its location. In group deregistration scheme information about the MTs is acknowledged by the HLR when it receives a registration request from a VLR. This paper compares group deregistration scheme over explicit deregistration scheme. Conclusion shows that group deregistration scheme is efficient over explicit deregistration scheme and it removes stale entries when mobility is very high.

Keywords: VLR, HLR, registration area, PCS network, explicit deregistration scheme, group deregistration scheme.

1 Introduction

Currently GSM (Global System for Mobile Communication) and IS-41 standard are using two-level database scheme to support the mobility of an MT. In this two level database scheme one entity is known VLR (Visitor Location Register) and another is called HLR (Home Location Register). In HLR information about the MT is kept on permanent basis and in VLR subset of this information is kept on temporary basis only when MT is residing in the RA (Registration Area) which is being served by the same VLR. Whenever MT moves from one RA to other RA and this move causes the change in the VLR, information of the MT is deleted from the VLR and provided to new VLR by the HLR when new VLR sends registration request. Entire steps are being summarized below as:

Step 1:The mobile terminal leaves an RA and enters into a new RA and sends a location update message to the nearby base station.

Step 2:The base station forwards this message to the new serving

MSC/VLR.

Step 3:The new MSC/VLR updates its associated VLR, indicating that the mobile terminal is now residing in its services area and sends a location registration message to the HLR.

Step 4: The HLR sends a registration acknowledgment message to the new MSC/VLR together with a copy of the subscriber's user profile.

Step 5: The HLR sends a registration cancellation message to the old MSC/VLR

Step 6:The old MSC/VLR removes the record for the mobile terminal at its associated HLR and sends a cancellation acknowledgment message to the HLR.

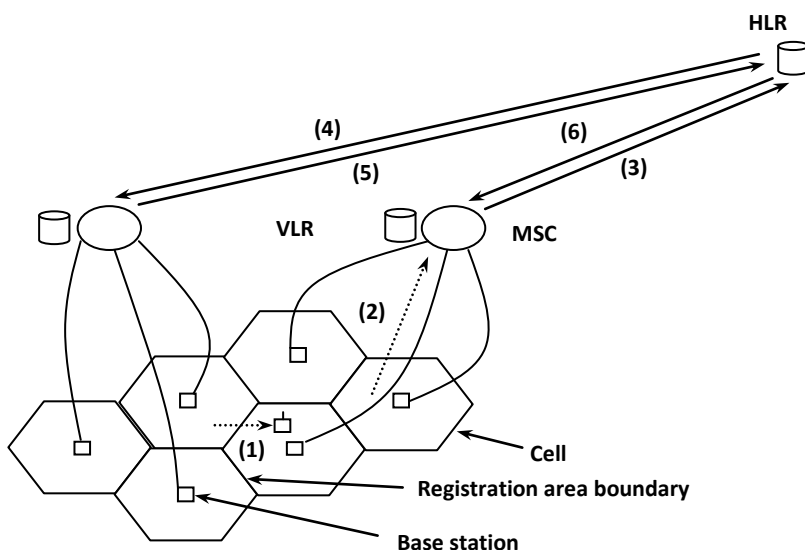


Fig. 1.Explicit de-registration

While changing an RA we have two possibilities: either VLR will be changed or not. Now we can differentiate the user's mobility in two ways.

1.1 Intra-VLR Move: This move occurs when MT's RA changed but VLR is same. In this case, MT's current location information is updated in the VLR. This move does not affect the HLR.

1.2 Inter-VLR Move: This move occurs when MT changes its RA and new RA is being served by the new VLR. This move is shown in fig1. It has six steps.

In Intra-VLR move HLR is not updated only VLR is updated by changing the location of current RA. In Inter-VLR move three databases are consulted

for updation.

In new VLR entry of MT is made, HLR updates the current VLR for MT and HLR sends deregistration request to the previous VLR to delete the MT information. In this way link is traversed four times and hence Inter-VLR move is more costly than the inter-VLR move [5], [6].

2 Group Deregistration

Whenever an MT leaves a VLR_A and comes under the RA of a new VLR_B, the VLR_B sends a location registration request to the HLR. The HLR knows that this MT is coming from the VLR_A, the HLR keeps the information of the MT in the OML (Old Mobile List) of VLR_A. On reception of location registration message, HLR updates the location of MT by replacing VLR_A by VLR_B and sends acknowledgement to VLR_B with the OML of VLR_B which was maintained by the HLR. When VLR_B receives acknowledgement from HLR, it registers the MT and removes all the entries from its database which are there in the OML. Note that for every VLR, the HLR maintains an OML and this OML is always sent to the VLR whenever HLR receives a location registration request.

- 1) When an MT performs inter-VLR move, the new VLR sends a location registration message to the HLR.
- 2) The HLR does not send any location cancellation message to the old VLR (from where MT is coming). The information that MT has changed its VLR is kept into an OML (for every VLR, the HLR maintains an OML separately).
- 3) The HLR sends acknowledgement to register the MT to new VLR along with the OML of this VLR.
- 4) On reception of this acknowledgement, the MT is registered at new VLR and this VLR deletes the profiles of all those MTs which are there in the OML.

In this way entries of all the left MTs are deleted from the VLR when an MT performs an inter-VLR move [1], [2], [3], [4]

3. Comparison of Explicit Deregistration Scheme & Group Deregistration Scheme

In explicit deregistration scheme, removal of left MT's profile takes place when an inter-VLR is experienced by the HLR. In group deregistration scheme, removal of this information is kept on hold till the HLR does not receive a location registration request from a VLR. In group deregistration scheme only two databases are affected and link is traversed in a pair only hence group deregistration scheme is less expensive than the explicit deregistration scheme. Important issue is removal of stale entries, in explicit

deregistration scheme old VLR is immediately informed by the HLR to cancel the location registration of the MT whenever inter-VLR move is performed but in case of group deregistration the profile of left MT is in the VLR till the HLR does not receive any location registration message from it. In other words, OML is sent to a VLR when location registration request is sent by VLR to the HLR. If mobility is very high and MT are very frequently performing the inter-VLR move then and only then group deregistration scheme can be applied instead of explicit deregistration scheme [1], [2], [3], [4].

4 Conclusion

Cost incurred in location management is generally measured by two factors, how many times links are being traversed and how many times databases are being consulted (for location read and location writes). These two factors are lower in number in case of group deregistration scheme than explicit deregistration scheme. Other important factor is removal of stale entry from the database. In real time communication a stale entry must be noticed and deleted when it is found. During a very high degree of mobility when inter-VLR move is being performed at VLR end only guarantees that network will be free from stale entry if group deregistration scheme is being used else not. Hence this paper recommends group deregistration scheme only in the scenario when mobility is very high else recommends explicit deregistration scheme.

References

1. Rajeev R. Kumar Tripathi, G.S. Chandel, Ravindra Gupta, "A Survey on De-registration Schemes in PCS Network," IJSER, Vol.4, Issue 6, June-2013.
2. Rajeev R. Kumar Tripathi, G.S. Chandel, Ravindra Gupta, "Profile Forwarding Scheme in PCS Network," IJSER, Vol.4, Issue 6, June-2013.
3. Rajeev R. Kumar Tripathi, G.S. Chandel, Ravindra Gupta, "Performance Analysis of Existing Location Management Scheme with its Variants in PCS Network," IJCA, Vol.68, No. 16, April 2013.
4. Rajeev R. Kumar Tripathi, G.S. Chandel, Ravindra Gupta, "Multi HLR Architecture for Improving Location Management in PCS Network," IJCA, Vol.51, No.20, August 2012.
5. Rajeev R. Kumar Tripathi, Sudhir Agrawal, Swati Tiwari, "Performance Analysis of De-registration Strategies in Personal Communication Network," IJCA, Vol.24, No.16, June 2011.
6. Rajeev R. Kumar Tripathi, Sudhir Agrawal, Swati Tiwari, "Modified HLR-VLR Location Management Scheme in PCS Network," IJCA, Vol.6, No.5, September 2010.

ICT implementation- GIS and Smartphone application for disaster evacuation

Dr.Bhoomi Gupta¹, Dr.Sachin Gupta²

¹Department of Information Technology,
Maharaja Agrasen Institute of Technology, New Delhi, India

²Department of Computer Science Engineering,
MVN University Haryana, India

¹guptabhoomi@gmail.com

²sachin.gupta@mvn.edu.in

Abstract

An application has been developed to handle the situations of disaster efficiently by use of Information and Communication Technology (ICT). This application shall aim at providing an effective solution during disaster and help in minimizing its impacts. Application not only incorporates the maps of safe zones and evacuation zones but also at same time shall be able to send push notifications in alerting users about disaster in advance which directly shall help in restricting the impact of a disaster. The application shall make use of Java language in android application development, XML for layout designing and PHP for making web pages. There shall be an application component called an activity that shall provide a user interface with which users can interact in order to do something, such as dial the phone, take a photograph, send an email, or view a map.

Keywords— Disaster, ICT, Evacuation.

i. Introduction

The unique climatic conditions cascaded with socio-cultural-economic and political diversities of India have made it exceptionally vulnerable to both natural and manmade disasters. India has a long history of coping with and recovery from disasters. The recurrences of natural disasters like cyclones have increased essentially over the last decade particularly in the coastal line of Odishain India which is affirmed with the impact of climate change. The state of Odisha on the east coast of India is called the disaster gateway as it is one of the most vulnerable states in the country with a very high probability of cyclonic hits.

The present research analyzes that the country has had a series of studies on the mathematical modeling of the cyclonic surges or social analysis of the human apathy. However, there are not many cases where a combination of information, technology and humanistic approach is being tested.

The research intends to bridge the gap between the Information Communication Technology (ICT) and the need. The permutations of social networks and spatial analysis is seldom tried with an outcome poised to aid

the community in being more aware and informed thereby leading to sustainable Disaster Risk Reduction(DRR) and resilience. Over the past decade, numerous analysts and decision makers have invested resources into developing geospatial (GIS) databases to support their risk management decision making. Most of these models generate results that are spatially explicit (mapped). Little work has been done to link the models with the GIS databases such as road networks, building locations and key utility databases in a decision support framework. Such initiatives can be brought into an integrated GIS decision support system.

Complete prevention of natural disaster is beyond human capabilities but the involvement of the state-of-the-art Information and Communication Technology (ICT) systems are panacea for implementing reliable disaster prevention and mitigational measures. ICT provides capabilities that can help people grasp the dynamic realities of a disaster more clearly and help them formulate better decisions more quickly.

The paper aims to address the role of ICT infrastructure as a framework for facilitating disaster management in the study area of Kendrapara, one of the most cyclone hit area of Odisha. The ubiquity of handheld computing technology has been found to be especially useful in disaster management and relief operations. A typical design and implementation of ICT based smart phone application has been executed comprising of communication, decision support and early warning components for enhancing the efficiency and effectiveness in all. This disaster alert and rescue management system has been implemented on Google's Android development environment as a smartphone application, which endows the decision makers, volunteers and the general community with options of optimal routes across various geographical areas leading to safe zones and pertinent information henceforth.

The application has been tested effectively for the pre-defined data set of coastal villages of Kendrapara. Recursive Decomposition Algorithm has been applied for optimization and different parameters have been evaluated to determine the most optimum routes. The results have been hosted on the server, tested with the district database and results verified.

The application exhibits emergency calls, reporting system, disaster alerts and geovisualization as its key features. The application allows field data reporting, sending geo-location SMS, cell broadcasting, viewing and retrieving location information on the mobile device. An attempt, through this research, has been made to realize the utility of the ICT in a disaster scenario and successfully achieved.

ii. Methodology

A. Evacuation Planning and Transportation Modeling

The methodology carried out is to map evacuation routes on the transport (road) networks of the identified villages of Kendrapara for massive

evacuation using optimization of the carrying capacity of the roads and the maximization of the capacity handling at the end/destination points. After identification of the village i to be evacuated from the stage of vulnerability analysis, a methodology is developed to identify the topological network of the routes/roads of village i based on the demographic features of that village. This result in the computation of an optimization model that can be used to identify and evacuate to the safer zones (say the cyclone shelters).

This model integrated with a GIS System can be used to make reference (ready real time scenario based) maps (road wise) for the evacuation risk minimization.

The research shall keep the scope limited to road networks only which have witnessed cyclonic destruction, the emergency response of evacuation being short term and it is assumed that the evacuation zones shall have been so carefully planned that each of them will have a sufficient demand after cyclone and that traffic originating from different evacuation zones can be destined to any safe zones. Subsequent to findings of the previous steps, a more realistic post cyclone travel model is formulated to analyze a realistic scenario. The traveler's behavior and routes change significantly, and it necessitates the travel demand modeling, to establish spatial distribution of the travel between the travel areas.

Evacuation From the village i (impact zone) to village j (safe zone) depend on a number of factors:

Number of people in need or in demand of the evacuation i.e., $demand_pop_evac (pop_{vi})$

Transport Carrying (vehicle load) for population pop_{vi} for village i to village j , i.e., $Trans_{(vij)}$

Rate at which the demand is fulfilled (R_D)

Rate at which the capacity is actually provided (for shelters), i.e., R_{cap}

Finally, $R_{cap} - R_D = \text{People at Loss (i.e., Human Loss)}$

e.g., during a cyclone, an area needs to be evacuated, even though the bulk capacity of all vehicles may be large enough to carry out all the people in danger out of that area, the maximum flow rate of the transport or road network (number of lanes available) may become a limiting factor. e.g., there may be no information about the shortest paths in the topological graph of the road network, e.g., or a shortest path (P_i) may not be practically feasible to be used because of some obstructions etc. For this reason, emergency planning zones (EPZ) also are sketched well in advance which help in evacuation planning . The procedure is two a step process:

Step 1: To build a topological graph for the road network

- a. Build up a topological (road based) network from the GIS map of the district available. Sketch a graph (directed) out of this network.
- b. Build up nodes and edges of that graph in such a way that the nodes are assigned as villages and edges represent total traveling distance on these roads. Now we find a shortest path /optimal path on that graph from node i to node j and also create a subset of shorter paths on every node or intersecting points.

Step 2: To calculate the Evacuation Risk Factor- The carrying capacity of the existing vehicles to carry the total population in demand is estimated .

B. Formulation of the problem as an Optimal Routing Problem (ORTP)

These set of problems deal with routing from a specific set of source nodes to a set of destination nodes through a transportation network. Denoting the symbols used in the equation for minimizing Z, where Z=linear combination of Dis_{min} and $T_{cl min}$.

i.e., $Z = \text{fn}(\text{linear}) \{Dis_{min}, T_{cl min}\}$

$$\text{where } Dis_{min} = \sum \text{pop}_{vi} \sum (x_{ijk} * d_{ijk})$$

for k = index for kth shortest paths;

if k= 1 (1st shortest path selected)

$$\Rightarrow Dis_{min} = \sum \text{pop}_{vi} \sum (x_{ij1} * d_{ij1})$$

\Rightarrow The evacuee will travel Dis_{min} only if he/she is assigned the first shortest route, therefore this Dis_{min} is the lower bound value for the evacuation paths set.

$$\Rightarrow \text{Similarly, } T_{cl (min)} = \sum t_1^{(1)} \sum \text{pop}_{vi} \sum (x_{ij1} * \alpha_{ij1})$$

where $\alpha_{ij1} = 1$ if the road link 1 is on the route present; 0 otherwise.

and $t_1^{(1)}$ = minimum expected travel time on the route selected.

C. Network reachability and reliability analysis of the transportation network.

The paper discusses the reliability of network reachability of transportation infrastructure systems. The paper explains not only where the evacuations originate from and where do they go, but also addresses what routes are to be followed. For transportation systems, the connectivity denotes the reachability of a random node-pair via at least one path. Connectivity is dependent upon the post-cyclone connectedness of a transportation network and hence it is a suitable metric for the case of immediate post-disaster humanitarian aid . The paper carries out the reachability analysis determining whether a path remains operational (or connected) between the given sources and destinations. If the path connects the selected node-pair following an impact, serviceability (or performance) analysis seeks

additional information on the remaining capacities that can be found mathematically.

The reachability analysis has been computed by Recursive Decomposition Algorithm (RDA) which is an analytical algorithm and which decomposes the topological network of Kendrapara area into sub graphs using DeMorgans' rule and the disjoint theorem till no path exists between the source destination pair in all sub graphs; In an emergency, it is critical to identify the passable ingress and egress routes for emergency response within a short time frame, e.g., to send search and rescue teams into the impacted area immediately after a disruptive cyclone. Immediately after a disruptive cyclone, emergency managers and rescue workers often face the problem of promptly identifying the emergency routes to send rescue teams and relief resources into the impacted area.

The analysis takes into consideration the key issues like road damage assessment and travel time estimations. After a disruptive event such as a major cyclone, the travel behavior (i.e., route choices) and travel demand could change significantly due to travelers' reaction to road damage, road closures, and congestions. Though it is still infeasible to obtain realistic behavior of traveler route choice and "real-time" travel demand, post-cyclone travel demand can be approximated with some general principles to capture the essential characteristics of post-cyclone travel patterns and effects of emergency facilities such as hospitals and emergency shelters. This approach approximates the "abnormal" travel demand by adopting several general principles.

The proposed methodology does not attempt to provide "real-time" post-cyclone traffic simulation. Instead, it aims at providing general principles and procedures for emergency training and planning purposes. The cyclone scenarios for demand modeling are created which specifically consider the occurrence time of day (e.g., morning and late-night period). In this paper, two hypothetical scenarios are developed to model the impact of a no-notice event on transportation systems—one occurring during morning rush hours (the day scenario), and another at late night (the night scenario). The hypothetical scenario cyclones will leave several roads (e.g., major river crossings) and essential facilities (e.g., schools) severely damaged. The research has been carried out with the following underlying assumptions:

- This study assumes that people will evacuate directly from their current locations immediately after cyclones. Social vulnerability to disasters such as race, gender, and social inequality has a crucial role in shaping the evacuation patterns, but is beyond the scope of this study.
- Contingent upon the vicinity of attractants (e.g., hospitals, dispensaries or emergency shelter) and repellents (e.g., harmed facilities); the TAZ (Traffic Analysis Zones) can be categorized into four zone types. The underlying suppositions are that: (i) if a

zone does not have damaged facilities, its trip production won't be influenced by the cyclone, while the trip production will increase in the influenced zones due to facilities damage and (ii) if a zone does not offer emergency shelters or hospitals, its trip attraction will stay unaltered; while the trip attraction will build in light of the vicinity of emergency shelters, medical dispensaries and hospitals.

- Emergency shelters and hospitals are assumed attractive sites to injured or displaced people. The attracted trips to shelters and hospitals are proportional to their capacities (e.g., the number of beds in a hospital).

D. ICT implementation - GIS for disaster evacuation modeling and rescue operation in Kendrapara district.

The plan of action for ICT implementation has been implemented on a QGIS (Quantum Geographic Information System) for efficient analysis and map building. Following are the steps undertaken for application building and further analysis:

- i. Map the mentioned areas on a QGIS as a vector layer. This can easily be done by entering the co-ordinates of the area or mapping them using an OSM (Open Street Map) layer plug in using the existing maps.
- ii. After we get the shortest route by analyzing the possible routes and using the algorithms mentioned in this documentation, we can map these routes by drawing appropriate lines.
- iii. Also, we can easily call the inbuilt function of distance matrix to get the minimum distances geographically between each point as shall be shown in the maps.
- iv. As we can see in these maps, the zones at potential risk are shown in red color and the evacuation zones are shown with a green color. Moreover in the same map layer in the QGIS, we can have a single database having all the relevant data like the nearest evacuation zone, its distance in kilometers, total population, number of males and females etc. Unlike before we don't need to refer different excel sheets and search for the relevant information corresponding to a relevant zone instead, we can just select the region in the QGIS layer and get all the required information at one glance.
- v. The use of QGIS can make data interpretation easier and can make analysis and creation of evacuation plans easier.

E. ICT implementation- Smartphone application for evacuation data dissemination in district (Kendrapara)

An application has been developed to handle the situations of disaster efficiently by use of Information and Communication Technology (ICT). This application shall aim at providing an effective solution during disaster and help in minimizing its impacts. Application not only shall have the maps of safe zones and evacuation zones but also at same time shall be able to send push notifications in alerting users about disaster in advance which directly shall help in restricting the impact of a disaster. The application shall make use of Java language in android application development, XML for layout designing and PHP for making web pages. There shall be an application component called an activity that shall provide a user interface with which users can interact in order to do something, such as dial the phone, take a photograph, send an email, or view a map. Each activity is to be given a window in which to draw its user interface. The required features are:

- i. Main Activity- It has to be the first screen that comes when application opens and shall have the options like About, Help, Gallery, Check Message, Register, Safe Zone, and Evacuation Plans.
- ii. Save Zones-It shall have a portable document format of maps and routes of safe zones which the user can access. It requires a PDF reader to be installed for reading the document.
- iii. Evacuation Plan- It has portable document format of maps and routes which the user will require for selecting the routes .It requires a PDF reader to be installed for reading the document.
- iv. Gallery-It contains the images of the various villages i.e. their geographical location.
- v. Register-It is a required activity which takes information from users and store it in online database hosted on a database server. It registers user for the push notification service.

iii. Results

- i. Evacuation map for Village ‘Baro’ in DistrictKendrapara

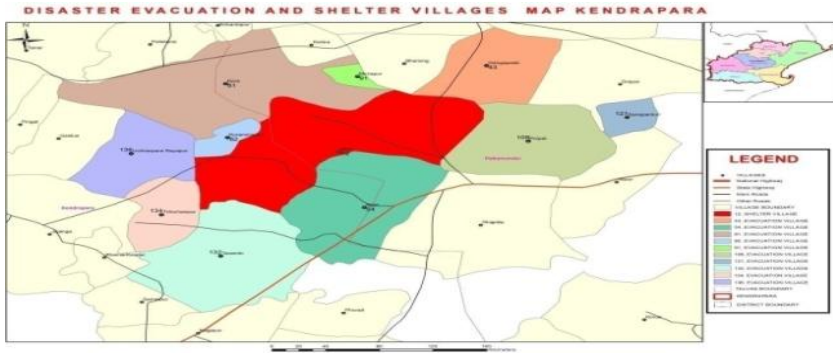


Fig1. Evacuation map: Village Baro (Shelter Zone)

Figure 1 illustrates the evacuation map depicting the shelter Baro(Village numbered 12 as depicted in map) and shows that the adjoining villages can be evacuated and rescued down to Baro. The color codes indicate the disastrous villages and the red color code indicates the shelter. This map has been created for reference in ICT applications such as web portals or smart phone devices; specific village views can help vulnerable and disaster hit people evacuate the place as soon as possible without any delay. If such kinds of maps are available on the ICT enabled devices like mobiles and smart phones, vulnerable people can be rescued and evacuated with much ease. A greater possibility lies in the hands of the rescue team to evade the disastrous places and to reach the most possible and safer shelter zones.

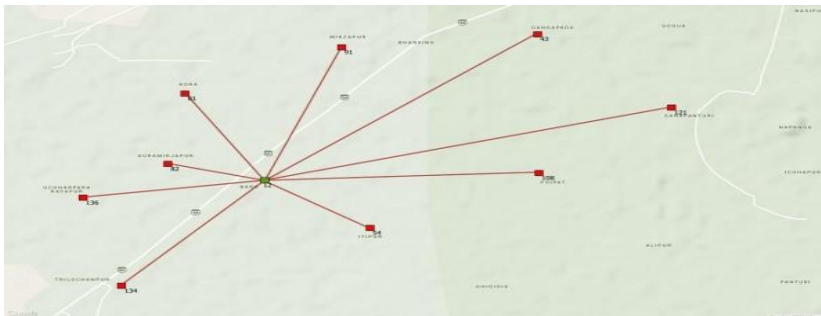


Fig 2. Evacuation Map (Public View: VillageBaro)

A simpler version of the map has been developed for the public in general as shown in figure 2. The analysis has been done with respect to the TAZ and developed in ArcGIS and QGIS platforms.

- ii. Evacuation Map with database view of vulnerability indices.

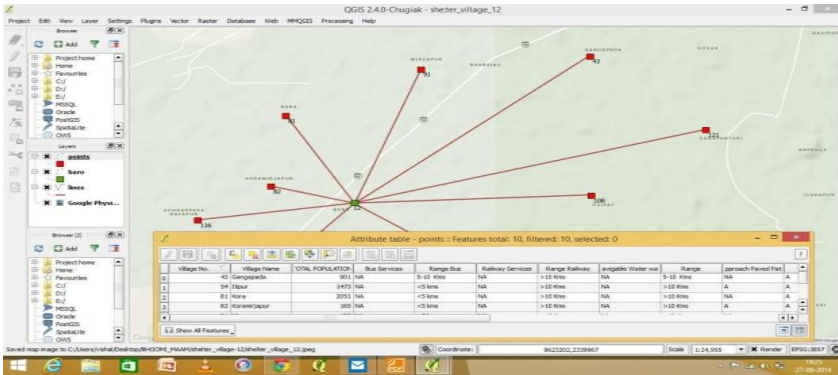


Fig 3. Database view of vulnerability indices in concern forevacuation with respect to village Baro.

Figure 3 shows the vulnerability indices, their values and all other social and demographic data related to the village Baro for quick reference to the administrators to take quick response and action regarding evacuation.

Table 1. Distance Matrix between Shelter Village (id: 12 indicated as InputID; Title: Baro) and Disaster hit Village(indicated as TargetID); Distance between shelter id and disaster id is given by Distance (measured in meters).

InputID	TargetID	Distance(m)
12	82	924.3
12	54	1173.1
12	81	1399.4
12	136	1701.0
12	91	1958.0
12	134	1962.7
12	108	2542.9
12	43	3225.1
12	121	3899.5

Table 1 is obtained and calculated from Google open street view and QGIS are found to be matchable with the distances computed from shortest path Recursive Decomposition Algorithmic analysis.

- ii. Smartphone application view: Disaster Alert.



Fig4. Smartphone application view of 'Disaster Alert' application.

iii. Registration procedure of the application 'Disaster Alert'



Fig 5. Application View and Menu Availability in 'Disaster Alert' Application.

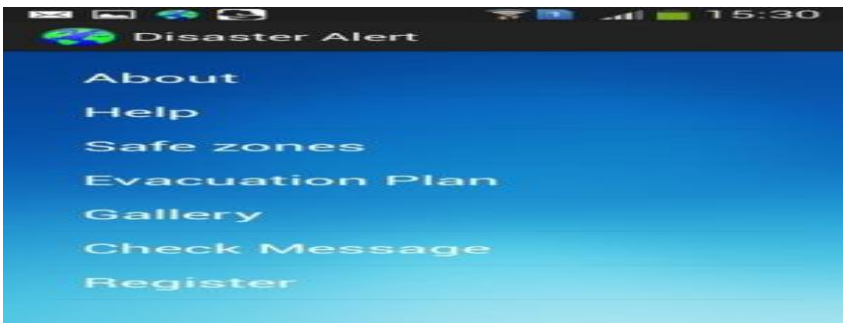


Fig 6. Menu of the 'Disaster Alert' application indicating options available for the user.

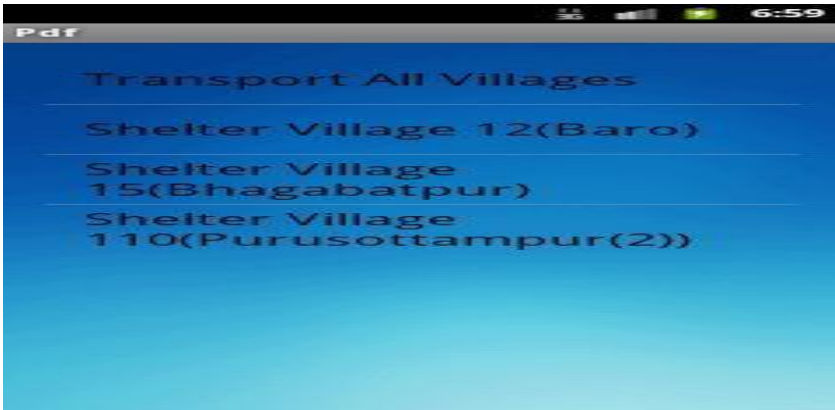


Fig 7. List of villages (example)available as shelter zones for evacuation.

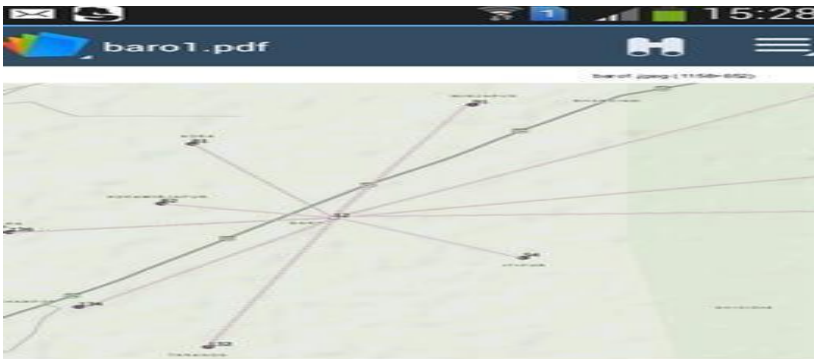


Fig 8. pdf file loaded into the gallery of 'Disaster Alert' for Village Baroas shelter zone and its routes to the evacuation villages.

Conclusions

The application developed facilitates the work of administrators and rescue teams to track their evacuation progress ceaselessly so that they can take immediate steps whenever needed. In brief, the developed methodologies shall allow state and local authorities; and emergency managers to perform:

- Risk Modelling and Vulnerability Analysis: Model risks, evaluate post-cyclone damage, and assess the performance and functionality of critical vulnerability indices; hence understanding disaster resiliencies for sustainable growth.
- Transportation Modeling and Reliability Analysis: Assess the contingency plans in a transportation network for emergency training and

response, secure the critical ingress and egress routes for emergency response as well as avoid excessive queues and delays.

- GIS Map Implementation: Implement Geographical Information System based maps for a ready reference to the evacuees and the administrators for quick decision making and self evacuation; hence causing optimization of the rescue operations.

- ICT Based Smartphone Application Development: Implement a smartphone application in the disastrous area for successful evacuation; results being presented in a handheld device like a smartphone for a ready reference to the evacuees.

References

1. Asian Disaster Preparedness Centre, <http://www.adpc.ait.ac.th>; Asian Disaster Preparedness Schemes.
2. Cyclone shelters and their locational suitability: an empirical analysis from coastal Bangladesh, Bishawjit Mallick, *Disasters*, Volume 38, Issue 3, pages 654–671, July 2014.
3. Das, Saudamini, “Storm Protection Values of Mangroves in Coastal Orissa”, in P. Kumar and B. Sudhakar Reddy (ed) *Ecology and Human Well-Being*, New Delhi, Sage Publications, (2007a).
4. Jayanthi, N., “Cyclone Hazard, Coastal Vulnerability and Disaster Risk Assessment along the Indian Coasts”, *Vayu Mandal*, 28 (1-4): 115-119 (1998).
5. “Near-Real-Time Analysis of Publicly Communicated Disaster Response Information”, Trevor Girard, Friedemann Wenzel, www.ijdrs.com, DOI 10.1007/s13753-014-0024-3

Implementing Cloud Based Approach to Maintain Household Routine Transactions

Bansi Khimani¹, Prof. Kuntal Patel²

¹School of Computer Science, RK University, Rajkot, India

²School of Computer Studies, Ahmedabad University, Ahmedabad, India

Abstract. Today's daily life involves many routine transactions. It is hard for us to remember transactions' amount; hence people use traditional pen-paper based methods to maintain their transactions related details. Some of the low cost daily transactions are with milkman, laundryman, local grocery and vegetable shops. Instead of maintaining such details on a paper for a month, if we use any application while performing such transactions, it would be very easy to summarize all such transactions at the end of month. This paper proposes implementation of Cloud Based Mobile Application which helps us in handling routine transactions with security. Proposed approach will also help us in solving any dispute between buyer and sellers. Implementation of proposed work will help society to simplify their routine activities.

Keywords: Routine Transaction, Cloud Services, Cloud Security, Cryptography

1 Introduction

According to the study, Indian users spend 3 hours 18 minutes on average everyday with their smartphones, of which one-third time is spent on apps. Also, there has been a 63 percent increase in app usage in the past two years, the study added [5]. Moreover, looking into statistics, Android users were able to choose 2.2 million apps in June 2016[6]. So, why should we not think about app which manages our routine transactions? It is very difficult to manage small transactions in routine life. It's important to keep accurate and complete records. For that, you need to organized, keep your records up-to-date and then hold on to them for few days or maybe for few years. Sometimes few records are not called important records but they are good records. Maintaining daily records are not easy to remember for long period of time. So, person maintains paper based records. It's a good idea to keep your paper records off the ground and in a dry place so they stay in good condition. For eliminating paper-based activity, one can go for electronic records. It can be backed up by person periodically but it is also the manual process to have the backup of your data. And backup of your

data must be placed in a secure place. So, to eliminate that redundancy, we can implement our transaction on the cloud.

Many businesses - large or small, use cloud computing today either directly (e.g. Amazon, Google) or indirectly (e.g. Twitter, Facebook). There are several benefits of cloud i.e. Reduction of costs, Universal Access, Flexibility, 24/7 Supportive Software etc...These benefits attract more and more people to use a cloud. Cloud providers provide various facilities to the user according to their needs. The main thing user don't compromise is – security of their sensitive data [3].

Here, we are focusing household activities or transaction which can be maintained by both (Customer and Service Provider). Initially, it would be convenient to implement proposed system in the local language.

We did the survey of various applications, where records are maintained about transaction done with the milkman, ironman, other pay-income related expenditure. Few applications provide better reports also. But, none of the application specifies authentication of the transaction. Few applications just work for customers only. In proposed system, we will introduce cloud security where data in transit, data at rest and Authentication of the user is also done by password facility. Moreover, from the particular date to other date, reports can also be generated by the user. Here, we provide facility to request extra service to a provider for the next day. So, there are more advantages compare to other existing application.

2. Literature Review:

2.1 Cloud Security:

Cloud computing can become more secure using cryptographic algorithm. Cryptography is the art or science of keeping messages secure by converting the data into non readable forms[4].To secure the data resided in the cloud, there are many cryptographic algorithms available in the market to encrypt the data. There are two main ways to encrypt the data - first one is classical way and other one is modern way. In classical approach, monoalphabetic and polyalphabetic way can be implemented. But now-a-days, modern approach is followed by most of the users. Here, again two classifications can be done of modern approach. Symmetric key and Asymmetric key are two modern approaches. Here, RSA, DSA, Diffie-Hellman are examples of Asymmetric key approach and DES, 3DES, AES, RC4 are examples of Symmetric key approach. Asymmetric encryption techniques are about 1000 times slower than symmetric encryption which makes it impractical when trying to encrypt large amount of data.[2]Following table 1 shows comparative analysis of various symmetric cryptographic algorithms.

[Table 1: Comparison of DES, 3DES, AES]

Factors	DES	3DES	AES
Created by	IBM	IBM	Vincent Rijmen, Joan Daemen
Year	1975	1978	2001
Key Length	56bits	168 bits(k1,k2,k3) 112 bits(k1 and k2)	128,192 or 256 bits
Round(s)	16	48	10(128 bit key),12(192 bit key), 14(256 bit key)
Block size	64bits	64bits	128bits
Speed	Slow	Very Slow	Fast
Security	Not Secure Enough	Adequate Secure	Excellent Security

AES is the new encryption standard recommended by NIST to replace DES in 2001. AES can support any combination of data and key length. The algorithm is referred to as AES-128, AES-192 or AES-256, depending on the key length.

S. Gurpreet et al. [2] Studied about the performance of different encryption algorithms such as RSA, DES, 3DES, AES. According to their research, AES algorithm is most efficient in terms of speed, time, throughput and avalanche effect.

B. Akashdeep et al. [1] Studied about Symmetric Algorithms for different encryption and encoding techniques, found AES to be good for key encryption and MD5 being faster when encoding.

In our proposed Model, We can encrypt our data using AES. AES is more secure than other symmetric algorithms and here, we also consider parameter as speed to encrypt data from client end.

3. Proposed Approach:

Proposed approach will perform 4 main functionalities. Currently scope is limited for specific transaction done between milkman and customers. Here, consider that one customer can have many milkmen and one milkman can have many customers. So, many to many relationship in the database can be done. Following are main the functionalities of proposed approach.

A. System Input:

The input of our system is simply password of customer. Password is taken by the system.

B. Authentication:

User's Authentication is a very important process. Here, entered Password has to be matched with the database and then the user is authenticated.

C. Encryption:

Entered Password will be encrypted with the data entered by user with AES encryption algorithm.

For encryption of information:

- Input password and other information
- Implement AES algorithm on information to generate Cipher Text.
- Store CipherText into Database.

For Decryption of information:

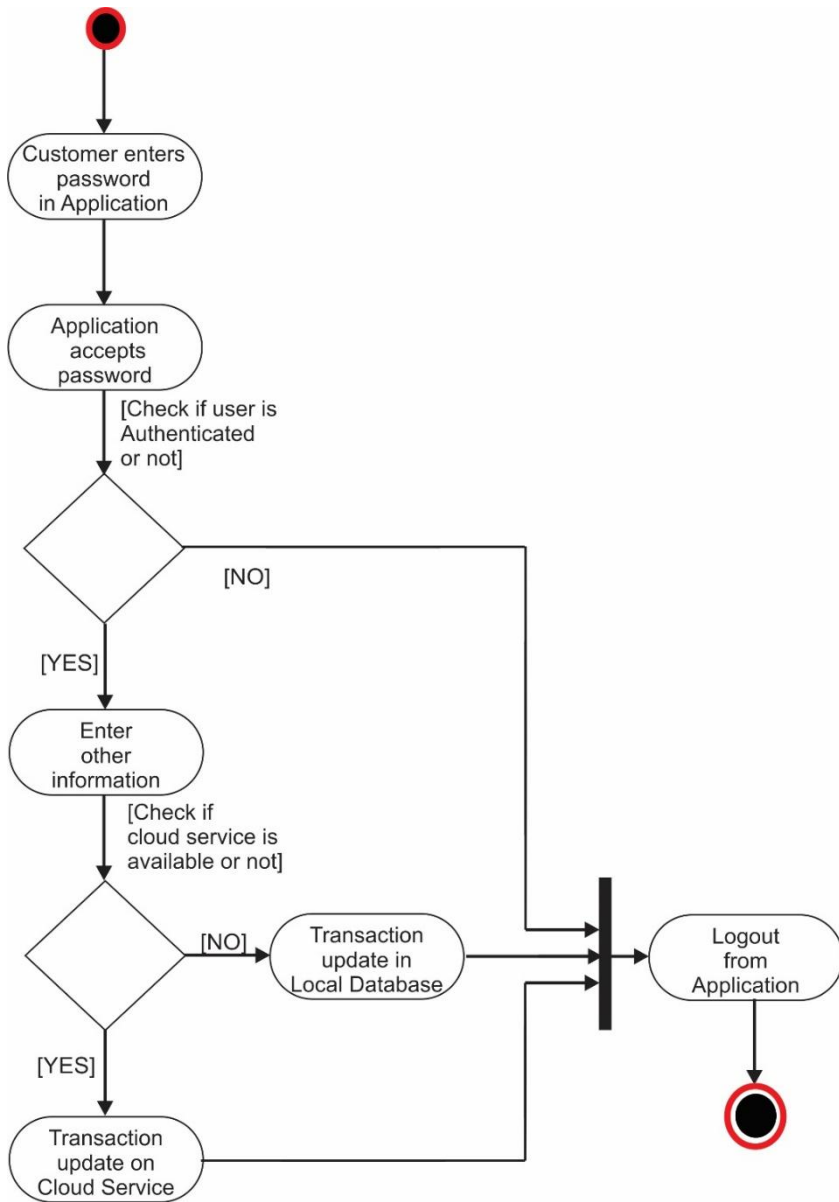
- Read Cipher Text from Database.
- Implement AES algorithm on Cipher Text to generate Plain Text.
- Display plain Text to user.

D. Transaction:

The user will do particular transaction or activity if Authentication is done successfully. E.g. User can enter the digit in the system about how many liters milk he got from the milkman.

E. Transaction Updating Method:

Milkman need to bring mobile device with installed application in it. If currently data pack is not available then at system level database will generate new entry of transaction. As and when Internet Service is available, the database will be modified in cloud.



[Fig 1: Activity Diagram for the Proposed System]

Above Activity Diagram is explained by Following steps which will give you the exact scenario of proposed system:

Step 1: Initially if information about your milkman is not added in your Application then add the details. Once you added the details, it will be saved permanently in the application. If required, you can add multiple milkman's details in the same system.

Step 2: Initially define the price of milk in the system. (There will be 3 categories of milk. Cow milk, Buffalo milk, other). Once you feed the price details in an application, then on daily basis you need not specify the price. You can change the price when the price of milk is increased or decreased.

Step 3: There will be two timings available for buying milk. Morning and evening are default time quantum for doing an entry in the system. If a user just buys milk in morning, then no need to enter a null entry for an evening.

Step 4: If you need more milk from next time, then you can send a request to milkman by application only. You can enter a number of liters milk extra you want in a request form.

Step 5: When milkman arrives, he will give you his mobile with the application where you need to input your password for the successful transaction. The password can be set and reset by you from your mobile device only where the application is installed. When typing the password, customer needs to check that number of liters he is buying milk from milkman is exactly the same as milkman's input in the application.

Step 6: Customer or milkman cannot give input of future days in the application system.

Step 7: Every transaction will be stored in cloud sooner or later. If Internet service is available on the spot then directly it will be stored in the cloud. But if on this specific place or time, Internet service is not available then System will store data locally and as and when application gets Internet Connectivity, it will be updated automatically.

Step 8: Not only authentication of a user, but also data at transit and data at rest are also secured by the application system. Cryptography is done on data.

Step 9: Whenever user installs the application at that time application will ask the security question's answer from the user. So, later on, if user forgets his password, then on the basis of security question and OTP sent on the registered mobile device, verification can be done.

Step 10: Various Reports can be generated by customer and milkman. E.g. Milkman can generate particular customer's report or report of every customer in one format like how much he earned from 10 Customers in the current month? And Customer can generate report about how much he need to pay to milkman for particular month (Here, from particular date to particular date will be considered)

To perform proposed architecture, there are following hardware and software requirements.

- Cloud service

- Internet Service or Data pack.
- Mobile device which support the access of cloud service.
- Mobile app installed in mobile.

3.1 Advantages of Proposed System:

- To remember every small transaction on a day to day life is cumbersome and tedious. Generally, people don't remember but try to note it down in a paper. But, it is also a time-consuming task. Each and every day you do the same transaction and it consumes time to note down that in a paper - is very difficult. So, here proposed system will save consumer's time.
- It is difficult to get the monthly or yearly report for the particular transaction manually. Here, every report is handy. E.g. How many liters of milk we bought in last month?
- Whenever there is no transaction, no need to remember about it.
- Any family member who know the password can do the transaction.

3.2 Disadvantages of Proposed System:

- If the Internet connectivity is a big issue in that particular area then this system is not proper for the user who wants to maintain the record of daily activity on the same time.
- Sometimes if Internet connectivity is not possible then transaction will be update later on when Internet connectivity achieved.
- The user needs to carry Mobile device to the customer's place.

3.3 Application of Proposed System:

- This system is not only helpful to the daily activities of the household, but also helpful to big companies. E.g. If a big dairy gets milk from so many milkmen then this system is also helpful to them.
- For Ironman, Vegetable man, Merchant or any other daily work can be added to this system. We can maintain regular records of such transactions with this system. (Here, we just need to update measurement of transactions. E.g. Ironman will get clothes from us in numbers not in liters.)

4. Conclusion:

To manage daily small but important transactions are tedious and time-consuming. To avoid paper-based activity here we proposed the model which focuses on daily transaction done between customer and milkman. This paper is one of the small steps from our side to manage a small activity that particularly focuses housewives who are engaged in these activities. This model is secure also because authentication using Password and Encryption on data is done.

References

- [1] B Akashdeep, S GVB, A Vinay, S Hanumat, "Security Algorithms for Cloud Computing", Proceedings of International Conference on Computational Modeling and Security(CMS 2016),535-542
- [2] S Gurpreet, Supriya, "A study of Encryption Algorithms (RSA, DES, 3DES and AES) for information Security", International Journal of Computer Applications(0975-8887), Vol.67,No.19,pp. 33-38(2013)
- [3] K Bansi, P Kuntal, "Analysis of Authentication Techniques Adopted by End Users in Real Life Cloud Implementation", Proceedings of International Conference on ICT for Sustainable development, Advances in Intelligent Systems and Computing, 99-107
- [4] K Shakeeba,R R Tuteja,"Security in Cloud Computing using Cryptographic Algorithms", International Journal of Innovative Research in Computer and Communication Engineering,Vol.3 , Issue 1,pp.148-154(2015)
- [5] Retrieved January 26, 2017, from <http://gadgets.ndtv.com/mobiles/news/india-has-higher-smartphone-usage-than-the-us-study-563208>
- [6] Retrieved February 13,217, from <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>

Performance Analysis of K-Nearest Neighbor and K-means clustering to predict the diagnostic accuracy

Kavita Mittal¹
Prerna Mahajan²

¹ JaganNath University, Bahadurgarh, Haryana, India

² Institute of Information Technology and Management, Janak Puri, India.

Abstract. The major challenge related to data management lies in the healthcare sector due to an increase in patients proportional to the population growth and change in lifestyle. Data analytics and big data are becoming trends to provide solutions to all analytical problems that can be obtained by using machine learning techniques. Today, breast cancer is evolving as one of the major attention-seeking phenomena in developed as well as in developing countries that may lead to death if not diagnosed at the early stage. Late diagnosis, and hence delayed treatment, increase the risk for survival. Thus, early detection to improve breast cancer outcome is very critical. This study is intended towards early diagnosis of breast cancer using more efficient analytical techniques. Moreover, accuracy plays an important role in prediction to improve the quality of care, thereby increasing the survival rate. For this study, the dataset has been extracted from the UCI Machine Learning Repository prepared by the University of Wisconsin Hospitals. For the diagnosis and classification process, K-Nearest Neighbor (KNN) classifier is applied using R Studio, one of the best Business Intelligence and Analytics (BIA) tools. Later, the performance of KNN is compared with K-Means clustering on the same dataset.

Keywords: Classification, K-Nearest Neighbor, K-Means, Breast Cancer Database, Performance Analysis, Diagnostic accuracy, Responsiveness, Relevance.

1 Introduction

In India, the rate of increase in breast cancer is becoming uncontrolled, that if action is not taken, then the healthcare sector will be in a major problem for the next two decades. About 231,840 breast cancer cases are diagnosed in India in 2015, out of which approximately 40,000 were died due to late diagnosis [1]. The factors contributing to late marriages, no children, improper diet, too much stress, family history, living styles trigger the occurrence of breast

cancer in women. It makes necessary to work on increasing awareness about the disease and early detection to increase the survival rate and treatment option for the patients. BC begins with fast and uncontrolled multiplication of a part of breast tissue which may be categorized as benign and malignant [2]. Benign represent an abnormal outgrowth but it may not lead to patient's death where as malignant type tumors are more serious and their timely diagnosis can save lives by contributing to better and successful treatment at lower costs. The classification techniques of data mining are one of the predictive modeling techniques that involves assigning of new records to predefined classes .The accuracy of a classification technique is the measure of test set tuples classified correctly to the class [3].Research shows that different methods and algorithms have been applied to cancer diagnosis among which KNN is one of the efficient machine learning technique for classification and K-Means an efficient clustering algorithm. Many research proved that KNN provided reliable accuracy in classification based on supervised learning methodology whereas K-Means based on unsupervised learning methodology. This paper uses (KNN) and K-Means classifier techniques based on different methodologies to diagnose breast cancer type as benign or malignant and perform performance analysis based on three parameters :responsiveness, relevance and validity for its accuracy.

2K- Nearest Neighbour (KNN)

KNN is non parametric machine learning algorithm that can be implemented without any assumption on the dataset. It uses the entire training data during the testing phase .KNN makes decision based on the entire training data. In this study KNN is implemented using Euclidean distance to find the nearest neighbor based on the mathematical equation $D(a,b)=\sqrt{\sum_k(a_k-b_k)^2}$ [4].

Euclidean distance treats each factor as equally important. Euclidean distance is affected by noise when useful features are less than noisy features. To smooth the estimate, the neighbor size K can be increased taking into account the large region [5]. The performance and accuracy of KNN classifier is based on the selection of K and the distance metric applied. In theory, for large sample size, the large K value is chosen for better classification results. Larger K gives smaller boundaries, better for generalization. We can choose K through cross validation.

KNN Pseudo code

- Determine K , nearest neighbor.
- Calculate Euclidean distance b/w query instances and all training tuples.
- Determine K ,ie smallest distance neighbors.
- Collect category Y value of the nearest neighbors.
- Predict the value of query instances using majority of nearest neighbors.

3 K-Means Clustering Algorithm

K-Means ,one of the unsupervised learning algorithm can be used to classify the data through a number of clusters say, k-clusters with an aim to minimize the squared error for optimized classification. The algorithm is advantageous in terms of simplicity, speed, and robustness. It is relatively efficient as compared to other clustering algorithms [15].and performs well with distinct and linear dataset. The algorithm works on A priori specification of number of clusters.

4 Materials and Methods

In this study, KNN and K-Means classifiers are used based on input variables, to predict the type of cancer .The proposed study relied on the available data of patients with breast cancer from UCI Machine learning repository. In this study, data mining and analytics is applied for diagnosis of cancer patients whose data exists in hospital databases and are used for prediction purposes. This study uses KNN based on input variables, to predict the type of cancer. This dataset consists of 699 samples and 10+1 attributes (one for class) listed as -Sample code number (id number) , Clump Thickness , Uniformity of Cell Size , Uniformity of Cell Shape , Marginal Adhesion , Single Epithelial Cell Size , Bare Nuclei , Bland Chromatin , Normal Nucleoli , Mitoses with their domain values lying between 1-10 and Class: (2 for benign, 4 for malignant)

The dataset contains actual 458 (65.5%) Benign cases and 241 (34.5%) Malignant cases . The aim to find the measure the performance of the algorithm by choosing the appropriate value of K for the accurate classification of the data set.

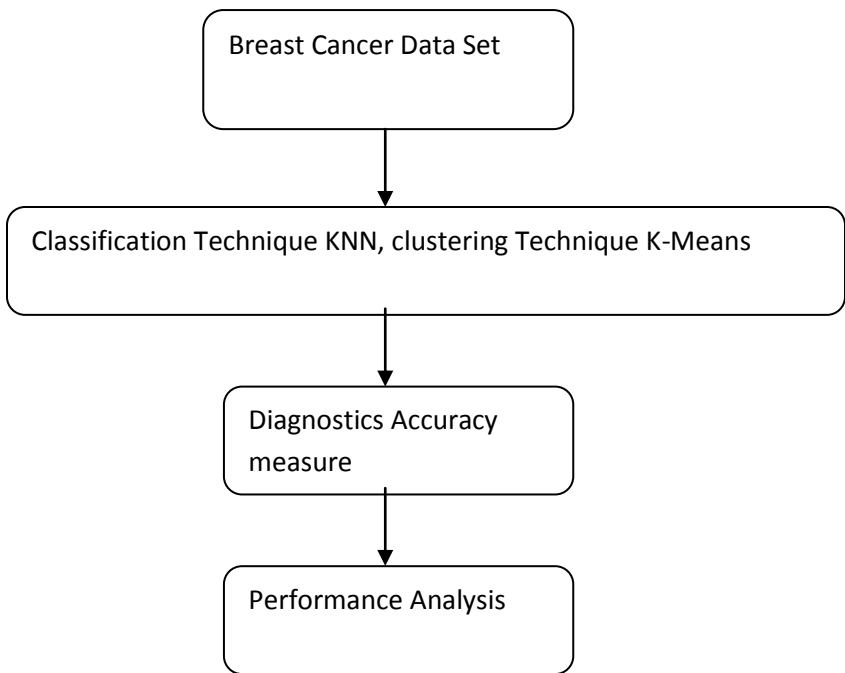


Figure1:Research Methodology

5Related Work

Researchers discovered the utility of machine learning techniques for diagnosis of many diseases by discovering patterns of medical data , improving the decision making process, reducing the cost and enhancing the quality of healthcare.

[6] have experimented three machine learning techniques namely Naïve Bayes, the back propagated neural network ,and the C4.5 decision tree algorithms and found C4.5 better in terms of performance. The study used breast cancer data set from SEER database and result showed that among computation time of Naïve Bayes is less as compared to neural network and C4.5 but C4.5 showed better performance.

[7] has applied ANN for survival analysis on two different Breast Cancer datasets to predict the recurrence probability and classify patients with good

(more than five years) and bad (less than 5 years) prognosis. The result indicate that ANN model can accurately classify as good and bad and predict the survival probability of each time period after a patient had surgery.

[8] demonstrated the performance of modern data processing methods ,suchas principal component analysis (PCA) and artificial neural networks(ANN) analysis for prediction of the recurrence of the disease in spite of being treated previously. The results indicate that using Principal component analysis (PCA) useful information can be extracted from huge data that can be valuable in treatment of breast cancer.

[9]used clustering data mining algorithm for early diagnosis of Breast cancer patients. This study used WEKA as open source data mining tool, but it claims that Orange, Tavera, Rapid Miner tools can provide more optimum outcomes. The study indicates that K-means clustering algorithm and Farthest First(FF) algorithm are most efficient in early diagnosis of breast cancer patients as compared to Hierarchical clustered method (HCM) and Expectation Maximization (EM).

[10]proposed an ideal sampling method with the traditional classifiers such as decision trees ,Naïve Bayes, K-nearest neighbor to enhance the care quality and the rate of survival. This study used SEER cancer dataset and Matlab tool .The results indicates rise in prediction accuracy of balanced stratified model by increasing the size of sample.

[Abhinn Pandey,2014]describes the use of clustering machine learning techniques namely to improve the efficiency of academic performance in educational institution using Rapid miner tool. The result of analysis help to identify the students who need special advising and counseling by the teachers to gain high quality of education.

[11] has analyzed the performance of 3 classifier algorithms namely Naïve Bayes, Random tree and Support vector Machine for accurate prediction of class labels of unknown records. The results proved Naïve Bayes better based on the factors such as classification accuracy and error rates.

[2] devised the decision support system (DSS) to classify the type of breast cancer in patients using probabilistic neural network(PNN) based on three performance indices namely sensitivity, specificity and accuracy. The result showed that the parameter of sensitivity, specificity and accuracy were

found to be 1,0.98, and 0.99 respectively implying that the network involved an acceptable level of reliability in classifying the cases.

[12] introduced a new modified K-nearest neighbor method to classify the data to enhance the accuracy in classification but the results show that it is less effective for large size dataset.

[13] used Instance based KNN(IBK) on three different datasets of breast cancer. The study proved that IBK performed better in fusion with Multi-layer Perception(MLP), Naïve Bayes (NB), decision tree (J48), Sequential Minimal Optimization(SMO).

[14] used KNN algorithm with different distance metrics and different classification rules and analyzed the performance based on classification accuracy rate and time and validated the results with different training and test sets. The results proved Euclidean distance more effective in terms of classification and performance but consuming much time.

6 Experimentation and Results

The KNN algorithm is implemented on the training data set and followed by the validation of results on the test dataset. To categorize the training data and test data the dataset is divided into two portions with ratio 499:200 (assumed as 70:30 approx) for the training and test data resp. The data supplied to the model were normalized through linear method to represent 1-10 values. The target matrix included two classes benign and malignant. The accuracy of the obtained results lies on the K value, chosen. The chosen value of K determines the efficient data utilization to generalize the results of KNN. Choosing average value of K reduces the variance due to noisy data and may provide reliable accuracy. Generally the value of K can be taken approximately as the square root of number of training data samples.

6.1 KNN with K=1

Initially the model is executed on K=1 or 3 and the results are recorded as follows in figure2

Cell Contents

	N
N / Row Total	
N / Col Total	
N / Table Total	

Total observations in Table: 200

wbc_d_test_labels	wbc_d_test_pred		Row Total
	Benign	Malignant	
Benign	156	0	156
	1.000	0.000	0.780
	0.987	0.000	
	0.780	0.000	
Malignant	2	42	44
	0.045	0.955	0.220
	0.013	1.000	
	0.010	0.210	
Column Total	158	42	200
	0.790	0.210	

Figure 2. Confusion matrix with K=1,3.

Then the model is executed with K=21 and the results are recorded as follows in figure3.

Total observations in Table: 200

wbc_d_test_labels	wbc_d_test_pred		Row Total
	Benign	Malignant	
Benign	156	0	156
	1.000	0.000	0.780
	1.000	0.000	
	0.780	0.000	
Malignant	0	44	44
	0.000	1.000	0.220
	0.000	1.000	
	0.000	0.220	
Column Total	156	44	200
	0.780	0.220	

Figure 3. Confusion matrix with K=21

6.1.1 Evaluation of the Performance

The quality of result relies on the value of K, the number of nearest neighbors chosen with respect to the dataset. To analyze the accuracy of the predicted samples in the test data we used cross tables. The test data consisted of 200 observations. In this study we have evaluated the performance on different values of K. To investigate the degree of effectiveness of disease diagnosis and classification, the confusion matrix is used that produces four results: True Negative (TN), False Positive (FP) and False Negative (FN), True positive (TP). The confusion matrix provides three parameters to assess the performance of classification: Responsiveness (sensitivity), Relevance (specificity), validity (accuracy).

Responsiveness indicates the algorithm's precision in diagnosing the malignant type. Relevance indicates the precision in diagnosing the benign type. Validity indicates the proportion of all cases truly diagnosed.

$$\text{Responsiveness} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Relevance} = \text{TN} / (\text{FP} + \text{TN})$$

$$\text{Validity} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

For K=1, 156 cases were accurately predicted (True Negative (TN)) as Benign (B) in nature, i.e. 78%, 42 were predicted accurately (True Positive (TP)) as Malignant (M) in nature that constitutes 22%. There were 2 cases of False Negative (FN) that means two cases were malignant in nature but they were predicted as benign. To improve the accuracy of the model FN's needed to be reduced. The cases of False Positives (FP) were 0 that means no case was false predicted as Malignant. The values of Responsiveness, Relevance and Validity were found to be .88, 1, .99 respectively.

6.1.2 Improving the Model Performance with K=21

By repeating the process and changing the value of K, i.e. K=21, that is not exactly but near to the square root of the count of training samples, then performance of the algorithm was analyzed. The K values may be fluctuated in and around to evaluate the improved accuracy of the algorithm. For K=21, out of 200 observations 156 cases were predicted as benign (TN) and 44 cases were predicted as Malignant (TP) with no FP and FN cases. The values of Responsiveness, Relevance and Validity were found to be 1, 1, and 1 respectively.

6.2 Experiment with K-Means clustering

	Benign	Malignant
Benign	447	16
Malignant	11	225

K-means clustering with 2 clusters of sizes 463, 236

Total Observations:699

Applying K-Means on the same dataset with 699 observations the values of the parameters: Responsiveness, Relevance, Validity are found to be 0.95, 0.97,0.96 respectively.

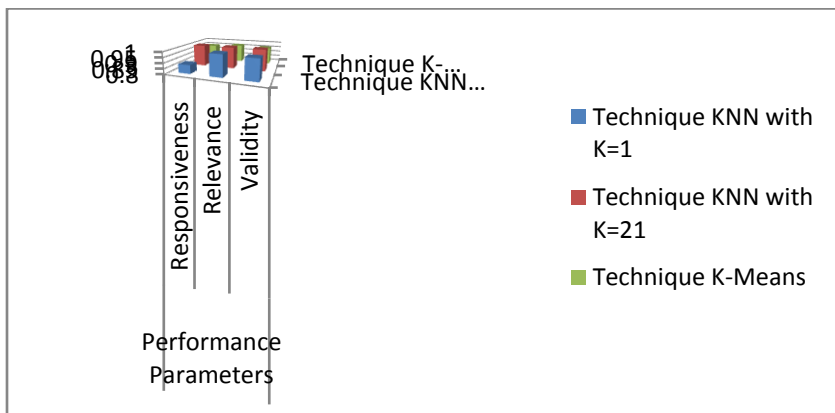


Figure 4. Comparison of KNN with values K=1,21 and K-Means

The figure 4 indicates that the performance of K-means is more fruitful if compared to KNN with K=1, and but less efficient with K=21 .Thus, KNN a classification technique performs more efficiently than cluster based classifier provided the value of k variable is chosen correctly.

6 Conclusion

The study has implemented the KNN classification algorithm and K – Means clustering algorithm for diagnosis of breast cancer type ,presented the results and performed the performance analysis with different values of

K based on three parameters for its accuracy. The study intends to find the right technique for classification of diverse datasets. In the literature KNN has shown remarkable utility in solving the classification problems. However, choosing K may be tricky and it needs large number of samples for accuracy. It requires no training phase, all the work is done during the testing phase. In this study the dataset is classified using KNN and K-Means as benign and malignant. The experiment is conducted by training the samples first and then testing through testing samples. The parameters of Responsiveness, Relevance and Validity are evaluated for different values of K and compared with the results of K-Means. Finally, it is concluded that the KNN model provides acceptable level of reliability in classification with best chosen K value. Thus, KNN seems to be more efficient and more reliable based on the above mentioned parameters.

References

1. <http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-046381.pdf>
2. Asieh Khosravianian, Saeed Ayat. Diagnosing Breast Cancer Type by Using Probabilistic Neural Network in Decision Support System. International Journal of Knowledge Engineering, Vol. 2, No. 1, March 2016.
3. S. Kalaivani, S. Gandhimathi. An Efficient Bayes Classification Algorithm for analysis of Breast Cancer Dataset using Cross Validation Parameter. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 10, October-2015.
4. Zehra Karapinar Senturk and Resul Kara1 Breast Cancer Diagnosis Via Data Mining: Performance Analysis Of Seven Different Algorithms. An International Journal Computer Science & Engineering, (CSEIJ), Vol. 4, No. 1, February 2014.
5. Imandoust, S.B., Bolandraftar, M., Application of K Nearest Neighbor (KNN) for Predicting Economic Events: Theoretical Background. International journal of Engineering Research and Application. Vol 3(5), pp 605-610, 2013
6. Abdelghani Bellaachia, Erhan Guven. Predicting Breast Cancer Survivability Using Data Mining Techniques. Journal of Society for Industrial and Applied Mathematics. 2003 Mar; 7(1):37-42. 2003.
7. Chih-Lin Chi, W. Nick Street, William H. Wolberg. Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. Proceedings of AMIA 2007 Symposium.
8. Adam Buciniński¹ACDEF, Tomasz Bączek^{2,3}DEF, Jerzy Krysiński⁴ADG, Renata Szoszkiewicz⁵BD, Jerzy Załuski⁵BD. Clinical data analysis using artificial neural networks (ANN) and principal

- component analysis (PCA) of patients with breast cancer after mastectomy. *Rep Pract Oncol Radiother*,; 12(1): 9-17, 2007.
9. Jahanvi Joshi , Rinal Doshi , Jigar Patel .Diagnosis of Breast Cancer using Clustering Data Mining Approach. *International Journal of Computer Applications*. Volume 101– No.10, September 2014.
 10. Saleema,J.S.,et.al.Cancer Prognosis Prediction using Balanced Stratified Sampling.*International Journal on Soft Computing,Artificial Intelligence and Applications(IJSCAI)*,Vol 3(1),2014
 11. Abhinn Pandey .Study and Analysis of K-Means Clustering Algorithm Using Rapidminer A case study on students' exam result. *Int. Journal of Engineering Research and Applications* .ISSN : 2248-9622, Vol. 4, Issue 12(Part 4), December 2014, pp.60-64.
 12. Parvin,H.,Alizadeh,H.,Minaei-Bidgoli,B. MKNN:Modified K-Nearest Neighbour.*Proceedings of World Congress in Engineering and Computer Science,USA,2008*
 13. Gouda I.Salama, Abdelhalim,M.B., Zeid,M.,A. Breast Cancer Diagnosis on three Different Datasets Using Multi-Classifiers.*International Journal of Computer and Information Technology* .Vol 1(1),2012.
 14. Madjahed,S.A.,Saadi,T.A.,Benyettou,A. Breast Cancer diagnosis bu using k-Nearest Neighbour with Different Diatances and Classification rules.*International Journal of Computer Applications*,Vol 62(1),2013.
 15. <https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm>

Perception and Utilization of Social Media Network among Adolescents In Samaru Community, Sabongari Local Government Area, Kaduna State, Nigeria

Hussaini Suleiman¹, Dr Rajeev Vashistha²

¹Research Scholar
Department Of Library And Information Science, Nims University,
Jaipur.India

hsuleimanabu@gmail.com

²Associate Professor,
Department Of Library And Information Science, Nims University Jaipur,
India.

Rajzz.V@gmail.Com

Abstract. The paper examines the perception and utilization of social media network among adolescents in Samaru community, Nigeria. This study is guided by four research questions. The design of the study is descriptive survey method and the sampling design used for the study is simple random sampling. The population consists of adolescents between the ages of 10-19 in Samaru community, Nigeria. The instrument used for data collection was questionnaire. Data generated was analyzed using frequency tables and percentages. It was found out that, majority of the adolescents in Samaru community prefer to use whatsapp and facebook as their most preferred social media. The findings further revealed that, adolescents in Samaru community perceive social media to be a kind of network where they can only send and receive messages and chat with friends. The study recommend that parents should checkmate the dailies activities of their wards, as this may change their mind setting.

Keywords: Perception, Utilization, Social Media Network, Adolescent, Samaru Community.

1 Introduction

Young people all over the world are the potential of a country's future, and if their needs particularly social media needs are not addressed, they have the real potential to jeopardize that future [1]. Nigeria being one of the

biggest country in African continent is located in West Africa with an estimated population of over 170 Million [2] 115 million mobile subscribers and 56 million internet users, and has been seen as the biggest mobile internet and mobile market subscribers [3]. Today, an average of 4 out of every 10 adolescents in Nigeria you find on the street has atleast one information communication technology device such as mobile phones, computers, iPad etc. that enables him or her to connect or communicate via social media networks. Nigerian is estimated to have over 22 million social media networks users using social sites, such as whatsapp, facebook, youtube, linkedIn, Imo, skype etc.[4].

However, social media forms an integral part of daily life among adolescents and has shown to benefit adolescent communicative capabilities and skills [5]. Where they can share and get firsthand information from each other. Adolescent constitutes majority of active users of social media network sites.[6]

1.1 Social Media

[7] sees social media as networks such as whatsapp, linkedIn, facebook etc., political blog like video sharing like youtube, audio sharing like podcast and mobile sites like 2go that have the capacity of boosting participation as a result of their open, conversational nature, connectedness and textual and audio visual characteristics. According to Shrestha cited in [8][9], social media means association among folks in that they produce, share, and exchange information and concepts in virtual communities and networks. [10] sees that, social media as inevitable for the least of majority of organizations globally. [11] posits that, social media is a network communication platform that may consume, produce and move with cluster of users generated happiness provided by their connections on the web site.

1.2 Adolescent

Adolescent is a powerful formative time of transition to adulthood, that's when kinsmen become most aware of their gender. It is a physical, psychological and social change from childhood to adulthood and falls within the ages of ten to nineteen years [12]. [13] defines adolescent as person aged 10 to 19 years. It has been estimated that, 16% of adolescent worldwide are found in Africa (Population Institute, population and account Generation: A guide to Action). Adolescent use social media more often for example, a survey in Nigeria 2009 finds that, 73% of internet users particularly adolescent use social network sites, which is an increase from 55% 3 years earlier Lenhart Purcell cited in [14]

However, according to Flanner in [15] on use and consumption of social media, said 93% of adolescents are active users of the internet (10-70%), 75% of adolescent own a cell phone, teens text per month (100/day), text messages has increased most dramatically, along with media multi tasking.

In study conducted on Social Network Addiction among Youths in Nigeria[16], concluded that majority of the respondents use much of their time on social networking sites, which affects their productivity negatively. The findings of this study also indicated that youths in Nigeria spend most of their time on social networking sites communicating at the impairment of various necessary things like their studies.

However, in a study conducted by [17] reveals that, online social network have made net users to become distinguished communication users, particularly in students community. It however allows them to connect with potential fellow friends and to deliver educational content.

Social Networking Sites are turning into omnipresent aspects of youth and young adult life. Students have become captivated with this lifestyle way more than older generations have in recent years, as this way of living is all they apprehend [18]

Similarly, Seiter in Ali [19] observes that young people magnificently use digital communications, instant electronic messaging, cell phone texting, and social networking websites to keep up their social capital, atleast with those peers who will afford to stay up with the expensive needs of those technologies.

[20] identified facebook as a key example of a social media site. It is observed from research that facebook is now the primary method of communication by college students in the developing countries like Nigeria. Users spend more than six hours on social networking sites [21]. Social media according to [22] social media has made most of the Nigerian youths to continue spending more time on social networks than any other category of sites.

1.3 Historical Background of the Study Area

Samaru community is a cosmopolitan community with major ethnic group Hausa and Fulani. It was created from Bomo community in 1922, formely known as Labour line by a group of inhabited staff of Institute of Agricultural Research Zaria (IAR), who came from different part of Nigeria. Their major occupation apart from public and civil service are; farming, herdsmen, blacksmithing, malamai (Arabic school) and butchers

(Mahauta). Samaru Community is also housed to a number of staff and students from the neighboring institutions of learning.

Samaru Community is located at Zaria-Funtua road. Geographically, it is located on latitude 11.0⁰N, longitude 7.37⁰E with an altitude of 239 feet (730 meters). It's about 12 km from the ancient city of Zaria. It is also located opposite the famous Ahmadu Bello University, Zaria, (ABU) main campus Samaru, Zaria. There had been four (4) herbal rulers namely; Sarkin Samaru Ibrahim, Sarkin Samaru Abubakar, Sarkin Samaru Idrissu Abdullahi and the present ruler called Sarkin Samaru Dauda Abubakar. Currently, Samaru community has two (2) additional kingdom namely: Sabon layin and Tsakiya all headed Sarki Saidu Abdullahi and Sarki Isa Ibrahim (Village head). Samaru is now a district in Sabongari Local Government headed by District head traditionally called Hakimi. The village heads is answerable to the District head and the ward heads (Masu anguwanni) are also answerable to Village Heads. Samaru community has eight(8) communities under it's rule namely: Samaru, Tsakiya, Sabon layi, Jama'a, Korea, Yelwa, Makera, kallon kura and Yardoruwa, all turbaned and posted by Zazzau Emirate Council. The Village head and District head are turbaned by His Highness, the Emir of Zazzau.

It has a population of over 60,000, with an estimated population of over 15,000 adolescents [23] (Village head), all these attributed to the immigration and emigration of the people in the community. The major ethnic group in Samaru community is Hausa and Fulani. Islam and Christianity are the main religious belief practiced. Thus, the community has a number of government primary schools, a number of private schools, 2 tertiary institutions (that is Ahmadu Bello University, Zaria and NILEST formerly called Leather Research situated in Danraka Samaru), a number of Centres such as Centre for Energy Research, Industrial Development Centre (IDC) and Centre for Automobile Research Design (CARD), a Division of Agricultural College (DAC), a number of private hospitals and clinics, a number of primary health care centers, a number of markets and a number of banks.

1.4 Negative effects of social media on adolescents

- i. **Fornication:** Most adolescents at this stage are in their active sexual period and are often exposed to sexual urge due to nature or through watching of sexual videos on social media networks. Most of them try to practicalize what they watch which result to unwanted pregnancy and it sequel.
- ii. **Gossip:** This is another medium that often spread false rumor thereby creating instability and anarchy in the community resulting to break down of law and order.

- iii. **Crime:** Today, social media networks have drastically increased the number of crimes committed in our society. Most adolescent often learn new methods of crime on social media. They learn the various tricks they can use to dupe people online since it is often posted online.
- iv. **Time wasting:** Adolescent often waste their time chatting irrelevant issues on social media, issues that are not important to their life, they spend the whole day without doing meaningful things that will be of benefit to their life.

1.5 Objectives of the Study

- i. To identify the various types of social media used by adolescents in Samaru Community, Sabongari, Kaduna State Nigeria.
- ii. To find out the perceptions of adolescent towards social media in Samaru community, Sabongari Kaduna State Nigeria.
- iii. To determine the level of utilization of social media among adolescent in Samaru Community, Sabongari, Kaduna State, Nigeria
- iv. To identify the problems of social media among adolescents in Samaru community, Sabongari, Kaduna State Nigeria.

1.6 Research Questions

- i. What are the various types of social media used by adolescents in Samaru community, Sabongari, Kaduna State, Nigeria?
- ii. What are the perceptions of adolescent towards social media in Samaru community, Sabongari, Kaduna State, Nigeria?
- iii. What is the level of usage of social media by adolescent in Samaru Community, Sabongari, Kaduna State Nigeria?
- iv. What are the problems of social media among adolescents in Samaru community, Sabongari, Kaduna State, Nigeria

2 Methodology

In this study, survey research method was adopted, as this was deemed appropriate for the study. Cresswell cited in [24] posit that, survey research represent a set of orderly procedures specifying what data is to be obtained and from whom. Adolescent in Samaru community constitute the target population and simple random sampling techniques was used to collect data directly from the respondents using self-administered questionnaire. Thus, survey research method is found appropriate for this study because it would facilitate gathering of data concerned with the perception and utilization of social media network among adolescents in Samaru Community, Sabongari local Government, Kaduna State, Nigeria. The data generated for the study was analyzed using frequency tables and percentages.

2.1 Sampling size

A sampling size of 15,000 adolescents representing the respondents is calculated using Yamane's formula.

$$n = \frac{N}{1+N(e)^2} = n = \frac{15000}{1+15000(0.05)^2} \quad n = \frac{15000}{1+15000(0.0025)}$$

$$n = \frac{15000}{38} = 394.7$$

(1)

Approximately 395 adolescents would be sampled in Samaru Community, Nigeria.

2.2 Response Rate

The below table shows that, of the 395 copies of questionnaire administered, only 348 responses were retrieved and returned completed for the study. 47 questionnaires were either incomplete or invalid.

Table 1. RESPONSE RATE

Respondents	Administered	Returned
Adolescents	395	348

Source: field work 2016

The above table shows that, a total number of 395 questionnaires were administered to the respondents but only 348 questionnaires were returned by the respondents. This shows that 47 questionnaires were invalid. 27 Questionnaires were not properly filled and 20 questionnaires were not returned by the respondents. The researcher was able to get more response from the respondents because the questionnaires were made easy for the understanding of the respondents.

Table 2. TYPES Of SOCIAL MEDIA USUALLY USED By ADOLESCENTS In SAMARU COMMUNITY

Type of Social Media	Frequency	Percentage %
Whatsup	123	35%
Facebook	101	29%
Youtube	38	10%
LinkedIn	17	5%
Skype	39	11%
Twitter	30	9%
Total	348	99%

Source: Field work 2016

Table 2 shows that, the mostly used social media by adolescents in Samaru community is whatup and facebook 35%(123) and 29% (101). 10%(38) of adolescent prefer using Youtube while others prefer using skype and twitter 11%(39) and 9% (30). Very few of them prefer using LinkedIn 5% (17). This implies that, majority of respondents are whatsapp and facebook users.

Table 3. PERCEPTION Of SOCIAL MEDIA Among ADOLESCENTS In SAMARU COMMUNITY

Perceptions of social media	Frequency	Percentage
Updating Knowledge	30	9%
Chatting with friends	108	31%
Watching movies/sports	50	14%
Sharing & discussing school information	30	9%
Uploading and downloading videos	39	11%
Posting & updating status on timeline	91	26%
Total	348	100%

Source:Field work 2016

Table 3 above shows that, majority of the adolescents perceive social media to be where they can be chatting with friends and posting & updating status on the their timeline 31%(108) & 26% (91). While very few of them perceive social media a place where they can upload & download videos and watch movies/sports 11% (39) & 14% (50). 11% (39) and 9% (30) were perceived by adolescents as a place where they can be updating their knowledge and sharing & discussing school information. This implies that, majority of respondents take social media to be a place where they can send and receive messages.

Table 4: LEVEL Of USAGE Of SOCIAL MEDIA Among ADOLESCENTS In SAMARU COMMUNITY

The below table shows that, majority of the respondents in Samaru community utilizes more social media like facebook 134(38.5%) and whatsapp 112(32.2%). It also shows the least used social media as linkedIn indicating 16(4.6%). Twitter Skype and Youtube indicating 27 (7.8%), 16(4.6%) and Youtube 40(11.5%) also shows the level of usage of social media among adolescents in Samaru community. This implies that facebook and whatsapp are more non and used by adolescents in the community.

Table 4: LEVEL Of USAGE Of SOCIAL MEDIA Among ADOLESCENTS In SAMARU COMMUNITY

Type of Social Media	Frequency	Percentage %
Twitter	27	7.8%
Skype	16	4.6%
LinkedIn	19	5.5%
Youtube	40	11.5%
Facebook	134	38.5%
Whatapp	112	32.2%
Total	348	100.1

Source: Field work 2016

Table 5: PROBLEMS Of SOCIAL MEDIA Among ADOLESCENTS In SAMARU COMMUNITY

Problems	Frequency	Percentage
Unwanted friendship request	93	27%
Power failure	138	40%
Lack of internet access	95	27%
Unwanted social advances from adult	22	6%
Total	348	100%

Source: Field work 2016

The above table shows that, power failure, unwanted friendship request and lack of internet access were the most likely problems encountered by students with 40% (138), 27% (93) and 27% (95). Unwanted social advances were the least problem encountered by adolescents in Samaru community. This implies that adolescents do experience hitches while using social media.

2.3 Summary of the findings

The majority of the findings from the analysis of the data of the study are summarized below:

1. The study shows that, majority of the adolescents' 35% (123) prefers to use whatsapp as their favorite communication medium and the least used type of social media is linkedln 5% (17).

2. The study further reveals that, adolescents usually waste their time on social media sending and receiving messages instead of utilizing it for developing themselves self academically.
3. The study reveals that, power failure which is a problem in our present day community, is limiting the usage social media by adolescent in our community.

Recommendations

1. Stakeholders in the community should encourage adolescents not to use social media during class period as it will not allow them to pay full attention to their teacher.
2. Stakeholders should checkmate the day to day interaction of their wards, as this may change their mind setting and may even expose them to health hazards.
3. Schools in the community should purchase devices that will lead to blockage of internet connection during school period or hours.
4. Stakeholders in the community should encourage adolescent to use social media for purely academic purposes rather than using it for social communication only.
5. Parents should make sure that, they monitor their wards from time to time so as to ensure they are not misusing social media to commit crimes and other ill activities.
6. Adolescents in Samaru community should use the social media in developing themselves academically.
7. Stakeholders in the community should also create awareness to adolescent living in their community so as to reduce the level of crime and other ills and it sequel in the community.

Conclusion

The emergence of social media has brought a dramatic shift from the way adolescents behave in the community and if this is not checked on time, it could jeopardize the life and future of the adolescents in the community.

References

1. Adesokan, F. Reproductive Health for all Ages. (3rded). Bosem Publishers Nigeria Ltd. 417 (2014)
2. National population Commission (2016)
3. Nigerian Communication Commission (2013)
4. Buhari, S.R. Ahmad, G.I., Hadi, A. Use of Social Media among Students of Nigerian Polytechnic. ICCMTD International Conference. Istanbul Turkey ; 24-26 April; 302-305.
5. Itom, M., Horst, H., Bittani, M. Living and Learning with Media: Summary of Findings from the Digital Youth Project, Chicago II: John D and Catherine T. Mac Arthur foundation reports on digital

media and learning: 2008. Available at <http://digitalyouth.school.berkeley.edu/files/report/digitalyouth>. Retrieved November 20, (2016).

6. June, A. The effects of Social Networks Sites on Adolescents Social and Academic Development: Current Theories and Controversies. vol. 2(8), pp.1435-1445. Journal of the American Society for Information Science and Technology. (2011)
7. Abubakar, A.A. Political Participation in Social Media During the 2011 Presidential Electioneering in Oladokun Omojola et al (eds) Media Terrorism and Political Communication in a Multi Cultural Environment: ACCE conference proceedings. Ota Nigeria. ACCE loc. Pp. 445-453. (2011)
8. Ghulam, S. Yousef, M., Yousef, H., Ghulam, S. The Impact of Media on Youth: A Case Study of Bahawalpur City. Vol. 3(4). Asian Journal of Social Sciences & Humanities. (2014)
9. Makwei, E.D., Appiah, D. The Impact of Social Media on Ghanaian Youth: A Case Study of the Nima and Maamobi Communities in Accra, Ghana. Vol. 2(2). Journal of Research on Libraries and Youth Adults. (2016)
10. Adesokan F. Reproductive Health for all ages. (3rdeds.). Bosome publishers Nigeria ltd.(2014)
11. World Health Organization (2015)
12. World Health Organization (2015)
13. Population Institute, Population and Account Generation: A guide to Action (nd)
14. Flannery, D.J., Semi, J., Ruth, B. Social Media and its Effects on Youth Retrieved 21/02/2017 from http://ja.cuyahogacounty.us/pdf_ja/en-US/SocialMediaEffects-DJFlanneryPHD.pdf
15. Ghulam, S., Yousef, M., Yousef, H., Ghulam, S. The Impact of Media on Youth: A Case Study of Bahawalpur City. Vol. 3(4). Asian Journal of Social Sciences & Humanities. (2014)
16. Ajewole, O.O., Fasola, O.S. Social Network Addiction Among Youths in Nigeria, Journal of Social Science and Policy Review, Volume Journal of Social Science and Policy Review 4. (2012).
17. Paul, J., Baker, H., Cochran, J., Effect of Online Social Networking on Academic Performance. vol. 1, pp. 2118-2119:2123. Elsevier, (2012)
18. Lewis, S. Where Young Adults Intend to get News in Five Years. Newspaper Research Journal, vol. 29(4), pp. 36-52. (2008)
19. Ali, FA., Farah., Aliyu, U.Y. The Use of Social Networking among Senior Secondary School Students in Abuja Municipal Area of Federal Capital Territory, Nigeria. Journal of Education and Practice. vol 6(15) pp. 15-22. (2015)

20. Debora, D. "Finding High Quality in Social Media". International Conference on Web Search and Data Mining. Calabar: WSDM. (2012)
21. Bassey, E. "The Influential Role of Social Media on Nigerians". Thisday Newspaper, September 25, p.21. (2002)
22. Jude, U. Nigerian Youths and Internet Exposure: Onitsha: Sofie Publicity and Printing company. (2011)
23. National Population Commission (2006)
24. Yusuf, A. Accessing Business Information for Enhanced Entrepreneurial Participation. Journal of Nigerian Library Association. vol.47(1). pp.16-39. (2014)

Error Detection by Checksum

Dr. Anil Kumar Singh,

Associate Professor,

Department of Information Technology

Jagran Institute of Management,

City: Kanpur (UP), India

E-mail: anil.sysadmin@gmail.com

Abstract– There are different methods are used in detecting the error of code word i.e. parity check, CRC etc. In this paper we will try to detect the error of codeword with the help of Checksum. This technique involves sum of data and its compliment. The sender performs an addition and 1’s compliment to calculate the checksum. Before sending the data, the sender appends the checksum at the end of the data word. In this paper we have taken two scenarios as 1 and 2. It will help in generating the codeword and checking the error.

Keywords: Checksum, codeword, 1’s compliment, data word, sender, receiver

1. Introduction

In checksum error detection method, the data is divided into “m” segments each of “n” bits. In the sender’s side the segments are added by using simple binary addition to get the sum as shown in Fig. 14. The sum is complemented to get the checksum as shown in Fig. 15. The checksum segment is sent along with the data segments as shown in Fig. 16. At the receiver’s end, all received segments are added using simple binary addition to get the sum. The sum is complemented. If the result is zero, the received data is accepted; otherwise discarded [1].

2. **Scenario 1:** Let us take an example of arithmetic data word with the help of following example we will try to reveal that how checksum works.

Data word:

9	12	7	2	0	5
4 bits Data word					

Fig. 1 Data word

3. Method

First of all sender does add all the digits

$$= 9 + 12 + 7 + 2 + 0 + 5 = 35$$

Sum = 35 so Checksum will be -35

Now sender will send the data word along with the checksum

9	12	7	2	0	5	-35
4 bits Data word						Checksum

Fig. 2 Codeword

When receiver receives the codeword it checks by addition of codeword.

9	12	7	2	0	5	-35
Data word						Checksum
Sum = (9 + 12 + 7 + 2 + 0 + 5 + (-35))						
35						-35
0 (Zero) it's indicate there is no error in received packet						

Fig. 3 error detection by receiver

4. How checksum works

The checksum is another method of error checking, like VRC, LRC and CRC it is also works on the basis of redundancy (means adding some extra bits). It has two parts i.e. checksum generator and checksum checker. These are works on sender side and receiver side respectively [2].

Now we will see in details that how checksum performs.

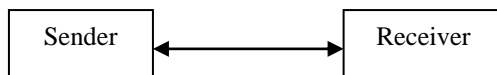


Fig. 4 Components of Communication

4.1 Sender's Side

9	12	7	2	0	5
Data word					

Fig. 5 Data word

Before sending the data word it follow the following steps.

Sender Calculates the Sum of Data word

9	12	7	2	0	5	= 35
Data word						Sum

Fig. 6 Sum of data word

Now sender converts the binary values of = 35 decimal number

32	16	8	4	2	1
1	0	0	0	1	1

Fig.7 Binary No. of sum value

The binary of 35 is 6 bits but data is 4 bits so we have to wrap the two bits.

1	0	0	0	1	1
Wrapped bits		←————→			
Wrapped Sum = 5		0	1	0	1
1's Compliment		1	0	1	0
(1010) binary = 10 (Checksum)					

Fig. 8 Calculating Checksum

Now Sender will send the following codeword to receiver

9	12	7	2	0	5	10
Data word						Checksum
Codeword						

Fig. 9 Codeword

4.2 Receiver's Side

The receiver adds the codeword

9	12	7	2	0	5	10
Data word						Checksum
Sum = (9 + 12 + 7 + 2 + 0 + 5 + 10) = 45						

Fig. 10 Sum of Codeword

Now calculates the binary of 45

32	16	8	4	2	1
1	0	1	1	0	1

Fig. 11 Binary number

Now wrap the two bits

1	0	1	1	0	1
---	---	---	---	---	---

Wrapped bits				1	0
Wrapped Sum = 15	1	1	1	1	
1's Compliment	0	0	0	0	
Checksum = 0 (Zero) It shows received packet has no error					

Fig. 12 shows checksum of received packet

5. Scenario 2: Let us take another example of binary data word.

Binary Data word has “m” segment and each segment has “n” bits.

1	0	1	0	1	0	0	1	,	0	0	1	1	1	0	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Fig. 13 Binary data word

5.1 Sender's Side: Sender first add the data word

	1	0	1	0	1	0	0	1
	0	0	1	1	1	0	0	1
Sum	1	1	1	0	0	0	1	0

Fig. 14 sum of data word

Now calculate the 1's compliment of sum (conversion of 1 into 0 and 0 into 1)

1's Compliment	0	0	0	1	1	1	0	1
	Checksum							

Fig. 15 Checksum

Sender sends the following codeword (data word and checksum)

1	0	1	0	1	0	0	1	,	0	0	1	1	1	0	0	1	,	0	0	0	1	1	1	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Fig. 16 Codeword

Suppose codeword corrupted during the transmission

1	0	1	0	1	0	0	1	,	0	0	1	1	0	0	0	1	,	0	0	0	1	1	1	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Fig. 16 shows the corrupted bit

5.2 Receiver's Side -The receiver sums the codeword

1	0	1	0	1	0	0	1
0	0	1	1	0	0	0	1

	0	0	0	1	1	1	0	1
Sum	1	1	1	1	0	1	1	1

Fig. 17 Sum of codeword

Now take the 1's compliment of sum

1's Compliment	0	0	0	0	1	0	0	0
	Checksum							

Fig. 18 Checksum

The value of checksum is nonzero, shows that there is some error during the transmission.

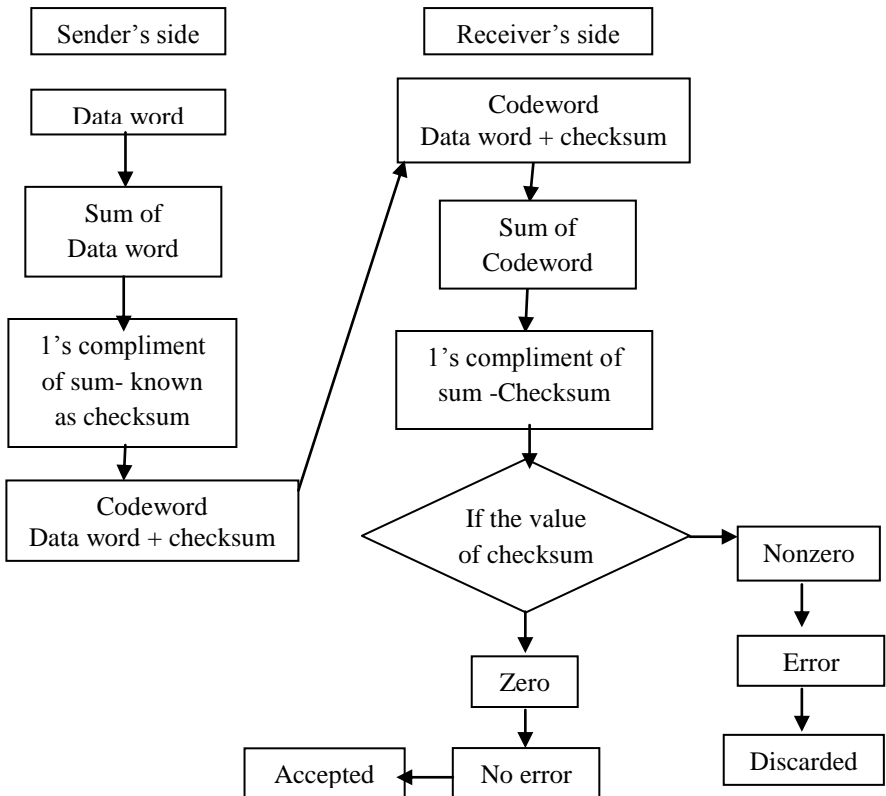


Fig. 19 Process of error detection

Conclusion

The sender before sending the data it generates the codeword, (which is combination of actual data and checksum) by calculating the sum and converting it into checksum (it is 1's complement of sum).

The receiver after receiving the codeword it tries to test the error by checking the value of checksum. If the value of checksum is zero indicate that received packet has no error and if the value of checksum is nonzero, indicates that there is an error during transmission.

It is observed that the error of received packet depends of the value of checksum. As Fig. 12 shows the value of checksum is equal to zero, means no error. The value of checksum is nonzero means there is an error as shown in Fig. 18.

References

1. Error Detection and Correction, Version 2 CSE IIT, Kharagpur.
<http://nptel.ac.in/courses/Webcourse-contents/IIT%20Kharagpur/Computer%20networks/pdf/M3L2.pdf>
2. Andrew S. Tanenbaum, Computer Networks, fourth edition, Pearson publication, ISBN 81- 7758-165-1

Desirable Features for an Effective Sentiment Analysis System

Sujata Rani and Parteek Kumar

CSED, Thapar University, Patiala (Punjab), India
{sujata.singla, parteek.bhatia}@thapar.edu

Abstract. In the era of digitalization, a huge amount of data is generated by people on online websites, social media platforms, blogs and forums, etc. in the form of comments, audio and visual content about different entities like products, services, and organizations. It's very difficult to crawl the enormous amount of data by a user. There sentiment analysis plays an important role in this case. To develop an effective sentiment analysis system, it is necessary to consider the desirable features of the system to be built. This paper discusses the desirable features of an effective sentiment analysis system for diverse modalities like text, audio and visual data. Also the visualization features of representing sentiments are discussed in this paper. The paper includes the different applications and challenges of sentiment analysis. The features elaborated in this paper can help in building a web based effective sentiment analysis, including visualization of sentiments on dashboards for real-time social media data in future.

Keywords: Features, Sentiment Analysis, Visualization, Sentiment Cloud.

1 Introduction

Subjectivity and Sentiment Analysis (SA) play an important role in identifying the state of mind of humans, *e.g.*, behavior, sentiment and opinion. Subjective analysis aims only on identifying whether the text is subjective or objective, whereas sentiment analysis focuses on identifying the sentiment polarity of text whether it positive, negative or neutral. Till now, most of the SA research work has been carried out on text data only due to availability of dataset and resources. With the growth of the Web, people are using social media platform to share their sentiments. They are gradually making use of videos, images and audios (*e.g.*, podcasts) to express their sentiments on social media platforms [1]. Therefore, it is very important for crawling users' sentiments from different modalities like text, audio as well as videos.

1.1 Types of Sentiment Classification

Sentiment classification can be done in two ways, *i.e.*, binary and multi-class sentiment classification. In binary sentiment classification, each sentence s_i in document D is classified as a label in a predefined category set C , where $C = \{positive, negative\}$ and $D = \{s_1, s_2, \dots, s_n\}$. In multi-class sentiment classification, each sentence s_i in document D is assigned as a label in C^* , where $C^* = \{strong\ positive, positive, neutral, negative, strong\ negative\}$. Since few years back, researchers are also working on classifying the document or text into different moods like happy, sad, anger, disgust, joy and anticipation etc.

This paper is organized into 6 sections as follows. Section 2 discusses the applications of sentiment analysis and challenges are discussed in Section 3. Section 4 presents the desirable features of diverse modalities for sentiment analysis. The desirable visualization features are represented in Section 5. Section 6 concludes the paper and presents future implications.

2 Applications of SA

The research work in the field of sentiment analysis is growing exponentially from the last few years. For example, it plays a vital role in tracking teacher's performance by analyzing students' comments in education field [5]. It also helps the government to know about their strength and weakness by analyzing the opinions of public. Sentiment analysis helps the companies to know about their reputation, services and products in the market from the opinions shared by people on the web [4]. Sentiment analysis plays an important role in almost every field. In the next section, a brief description about the challenges which arise while performing SA is presented.

3 Challenges of SA

Researchers are working in the field of sentiment analysis all over the world. Still, there raise many challenges while performing sentiment analysis of text, audio and visual data. In case of text data, sentiment of word changes from domain to domain. Sarcastic sentences are also difficult to identify. Conference resolution is the major problem while performing

aspect based sentiment analysis. It is also difficult to identify the synonyms in case of reviews because sometimes people use different words for the same feature, For example, “voice” and “sound” represents the same feature for domain of phone reviews [4]. In case of audio data, it is very difficult to differentiate the sentiments like anger and surprise using acoustic features only [3].

4 Desirable Features of SA for Diverse Modalities

Features are an important part of an effective sentiment analysis system. There exist different features for different formats of data. The brief description about the desirable features of diverse modalities is given as follows.

4.1 For Text Data

Due to the use of ambiguous and complex words in text; writing style; politeness and variability of language from person to person and from culture to culture, identification of sentiments from text become challenging [1]. Sentiment analysis can be performed at three levels, *i.e.*, phrase level, sentence level and document level. For example, consider the hotel review, “We stayed at the hotel last Christmas and I loved it! The rooms were clean and the food was great, but if I had one complaint, it would be that the hotel was very expensive and the staff was a bit rude.”

In this review, the sentences “I loved it”, “The rooms were clean” and “the food was great” represent the positive sentiment while “the hotel was very expensive” and “the staff was a bit rude” represent negative sentiment about the hotel. The overall sentiment about the hotel is positive in the case of this review. If we talk about a particular topic or sub-topic like “food” and “staff” of hotel, then above review represents the positive sentiment about “food” and negative sentiment about “staff”.

Some of the desirable features for performing effective sentiment analysis of text data are briefly explained as follows.

(i) Concept Extraction

Concepts are formed on the basis of the syntactic structures of the sentences. These help in performing SA at the feature level.

For example: The movie is boring. (1)

In (1), 'movie' is in a subject related to 'boring'. Here the concept (boring-movie) is extracted.

(ii) Language Detection

Language detection is an important feature in sentiment analysis. For example, India is a land of many languages. People from different regions express their sentiment in different forms or different languages. The same words may have different meanings in different languages.

(iii) Part of Speech (POS) Tagging

POS tagging of text may also help in identifying sentiment from text. POS tagging of text may vary depending upon the domain for which SA is being performed.

For example: "Fear and Loathing in Las Vegas." (2)

"I was absolutely loathing every minute of it." (3) In case of movie reviews, 'loathing' acting as a gerund in sentence (2) has fairly neutral sentiment while 'loathing' acting as a verb in sentence (3) is overtly negative.

(iv) Theme Extraction

In this, mainly the topic is identified from the text. In sentence (4),

Law_government_and_politics is the main theme of the sentence.

For example: Obama is the president of the United States. (4)

(v) Entity Recognition

Since opinions are usually targeted at an entity, so entity recognition also plays an important role in sentiment analysis. Entity may be the name of person, company, product or organization. For example, Obama, President and United States are the three entities in the sentence (4). Entity Recognition may help also in performing feature-based SA by identifying different features of an entity. Consider the phone review given in (5), "picture quality" and "battery life" represent different features of entity "Nokia".

For example: "The picture quality of the Nokia is great, however, the battery life is disappointing." (5)

In sentence (5), sentiment about "picture quality" of phone is positive and sentiment about "battery life" is negative. The overall sentiment about phone is neutral as review consists of both positive ("great") and negative ("disappointing") words. Including above features, some of the features that play an important role in case of sentiment shared by people on social media are given as follows.

(vi) Short Forms and Slang

People use short forms or abbreviations on web while writing their opinions. Sometimes, these short forms and slang words indicate the sentiment expressed by people. In sentence (6), “LOL” represents the positive sentiment of the person.

For example: Spent the whole day just playing video games, LOL.(6)

(vii) **Hashtag**

Hashtags also represent the sentiment about the user or the text. The current trends on Facebook by users are particularly inclined towards the expression of their sentiment in the form of hashtags while posting their pictures or status, e.g., #awesome weather, #enjoyed etc. On Twitter, people also use hashtag at the end of the text about the current trending topic. In sentence (7), “#theseguysarecheaters” represent the negative sentiment.

For example: Oh come on referee that was a terrible call #theseguysarecheaters. (7)

4.2 For Audio Data

With the growing amount of music and the demand of human to access the audio information retrieval, audio sentiment analysis is emerging as an important and essential task for various system and applications. Sentiment analysis of audio data can be performed by extracting its audio features [2]. Audio data include a set of features like prosody, temporal, spectrum, harmonics, tempo and chroma, which are briefly explained as follows.

Prosody features include intensity, loudness and pitch that describe the speech signal.

Temporal features also called as time domain features which are simple to extract like the energy of the signal, zero crossing rate.

Spectral features also called as frequency domain features like fundamental frequency, spectral centroid, spectral flux, spectral roll-off, spectral kurtosis, spectral skewness. These features can be used to identify the notes, pitch, rhythm, and melody.

Harmonic tempo is the rate at which the chords change in the musical composition in relation to the rate of notes.

Chroma features are the most popular feature in music and is extensively used for chord, key recognition and segmentation.

4.3 For Visual Content

Facial expressions play an important role in the identification of sentiments in case of visual content. On social media platforms, people use different smileys to represent their different moods or facial expressions in text. A facial expression analyzer can be used to identify the sentiments associated with facial expressions and then can be classified into different sentiment classes [1]. Some of the visual features related to sentiment analysis are discussed as follows.

(i) Emotion Detection or Emoji

From the past few years, people mostly use emotions or emoji on social media platforms like Facebook, Twitter instead of writing their opinion or feeling in the form of text only. They use special symbols for expressing their different moods like smile, anger, sadness, etc.

For example: Got the new iPhone for my birthday!♥♥♥♥ (8)

“♥♥♥♥” symbols in the sentence (8) represent the “happy” mood of the person.

(ii) Image Processing

Sentiment analysis also can be performed from images, e.g., face detection, image tagging and image link extraction. Face detection includes detection of gender, age, height, width, identity, position of the person.

5 Desirable Features for Visualization of Sentiments

An effective visualization of sentiments is an important characteristic for an SA system. It helps the users to easily understand and analyze the sentiments about the data. There are many ways of visualizing the sentiments of a sentiment analysis system which are given as follows.

5.1 Coloring

For visualization, sentences in the document or text can be colored to represent the different sentiments. For example, in the review given in (9), the sentences with negative sentiment are red colored and sentences with positive sentiment are colored green.

“We stayed at the hotel last Christmas and I loved it!The rooms were clean and the food was great but if I had one complaint, it would be that the hotel was very expensive and the staff was a bit rude.” (9)

5.2 Sentiment Word Clouds

Word clouds can also be created to represent the different sentiments. For example, the positive word cloud may represent the important positive words and Negative word cloud may represent the negative words used by people. Figure 1 represents the sentiment word cloud of positive words.



Fig. 1: Sentiment Word Cloud of Positive Words

5.3 Plotting

The different bar charts, line charts and pie charts can be plotted to represent the sentiment of the text. Tweets can be grouped based on the day or time of their creation and the classified sentiment for each day can be then plotted on a timeline. It can also be plotted on the bar charts that at what date and time, people express what kind of sentiment about any entity. Figure 2 represents the day-wise or time-wise sentiment timeline can be plotted that shows how the sentiment towards an entity changes over time.

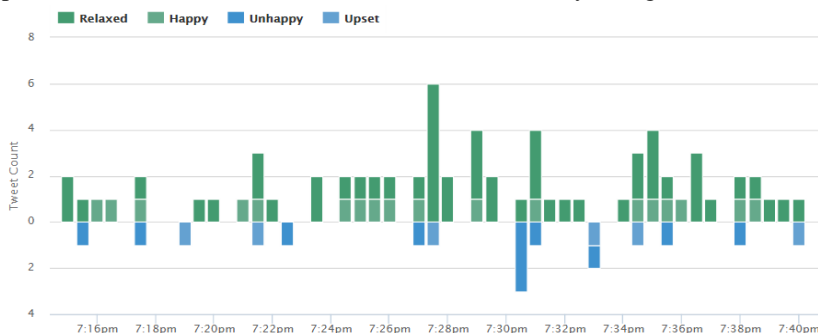


Fig. 2: Time-wise sentiment timeline about “Barack Obama”

5.4 Mapping of Sentiments on Google Maps

In case of data extracted from social media, tweets and Facebook posts can be mapped on Google maps to identify the origin of them using the geo-location feature. Also, tweets or Facebook posts can be colored to analyze the positive and negative sentiment of people about any entity from the region.

6 Conclusions and Future Scope

This paper presents the desirable features to build an effective sentiment analysis system. The important features for diverse modalities like text, audio and visual data are discussed in brief. Also the visualization features like coloring of different sentiments, day-wise or month-wise plotting of sentiments on bar charts, creation of sentiment word clouds and mapping of data on Google map helps the users to easily grasp the sentiments contained in the data. Some of the important applications and challenges are also briefly explained. In the future, an effective web based sentiment analysis can be developed by using features of different modalities and visualization features for real-time social media traffic which can be helpful for the benefits of society.

References

1. Poria, S., Cambria, E., Howard, N., Huang, G., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174, pp. 50-59, (2016).
2. Abburi, H., Akkireddy, E. S. A., Gangashetty, S.V., Mamidi, R.: Multimodal Sentiment Analysis of Telugu Songs. In: 25th International Joint Conference on Artificial Intelligence, pp. 48-52, (2016).
3. Bhaskar, J., Sruthi, K., Nedungadi, P.: Hybrid approach for emotion classification of audio conversation based on text and speech mining. In: *Procedia Computer Science* 46, pp. 635-643, (2015).
4. Vohra, M.S., Teraiya, J.: Applications and challenges for sentiment analysis: A survey. In: *International Journal of Engineering Research and Technology*, 2(2), ESRSA Publications, (2013).
5. Altrabsheh, N., Gaber, M., Cocea, M.: SA-E: sentiment analysis for education. In: 5th KES International Conference on Intelligent Decision Technologies, (2013).

An Efficient Algorithm for Data Field Extraction and Data Cleaning to improve performance of Web Usage Mining

Preeti Rathi¹, Dr (Mrs) Nipur Singh²

¹Research Scholar,

Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, India,

mcapreeti.rathi@gmail.com

²Professor,

Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, India,

nipursingh@gmail.com

Abstract: In Today's scenario web mining is the one of the wide research area. Web usage mining is the one of application of web mining. It contains user information in log files, and it is also called web log mining. Web log mining extract useful pattern or information from log files and it help to determine user behaviour and according to behaviour clustered the data after cleaning.

In this paper we proposed an algorithm for data field extraction and data cleaning and analyze the log files (data) using analyzer tool after analysis the log files we examine the user behavior and cleaning of unwanted data and design a cluster, also proposed an future approach to improve the performance of web usage mining according to user behavior and optimize the result for personalization using clustering techniques implemented in MATLAB.

Keywords: Data Cleaning, Data field Extraction, Cluster, Web log Analyzer.

1. INTRODUCTION

The process involved in the extraction of information pertaining from structured to unstructured or semi-structured in the form of web source is referred to as the web data mining. The information extracted from the web is also called as the web mining.[9] With the help of web data mining one can connect to a website's web pages and request information or pages, exactly as one's browser would do. In turn, the task of the web server is to send the html web page whose sole purpose is to extract the particular information from that web page. [6]

The growth of web is tremendous as approximately one million pages are added daily. Web Applications are increasing at giant speed and its users

are increasing at exponential speed. Users' accesses are recorded in the web log files. [1]In today's period it has become important to know the user access mode. Because of the terrific usage of the web, the web log files are growing at a faster rate and the size is becoming huge. So to have a relevant data being resulted or analyzed we can take help of the concept which is known as Web Mining.[7]Web mining involves exploration of web server logs of a website whereas data mining involves techniques to find relationships in huge amount of data (server logs)[5]. There are three category of web mining-

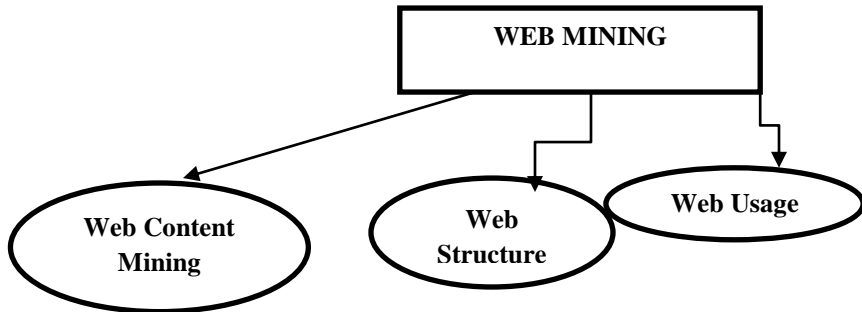


FIGURE-1: CLASSIFICATION OF WEB MINING

1.1 WEB CONTENT MINING

Extracting the knowledge behavior from the contents present in the documents.

1.2 WEB STRUCTURE MINING

Obtaining the knowledge from Internet links.

1.3 WEB USAGE MINING

Obtaining the patterns which are of interest from web log access. The technique of web usage mining involves the mining of data that extract usage patterns and identify the behavior from Web log data. As a whole, web usage mining is divided into pre-processing, discovery of pattern, and analyzing the same for pattern identification. [12]This process is used to find interesting pattern from log files and it helps to access information from log files.[2]The task of pre-processing involves the processing of site files that are untreated and convert the profile of the user data into page classification, site location and server session files [3].The pattern discovery treats a server session file into session rules, patterns, and statistical information. The analysis of pattern identifies the rules, patterns, and statistical information obtained from the pattern discovery process. [4] In data mining process, learning can be categorized as supervised and unsupervised learning technique. In supervised learning a trainer is available, mean to say the training data includes the attributes and their

outcomes. On the other way in unsupervised classification the data contains only attributes there are not any class labels exist. [11]In clustering data can be divided into different cluster and design of cluster according to user preference. [10]

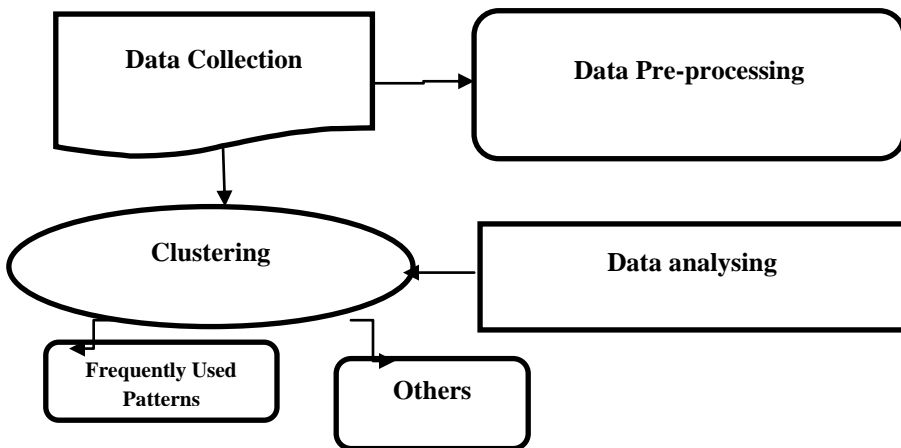
ARCHITECTURE OF PROPOSED WORK

In web usage mining first step collection of data, after data is collected from server logs we pre-processing of data after pre-processing pattern discovery and analyzing performed.

In this paper we proposed as an approach to improve the performance of web data. This is the architecture of our proposed work. We divide this architecture into following steps-

1. Data Collection
2. Data Pre-processing
3. Data Analyzing
4. Data Clustering

FIGURE-2: ARCHITECTURE OF PROPOSED WORK



PROPOSED WORK DATA COLLECTION

We have collected log data from website server. Web log data contains information about website visitors, IP-address, host name, Username, timestamp, method, path, protocol, status code and user information.

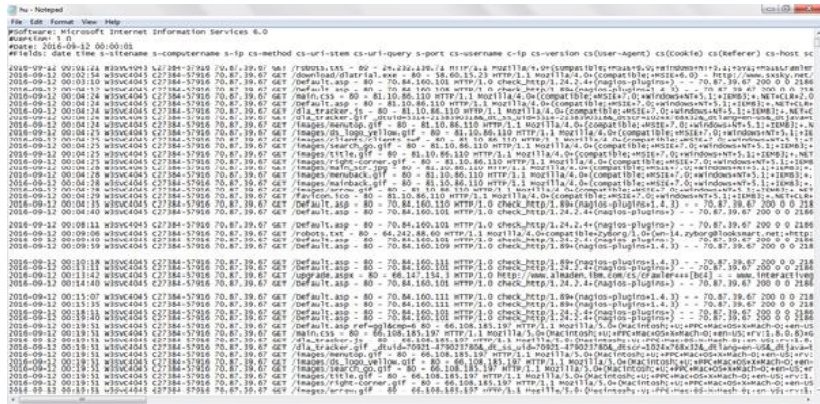


FIGURE -3:SAMPLE OF LOG FILE

The sample web log file has recorded information for each access as:

- a) IP Address: Remote hostname or IP address number.
- b) Remote User: The remote login name of the user. (If Remote user name is not present the – sign is normally used)
- c) User: The username as which the user has authenticated himself. This is available when using password protected WWW pages. (If not exists - sign is normally used)
- d) Timestamp: Date and time of the request.
- e) Request: The request line exactly as it came from the client.
- f) Method: The method is used to retrieve from request line.
- g) Status code: The HTTP response code returned to the client. Indicates whether or not the successfully retrieved, and if not, what error message was returned. (i.e. 200,400)

- h) Bytes: The number of bytes transferred.
- i) Referrer: The URL, the client was on before requesting your URL. (If it not exists then – sign be placed for that place)
- j) User Agent: The User Agent is whatever software the visitor used to access this site. (i.e. Mozilla)

LOG DATA FIELD EXTRACTION

A log data file (LDF) consists of various data fields. Data field extraction separates data fields before cleaning process. This process of separating different data fields from single server log entry. The implementation of the Data field extraction algorithm is in JAVA language. We collected log file data from server log files. The separated fields of the log file are saved into another file. We have calculated log file size and counted number of records and execution time of file.

ALGORITHM-1

Input: - Log Data File (LDF)

Output: - Extracted Log File (ELF)

Step 1: Read text data from log data file in read mode

Step 2: Open another file named Extracted Log File in write mode to write the extracted data.

Step 3: While {Read log data from log file until end of file.

If next line {

Read one line and write in extracted log file.}

}

Step 4: Calculate size of extracted log file and number of records.

Step 5: Close both log and extracted files.

TABLE-1: SUMMARY OF LOG FILE

	Extracted Log File
Size of File	3,77,65,942 bytes
No of Records	142568
Time Taken(Execution)	2 min 8 sec

3.2 DATA PRE-PROCESSING

Data pre-processing is important phase of Web usage mining. It is very complex process and takes 75% of total mining process. Pre-processing is necessary, because log file contain noisy, missed, irrelevant and unambiguous data which may affect result of the mining process. It is an important step to filter and organize appropriate data before applying any

web mining algorithm. The objective of data pre-processing is to improve the data quality and increase the accuracy and reduce the time in the mining process. In our proposed data cleaning algorithm of log files reduced the size of file 36.0MB to 15.4 MB. We have determined different type of status code, methods, and suffix in log files after cleaning we clean this record from log file and write into another log file. In pre-processing reduced the access time and reduced the memory consumption. Before cleaning access time is more in compare to after cleaning.

Log file cleaning algorithm contains those data entries in the log file whose status code is either 400 or 200, method is GET or POST or File_Ext is except from js, xml, txt, gif,jpg,png and css.

ALGORITHM-2

Input: Extracted Log File (ELF)

Output: Summarized Log File (SLF)

Step 1: Read data from server log file i.e. extracted log file

Step 2: While(read until EOF) {

 Read data.status_code

 Read data.method

 Read data.File_Ext

 If (data.status_code=400||200&&data.method= GET||POST && data.File_Ext! = css || xml|| js|| png||jpeg||gif)

 {
 Write data in Summarized log file;

 }
 Else remove another records from extracted log file;

 }
Step 4: above two step repeat until end of file

Step 5: Calculate size of CLF and number of records.

Step 6: Close Summarized log file.

TABLE-2: PARAMETER DESCRIPTION

S.No	Parameter	Value	Consumption in %
1	Status Code	400 or 200	0.04%
2	Method	GET or POST	-
3	File_Ext	jpeg,png,bmp,mp3,xml,js,cgi	0.56%

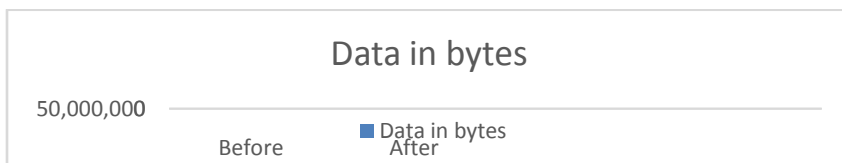


FIGURE-4: DATA CLEANING

3.3 WEB LOG ANALYSIS

In this step we analysis of data using Deep Analyzer Tool. It is a fast and powerful web access log analyzer tool used for predictive analytics[8]. It also generate information about site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. It helps to produce easy to read reports including text information and charts. Web log analyzer is used to analysis the website information.



FIGURE-5: VISITOR HISTORY

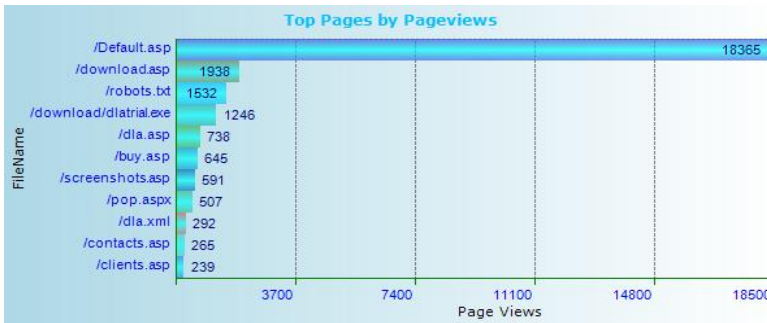


FIGURE-6: PAGE VIEW ACCORDING TO FILE NAME

3.4 DATA CLUSTERING

Clustering the user frequent access data to determine the user behavior. In this step we clustered the data according to the no of hits of a popular webpages. After clustering next step to personalized the data according to user behaviour. For clustering various algorithm are used like fcm, kmean, Genetic algorithmic. In figure- show the sample of cluster.

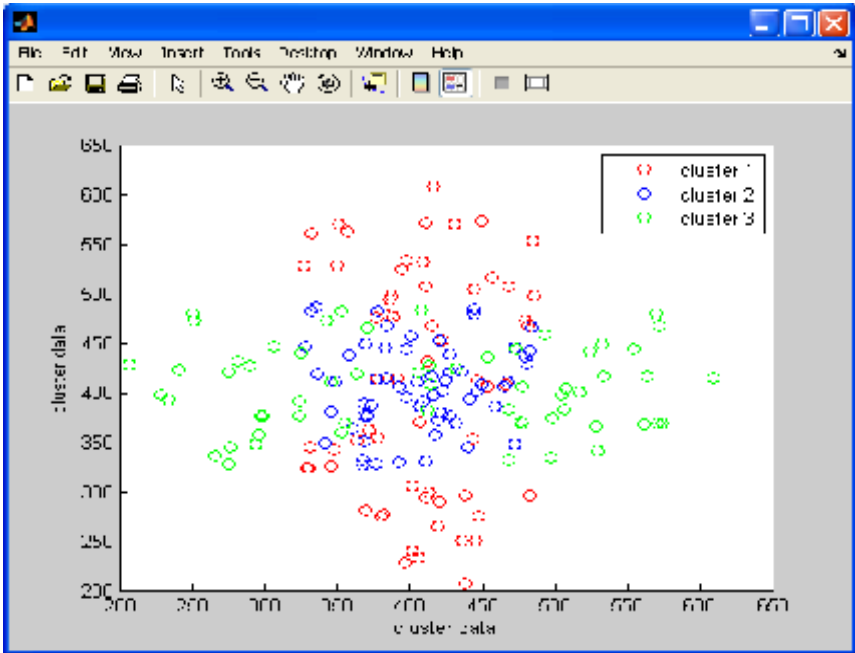


Figure- 7: Sample Of Data Clustered Observation

We have collected data from server logs. To implement field extraction and log cleaning algorithm in Java language as it is good for text processing. For Simulation MATLAB tool is used. For field separation regular expression is used. In Algorithm 1 data field extraction from log files, data size is 37765942 bytes and number of records are 142568. Algorithm 2 reduced the data size from 36 MB to 15.4 MB. We analysis status code (200) for successful requests is 84.25%, status code (400) is 19.05%, above 500 request is 4.5%. Highest number of requests containing GET (99.84%) method. Other method contains POST method and other else is 0.16%. The number of request of image files (it contains jpeg,jpg.png format) (62.48%), css file (10.59%), jss file is4.5% and other files is 0.08%.In below figure show the status code and images files in graphical format.

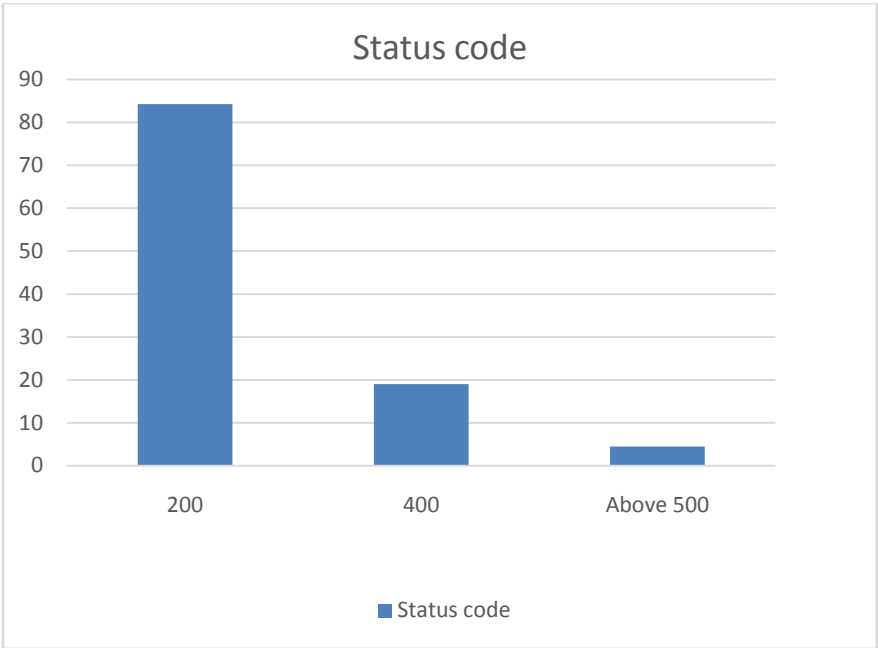


Figure-8: Summary Of Status Code

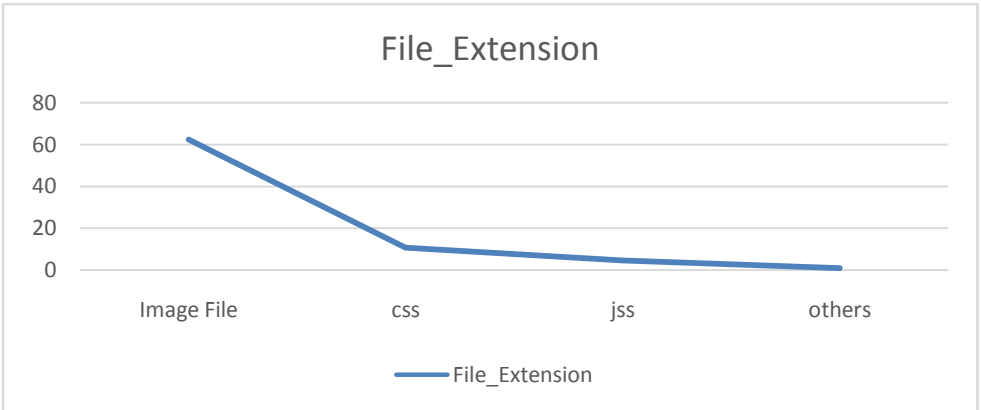


Figure- 9: Summary Of File Extension

CONCLUSION

Due to huge amount of log data, it is necessary to perform pre-processing of data applying mining algorithm on web log data. The goal of data pre-processing is to prepare structured data. Our main focused on data field extraction and data cleaning algorithm. We proposed an algorithm for data

cleaning and data field extraction, and extract useful patterns from log files. In this paper we also analysis the log files using log analyzer to clustering the data. We clustered the data according to user behavior and reduce the memory consumption and time.

FUTURE WORK

Future direction is to propose an algorithm for clustering to better performance to Web Usage Mining and according to user behavior personalized it. For clustering genetic algorithm is used and these algorithm simulate in MATLAB and comparison result to previous running algorithm.

REFERENCES

1. Bhupendra Kumar Malviya, Jitendra Agrawal, "A Study on Web Usage Mining: Theory and Applications", Fifth International Conference on Communication Systems and Network Technologies, IEEE, Page: 935-939, April 2015, ISBN (Print) 978-1-4799-1797-6/15
2. Dr. Girish S. Katkar, Amit Dipchandji Kasliwal, "Use of Log Data for Predictive Analytics through Data Mining", Current Trends in Technology and Science, page-217-222, ISSN: 2279-0535. Volume: 3, Issue: 3 (Apr-May. 2014).
International Journal of Computer Applications (0975 – 8887) Volume 103 – No.6, October 2014
3. M.Praveen Kumar, "An Effective Analysis of Weblog Files to improve Website Performance", International Journal of Computer Science & Communication Networks, Vol. 2(1), Page: 55-60, 2011, ISSN: 2249-5789.
4. Mr. Jitendra B. Upadhyay, Dr. S. V. Patel, "A Review Analysis of Preprocessing Techniques in Web usage Mining", International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 04, April-2015, page -1160-1166,ISSN: 2278-0181
5. Nehal G. Karelia, Prof. Shweta Shukla, "Data Preprocessing: A Pre requisite for Web Log Files", International Journal of Engineering Research & Technology (IJERT), page-1571-1574, Vol. 3 Issue 4, April – 2014, ISSN: 2278-0181
6. Oren Etzioni, "The World-Wide Web: Quagmire or Gold Mine?" ACM, Vol. 39, No. 11, November 1996, Page: 66-68.
7. Sameer Dixit, Navjot Gwal, "An Implementation of Data Pre-Processing for Small Dataset",
8. Saurabh Choudhry, Prof A. K Solanki "Errors in Internet Log files for Website Improvement and Interaction", International Journal of Advanced Research in Computer Science and Software Engineering, Page-365-371, Volume 4, Issue 10, October 2014, ISSN- 2277 128X
9. Shakti Kundu, "An Intelligent approach of web data mining", International Journal on Computer Science and Engineering, page-919-928, Vol. 4 No. 05 May 2012, ISSN: 0975-3397.

10. Sheetal A. Raiyani, Rakesh Pandey, Shivkumar Singh Tomar, "Performance Enhancement of Web Server log for Distinct User Identification through different Factors", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 6, June 2014, Page: 7262-7267, ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940.
11. Shivaprasad G., N.V. Subba Reddy, U. Dinesh Acharya," Knowledge Discovery from Web Usage Data: An Efficient Implementation of Web Log Preprocessing Techniques", International Journal of Computer Applications (0975 – 8887) Volume 111 – No 13, February 2015
12. Surbhi Anand , Rinkle Rani Aggarwal "An Efficient Algorithm for Data Cleaning of Log File using File Extensions ", International Journal of Computer Applications (0975 – 888)Volume 48– No.8, June 2012
13. V.Chitraa, Dr.Antony Selvadoss Thanamani ," A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011

Comparative Approach of Various Image Denoising Techniques Using Filters

Amanpreet kaur

Dept.of Computer Engg. & Technology
Guru Nanak Dev University, Amritsar, (Pb.) India
amanpreet.kaur051993@gmail.com

Abstract:Image denoising is a major problem in the image processing field. Numerous algorithms are implemented which has its own advantages and disadvantages. Having knowledge regarding noise type present in the image helps in choosing an best denoising algorithm. Various types of noise are present and different denoise filters are developed to remove noise from degraded images and improve quality of image by preserving edges. This paper presents the different methods for removal of noise and it also shows the deep incite about the methods that offer reliable and appropriate estimation regarding original image when its degraded form is given.

Keywords: Image denoising, median filter, mean filter, weiner filter, Independent Component Analysis(ICA).

1 Introduction

Nowadays Digital Images have a crucial importance in daily routine applications like MRI, Satellite TV as well as in field of research and technology. Noise is an unwanted signal which adds up in the original image and degrade it. The main sources of noise are improper equipments, incorrect data acquisition approach, transmission as well as compression [1]. Image denoising is intended for removing noise that damage an image at the time of acquiring it or perhaps transmission, while preserving its visual quality. Hence image denoising is an essential part of image analysis. Therefore, its important to use several effective image denoising approaches in order to avoid degradation. Noise modelling depends upon various aspects like data capturing devices, transmission media, or image quantisation. Various algorithms are often utilized depending on the noise model. In ultrasound images, noise which is detected is speckle noise in contrast to MRI images in which rician noise [3] is detected.

1.1 Various Models of Noise

Noise is actually contained in image in the form of additive or in the multiplicative[4].

1.1.1. Additive Noise Model

Additive noise is combined to the original signal and deliver a corrupted signal that is noisy which is written as:

$$d(u,v) = f(u,v) + n(u,v)$$

1.1.2. Multiplicative Noise Model

Here, noise signal is multiplied with original and written as:

$$d(u,v) = f(u,v) \times n(u,v)$$

Here $f(u,v)$ is the intensity of original image and $n(u,v)$ is the noise imported which results in corrupted signal $d(u,v)$

1.2 Types of Noise

Different noise have different traits and therefore are included into images in several manners.

1.2.1 Gaussian Noise

It is uniformly spread out across the signal. Every single pixel within noisy image would be the total of true pixel value along with random gaussian distributed noise value. It carries probability density function [pdf] of the normal distribution. This is generally called Gaussian distribution [5].



Fig. 1. Image showing Gaussian Noise

1.2.2. Salt and Pepper Noise

Also referred as shot noise, impulse noise or even spike noise which is generally a result of defective memory locations, improper functioning of pixel components in the camera sensors, or due to timing errors within

digitization process. In this noise pixels takes gray level 225 for salt noise and 0 for pepper noise and it appears as black white spots on the images. The salt and pepper noise will have a noise density of $p/2$ if p is the total noise density.



Fig. 2. Image showing Salt & pepper Noise

1.2.3. Speckle Noise

This [7] [8] is a multiplicative noise. It arise in majority of systems like Ultrasound images, SAR images etc. This noise basically originates because of random interference among the coherent returns.

1.2.4. Amplifier Noise

This is additive noise, and is not dependent on every pixel as well as signal intensity. It contains a Gaussian distribution that contains bell shaped probability distribution function (pdf) written in equation as:

$$F(g) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(g-m)^2 / 2\sigma^2}$$

Some of the types of models for the noise:

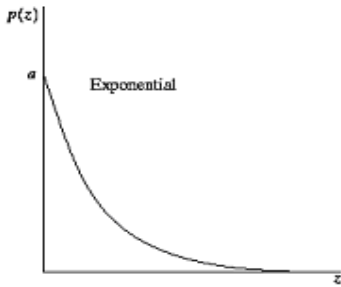
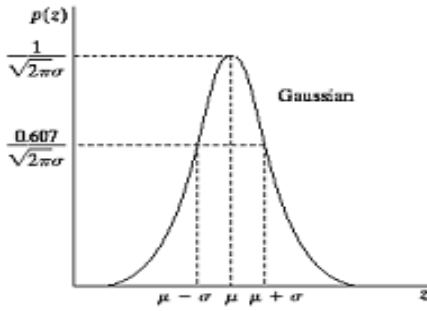


Figure 3 Gaussian Noise **Figure 4** Exponential noise

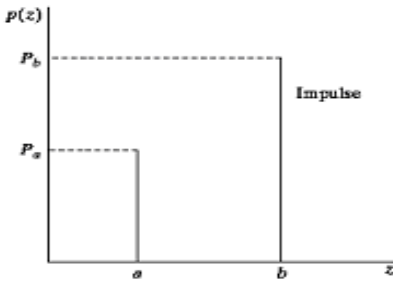
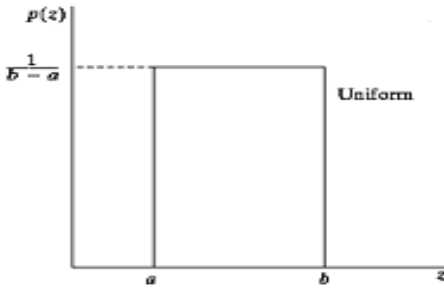


Figure 5 Uniform Noise **Figure 6** Impulse (Salt and pepper) Noise

1.3 Various Denoising and Filtering Techniques:

Various denoising techniques exist today and their usage varies according to image type and noise present. Image denoising is categorized in three types explained below. Filtering approach have various objectives like:

- To suppress the noise effectively in uniform regions.
- To preserve edges and other similar image characteristics.
- To provide a visually natural appearance [9].

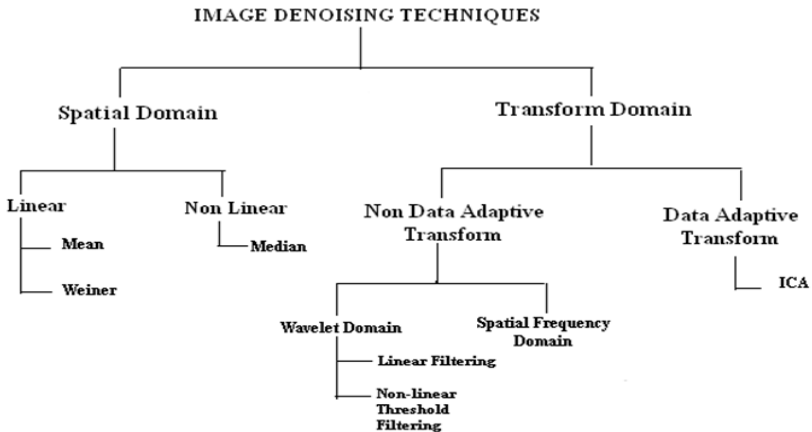


Fig. 7. Classification of Image Denoising Techniques

1.3.1 Spatial filtering

It is a conventional approach to eliminate noise out of images. It is of two types linear and non-linear filter [10].

1.3.1.1 Linear Filters

These filters destroy lines and other details in a picture, blur the sharp edges. Types of linear filter are:

a) Mean Filter

It works upon an image by means of smoothing it. It diminishes intensity modifications among the neighborhood pixels [11]. It is an averaging linear filter. Here within a predefined area the average value of image having noise is calculated and the intensity value of center pixel is replaced by means of average value of pixels in the neighborhood. This method will be done again for all pixel values within whole image.

b) Weiner Filter

This approach[12] need the details regarding noise spectra along with original signal and works properly if the underlying signal is continuous. This approach utilizes spatial smoothing. $H(u',v')$ would be degradation function, $H(u',v')^*$ will be conjugate complex. $G(u',v')$ will be corrupted image.

$$f(\mathbf{u}', \mathbf{v}') = \left[\frac{H(\mathbf{u}', \mathbf{v}')}{\left(H(\mathbf{u}', \mathbf{v}')^2 + \left[\frac{Sn(\mathbf{u}', \mathbf{v}')}{Sf(\mathbf{u}', \mathbf{v}')} \right] \right)} \right] G(\mathbf{u}', \mathbf{v}')$$

1.3.1.2 Non- Linear

Nowadays several different non-linear filters like, weighted median, relaxed median rank conditioned, rank selection have come up to conquer the drawbacks of linear filter. Using non-linear filter, noise is eliminated with no efforts to identify it. Spatial filters eliminate the noise to an acceptable degree but it leads to blurring that will hide the edges in the image.

a) Median Filter

It is a best order static filter in which feedback depends upon location of pixel values based upon rank enclosed within the region of filter. This filter works well in case of salt and pepper noise. This filters act like smoothers in image and signal processing. It has the ability to remove the consequence of input noise values with huge degree which is its advantage[13].

1.3.2 Transform Domain

This filtering approach is categorized based on the functions. These functions are divided into (i) Non-data adaptive functions (ii) Data adaptive functions.

1.3.2.1 Non-Data Adaptive Transforms

(a) Spatial-Frequency Filtering

It uses Fast Fourier Transform (FFT) with low pass filters (LPF). In Spatial-Frequency method, denoising is achieved by deciding a cut-off frequency.

(b) Wavelet Domain

This is categorized into two methods called (i) linear (ii) non-linear.

(1) Linear Filters: Wiener filter is the generally used linear filtering method which yields most valuable results in the wavelet domain. It is utilized when data degradation is represented as Gaussian procedure and when accuracy measured as mean square error (MSE) [14], [15].

(2) Non-Linear Threshold Filtering: It uses wavelet transform that maps noise in signal domain to the that in transform domain. Two different types of thresholding functions are utilized i.e Hard threshold [16] and Soft threshold. If the input is bigger than the threshold, then it is kept as a Hard-Thresholding function, it is set to zero otherwise. The input arguments are reduced toward zero by the threshold, called Soft-thresholding function [17]. The result may still be noisy. Signal with large number of zero coefficients is produced by large threshold. This leads to a smooth signal. Selection of an optimal threshold is done with great attention.

1.3.2.2 Data-Adaptive Transforms

ICA is the most widely implemented approach in blind source partition problem. Its benefit is that signal is assumed to be Non-Gaussian that allow image denoising using Non-Gaussian and Gaussian distribution

Table 1: Comparison between various filtering techniques

Sr.No	Filtering Techniques	Features	Advantages	Disadvantages
1	Homomorphic Wavelet[3]	Threshold can be extended that gives better result	Reduce speckle noise	Complex technique
2	Soft Thresholding [4]	“Optimal recover model and Statistical inference “	Reduce as well as smooth the noise	Large threshold cuts the coefficients
3	Non Homomorphic [5]	Relies on characterization of marginal statistics of signal	Reduce the computational complexity of filtering method	Not a robust method for estimation distribution parameters
4	Adaptive wavelet domain Bayesian processor [6]	Combines the MAP estimation with correlated speckle noise	Speckle noise suppressed and remaining structure of image is not effected	Not effective technique
5	Wavelet based statistical [7]	Use realistic distribution of wavelet coefficients	Feature preserve , better for medical images, fast computation	Highly complex

6	Versatile technique for visual enhancement [8]	Combining MAP and speckle and signal wavelet coefficients	High correlation and structure similar and quality index	Cover only medical images not other
7	Wavelet thresholding (normalshrink) [9]	Sub band adaptive threshold	Normalshrink is faster as compare to bayeshrink	Need to reduce the number of bits while using normalshrink
8	Joint optimization quantization and wavelet packets[10]	Covers both us images and natural images	Highly compressed approach	Cost function is high
9	Curvelet and contourlet[11]	Noise improvement rectangle	High PSNR can be achieved	Consider only Gaussian noise not other noises

2 Literature Survey

Table 2: Comparison of State of Art Techniques

Ref No	Title	Classes	Tools/Techniques	Advantages
[3]	Homomorphic wavelet thresholding technique for denoising medical ultrasound images	Image denoising	Novel Homomorphic Wavelet Thresholding	It outperform the most effective wavelet based denoising

[4]	De-Noising by Soft-Thresholding	Image denoising	Abstract De-Noising Model	Increases statistical Inference
[5]	Robust non-homomorphic approach for speckle reduction in medical ultrasound images	Speckle reduction	Non-Homomorphic technique	Low complexity
[6]	Locally adaptive wavelet domain Bayesian processor for denoising medical ultrasound images using Speckle modeling based on Rayleigh distribution	Speckle reduction	Discrete Wavelet Transform, MAP estimator	Suppresses speckle noise effectively
[7]	A Wavelet Based Statistical Approach for, Speckle Reduction in Medical Ultrasound Images	Speckle reduction	Novel Multiscale Nonlinear for Speckle Reduction	Fast computation and better diagnosis
[8]	A versatile technique for visual enhancement of medical ultrasound images	Visual enhancement of image	Versatile Wavelet Domain despeckling	Provide better performance in speckle smoothing and edge preservation
[9]	Wavelet-based statistical	Speckle reduction	Novel Speckle-Reduction	Fast computation

	approach for speckle reduction in medical ultrasound images			and Despeckling
[10]	Medical ultrasound image compression using joint optimization of thresholding quantization and best-basis selection of wavelet packets	Image denoising	Image Coding Algorithm	Performance of JTQ-WP coder is concluding better
[11]	Performance evaluation of wavelet, ridgelet, curvelet and contourlet transforms based techniques for digital image denoising	Image denoising	X'let transform	Provide effective denoising
[14]	Denoising Of Medical Ultrasound Images In Wavelet Domain	Image denoising	Wavelet Transformation, Wavelet Thresholding	Preserves image and visual quality
[15]	Image Denoising using Wavelet Thresholding	Image denoising	Adaptive Threshold Estimation	Provide smoothness and Effective edge preservation
[16]	Image denoising using curvelet transform: an approach for edge	Image denoising	Soft Thresholding Multiresolution	Improve smoothness

	preservation			
[17]	Ideal spatial adaptation by wavelet shrinkage	speckle reduction	Signal-dependent Multiplicative Speckle Noise Model, DWT	Smoothness increases

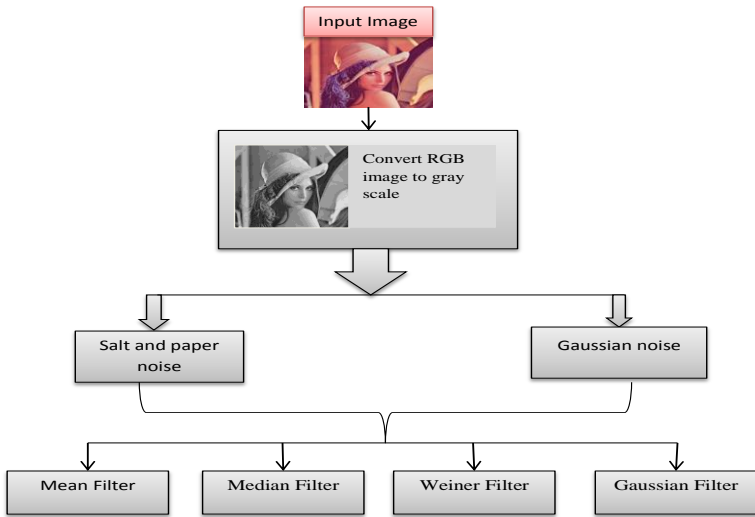
Table 3: Exploring Medical Denoising Techniques

Title and References	Dataset	Features	Tools/Techniques Used	Classification Approach
Title: “An Efficient Denoising Technique for CT Images using Window based Multi-Wavelet Transformation and Thresholding [32]	CT images of size 256X256	PSNR values computed, Additive White Gaussian Noise removed	Window based Multi-wavelet transformation , band pass filtering method	“Multi-wavelet classification on windows based”
Title: “A GA-based Window Selection Methodology to Enhance Window-based Multi-wavelet transformation and thresholding	industrial CT volume data sets	Number of window selected, Gene length, Mutation Rate, PSNR values	Window based Multi-wavelet transformation , Genetic algorithm	Window Based Multi-wavelet classification

aided CT image denoising technique.”[33]				
Title: “Qualitative and Quantitative Evaluation of Image Denoising Techniques” [34]	Standardised Images	CoC, PSNR and S/MSE	Median Filter, Lee Filter, Kuan Filter, Wiener Filter, NormalShrink, BayesShrink	Image Denoising Using Spatial Filters
Title: “Multilevel Threshold Based Image Denoising in Curvelet Domain” [35]	5000 images of sizes: 64 × 64, 128 × 128, 256 × 256, 512 × 512, 1024 × 1024	Mean and median of absolute curvelet coefficients	Curvelet Transformation and Cycle spinning	Curvelet based Thresholding
Title: “Digital Image Denoising in Medical Ultrasound Images: A Survey” [36]	Ultrasound images	Scattered density, Texture based contrast, MSE, RMSE, SNR, and PSNR	Multi-scale thresholding, Bayesian Estimation and Coefficient correlation, Application of Soft Computing like ANN, Genetic Algorithms and Fuzzy Logic	designing better algorithms correlating the Ultrasound image formation concepts

Title: “Mixed Curvelet and Wavelet Transforms for Speckle Noise Reduction in Ultrasonic B-Mode Images”[37]	Six ultrasonic B-mode images (Liver, Kidney, Fetus, Thyroid, Breast and Gall	PSNR value, Coefficient of Correlation(CoC)	Wavelet and curvelet transform	Wavelet transform handles homogeneous areas and Curvelet transform handles areas having edges
Title: “Image Denoising Method based on Threshold, Wavelet Transform and Genetic Algorithm”[38]	Images of Lena and Saturn Planet	Hard Threshold Function, Soft Threshold function	Wavelet Transform, Genetic Algorithm	Genetic Algorithm

4. Methodology



5.Simulation results

Figure 8 represents the implementation of mean, median , wiener and gaussian filter on different images corrupted due to salt and pepper noise and similarly Fig. 9 represents the implementation corrupted due to gaussian noise.

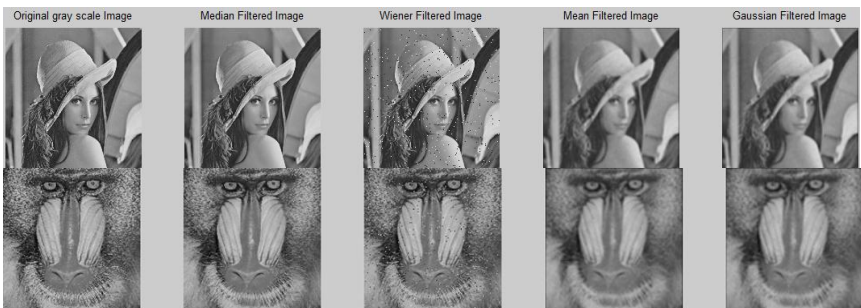




Fig 8. Filtered images of lena, bamboo, cameraman, and pirate corrupted by salt and pepper noise

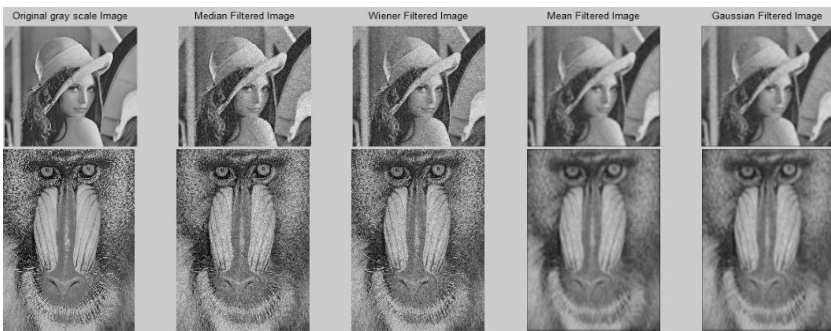


Fig 9. Filtered images of lena, bamboo, cameraman, and pirate corrupted by gaussian noise



by gaussian noise

Performance analysis of different filter for different types of noises is quantized through Mean Square Error (MSE) value and Peak to Signal Noise Ratio (PSNR) value.

TABLE 4: PSNR and MSE Values For Gaussian Noise Of Different Images

Image Name	Mean Square Error	PSNR
Lena	195.26	25.26
Bamboon	193.88	25.29
Cameraman	183.67	25.52
Pirate	195.59	25.25

6. Conclusion

In this paper, various noise models and filtering techniques like linear, nonlinear filtering have been discussed. There experimental analysis shows the original input image being passed through different noises giving the corrupted image which is then passed through filters to get the enhanced high quality image. In future various hybrid filtering techniques can be used to enhance the result

References

- [1] Kenneth, R. 1979, Digital Image Processing, Prentice Hall, New Jersey.
- [2] Stefan Schulte, Mike Stefan Schulte, Mike Nachtegaele, Valerie De Witte, Dietrich Van der Weken, Etienne E. Kerre, "Fuzzy Random Impulse Noise removal from Image sequences"
- [3] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing", Pearson Education, Second Edition, 2005.
- [4] Matlab6.1 —Image Processing ToolboxII, [http://www.mathworks.com/access/helpdesk/help/toolbox /images/](http://www.mathworks.com/access/helpdesk/help/toolbox/images/)
- [5] Umbaugh, S. E. 1998, Computer Vision and Image Processing, Prentice Hall PTR, New Jersey.
- [6] Priyanka Kamboj, Versha Rani, "A Brief Study Of Various Noise Model And Filtering Techniques", Journal of Global Research in Computer Science, 4 (4), April 2013, 166-171
- [7] Gagnon, L. 1999. Wavelet Filtering of Speckle Noiseome Numerical Results, Proceedings of the Conference Vision Interface, Trois-Reveres.
- [8] Goodman, J. W. 1976. Some fundamental properties of Speckle, J. Opt. Soc. Am., 66, pp. 1145–1150.
- [9] Yousef Hawwar and Ali Reza, "Spatially Adaptive Multiplicative Noise Image Denoising Technique", IEEE Transaction On Image Processing, December 2002, Vol.11, No. 12.
- [10] Motwani, M.C., Gadiya, M. C., Motwani, R.C., Harris, F. C Jr. Survey of Image Denoising Techniques.

- [11] Windyga, S. P. 2001, Fast Impulsive Noise Removal, IEEE transactions on image processing, Vol. 10, No. 1, pp.173-178.
- [12] Kailath, T. 1976, Equations of Wiener-Hopf type in filtering theory and related applications, in Norbert Wiener: Collected Works vol. III, P.Masani, Ed. Cambridge, MA: MIT Press, pp. 63–94 .
- [13] Vikas Gupta, Dr. Vijayshree Chaurasia Dr. Madhu Shandilya, “A Review on Image Denoising Techniques”, International Association of Scientific Innovation and Research (IASIR) IJETCAS 13-340; © 2013, IJETCAS
- [14] V. Strela. “Denoising via block Wiener filtering in wavelet domain”. In 3rd European Congress of Mathematics, Barcelona, July 2000. Birkhäuser Verlag.
- [15] H. Choi and R. G. Baraniuk, "Analysis of wavelet domain Wiener filters," in IEEE Int. Symp Time- Frequency and Time-Scale Analysis, (Pittsburgh), Oct.1998.
- [16] Pan, Q. 1999, Two denoising methods by wavelet transform, IEEE Trans Signal Processing, vol. 47, pp. 3401-3406.
- [17] D. L. Donoho, “De-noising by soft-thresholding”, IEEE Trans. Information Theory, vol.41, no.3, pp.613-627, May1995.
- [18] Florian Luisier, Thierry Blu, Brigitte Forster and Michael Unser, “Which Wavelet bases are best for Image Denoising”, SPIE Proceedings, Vol. 5915,
- [19] Sept.17,2005. S. Gupta, R.C. Chauhan, S.C. Saxena, Homomorphic wavelet thresholding technique for denoising medical ultrasound images, Taylor & Francis Int. J. Med. Eng. Technol. 29 (5) (2005) 208–214.
- [20] Donoho, D. L., 1995, De-noising by soft-thresholding. IEEE Transactions on Information Theory, 41, 613–627.
- [21] S. Gupta, R.C. Chauhan, S.C. Saxena, A robust multi-scale non-homomorphic approach to speckle reduction in medical ultrasound images, IEE J. Int. Fed. Med. Biol. Eng. 152 (1) (2005) 129–135.
- [22] S. Gupta, R.C. Chauhan and S.C. Saxena, “Locally adaptive wavelet domain Bayesian processor for denoising medical ultrasound images using Speckle modelling based on Rayleigh distribution”, IEEE Proc. On Vis. Image Signal Proces., Vol. 152, No. 1, February 2005, pp 129-135.
- [23] Savita Gupta, L. Kaur, R.C. Chauhan and S. C. Saxena, “A wavelet based statistical approach for speckle reduction in medical ultrasound images, Medical Image Processing, TENCON 2003, pp 534-537.
- [24] Gupta S., Kaur L., Chauhan R.C and Saxena S.C. (2007), “A versatile technique for visual enhancement of medical ultrasound images”, Digital Signal Processing, Elsevier, Vol. 17, pp 542-560.
- [25] S. Gupta, R.C. Chauhan, S.C. Saxena, A wavelet based statistical approach for speckle reduction in medical ultrasound images, IEE J. Int. Fed. Med. Biol. Eng. 42 (2004) 189–192.
- [26] L.Kaur, S.Gupta, R.C.Chauhan, S.C.Saxena , “Medical ultrasound image compression using joint optimization of thresholding quantization and

best-basis selection of wavelet packets”, Digital Signal Processing,2007, vol 17,pp 189-198.

[27] Vipin Milind Kamble, Pallavi Parlewar, Avinash G. Keskar, Kishor M. Bhurchandi, “Performance evaluation of wavelet, ridgelet, curvelet and contourlet transforms based techniques for digital image denoising”, *ArtifIntell Rev* (2016) vol 45 pp 509-533. doi: 10.1007/s10462-015-9453-7

[28] Amit Jain , “Denoising Of Medical Ultrasound Images In Wavelet Domain”, *International Journal Of Engineering And Computer Science* ,Volume 4 Issue 5 May 2015, Page No. 11871-11875.[29] Lakhwinder Kaur, Savita Gupta, and R. C. Chauhan, "Image Denoising Using Wavelet Thresholding," in *Indian Conference on Computer Vision, Graphics and Image Processing*, December 2002, pp. 1-4.

[30] Anil A. Patil and Jyoti Singhai (2010), “Image Denoising using Curvelet Transform: an approach for edge preservation” *Journal of Scientific & Industrial Research*, Vol. 69, pp 34-38, Jan 2010.

[31] D. L. Donoho and I. M. Johnstone, “Ideal spatial adaptation via wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425-455, 1994.

[32] Syed Amjad Ali, Srinivasan Vathsal and K. Lal kishore ,“An Efficient Denoising Technique for CT Images using Window based Multi-Wavelet Transformation and Thresholding”,*European Journal of Scientific Research*, Vol.48 No.2 (2010), pp.315-325

[33] Syed Amjad Ali, Srinivasan Vathsal and K. Lal kishore, “A GA-based Window Selection Methodology to Enhance Window-based Multi-wavelet transformation and thresholding aided CT image denoising technique”, *International Journal of Computer Science and Information Security*, Vol. 7, No. 2, February 2010, pp 280-288

[34] Bedi C.S. and Goyal H. (2010), “Qualitative and Quantitative Evaluation of Image Denoising Techniques”, *International Journal of Computer Applications*, Vol.8 No.14, pp 31-34, October 2010

[35] Binh N.T. and Khare A. (2010), “Multilevel Threshold Based Image Denoising in Curvelet Domain”, *Journal of Computer Science and Technology*, Springer, Vol. 25 No.3, pp 632–640, May 2010.

[36] N. K. Ragesh, A. R. Anil, Dr. R. Rajesh, “Digital Image Denoising in Medical Ultrasound Images: A Survey”, *ICGST AIML-11 Conference*, Dubai, UAE, 12-14 April 2011, pp 67-73.

[37] A.A. Mahmoud, S. El Rabaie, T.E. Taha, O. Zahran, F.E. Abd El-Samie and W. AlNauimy(2015), “Mixed Curvelet and Wavelet Transforms for Speckle Noise Reduction in Ultrasonic B-Mode Images”, *Information Sciences and Computing*, pp 1-21

[38] Yali Liu, “Image Denoising Method based on Threshold, Wavelet Transform and Genetic Algorithm”, *International Journal of Signal Processing, Image Processing and Pattern Recognition* Vol. 8, No. 2 (2015), pp. 29-40

A Comparative study on various retinal vessel segmentation techniques

Naina Singh¹, Aarti²,

¹ M-Tech Student, Department of Computer Engineering & Technology, Amritsar College of Engineering and Technology, Amritsar, Punjab, India-143001

naina.raj92@gmail.com

² Assistant Professor, Department of Computer Engineering & Technology, Amritsar College of Engineering and Technology, G.T. Road, NH-1, Amritsar, Punjab, India-143001

aarti.acet@yahoo.com

Abstract. This paper represents the retinal images exact detection of the vessel is an important and hard task. In pathological images, detection is difficult with the presence of exudates and abnormalities. The motive of segmentation is to really make the illustration of the image easier so that it can be more smoothly analyzed. Many ways of retinal vessel segmentation are planned which could identify the exudates in fundus images in more encouraging manner. The overall objective of this paper is to utilize Neighborhood Estimator before Filling (NEBF) called inpainting filter which is used to inpaint exudates so that false positive are reduced during vessel enhancement.

Keywords: Image Segmentation, retina, vessel segmentation, exudate, in painting.

1 Introduction

1.1. Image Segmentation

Image segmentation is process connected with partitioning an electronic image in to several segments. The prospective connected with segmentation is usually to simplify as well as to modify the rendering connected with an image in to a little something that is definitely additional significant and easier in order to analyze. Image segmentation is generally utilized to find out materials and also limitations inside images. Image segmentation is definitely the procedure connected with delegating the indicate to each pixel inside a picture along with exactly the same content label share specified visible traits In computer perspective, graphic segmentation is normally debris partitioning an automated a digital photo in to many portions generally known as super-pixels. The intention of segmentation should be to make simpler and to alter the manifestation associated with a photo into one thing that's more special and much easier in order to analyze. Image segmentation is commonly employed to seek out things and restrictions within images. Image segmentation is usually particles determining a brand

to each pixel in the photo techniques pixels with the exact same brand talk about specified characteristics.

1.2 Retinal Vessel Segmentation

Retinal vessel segmentation algorithms are often the middle little bit of developed retinal sickness diagnosis systems[4]. Segmentation strategy by means of two-dimensional retinal drawings acquired coming from a retina photographic camera in addition market research involving approaches will probably be presented. By means of examining and revealing of vasculature components within retinal illustrations, we could very early diagnose a type 2 diabetes within sophisticated levels in comparison of the declares of retinal bloodstream vessels. Segmentation of veins within retinal illustrations will allow very early diagnosis of disease, using this method gives many benefits. A vascular system usually performed manually is the time-consuming method demanding coaching as well as knowledge. Automating task will allow consistency, above all, saving plenty of time of which an expert professional or even medical doctor would commonly apply regarding information screening.

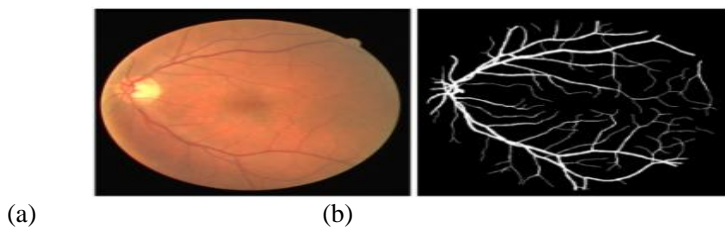


Figure 1 (a) Color Retinal Image (b) Blood Vessel Segmentation

2Related Work

Koozekanani et al. [1] has showed unsupervised iterative vein segmentation algorithm employing fundus images. Very first, some sort of exceptional image can be made by tophat renovation for the nonpositive environment friendly planes image. A estimation towards segmented vasculature is certainly utilized through foreign thresholding the highest image. Akramet al. [5] has presented a story way of correct detection with drusen within coloured retinal images. The unit makes use of filtration system traditional bank to get all achievable drusen regions kind retinal image as well as eradicates fake pixels which usually seem as a consequence of likeness with drusen along with optic disc. This technique signifies each and every district along with several attributes and is applicable assist vector machine to categorize these types of regions because drusen as well as non-drusen. Giachetti et al. [9] has presented an account way of the automated site and segmentation on optical disk in retina images. It is actually judging by involving charter vessel and also historical past info acquired together with morphological segmentation plus painting. Hector et al. [10] presented

almost any statistical impression taking on strategy of point drusen with all the finest intention involving characterizing a AMD further development in a really details band of longitudinal images. The tactic characterizes retinal structures through a math form for the colorings inside the retina image. Yin et al. [11] presented a way that fuses edge detection, the actual Rounded Hough Change plus a mathematical deformable unit to be able to diagnose the actual optic blank disc through retinal fundus images. The result suggests an increased link together with floor truth segmentation thereby exhibits an outstanding potential for using this method being integrated compared to other retinal CAD systems. Guoliang et al. [12] has presented a robust framework to on auto-pilot segment tricky exudates throughout fundus images. The framework is definitely based on a course-to-fine tactic, because they initially purchase a harsh result permitted connected with a number of adverse examples. The major element donations from the cardstock are generally: (1) recommend a proficient and sturdy framework regarding automatic HEs segmentation; (2) provide your raised smooth segmentation to mix multiple channel data; (3) setting a double band filtration system to segment an OD region. Lee et al. [14] has presented the kind of segmentation opportunity for quantification in GA simply by hypo fluorescent GA locations business enterprise interfering retinal vessel structures. Firstly, they are going to make use of history lamination static correction applying some kind of non-linear adaptable removing operator. The actual technically obvious parts of hypo-fluorescence had been sketched by a specialist Salemand et al. [15] has presented the sectioning of retinal blood stream with a coloring fundus images which comprises of attribute vector consists of not one but two scale-space attributes - the most significant eigenvalue as well as gradient value - in the high-intensity impression, that represent both the attributes of every ship and also similar perimeters, and also the natural sales channel impression intensity. Walter et al. [16] has proposed the latest formula pertaining to recognition with exudates. Good exudates within the macular place will be the main trademark with suffering from diabetes macular edema and also allows their recognition by using a higher sensitivity. Therefore, recognition with exudates is a crucial diagnostic job, through which laptop support may well perform a primary role. Exudates are only using their higher grayish amount alternative, in addition to their curves tend to be motivated by using morphological reconstruction techniques.

Table 1. Comparison table.

Ref no.	Author	Year	Technique	Features	Limitations
[1]	Koozekanani, et al.	2015	Iterative Vessel Segmentation	Completely new vessel pixel are generally	modest micro aneurysms at the

				identified by adaptive thresholding	vasculature, will likely be involved because an element of the vasculature due to the district improve business
[2]	Koozekanani, et al	2014	Blood Vessel Segmentation	Novel and natural interaction modality, large-scale quantitative evaluation.	Owing to many architectural challenges the option would be never scalable plus efficient.
[3]	Kafieh, et al.	2013	3-D vessel segmentation method	Rotation invariant parts-based model to detect objects	The solution is definitely not thought to be efficient
[4]	Odstrcilik et al.	2013	Retinal vessel segmentation	Use of the powerful Bag-of-Words model for recognition	Evolutionary technique has not been considered.
[5]	Akram et al.	2013	Automated drusen segmentation	Some sort of multistage convolution circle trained through fresh pixels so that you can extract thick feature vectors	A Meta heuristic technique has not been considered.
[7]	Rozlan et al.	2012	blood vessels segmentation	Computerized Targeted Detectors in	A number of the executive difficulties

			n	High-Resolution Remote control Stinking Images	accepted as overlooked
[9]	Giachetti et al.	2011	Multiresolution localization	Easy, rapidly, and also premium quality objectness evaluate	Bounding package might not exactly localize the item circumstances when properly as being a segmentation place.
[10]	Hector et al.	2011	Drusen segmentation	Detects further advancement in our longitudinal details established	Insufficient precise style in addition to stretching the item to characterize different retinal wounds

3Retinal Vessel Method

Detecting blood vessel stream in retinal illustrations or photos is actually a step by step procedure which will if not followed will not contribute to quality segmentation. Ahead of the segmentation procedure, image preprocessing is usually important to get rid of the consequence connected with improper brightness along with noise items aside from enhancing the compare regarding the background the retinal our blood vessels. Research has shown which the products image segmentation will depend on the products retinal image. A strongly elevated image is usually segmented having a segmentation strategy which can be additionally threshold to have the side map. Lastly, post handling of the threshold image is usually done to get rid of incorrect pixels or cut off pixels.

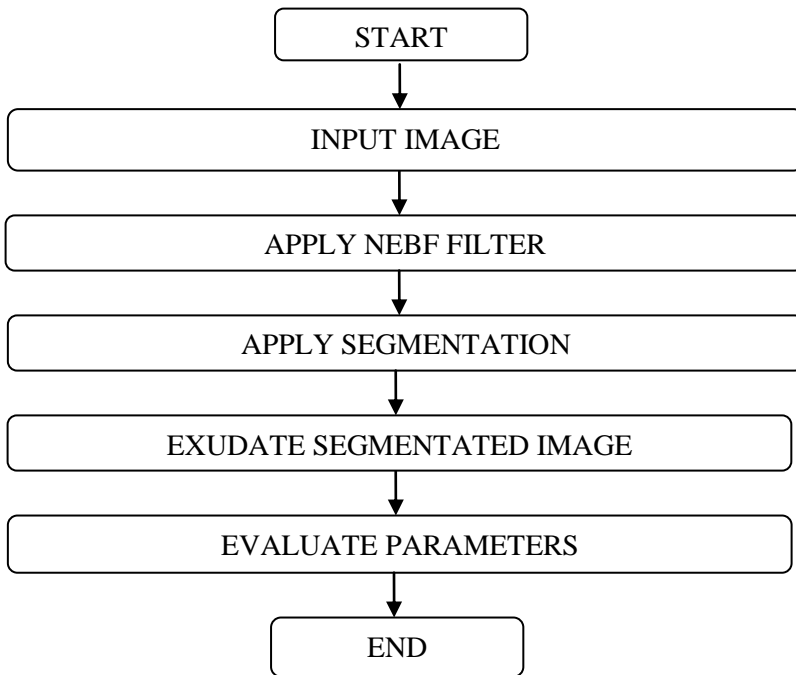


Figure 2 Methodology of retinal vessel segmentation

4.Research Gap

Roberto Annunziata, et al. [1] shows that neighborhood Estimator Before Filling (NEBF), is suggested to inpaint exudates inside regional false positives usually are significantly lower in the course of vessel enhancement. By conducting the survey it has found that the majority of existing exudate segmentation techniques suffers from the following issues.

1. The issue of noise in retinal vessel images is ignored in the majority of existing literature.
2. Poor computational speed can also be considered as major gap found in most of the retinal images.
3. Transform based methods which can be used to improve the speed of segmentation which was ignored in this.

5.Mathematical Formulation

A. Image Preprocessing for Exudate Detection

Normally, exudates search significantly better from boats. Even so, nonuniform illumination plus in homogeneities generate unfeasible a straightforward gray-level thresholding pertaining to discovering them. Indeed, exudate p in the retina photo which share the similar gray[4] degree

of terribly poor constrained. To cope with these complaints, a new preprocessing point comparable to is used.

This phase comprises of:

1. Non uniform light correction;
2. image homogenization;

We work with a massive mean narrow with regard to qualifications estimation. The hepa filter may be determined because it's specially able to close to getting rid of arteries and without clouding tips connected with larger regions inside background. After that, this believed qualifications I actually is actually subtracted through saving money station connected with the main graphic I actually to get the big different image.e. D:

$$D(x, y) = I(x, y) - I_{med}(x, y)$$

A lighting effects remedied impression I_C is usually obtained simply by levels of gray of D image linearly to protect main choice of feasible strength values[9]. A histogram from the brightness-corrected impression I_C is usually homeless to the core of the gray range, simply by changing p strength based on

$$g_{Output} = \begin{cases} 0, & \text{if } g < 0 \\ 255, & \text{if } g > 255 \\ g, & \text{otherwise} \end{cases}$$

Where

$$g = g_{Input} + 128 - g_{Input_M}$$

and g_{Input} and g_{Output} states the gray-level valuations on the input and also the productivity photos(I_C and I_H). The particular homogenization move is definitely in line with the proven fact that the backdrop is composed of great importance and more pixels versus forefront, therefore, the intensity price affiliated for the mode with the histogram presents the backdrop value.

B. Inpainting of Exudate

A novel inpainting filter (Algorithm 1), named neighborhood estimator before filling (NEBF) to assist in finding the exudates and filling them up.

Algo 1 NEBF
 $ExMk \leftarrow \text{dilate}(ExMk);$
 $TmInp \leftarrow \text{OrIm}(ExMk \neq 0)=0;$
while exudates \neq inpainted **do**
 $ExMk \leftarrow \text{erode}(ExMk);$
 $TInp \leftarrow \text{call } ExInp(TmInp, ExMk);$
end while

$$\text{ImInp} \leftarrow \text{TmInp};$$

Where ExMk(Exudate Mask), OrIm (Original Image), ExInp(Exudate Image), TmInp(Temporary Input).

Algo 2 $I = \text{ExInp}(I, \text{ExMk})$
 $\text{PxToFill} \leftarrow \text{ExMk} - \text{erode}(\text{ExMk});$
 $\forall \mathbf{p} \in \text{PxToFill} \text{ ill}|\text{PxToFill}(\mathbf{p}) \neq 0$
 $I(\mathbf{p}) = \text{mean } I(\mathbf{q})$
 $\mathbf{q} \in N_{\mathbf{p}}$
 $I(\mathbf{q}) = 0$
 $N_{\mathbf{p}} = \{\mathbf{q} \in N, |\mathbf{q} - \mathbf{p}| \leq r\}$

The criteria proceed iteratively into the core of exudate [1]. With a careful limit in order to diagnose exudates typically under segments every exudate, causing a narrow boundary with unseen pixels. The reconstruction leading o linear opening is employed to clear away scaled-down and inadequately contrasted exudates certainly not found in previous steps [2]. It can be stated as

$$\min_{\gamma B} \gamma B(I, I)$$

where $\gamma B(I)$ is described as the particular morphological launching regarding I along with B as structuring ingredient

C. Multiscale Hessian Eigen value Analysis for Vessels Enhancement

Hessian-based methods possess highly efficient with retina vessel enhancement. The main element strategy would be to acquire main directions when the community second-order structure from the photograph might be decomposed. Analyzing the boat, the largest eigenvalue (λ_2) from the Hessian matrix will be in accordance with the tiniest eigenvector. On the other hand, the tiniest eigenvalue (λ_1) affiliated for the premier eigenvector will be in-line with the entire vessel.

The largest eigenvalue, λ_{max} is attained as

$$\lambda_{max} = \max_s \frac{\lambda_2(s)}{s}$$

All of us discovered that only using the greatest eigenvalue is sufficient to get charter boat advancement, in the past revealed solutions which usually use of both λ_1 and λ_2 .

6. Performance Measures

To quantify performance, we use Sensitivity (S_n), Specificity (S_p), Positive Predictive Value (PP_v), Negative Predictive Value (NP_v), and Accuracy (Ac_c). These measures are:

$$S_n = \frac{T_p}{T_p + F_n}$$

$$S_p = \frac{T_n}{T_n + F_p}$$

$$PP_v = \frac{T_p}{T_p + F_p}$$

$$NP_v = \frac{T_n}{T_n + F_n}$$

$$Ac_c = \frac{T_p + T_n}{T_p + F_n + T_n + F_p}$$

where T_p (true positives), F_p (false positives), F_n (false negatives), and T_n (true negatives) tend to be bought by considering exclusively pixel. S_n and S_p will be measured precisely well-classified pixel as well as the non-vessel pixel, correspondingly [11][12]. PP_v is definitely the ratio of correctly categorized vessel pixel. NP_v is precisely properly labeled nonvessel pixel. Ac_c is the proportion associated with real ends in the collection of pixel.

7 Conclusion

Accurate vessel recognition performed in retinal images is a significant and tedious process. Automated segmentation of fundus image represents a significant role in detection of eye diseases. Lately, a few types of retinal vessel segmentation are planned which could detect the exudates in fundus images in more promising manner. The NEBF filter has shown great usefulness in inpainting exudates Detection regarding vessel as well as Retinal constructions mixed collectively may resolve the issue regarding highly accuracy in segmentation strategy. In this paper it has discussed the comparison of various techniques based on vessel segmentation and also discusses the performance measures by taking parameters. The review has shown that the issue of noise in retinal vessel image is ignored in fundus images. In future we will evaluate the effectiveness of Adaptive NEBF filter based retina image segmentation technique in cellular domain.

References

1. Roberto Annunziata, Andrea Garzelli, Lucia Ballerini, Alessandro Mecocci, and Emanuele Trucco. "Leveraging Multiscale Hessian-based Enhancement with a Novel Exudate Inpainting Technique for Retinal Vessel Segmentation"(2016)
2. Roychowdhury, S., D. D. Koozekanani, and K. K. Parhi. "Iterative Vessel Segmentation of Fundus Images." (2015).

3. Roychowdhury, S., D. D. Koozekanani, and K. K. Parhi. "Blood Vessel Segmentation of Fundus Images by Major Vessel Extraction and Sub-Image Classification." (2014).
4. Kafieh, Rahele, Hossein Rabbani, FedraHajizadeh, and MohammadrezaOmmani. "An accurate multimodal 3-D vessel segmentation method based on brightness variations on OCT layers and curvelet domain fundus image analysis." *Biomedical Engineering, IEEE Transactions on* 60, no. 10 (2013): 2815-2823.
5. Odstreilik, Jan, Radim Kolar, Attila Budai, Joachim Hornegger, Jiri Jan, Jiri Gazarek, Tomas Kubena, Pavel Cernosek, Ondrej Svoboda, and Elli Angelopoulou. "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database." *IET Image Processing* 7, no. 4 (2013): 373-383.
6. Akram, M. Usman, SundusMujtaba, and Anam Tariq. "Automated drusen segmentation in fundus images for diagnosing age related macular degeneration." In *Electronics, Computer and Computation (ICECCO), 2013 International Conference on*, pp. 17-20. IEEE, 2013.
7. Muthu Rama Krishnan, M., U. Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, Eddie Yin-Kwee Ng, Milind M. Mushrif, and Augustinus Laude. "Application of intuitionistic fuzzy histon segmentation for the automated detection of optic disc in digital fundus images." In *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, pp. 444-447. IEEE, 2012.
8. Rozlan, Ahmad Zikri, N. S. Mohd Ali, and HadzliHashim. "GUI system for enhancing blood vessels segmentation in digital fundus images." In *Control and System Graduate Research Colloquium (ICSGRC), 2012 IEEE*, pp. 55-59. IEEE, 2012.
9. Yin, Fengshou, Jiang Liu, Damon Wing Kee Wong, NganMeng Tan, Carol Cheung, Mani Baskaran, Tin Aung, and Tien Yin Wong. "Automated segmentation of optic disc and optic cup in fundus images for glaucoma diagnosis." In *Computer-based medical systems (CBMS), 2012 25th international symposium on*, pp. 1-6. IEEE, 2012.
10. Giachetti, Andrea, Khai Sing Chin, EmanueleTrucco, Caroline Cobb, and Peter J. Wilson. "Multiresolution localization and segmentation of the optical disc in fundus images using inpainted background and vessel information." In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 2145-2148. IEEE, 2011.
11. Santos-Villalobos, Hector, Thomas Paul Karnowski, DenizAykac, Luca Giancardo, Yaquin Li, T. Nichols, K. W. Tobin, and Edward Chaum. "Statistical characterization and segmentation of drusen in fundus images." In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 6236-6241. IEEE, 2011.
12. Yin, Fengshou, Jiang Liu, SimHeng Ong, Ying Sun, Damon WK Wong, NganMeng Tan, Carol Cheung, Mani Baskaran, Tin Aung, and Tien Yin Wong. "Model-based optic nerve head segmentation on

- retinal fundus images." In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 2626-2629. IEEE, 2011.
13. Fang, Guoliang, Nan Yang, Huchuan Lu, and Kaisong Li. "Automatic segmentation of hard exudates in fundus images based on boosted soft segmentation." In *Intelligent Control and Information Processing (ICICIP), 2010 International Conference on*, pp. 633-638. IEEE, 2010.
 14. Kong, Lingwang, Qiong Li, and Shanhu Huang. "Color Image Segmentation Scheme for Retinopathic Fundus." In *Computer Science and Software Engineering, 2008 International Conference on*, vol. 6, pp. 237-240. IEEE, 2008.
 15. Lee, Noah, Andrew Laine, and R. Theodore Smith. "A hybrid segmentation approach for geographic atrophy in fundus auto-fluorescence images for diagnosis of age-related macular degeneration." In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 4965-4968. IEEE, 2007.
 16. N.M.SalemandA.Nandi, "Segmentation of retinal blood vessels using scale-space features and k-nearest neighbor classifier," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings, May 2006*, vol. 2, pp. II-II.

Blood Mate – An Android Application for blood donors and receptors

Kavita Pabreja

Ph.D., Associate Professor, Head of Department of Computer Science
Maharaja Surajmal Institute, New Delhi-110058,
kavita_pabreja@rediffmail.com

Akanksha Bhasin

Student, Bachelor of Computer Applications
Maharaja Surajmal Institute, New Delhi-110058,
akankshabhasin36890@gmail.com

Abstract: Blood is the supreme element in any individual's life support system. It is the elementary unit in the existence of mankind without which the living being's survival in this world is impossible. Additionally, the process of blood formation takes place only by natural means i.e. it can neither be synthesized nor be produced artificially in any laboratory. Hence, the only alternative left for the fulfillment of various blood needs and requirements is through blood donation. Also, people are less aware about the characteristic properties of their blood. The need of this application is felt in an attempt to generate awareness among people and discuss the various traits that are specific to each blood group. As the name says, "Blood Mate", it definitely explains to the user every possible detail about his blood group largely in simple and basic terms that can be well understood by any individual. The major features of this app embraces details about each blood group based on diet, personality and career choices that can be taken by an individual, supplying various tips for blood donation and predicting the blood group of an unborn child as well. The present paper is based on an attempt to investigate the various features of the android application named "Blood Mate" developed by authors.

Keywords: Blood groups, donor, receptor, Android, Java Development Kit, Diet, Map activity

I. Introduction

In India, October 1st is observed as the ‘National Blood Donation Day.’ The red stuff that oozes out of a person’s body after a paper cut or logically speaking, a fluid that circulates constantly in our body, providing us with the basic nutrition, oxygen and waste removal, is called as Blood. As explained by the author in [1], it is not just a simple fluid, but a group of cells suspended in it along with proteins that not only support our life system but also enables us to fight diseases and infections. Blood can neither be manufactured nor there do any substitute available for it. Hence, Blood Donation is the only source through which blood can be received. As stated by the author [2], blood group system was discovered in 1901 by Karl Landsteiner. So far 19 major groups have been identified of which “ABO” and “Rhesus” groups are of major importance. The genetics of blood groups is proved by the fact that specific diseases are common in particular blood group; for example: duodenal ulcers in ‘O’ blood group, gastric cancer in ‘A’ blood group.

Analysis of the problem to create one mobile application that is complete in itself, has been done thoroughly which can be defined as breaking up of any whole so as to find out their nature, function etc. It defines design as to make preliminary sketches of; to sketch a pattern or outline for plan, to plan and carry out especially by artistic arrangement or in a skilled manner. System analysis and design can be characterized as a set of techniques and processes, a community of interests, a culture and an intellectual orientation. The various tasks in the system analysis include the following.

- Understanding application.
- Planning.
- Scheduling.
- Developing candidate solution.
- Performing cost benefit analysis.
- Recommending alternative solutions.
- Supervising, installing and maintaining the system.

This system provides the information about the diet specific to each blood group along with the ability to distinguish between the various personality traits amongst the blood groups, predicting the possible career choices as well as finding the blood group of an unborn child. The system also aims at

the inclusion of various data mining results so as to generate results that can further be added to the success of the app. Along with this; general blood donation tips have also been provided. The system is interactive, and has an easy to use interface.

II Literature Review

The existing and available applications on the Google Play Store deal with only a single or at most two of the factors. This leads to the problem for the user as he/she has to download a variety of applications which can lead to a lot of memory wastage over and above filling the mobile phones with a lot of useless material. But apparently, there is no common platform for users to access all facilities together. This creates a need for the android users to download multiple apps and surf each one separately for each of their functionality. In conjunction, people have to consult various dieticians and nutritionists so as to maintain a healthy diet as well as sustaining their well-beings which steers to wastage of time and money. This is the chief shortcoming of the existing systems.

Following is the list of reviewed applications-

- 1) **Blood Groups and You-** This application [3] supplies information about only two aspects *viz.* the blood donation tips along with prediction of the blood group of the unborn child based on the input from the parents i.e. taking their blood groups' as an input.
- 2) **Diet-** This Android application [4] imparts details about only one attribute i.e. about the diet that is specific to each blood group. It does not support a good user interface as it consists of a slide show comprising of 11 screens that describe about the diet pertaining to each blood group.
- 3) **Blood Type-** It also incorporates [5] information about the diet for each group besides letting the user know from which group he can receive blood and to which group he can donate blood i.e. the possible donor-receptor groups.

As evident, this creates a need for the android users to download multiple applications and surf each one discretely and exclusively for each of their functionality that shams as the major drawback of the prevailing system.

III Proposed System

To overcome the weaknesses and problems that lie in the existing android applications, this project has been evolved. Our app **“Blood Mate”** caters to this pre-requisite by offering a unified platform to overcome the abstruseness and deficiencies of the above mentioned applications accompanied with additional features in details. It aims to reduce the time and money wastage by providing an integrated platform for all the modules. It also targets to provide precautionary measures to be taken, by advising the blood donation tips through a simple application. The system or app provides an easy to use interface. The various other advantages of the proposed system are:

- Trouble free to use
- Relatively faster than traditional methods
- Highly reliable
- Easy and interactive GUI
- Easy to operate and maintain
- Use of latest technological means

Instead of downloading multiple applications and wasting space, our app **“Blood Mate”** aims to deliver a single integrated platform for all facets related to blood.

IV Methodology Used

The software(s) needed to develop this application are as follows:

1) Android Studio -

Android Studio provides the fastest tools for building apps on every type of Android device. World-class code editing, debugging, performance tooling, a flexible build system, and an instant build/deploy system all allow you to focus on building unique and high quality apps. Android software development is the process by which new applications are created for the Android operating system. Applications are usually developed in Java programming language using the Android software development kit (SDK), but other development environments are also available. The Android SDK includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on Quick Emulator, documentation, sample code, and tutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows XP or later. Following are the features of Android Studio:

- Instant Run

- Intelligent Code Editor
- Fast & Feature Rich Emulator
- Robust & Flexible Build System
- Designed for Teams
- Optimized for all Android Devices

2) Java Development Kit -

The Java Development Kit (JDK) is an implementation of either one of the Java Standard Edition, Java Enterprise Edition or Java Micro Edition platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris, Linux, Mac OS X or Windows. The JDK includes a private Java Virtual Machine and a few other resources to finish the development of a Java Application. Since the introduction of the Java platform, it has been by far the most widely used Software Development Kit.

3) Adobe Photoshop -

Photoshop is a graphics-editing program that is used to create and manipulate images. The program's versatile nature makes it useful for a huge range of imaging tasks. This software has been used in this project to creatively design a simple but attractive and efficient user interface design for the users. All tools present on the app screen have been designed using Adobe Photoshop.

V. Project Insights

‘Blood Mate’ android application deals with postulating a cohesive and amalgamated platform for various applications that are present on Google Play Store and incorporating the features of all those applications into a single component so as to ease the glitch of numerous downloads of diverse applications thereby condensing the wastage of memory space and time. Likewise it stipulates the details in a friendly and simple language so that any individual can understand it in an utmost convenient and competent manner. The application can currently be divided into five modules. The authors have created a dashboard which includes the assimilation of the following modules:

1) Blood Diet

The module named as ‘Blood Diet’ which provides information on the diet to be followed by a particular blood group.

2) Donor- Receptor Groups

The module named as 'Donor-Receptor Groups' which provides the information on the possible recipient and donor blood groups i.e. to which blood group the user can donate the blood and from which blood group he/she can receive blood.

3) **Explore yourself**

A module named as 'Explore Your Self' that discusses about the various personality traits and career choices that are usually similar in characteristic blood groups.

4) **Your Baby's Blood Group**

The module named as 'Your Baby's Blood Group' which helps in predicting the blood group of the unborn child by taking the father's and mother's blood groups as an input.

5) **About Donating Blood**

The module named as 'About Donating Blood' which provides tips for effective blood donation as well as precautions that can be taken.

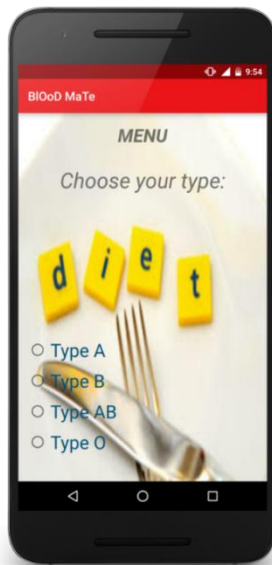


Fig. 1 demonstrating the “Diet” activity which includes a set of radio-buttons based on the blood group of a person

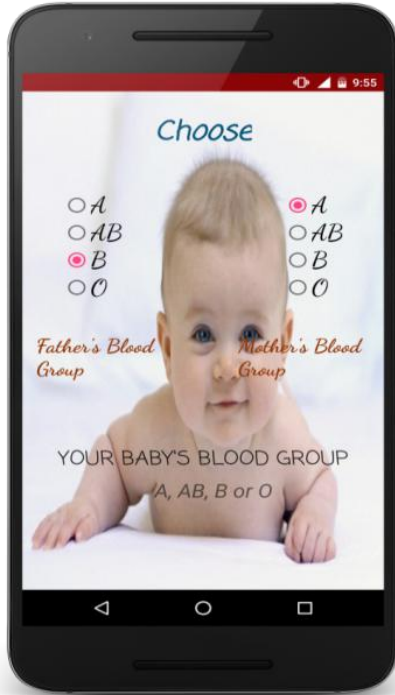


Fig.2 representing the “Baby Blood Group”activity/module

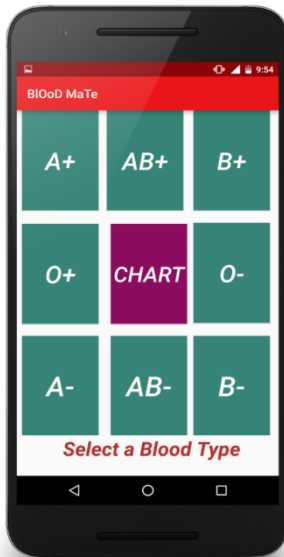


Fig. 3
“Donor-
module.



depicting the
Receptor Groups”

Fig. 4 illustrating the “About Donating Blood” module that provides a list of blood donation tips



Figure 5 :Explaining the "Nearby Places" module.

6) Nearby Places

The module named 'Nearby Places' which provides the location and routes of nearby hospitals and pharmacies that aid the user to trace them conveniently.

VI. Results And Discussions

The screenshots of the application demonstrates its various modules:

- i) The "Diet" module consists of a set of radio-buttons which on clicking displays the list of food items that a particular blood group should follow, as shown in Fig.1.
- ii) The "Baby Blood Group" module entails 2 radio-groups which on selecting a blood group from each group, displays the probable blood group of the unborn child in a text view, as shown in Fig.2. It uses a set of nested if-else constructs where the input is taken from the 2 radio groups; in which each radio group contains the blood group of the child's parents i.e. the mother and father. The input obtained is then used for finding the blood group of the child and is then displayed in a text view.
- iii) The "Donor-Receptor Groups" Module shown in Fig.3 consists of a table of buttons that when clicked, produces the output as the blood groups to which the user can donate and from whom they can receive.
- iv) The "About Donating Blood" module that comprises of a list-view which provides various blood donation tips in the form of a list, as shown in Fig.4.
- v) The "Nearby Places" module, as shown in Fig. 5, comprises of the map activity which consists of 2 buttons namely: Nearby Hospitals and Nearby Pharmacies to aid the users to locate the fore-mentioned

places near them. Each location is marked with a pointer which when clicked, redirects the application to Google Maps where the user can look for the best possible route accordingly.

VII Conclusions And Future Scope

To conclude, “Blood Mate” has a very wide scope in terms of generating awareness among people for understanding the most important aspect of their health – ‘Blood.’ In consort with its various modules, it also delivers an easy implementation environment and generates responses flexibly.

In the future, the user would be able to use Global Positioning Systems (GPS) services so as to aid him/her in locating the nearest clinics and blood banks and also provide them lab facilities at their homes itself. This facility would be chargeable and would offer rewards and perks to donors.

Also, the app aims to further integrate the finger-print sensing technology with which the blood group of an individual can be predicted based on the fingerprint’s pattern.

It is also proposed to extend this study by collecting data from students of various under-graduate courses so as to get a real picture of blood donations and associated beliefs. Data, hence collected, would be mined for extracting hidden facts related to blood donation.

VIII. References

[1] Haniff, F. "Blood and its importance." www.kaieteurnewsonline.com, 22 March 2009.

[2] Narkhede, P. “An empirical study on Blood Types and Personality.” IJSSBT, 2 June 2015.

[3] Android Application present on Google Play Store: <https://play.google.com/store/apps/details?id=blood.ciencia.technologies&hl=en>

[4] Android Application present on Google Play Store: <https://play.google.com/store/apps/details?id=it.betterdays.dietagrupposanguignoandroid&hl=en>

[5] Android Application present on Google Play Store: <https://play.google.com/store/apps/details?id=malix.diet&hl=en>

[6]Android Studio review

from:<https://developer.android.com/studio/index.html>

[7]Java Development Kit review

from:https://en.wikipedia.org/wiki/Java_Development_Kit

[8]Adobe Photoshop review from:<http://kivaindia.com/Photoshop-Institute-In-Ahmedabad-Photoshop-is-a-Graphics-Editing-version-program-that-is-used-to-create-and-manipulate-images-The-program-s-versatile-/b961#>

Survey on big data analytics for cleaner manufacturing and maintenance processes

Pawandeep kaur¹, Dr. Pankaj Deep Kaur²

¹ Research Scholar, Department of Computer Engineering & Technology,
Guru Nanak Dev University Regional Campus ,Jalandhar, Punjab,
143001,India
pkmangat32@gmail.com

² Assistant Professor, Department of Computer Engineering & Technology,
Guru Nanak Dev University Regional Campus ,Jalandhar, Punjab,
143001,India

Abstract. The paper presents that Cleaner production (CP) is considered as the essential means for manufacturing enterprises to achieve sustainable production. It also improves sustainable competitive advantage. The different types of architecture of big data has been discussed i.e. one of the big data based analytics for product lifecycle (BDA-PL) is discussed in this paper, which can be used to offer better help on decision-making of control and optimization on the product lifecycle management (PLM) and the entire CP method. The overall objective of big data architecture is that benefited customers, manufacturers, environment and even all stages of PLM, and effectively. It gives a theoretical and realistic base for the sustainable development of production enterprises.

Keywords: Big data analytics, Data mining, cleaner production, Product lifecycle management, Architecture of BDA-PL, Manufacturing, Maintenance.

1 Introduction

Big data analytics (BDA) is described as the process of collecting, organizing and analyzing of huge sets of data to find the patterns and other meaningful information. The most important characteristics of big data have volume, velocity and variety. Volume describes enormous amount of data, velocity defines the high speed of the data processing and variety identifies the big number of types of data including audio, video, geographical place, etc. Veracity and value both are the two additional characteristics of big data. The two additional characteristics of big data are veracity and value. As manufacturing processes start to use the advanced information

technology to carry out the management, a huge amount of data related to product lifecycle is produced. In field of manufacturing and maintenance, big data includes large amount of heterogeneous, multi- source and real-time data, which is produced during the three stages i.e. manufacture, operation and maintenance stages. Nowadays, manufacturing enterprises prefers to manufacture environmental-friendly products to avoid the pollution threats. So, to achieve the sustainable production, Cleaner Production (CP) is used for manufacturing products in the manufacturing industries. Cleaner production can provide various benefits such as economic, environmental and the social benefits. There are lot of issues related to the implementation of the CP program which includes the lack of information about the clean technologies, insufficient information, less supply of equipment, poor communication systems, available procedures, lack of availability and accessibility for the useful information relevant to product and lack of skills. CP and product lifecycle management (PLM) both strategies are used to improve the sustainable competitive advantage of the enterprises. The main three things that allow the CP strategy to be implemented successfully are capture lifecycle data, discover knowledge from raw data and share knowledge among all lifecycle stages. Manufacturing and maintenance process (MMP) mainly includes Research and Development and Manufacture (RDM) and Operation and Maintenance (OM), respectively.

1.1 Data mining

The whole product lifecycle in the manufacturing enterprises produce huge amounts of data which are collected and accessed by the database management systems. Data mining process plays a vital role to extract knowledge from the manufacturing databases. Knowledge Discovery in Database (KDD) is the name given to this whole process. The KDD process includes various steps such as cleaning of data, integration of data, selection of data, transformation of data, mining of data, evaluation of patterns and then present the knowledge as shown in Fig.1.

Data mining is a step of KDD process for extracting hidden patterns from the transformed data. The various issues of data mining are efficiency, scalability, handling noisy data and mining heterogeneous data. There are lots of data mining methods are used for extracting the useful information from huge amount of raw data. The various models of data mining techniques are clustering, regression, prediction, neural networks

association analysis and classification. Data mining in the manufacturing mainly deals with the two types of the applications. The various applications of data mining process emphasize on the single stage of the product lifecycle and some applications of data mining process focus on the multiple or all stages of the manufacturing and maintenance process (MMP).

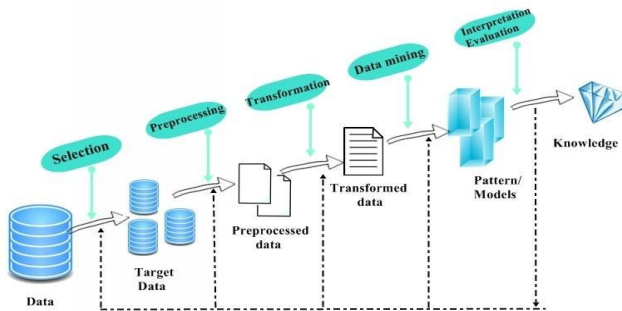


Fig. 1. Knowledge discovery in database process.

1.2 Product Lifecycle Management (PLM)

Product lifecycle management (PLM) introduced for handling the useful information rigorous process consisting of item design, product manufacturing, and product in use and product recycling. PLM can enhance the production of new products and reduce steadily the production cost by managing the things through their lifecycle. A product's lifecycle contains design, manufacturing, utility, maintenance and recycle. These all phases can be divided into three stages: beginning of life (BOL), middle of life (MOL), and end of life (EOL). BOL is that stage of PLM in which concept of merchandise is generated and designed. MOL is that stage in which the distribution of manufactured merchandises takes place, and then these products are utilized by consumers and maintained by developers. EOL is that stage in which the used products are recycled or disposed by customers.

1.2.1 Big data in BOL, MOL, and EOL

For the better efficiency of PLM, finding out which type of data is used in PLM is an important work. It helps the Big Data techniques make conclusions on the basis of huge amount of data at the various stages. The whole framework of big data in PLM is shown in Fig.2.

1.2.1.1 BOL

In BOL phase, the most two important steps are: marketing evaluation and the merchandise design. The most essential task in marketing evaluation is meeting the demands of the customers. The demands of the customers may exist in remarks on websites, videos on the Net and those sites they mark. In

the phase of merchandise design, the data used may be tracked from the explanation of needs to the particular product information and ultimately to the comprehensive design specifications. It is necessary to improve the merchandise design constantly. The current products can be produced more reliably and efficiently.

1.2.1.2 MOL

In MOL phase of PLM, as merchandises and services have endured in the ultimate form, problems related the service have become the significant and must be paid high concentration.

In the logistics phase, effective decision approaches are required to resolve the complicated problems in the management of warehouse or optimization of transport. The most crucial work in this phase is how exactly to convert the order data from consumers into the intelligent arrangements with the worldwide view. In the phase of utility, the consumer can use the product normally, on the basis of the instructions from consumer manual .To give the directions for maintenance stage, the information related to the usage environment will also be recorded and monitored. In the maintenance phase, by adding the maintenance knowledge with the merchandise status information, large deal of the failures may be predicted and stopped before occurring.

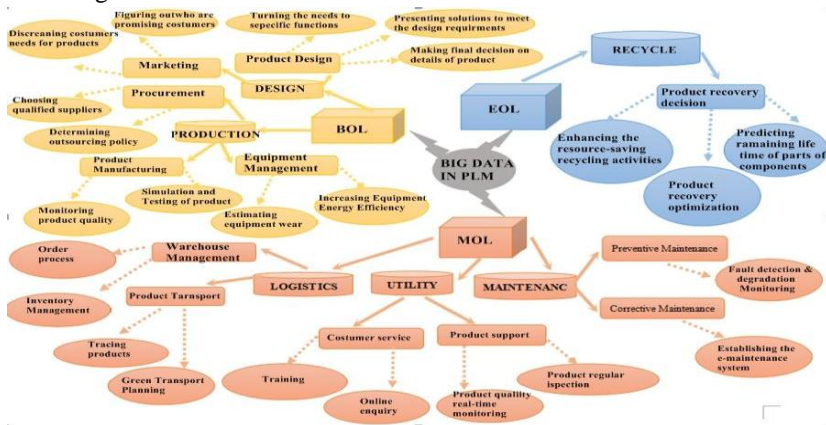


Fig. 2. Big data in PLM

1.2.1.3 EOL

In EOL phase volume conclusion must be taken which matter the EOL item disposable and recycle. By using maintenance history information, from MOL phase the usage environment information, product status information, degradation position and the rest of the value of the parts could be calculated. With aim to increase the use of EOL products, then suitable

EOL recovery choices such as recycle, removal, reuse and remanufacturing should really be decided considering manufactured products status.

2. Architecture of BDA- PL

An overall architecture of big data – based analytics for product lifecycle (BDA-PL) is proposed [23], for making the better CP and PLM decisions based on the large amount of the multi-source and real-time lifecycle big data. This architecture simply integrated the big data analytics and the service- driven patterns that helped to overcome the barriers such as lack of the complete data and the valuable knowledge. This architecture benefited manufactures, environment, customers and all the stages of PLM, and effectively supported the implementation of CP. An overall architecture of the BDA-PL is proposed in Fig. 3.

2.1 Application services of PLM

In this layer, the objectives of the PLM are used by the manufacturing enterprises (i.e. design improvement, maintenance, environment protection and energy conservation, etc.). PLM and CP insist on maximizing the coordination between environmental benefits (high energy efficiency and high environment efficiency) and the enterprises benefits (product design improvement, maintenance and high profits). Maintenance service, Remote monitor service, Recycling service, Operation service, Spare parts service and Integration with EISs are the six types of the services (shown in the Fig. 3) in the architecture.

2.2 Acquisition and integration of big data

In this phase, on the basis of arrangement of smart devices like smart sensors in the manufacturing resources and product the complete and accurate multi- source heterogeneous big data can be collected and transferred during the whole lifecycle. Integrating the data mining results with the other enterprise information systems like EISs is designed to make bridge for the processing and then exchanging the information between heterogeneous management systems.

2.3 Big data processing and storage

This phase of architecture mainly handles the processing of data and then stores the processed data. Product lifecycle data is of three types i.e. structured, semi- structured and unstructured data. Different computing frameworks are used for various kinds of data as some data need a very high real-time processing ability and other data need a very low real-time processing ability. Storm a real-time computing framework which is used to process data which need a very high real-time processing ability. For reliability, huge number of non-real-time is stored .On the other hand, Hadoop a computing framework which is used to process data which need a

very low real-time processing ability . To store the heterogeneous big data Hadoop distributed file

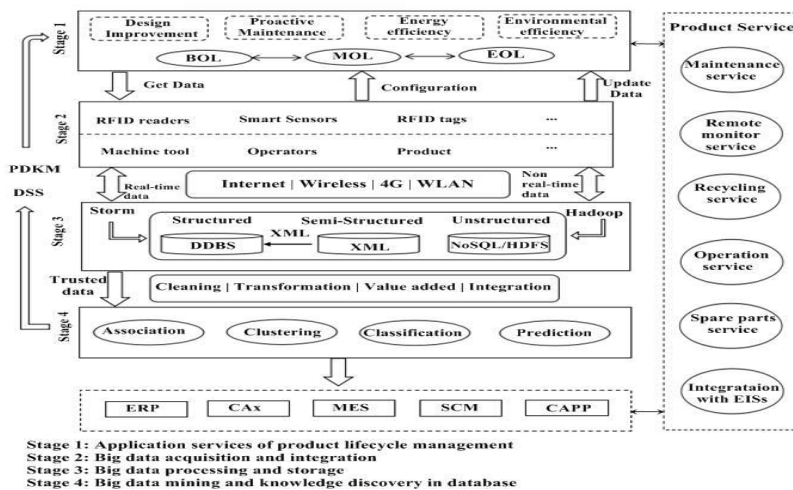


Fig. 3. Overall architecture of big data-based analytics for product lifecycle. System (HDFS), Distributed database system(DDBS) and Structured query language(SQL) data management system are used.

2.4 Big data mining and knowledge discovery in databases (KDD)

By using various techniques of big data analytics and also of data mining, useful information and knowledge can be extracted from the big data of the product lifecycle. by combining the big data mining result with the product data and knowledge management (PDKM) system and the decision support system(DSS), then a closed –loop procedure of knowledge share and the feedback is produced among all the lifecycle stages. For achieving the lifecycle optimization and cleaner production (CP) for manufacturing process, the knowledge sharing must be realized in the all individual phases of product lifecycle. There are a large number of models i.e. clustering, classification, regression and association analysis are introduced to extract the knowledge from the data.

Manufacturing and maintenance process (MMP) big data plays a vital role in PLM. So, to get the real-time and the complete MMP data, there is a framework for the real-time as well as for multi-source heterogeneous big data acquisition and integration of MMP. Big data –based analytics for MMP steps which are MMP acquisition and integration of big data, MMP mining of big data and MMP knowledge sharing mechanism. Fig. 4. Shows how data flows from BOL to MOL and MOL to BOL.

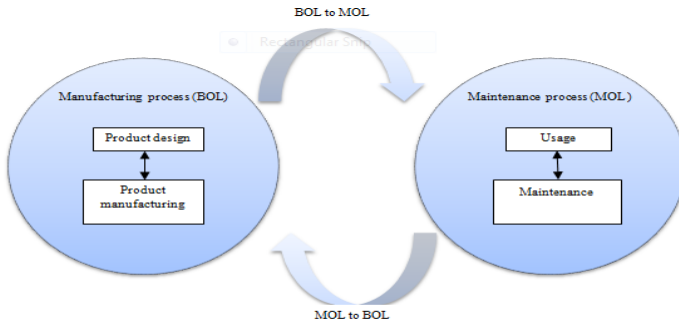


Fig. 4. Knowledge sharing mechanism of the MMP

3. Related work

Laney et al. (2001) [14] introduced the most important characteristics of the big data as 3Vs theory: velocity, volume and variety. According to this theory, big data includes huge amount of heterogeneous, multi-source and real-time data, which is produced during the manufacture and maintenance stages. All these data is characterized by the 3Vs theory and increasing at the exponential speed. Galletti et al. (2013) [8] focused on Big Data Analytics (BDA) and introduced how BDA may be observed and act as a driver for the industries' competitive advantage. Dai et al. (2011) [7] described a approach in which big data applied in the cloud, for building a simple, lightweight and the extremely scalable performance analyzer used for the dataflow-based evaluation. Rabl et al. (2012) [18] discussed about the analysis of challenges of big data for company application efficiency management and based on this experience and lessons learned from the investigation, big data applications in enterprise could be promoted. Jun et al. (2009) [11] introduced a standard structure for radio frequency identification (RFID) for PLM, which gives a better environment for gathering and analyzing product lifecycle information and provides decisions without any constraints. Also discussed about the application issues and the overall approaches for each lifecycle stage. Kiritsis et al. (2003) [12] discussed about the RFID technology that can bring new options for accessing, handling, and then controlling merchandise data and useful knowledge over full product lifecycle. To reduce the use of water, handle pollution at low cost, proper use of resources & reduce waste material, Kupusovic et al. (2005) [13] proposed a project for slaughterhouse industry. Wang et al. (2010) [21] presented a warehouse design management system for the tobacco industry, which is based on the RFID technology. By using the RFID technology, the system produced a digital

warehouse to get the maximum capacity, automatic storage, better visualization and high accuracy of inventory control. For selecting the dispatching rules, Metan et al. (2010) [16] introduced a new scheduling system. And this system is obtained by combining the various methods of statistical process control charts simulation and data mining. By constructing a decision tree, this proposed system obtains knowledge from manufacturing data and updates the decision tree dynamically whenever the conditions change and hence improves the quality of the decisions. Vinodh et al. (2011) [20] reported the utilization of the fuzzy association rules mining method that allowed the developers to take the efficient decisions by using the various attributes like quality, cost, pro-activity, robustness, innovativeness and flexibility for evaluating agility in the supply chains and also indicated that there is no need of any constraints for decisions for the processing of evaluation of agility. Choudhary et al. (2009) [5] discussed about the knowledge discovery in database and applications of data mining in the vast range of production with a specific focus on the kind of methods to be applied on data and the methods are association, characterization and description, clustering, prediction and evolution analysis. Also, exposed the progressive applications and existing gaps occurred in the context of data mining in production. CP has several social, economic and environmental benefits. So, for the successful implementation by overcoming barriers of it, Silva et al. (2013) [19] proposed a CP technique in which the various Quality Tools (QTs) are integrated. To obtain the emission reduction and energy conservation for a medium-scale ceramic tile plant, Huang et al. (2013) [10] presented an extensive application of cleaner production. Corominas et al. (2013) [6] proposed a tool i.e. Life Cycle Assessment (LCA) to enhance the efficiency of wastewater treatment plants by choosing the best strategy. Cheung et al. (2015) [4] investigated the disposal costs by using original Equipment Manufacturer (OEM) and then on the basis of this cost, the decision will be taken whether the EOL parts to be recycled or destroyed. Zhang et al. (2016) [23] proposed an overall architecture i.e. Big data-based analytics for product lifecycle (BDA-PL), for making better product lifecycle management (PLM) and cleaner production (CP) decisions based on data. This architecture integrated the big data analytics and service-driven patterns that helped to overcome the incomplete data and valuable knowledge barriers.. Also, discussed about the manufacturing and maintenance process (MMP) of product lifecycle and key technologies were developed to implement the big data analytics and this proposed

architecture benefited customers, environment, customers and even all stages of the PLM and effectively promoted the implementation of CP.

4. Comparison table

Reference	Title	Objective	Material recycling	Production cost	Energy consumption	Minimize waste generation
Kupusovic et al. [13]	Cleaner production measures in small-scale Slaughterhouse industry case study in Bosnia and Herzegovina Slaughterhouse industry case study in Bosnia and Herzegovina Cleaner production measures in small-scale Slaughterhouse industry case study in Bosnia and Herzegovina.	To minimize use of water, handle pollution at low cost, efficient use of resources & reduction of waste material at source.	✓	Low	Less	✓

Huang et al. [10]	Application of cleaner production as an important sustainable strategy in the ceramic tile plant a case study in Guangzhou, China.	To reduce energy consumption by 4.3% and water by 22.33%, use a cleaner production application in tile plant.	✓	Low	Less	✓
Corominas et al. [6]	Including Life Cycle Assessment for decision-making in controlling wastewater nutrient removal systems.	Life Cycle Assessment (LCA) cost effective tool is used to handle the wastewater removal systems.		Low	Less	✓
Li et al. [15]	Big Data in product lifecycle management .	Provide detailed information about the concepts of big data and three main phases of PLM in the manufact	✓	Low		

		uring process.				
Auschi tzky et al.[1]	How big data can improve manufacturing.	Advanced analytics proposed a approach in manufacturing process to minimize waste and improve product quality.		Low	Less	✓
Silva et al. [19]	Quality Tools Applied to Cleaner Production Programs: A First Approach Towards a New Methodology.	A systematic integration of QTs is proposed , to implement and overcome the barriers of CP process ,	✓	Low		✓

Cheung et al. [4]	Towards cleaner production: a roadmap for predicting product end-of-life costs at early design concept.	To predict disposal cost by OEM, which will help to get a solution whether the EOL part to be recycled or destroyed.	✓	Low		✓
Zhang et al. [23]	A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products.	An architecture BDA-PL is proposed, to make the better PLM and CP decisions.	✓	Low	Less	✓

5. Gaps in Literature survey

Zhang et al. [23] proposed an overall architecture i.e. discussed in the section 2, for making better product lifecycle management (PLM) and cleaner production (CP) decisions based on data. This architecture integrated the big data analytics and service-driven patterns and overcome the various issues for the proper implementation of cleaner manufacturing and maintenance. So, this study showed that this proposed architecture benefited customers, environment, customers and even all stages of the PLM and effectively promoted the implementation of CP. But, the review has shown that the not much work is done for software reuse analytics. The other major issue is that computational speed is still found to be challenging in big data analytics and the use of data preprocessing is also ignored by existing researchers.

To handle all the above stated issues a new integrated Random Forest and Neural Networks based analytics technique will be proposed. The proposed technique will utilize unsupervised filtering and Random forest based machine learning technique.

6. Conclusion

In this paper it focusing on manufacturing and maintenance process of the product lifecycle, is facing many problems. As, the Cleaner Production (CP) strategy was facing barriers, such as the lack of complete data and valuable knowledge that can be employed to provide better support on decision-making of coordination and optimization on the product lifecycle management (PLM) and the whole CP process. It has also discuss the comparison of various techniques based on different parameters which shows that energy consumption and production cost is less to take deep drive in real time data. The review has shown that the not much work is done for software reuse analytics and also computational speed is still found to be challenging issue in big data analytics. To overcome these issues in future we will propose integrated Random Forest and Neural Networks based analytics for cleaner manufacturing and maintenance processes.

References

- [1] Auschitzky, E., Hammer, M., Rajagopaul, A., 2014. How big data can improve manufacturing. McKinsey Glob. Inst. Available at: http://www.mckinsey.com/insights/operations/how_big_data_can_improve_manufacturing (accessed 25.05. 2015).
- [2] Bennane , A., Yacout,S., 2012. LAD-CBM; new data processing tool for diagnosis and prognosis in cindition-based maintenance. *J. Intell. Manuf.* 23(2), 265-275.
- [3] Chen, Y. S., Cheng, C. H., Lai, C. J., 2012. Extracting performance rules of suppliers in the manufacturing industry: An empirical study. *J. Intell. Manuf.* 23(5), 2037-2045.
- [4] Cheung, W.M., Marsh, R., Griffin, P.W., Newnes, L.B., Mileham, A.R., Lanham, J.D., 2015. Towards cleaner production: a roadmap for predicting product end-of-life costs at early design concept. *J. Clean. Prod.* 87, 431-441.
- [5] Choudhary, A. K., Harding, J. A., Tiwari, M. K., 2009. Data mining in manufacturing: a review based on the kind of knowledge. *J. Intell.Manuf.* 20(5), 501-521.
- [6] Corominas, L., Larsen, H.F., Flores-Alsina, X., Vanrolleghem, P.A., 2013. Including life cycle assessment for decision-making in controlling wastewater nutrient removal systems. *J. Environ. Manag.* 128, 759-67.
- [7] Dai, J.Q., Huang, J., Huang, S.S., Huang, B., Liu, Y., 2011. Hitune: data flow-based performance analysis for big decision-making in controlling wastewater nutrient removal systems. *J. Environ. Manag.* 128, 759-67.
- [8] Galletti, A., Papadimitriou, D.C., 2013. How big data analytics are perceived as a driver for competitive advantage: a qualitative study on food retailers. Master thesis, 1-58.

- [9] Hadaya, P., Marchildon, P., 2012. Understanding product lifecycle management and supporting systems. *Ind. Manage. Data. Syst.* 112(4), 559-583
- [10] Huang, Y., Luo, J., Xia, B., 2013. Application of cleaner production as an important sustainable strategy in the ceramic tile plant e a case study in Guangzhou, China. *J. Clean. Prod.* 43, 113-121.
- [11] Jun, H. B., Shin, J. H., Kim, Y. S., Kiritsis, D., Xirouchakis, P., 2009. A framework for RFID applications in product lifecycle management. *Int. J. Comp. Integ. M.* 22(7), 595-615.
- [12] Kiritsis, D., Bufardi, A., Xirouchakis, P., 2003. Research issues on product lifecycle management and information tracking using smart embedded systems. *Adv. Eng. Inform.* 17(3), 189-202.
- [13] Kupusovic, T., Midzic, S., Silajdzic, I., Bjelavac, J., 2005. Cleaner production measures in small-scale slaughterhouse industry e case study in Bosnia and Herzegovina. *J. Clean. Prod.* 15, 378-383.
- [14] Laney, D., 2001. 3D data management: controlling data volume, velocity and variety. META Group Research Note. Available [http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety .pdf](http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety.pdf)
- [15] Li, J.R., Tao, F., Cheng, Y., Zhao, L.J., 2015. Big Data in product lifecycle management. *Int. J. Adv. Manuf. Tech.* 84(1-4), 667-684.
- [16] Metan, G., Sabuncuoglu, I., Pierreval, H., 2010. Real time Selection of Scheduling Rules and Knowledge Extraction Via Dynamically Controlled Data Mining. *Int. J. Prod. Res.* 48(23): 6909-6938.
- [17] Ngai, E. W., Xiu, L., Chau, D. C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert. Syst. Appl.* 36(2), 2592-2602.
- [18] Rabl, T., Gómez-Villamor, S., Sadoghi, M., Muntés-Mulero, V., Jacobsen, H.-A., Mankovskii, S., 2012. Solving big data challenges for enterprise application performance management. *Proc. VLDB Endow.* 5 (12), 1724-1735.
- [19] Silva, D.A., Delai, I., Castro, M.A., Ometto, A.R., 2013. Quality Tools Applied to Cleaner Production Programs: A First Approach Towards a New Methodology. *J. Clean. Prod.* 47, 174-187.
- [20] Vinodh, S., Prakash, N.H., Selvan, K.E., 2011. Evaluation of agility in supply chains using fuzzy association rules mining. *Int. J. Prod. Res.* 49(22): 6651-6661
- [21] Wang, H. W., Chen, S., Xie, Y., 2010. An RFID-based digital warehouse management system in the tobacco industry: A case study. *Int. J. Prod. Res.* 48(9), 2513-2548.
- [22] Wei, F. F., 2013. ECL Hadoop: “Big Data” processing based on Hadoop strategy in effective e-commerce logistics. *Comput Eng Sci* 35(10):65–71.
- [23] Yingfeng Zhang, Shan Reh, Yang Liu, Shubin Si., 2016. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *J. Clean. Prod. JCLP* 7965.

A Comparative Study of Classification Approaches for Entity Linking in Semantic Web

Amit Singh¹, Aditi Sharan²

School of Computer and Systems Sciences,
Jawaharlal Nehru University
New Delhi, India

¹singhamit1320@gmail.com

²aditisharan@gmail.com

Abstract. In Recent years we have witnessed a rapid growth in Semantic Web Data Sources. These Data Sources follow Linked Data principles to facilitate efficient information retrieval and knowledge sharing. These data sources may provide complementary, overlapping or contradicting information. In order to integrate these data sources, we perform Entity Linking. Entity Linking is an important task of identifying and linking entities across data sources that refer to the same real-world entities. In this work, we have compared supervised machine learning based classification approaches for Entity Linking. We have evaluated different classifiers on standard datasets and presented results to measure efficiency of these classifiers.

Keywords: Semantic Web, Linked Data, Classification, Entity Linking.

1. Introduction

Over the past years, the World Wide Web is witnessing a rapid change from document-oriented web to a distributed, proliferated machine readable Semantic Web (SW). The SW contains data sources on various domains such as people, publications, media, government organizations and social web etc. These data sources are independent of each other and geographically distributed. In order to harness real value from these sources, we have to integrate these sources and build an interlinked SW. The main aim behind interlinking of data sources is to facilitate interoperability of data, better utilization, and consumption of data. These links play an important role in processing federated queries and complex question answering by retrieving information available across data sources.

The SW data sources follow linked data principles that provide standards for low barrier publication of information and interlinking of information across data sources. Linked Open Data cloud [1] is considered as a largest available source of structured information following Linked Data Principles. This Linked Open Data currently consists of 300 interlinked

datasets which consist of billions of facts as RDF triples¹. This Linked Open Data is growing rapidly in volume. As per LODStats² LOD recorded around 1 billion triples till 2011 which had increased to 85 billion by 2015. According to some studies [2], 44% of the Linked Open Datasets are not connected to other datasets at all and there exist less than 400 million links. Over 30 billion triples published as Linked Open Data. In order to establish more links between datasets, we need to identify different entities that refer to the same real-world resource. So, entity linking is a fundamental task in semantic web data integration. The main problem behind this lack of links is that manual creation of links is a very tedious process and infeasible in the case of large data sources like DBpedia³ (4.5 million entities) and Linked GeoData⁴ (1+million entities).

In this work we have presented an entity linking pipeline for systematic evaluation of different approaches. We have compared seven classifiers in terms of their precision, recall and F-measure on the proposed evaluation pipeline. In this work we aim to bridge the current research gap by providing a comprehensive study of performance of different classifiers on real world datasets.

Rest of the paper is structured as follows: in section 2 we have formally defined the problem. In section 3 we have presented a literature survey of different entity linking systems. In section 4 we have explained our evaluation pipeline of entity linking. In section 5 we have shown our experimental setting and reported results obtained from our work. The last section covers our concluding remarks and future scope.

2. Entity Linking Problem

The general problem definition of entity linking can be defined as follows[3]: Entity linking problem between two data sources S (source) and T (target) aims at determining pairs of entities in S and T that refer to same real world object. Each entity $e \in S \cup T$ consists a set of properties $e.p_1, e.p_2, e.p_3, \dots, e.p_n$. Given a relation R entity linking finds a subset $M \subseteq S \times T$ of a pair of entities $(s, t) \in S \times T$ such that $R(s, t)$ holds.

$$M = \{(s, t): R(s, t), s \in S, \text{ and } t \in T\}$$

The relation R relates all entities which represent same real world object. Similarly, we can define the set of all pairs for which R does not hold as:

$$U = (S \times T) \setminus M$$

¹ Information taken from <http://lod-cloud.net>

² Online available at <http://stats.lod2.eu>

³ Online at <http://dbpedia.org>

⁴ Online at <http://linkedgeo.org>

3. Related Work

We have broadly categorized approaches as manual and automatic approaches. Manual approaches rely on an expert user's domain knowledge to establish rules to generate links. In manual approaches, user's domain knowledge is a limiting factor.

SILK, LIMES, and Zen Crowd are famous linking system based on manual linking approaches where the user has to specify linking rules. In Silk[4] and LIMES [5] framework for establishing links between two data sources the user needs to specify entities to link, similarity measures to be used, pre-processing transformations and operator (MIN, MAX, AVG etc.) to combine the result of matching algorithms in a specific link specification language. Zen Crowd[6] uses probabilistic reasoning and crowd sourcing to perform entity linking. Zen Crowd uses an inverted index to identify possible matches and then uses various matchers to evaluate the similarity between entities. Non-promising results are fed into Micro task manager for crowd sourcing. Finally, results from different matchers and crowd source are fused to link entities.

With the advent of machine learning techniques researchers have developed learning-based automatic approaches. Learning-based approaches can be further categorized into supervised, unsupervised and semi-supervised approaches.

Active Atlas, TAILOR, MARLIN and GenLink are remarkable linking systems based on supervised learning models. In Active Atlas[7] and TAILOR [8] authors have designed a decision tree based classifier system to identify matches and non-matches to link entities. Active Atlas uses a C4.5 algorithm to build a decision tree while TAILOR uses the ID3 algorithm. MARLIN (Multiply Adaptive Record Linkage with Induction)[9] learns string similarity measure to calculate the similarity between entity properties and uses a support vector machine[10] to combine learned similarity measures to design a classifier model.

In supervised learning based approaches GenLink[11] and EAGLE [12] systems use a Genetic programming to learn linkage rules from the existing set of links. The developed systems are capable of generating linkage rules which select discriminative properties for comparison, select data transformations to normalize property values, choose appropriate distance measures and thresholds and combine the results of multiple comparisons using non-linear aggregation functions.

Semi-supervised learning approaches utilize both labeled as well as unlabeled data. They alleviate the problem of labeled training data by using a small set of initial labeled data and a large amount of unlabeled data. ObjectCoref[13] is a semi-supervised learning based linking system. It uses existing owl: sameAs links to find properties which could act as a

discriminator. Then system uses these properties and performs property value pair analysis to find new properties which have the same value for entities available in the training data. The System further improves by using functional, inverse functional properties and cardinality restrictions.

SERIMI, RiMOM, Zhishi.links and LN2R are linking systems based on unsupervised learning models. In SERIMI[14]and RimOM[15]authors have used string based similarity measures to get initial results which were improved by taking into account structural similarity. They have used string similarity measures to identify possible matches for an entity and later to predict a link they have used structural similarity measures over RDF. In Zhishi.links[16] authors have used the indexing technique to identify possible matches for an entity, and later they have used distance based similarity measures to identify equal entities. LN2R [17]is a combination of a logical (L2R) and a numerical approach (N2R). The logical approach exploits the semantics of the underlying ontologies by using inverse functional properties and disjoint axioms. Matches and non-matches provided by L2R are fed into N2R. N2R System uses properties provided in the above step and produce equations modeling dependencies between the similarities of entities. An iterative algorithm is used to compute the similarity of entities based on attribute similarities.

4 ENTITY LINKING PIPELINE

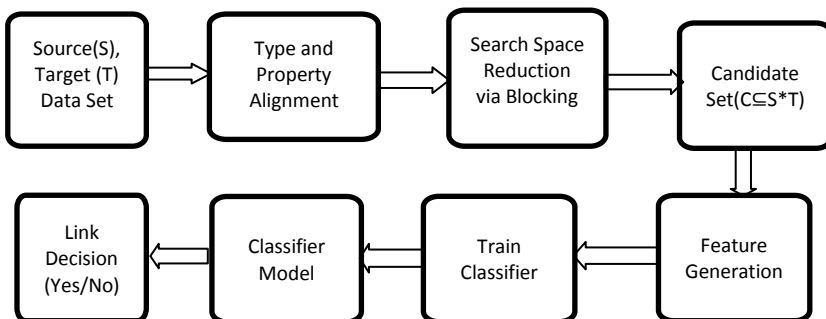


Fig. 1.Entity Linking Pipeline

Entity linking process starts by loading source and target datasets and preprocessing these datasets to remove any inconsistency. In order to make a linking decision, we have to explore total search space created by the Cartesian product of instances of both data sets. Exploration of this whole search space is practically infeasible for real world datasets. We can reduce search space by applying blocking to eliminate sure non-matches and obtain a candidate set having potential matches. In order to make a linking decision, we extract features from candidate sets and convert these tuples in candidate sets into feature vectors. Linking Decision is taken by training a classifier and applying on unseen data. We have described every step of entity linking pipeline in detail in coming subsections-

4.1 Inferring Type and Property Alignment

Type alignment helps in reducing the overall search space by guiding us to compare only instances of similar types. Similarly, Property alignment helps us in comparing similar properties of similar type instances. Type and property alignment can be computed by using state of the art ontology matching algorithms.

4.2 Blocking Step

In entity linking problem pairwise comparison of entities between two datasets is computationally expensive. In pairwise comparison, we have to compare each entity in one dataset against all entities in target dataset. The no of comparisons is quadratic to the size of datasets. The main objective of the blocking step is to reduce this search space by eliminating obvious non-similar entities and grouping possibly similar entities together.

The Blocking step uses type and property alignment to generate blocks having similar entities together. The main advantage of having similar entities together in a block is that we need to only compare entities present in the same block. In this work, we have used TYPiMatch[18] blocking technique that groups' entities of similar types together with the help of type alignment inferred in the previous step and further improve it by applying token blocking on the attribute values of entities of similar type.

4.3 Feature Generation:

In order to perform classification, we need to have a set of features of entities. In this work we have used property alignment (A) generated in the previous step, a set of tokenizers (T) and similarity functions (F) to generate features. These features are used to generate a feature vector for each pair of the entity. Our approach is pretty straightforward, for each property pair in property alignment we have generated $|T|*|F|$ features by combining each tokenizer and a similarity function.

4.4 Classification.

Blocking step reduces our search space and generates a set of candidate entity pairs C. In order to train a classifier we must have a training data containing labels for similar and non-similar entities. In this work, we take samples from the candidate set C and label them to train a classifier. In order to generate a balanced training data C_{Labeled} we adopt iterative process. We initially take n_1 samples from C and label them, if number of matches and non-matches in n_1 have a reasonable ratio then we randomly pick remaining n_2 samples and label them else we need to again randomly choose n_1 samples and repeat this process.

Labeled data C_{Labeled} generated in as per above-defined strategy is converted into feature vector (V) as per feature extraction strategy given in section 4.3. We divide V into two parts namely- development set (I) and test set (J). We have trained seven classification models namely- SVM, Decision Tree,

Random Forest, Multi-layer Perceptron, Naive Bayes, Linear Regression and Logistic Regression using development set (I) and predicted the accuracy of the model using test set (J).

5. EVALUATION

In this section we evaluate our entity linking pipeline defined in section 4 experimentally. Section 5.1 describes the used data sets while section 5.2 describes our experimental setup. Section 5.3 introduces evaluation metrics while overall linking results for several real world data sets are presented in Section 5.4.

5.1 DataSets

In this work, we have considered four datasets- Cora⁵, Restaurants, Persons1 and Persons2. The First dataset is very popular dataset for deduplication in relational databases and later three are released by Ontology Alignment Evaluation Initiative⁶ (OAEI) in 2010 as part of Instance Matching in Data Interlinking Track. Cora dataset is a collection of bibliographic data of 1,295 citations of 122 computer science papers. Restaurants dataset is a collection of 864 restaurant details curated by combining information from Fodor and Zagat’s guidebooks. Persons1 and Persons2 dataset contain information about Persons such as name, surname, street number, address, suburb, postcode, state, date of birth, age, phone number, and social security number etc. The persons1 dataset contains two files each having 500 records. The persons2 dataset contains two files having 600 and 400 records respectively.

5.2 Experimental Setup

Processor	Intel(R) Core(TM) i7-4790 CPU @3.60 GHz
Operating System	Windows 10Pro
RAM	16 GB
System Type	64-bit Operating System
Programming Language	Python
API's	Scikit-learn

⁵ Onlineavailable at: <https://people.cs.umass.edu/~mccallum/data.html>

⁶ Online available at: <http://oaei.ontologymatching.org/2010/im/index.html>

Table 2 Experimental Environment

5.3 Evaluation Metrics

In genetic programming, we need to calculate fitness of each (linkage rule in this case) which in turn depend on the following indicators calculated from training data containing reference positive and negative links -

- True Positives (TP) is the number of positive reference links generated by the system whose linking decision is correctly predicted by the system.
- False Positives (FP) is the number of negative reference links generated by the system whose linking decision is incorrectly predicted by the system.
- False Negatives (FN) is the number of positive reference links generated by the system whose linking decision is incorrectly predicted by the system.
- True Negatives (TN) is the number of negative reference links generated by the system whose linking decision is correctly predicted by the system.

Based on these indicators we can define precision, recall, and F-measure of a linkage rule-

Definition (Precision): Precision denote the correctness of a rule. We can define precision as the fraction of linking decisions that are correct

$$\text{Precision (P)} = \text{TP} / (\text{TP} + \text{FP})$$

Definition (Recall): Recall denote the completeness of a rule. We can define recall as the fraction of correct linking decisions from total positive reference links.

$$\text{Recall (R)} = \text{TP} / (\text{TP} + \text{FN})$$

Definition (F-measure): F-Measure can be defined as a harmonic mean of precision and recall defined above.

$$F1 = (2 \cdot \text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$$

5.4 Experimental Results

Dataset Classifier	Cora	Restaurant	Persons1	Persons 2
	F-measure	F-measure	F-measure	F- measure
Decision Tree	90.23	91.47	94.73	88.98
SVM	88.27	95.01	88.89	82.75
Random Forest	96.37	100	100	92.04

Logistic Regression	91.90	94.957	91.47	85.07
Linear Regression	89.02	92.95	92.90	83.49
Naive Bayes	88.49	83.82	93.49	55.68
Multi-layer Perceptron	92.43	94.68	100	97.17

Table 3. Results for Seven Different Classifiers on Four Datasets

6. Conclusion

In this work we have given a systematic study of supervised learning based classification approaches. We have presented an entity linking pipeline to cover over all process. The results presented in this work give a better understanding of different classifiers efficiency for entity linking task. In future we can explore the possibility of using Semi Supervised and Ensemble based learning on entity linking task.

7.References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *Int. J. Semant. Web Inf. Syst.* 5, 1–22 (2009).
2. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: *International Semantic Web Conference*. pp. 245–260 (2014).
3. Fellegi, I.P., Sunter, A.B.: A Theory for Record Linkage. *J. Am. Stat. Assoc.* 64, 1183–1210 (1969).
4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - A Link Discovery Framework for the Web of Data. *CEUR Workshop Proc.* 538, (2009).
5. Ngonga Ngomo, A.-C., Auer, S., Ngomo, A., Auer, S.: Limes-a time-efficient approach for large-scale link discovery on the web of data. In: *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*. pp. 2312–2317 (2011).
6. Demartini, G., Difallah, D.: ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: *Proceedings of the 21st international conference on World Wide Web*. pp. 469–478. ACM Press (2012).
7. Tejada, S., Knoblock, C., Minton, S.: Learning object identification rules for information integration. *Inf. Syst.* 26, 607–633 (2001).
8. Elfeky, M., Verykios, V.: TAILOR: A record linkage toolbox. In:

- 18th International Conference on Data Engineering, 2002. (2002).
9. Bilenko, M., Mooney, R.J.: Adaptive Duplicate Detection Using Learnable String Similarity Measures. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003). pp. 39–48. ACM Press, New York, New York, USA (2003).
 10. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* 20, 273–297 (1995).
 11. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. *J. Web Semant.* 23, 2–15 (2013).
 12. Ngomo, A., Lyko, K., Ngonga Ngomo, A.-C.C., Lyko, K., Ngomo, A., Lyko, K., Ngonga Ngomo, A.-C.C., Lyko, K., Ngomo, A., Lyko, K., Ngonga Ngomo, A.-C.C., Lyko, K.: EAGLE: Efficient active learning of link specifications using genetic programming. In: Extended Semantic Web Conference. pp. 149–163 (2012).
 13. Hu, W., Chen, J., Qu, Y.: A self-training approach for resolving object coreference on the semantic web. In: Proceedings of the 20th international conference on World wide web - WWW '11. p. 87. ACM Press, New York, New York, USA (2011).
 14. Araujo, S., Vries, A., Schwabe, D.: Serimi results for OAEI 2011. In: Proceedings of the 6th International Conference on Ontology Matching (2011).
 15. Li, J., Tang, J., Li, Y., Luo, Q., Juanzi Li, Jie Tang, Yi Li, Qiong Luo: RiMOM: A dynamic multistrategy ontology alignment framework. *IEEE Trans. Knowl. Data Eng.* 21, 1218–1232 (2009).
 16. Niu, X., Rong, S., Zhang, Y., Wang, H.: Zhishi. links results for OAEI 2011. In: CEUR Workshop Proceedings (2011).
 17. Saïš, F., Niraula, N., Pernelle, N., Rousset, M.C.: LN2R - A knowledge based reference reconciliation system: OAEI 2010 results. In: CEUR Workshop Proceedings. pp. 172–179 (2010).
 18. Ma, Y., Tran, T., Bicer, V.: Typimatch: type-specific unsupervised learning of keys and key values for heterogeneous web data integration. In: Proceedings of the sixth ACM international conference on Web Search and Data Mining (2013).

New Paradigm for Software Design: ADML

Namrata Sharma¹, Prerna Tyagi², Vaibhav Vyas³, Rajeev G
Vishvakarma⁴,

^{1,2,3} Aim & Act-Department of Computer Science
Banasthali University, Rajasthan,

⁴Indore Institute of Technology and Science, Indore

{¹nsbhardwaj11154@gmail.com, ²Prernatyagi022@gmail.com,

³Vaibhavvyas4u@gmail.com, ⁴Rajeev@gmail.com}

Abstract. They are several applications present in software engineering field that are difficult to understand and complicated to solve and also having different concerns that crosscut in each functionality of software some approaches are being proposed to deal with them. One of the approaches named Aspect-Oriented Software Development provides new insights with tools that aid in modular development of software in complex system by explicitly addressing crosscutting concerns. This paper focuses on Aspect-Oriented Software Development that evolves a designing language Aspect Design Modeling Language (ADML), to represent, document and design the aspectual elements.

Keywords: Aspect-Oriented Software engineering, Aspect Design Modeling language, Aspect-Oriented Programming.

1 Introduction

Aspect-Oriented Software Development provides wide objective of modular programming having more focus on crosscutting concerns. A crosscutting concern is one which crosscut in each module of the system. For example, error logging policy in which the requirement is that all errors of the system are logged in standard format. Due to this, the addition code available in the scattered form throughout the whole system code [1]. The crosscutting concerns explicate a significant portion of volume of the code with interdependencies retain among various modules of the software. Any modular interdependency makes difficulty in understanding, evolution and development of complex system which have broader application in real world.

Aspect-Oriented Software Development (AOSD) was introduced first time at the programming phase but later developed at the other level of software development as the time evolved. And now, it makes its way to all phases in the software development namely requirement engineering, design with architecture up to the implementation phase. At the implementation phase, Aspect-Oriented Programming (AOP) was developed for addressing the problem occurring at the implementation level. This finds a way to modular

the crosscutting concerns in the independent modules. These concerns show implementation independently from other concerns and linked automatically at the execution phase [2]. Concerns having redundant implementation is removed which makes the system more modular with comprehensible in nature. Concerns which are identified are called aspects. The implementation of these aspects having in form of pieces of code (advice) that run to the predefined points (joinpoints) when the execution is performed. During execution the above described aspects combine with the base modules of the system through a process known as weaving that defined in respect of composition techniques which are specific to the currently used aspect-oriented technique.

This presented paper developed a designing language by keeping in view Aspect-Oriented Software Development (AOSD). The new language known as Modeling Language for Aspect Design Modeling Language (ADML) has been evolved to implement aspectual elements with non-aspectual elements. The proposed language has some set of symbols and design notations which help in designing of aspect. For every aspectual element there is a distinct notation with the collection of diagrams that depict structural with behavioral crosscutting.

Systems like object-oriented system are designed commendably with the use of Unified Modeling Language (UML). As UML are having different designing notations with diagrams that help in modeling system by identifying, representing, designing and implementing objects with data entities. AspectJ which is the implementation tool allows to implements aspect with the object in the base system [3]. If implementing these kind of system that has aspects with objects together to implement then there is a need of proper designing technology which can effectively specify and design respective artifacts in same location. UML is a modeling language which supports object-oriented concepts rather than non-object based concept.

There are following two ways to design aspects with the objects in software engineering.

1. By extending the capabilities of the UML in the designing approach.
2. To develop a language that can represent both aspects and object based constructs and demonstrate the mutual relationship.

1.1 Crosscutting Concern

There are some concerns present in system which are linked among each other or in other words they are dependent on other concerns. For example, in system such as bank each transaction must be logged to logger that has to verify for purpose of security. In each transaction tracing is also done. If

one person implements the mentioned system by using technology of object-oriented he has to implement tracing with logging and security concerns with implementation to all transaction. In this manner, the logic gets implemented in place where it does not belong and it is clearly visible that it violates separation of concern principle with encapsulation also. The above described concern is crosscutting concern as it crosscut in various implementation module of system [4].

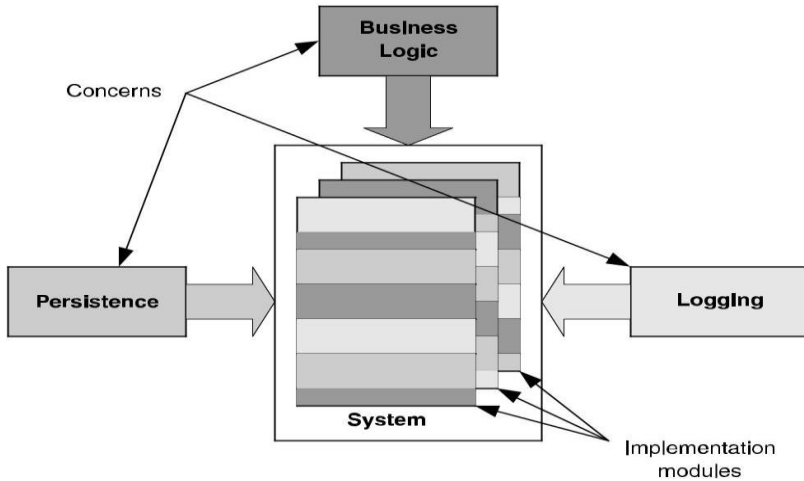


Fig.2. The system is viewed as consist of various concerns which are needed to be addressed.

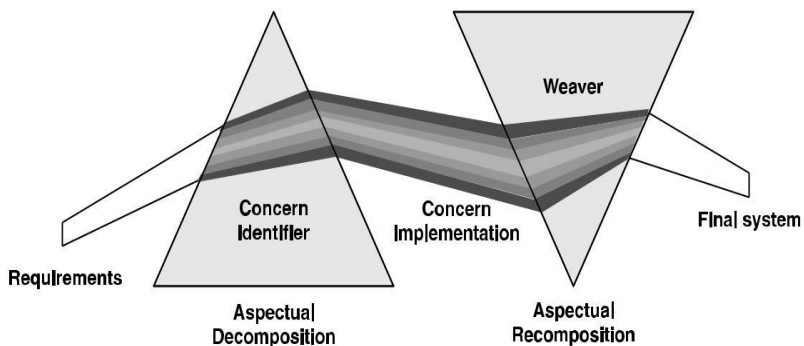


Fig. 2. Demonstrates concern which decomposes from requirements due to the need of developing an application to final system and in between performed concern implementation.

1.2 Problem of Tangled and Scattered Code

A single component may include elements of multiple requirements which causes tangling problem of code. The other problem arises when a single requirement may be implemented by several modules of software that results into code scattering. Scattered code provide anomaly with inconsistencies and maintenance problem, of the new developed software. Fig.3. depicts code tangling.

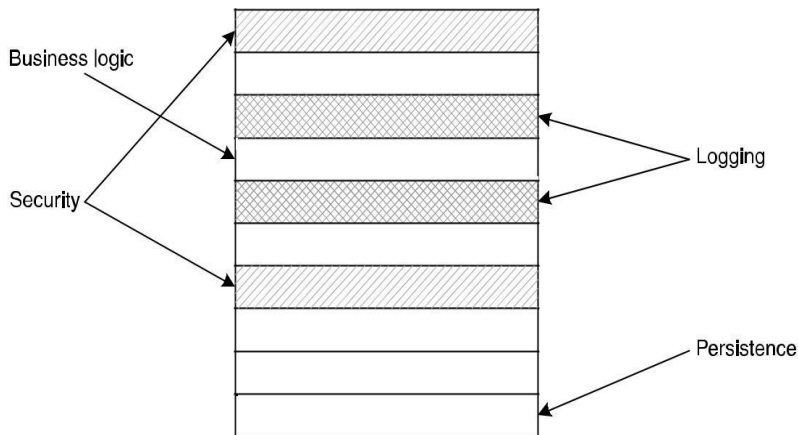


Fig. 3. Multiple simultaneous implementation of several concerns leads to code tangling. The above figure depicts how a single module handles a part of several concerns.

1.3 Design Principles

Both the ways of design aspect follow the basic and important principles of software design stated as:

- **Separation of Concerns-** In 1972 Parnas and in 1976 Dijkstra stated separating the concerns among other constructs of the system is the key principle in managing the complexity related to the system [5]. For solving the complexity, the best idea is the division of the complicated system to small independent units. As these units do not have information related to each other but they are first design and then implement separately. After that combination of them are done to provide a whole system. The new designing language of Aspect-Oriented would follow the following approach not just because concerns separation is an important step to every software language but because aspect programming proposed and perceived related to this principle.

- **Comprehensibility**-As stated by Parnas in 1972, comprehensibility means the power of understanding one specific part of the system completely at a particular span of time. Tangling nature of code shown by aspects, and this tangled code related to other modules of specific system so to understand aspect along with its behavior is not an easy job if one does not have information of other modules of system. So the new designing technology for AO based systems should have the capability for the representation of aspect in well - defined form also having separation with other system modules but relationship of aspects with other modules should be independently maintained.

- **Loose coupling**- Tight coupling shown by aspects to the other system units because of direct implementation of AOP approach is based on the principle of separation of concerns that means aspect is represented as a separate unit from other units of system which helps in reducing the coupling upto many extends. This above approach should be followed in the design phase. A new designing paradigm should have the ability to design aspects with minimal dependency in separation with other units. Tight coupling nature of aspects highlight some problems in [6].

- **Maintainability**-Aspects having the comprehensible design are easier to maintain in the code. But what if tangling of aspects with other units of the complicated system come across, also having addition, deletion and modification operations will perform on aspects result into a high overhead with inconsistencies. When designing good software it should be kept in mind to separate aspects with other modules which will provide an easy to maintain software in an effective manner.

- **Reusability**- Aspect of a system is modularized which helps in other systems by reusing the modularize aspect. This is one of the key objectives of AO approach. The reusable modules help in seeking maintainability. However, the following aspect property is difficult to get as aspect has the problems related to tight coupling and high cohesive nature. There should be minimal dependency and reference of aspects with rest of the units so as to develop an ideal design.

2 Literature Review

The following are aspect-oriented design and modeling approach.

Theme/UML- show implementation on identified themes of system with aid of Theme Doc.

Motorola Weaver Approach- This technique provides implementation of semantics and design technique.

3 Aspect-Oriented Design (AOD)

Aspect-Oriented Design is apply to Aspect-Oriented techniques to models with the aim of modularizing crosscutting concerns. The distinct level of abstraction is used in modeling software by using different notations. Aspect-Oriented Design is applied during design using Aspect-Oriented Modeling, and using Aspect-Oriented Programming for implementation. There are several AOM approaches which are used to model software by using Aspect-Oriented Design (AOD) process. Each of those AOD approach has different origin with different view supporting the distinct between aspect and base [7].

2.1 Aspect Design Modeling Language (ADML)

2.1.1 Objective of Modeling language

Modeling Language makes Analysis and Design method easier, by providing specifying, documenting and designing of aspects. The prime objective is to develop a secure software that addresses concern. Now, the software has been developed to show association among the aspectual elements and core elements of the system. It main work is the composition of base design with the aspect.

In the first stage of identifying the concerns, aspect are captured at the time of requirement engineering along with analysis phase. Over some year, there are number of approaches based on requirement engineering have been projected to identify aspects. This existing report does not account a particular aspect capturing method but aspects are captured by having an appropriate methodology. In this report, the language used to design, specify and represent aspects known as Modeling Language, not only provides design techniques and notations along with methods but also shows the association and relationship of aspects with the base concerns. There are some of the following objectives that are achieved at the span of development of language.

- Unify aspects with the objects in the design phase in one framework.
- Designing notations for aspectual elements and basic elements should be followed.
- Representation of aspect having structural characteristic along with behavioral in the diagram form should be followed.
- Developing a language that provides a designing solution which is comprehensive to aspectual elements should be maintained.

2.2.2 Concept of Aspect Design Modeling Language (ADML)

The important concept of standardized design language which has specified design notations is highlighted by many researchers. It is in the context of

Aspect-Oriented approach of software development over some years. UML is effectively used for the object design notations so there is a need of language arises that proven to be a de-facto standard that can commendably combine aspects and objects together. Earlier many design approach came up but no one provides a comprehensive design solution to aspect. Even there is one issue that is left unaddressed in most of the design approach which is how we unified aspect and object in single design framework. As far as it is tried, aspect may not be separated completely from core object constructs because of tight coupling among the pointcuts in base program behavior along with structure. The designing of aspect should be done in the way that is modeled with interacting base object. But most of the design methodology proposes separate technique of design to aspect and object which ultimately arises inconsistencies to system design.

There is a need of a language which shows similarity to UML so as to specify and represent aspect with their constituents. An extended UML is selected as a language for modeling aspect. There are some reasons and one of the basic reason is that it is best suited as a modeling language. Also it is object-oriented modeling language and aspect needs to be implemented with object constructs. So, for designing aspect an extended version of Unified Modeling Language is required. If a new design language is adopted then it will make difficult for designer to embraced it and will compel designer to work on two different languages one for object and other for aspect. Other reason for the extensibility of UML is that new notations can be easily introduced. So, in the ADML, there is freedom of introducing different new notations for aspectual elements. AspectJ forms the basis of ADML which provides design notations for the AspectJ constructs such as advice, join point, pointcut and aspect. To provide better understanding of aspect, it is essential to have design notations to follow which explains the structure and also the behavior of it.

Each and every aspectual element gets the structural along with behavioral support by ADML. When modeling different aspectual components, then designing the diagrams help in acquiring an alternate perspective. It is up to the designer to select an appropriate diagrams for the required model. At the behavioral level, internal flow of the aspect components and the composition of aspect components with the base constructs are shown preferably by behavioral diagrams. The above diagrams shown in context of behavioral UML diagrams these are like collaboration diagrams and also activity diagrams. Likewise, structural representation of the aspect elements and their composition with base elements can be captured by the structural diagrams of system. For example, Diagrams like Aspect Design shows a structural design in which association with all features of aspect is available in structural notations. The second type of structural diagrams such as Aspect Class which presents a good view of relationship that exists between aspects and core objects.

There are some systems present which are more acute and security needed systems require extra behavioral representation to get the good test

generation and on the other hand, there are systems that need more structural representation to have a better understanding of relationship among aspects and core components. So, it depends on the designer to select the most appropriate diagram for modeling of a system.

2.2.3 ADML Design Diagram:

Aspect-Class Diagrams

There are two types of following Aspect-Class Diagram:

1. Aspect-Class Static Diagram
2. Aspect-Class Dynamic Diagram

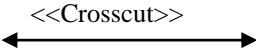
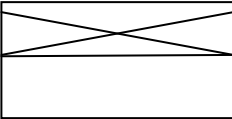
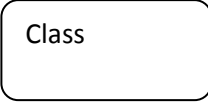
Aspect-Class Static Diagram


Description

In a static model of ADML, the aspect-class relationship is designed. Modeling of aspect-class diagram represents that in single diagram both class and aspect interface with each other which aid in the identification of entity involve in weaving process. The stereotype in diagram <<crosscut>> depicts the crosscutting concern. The stereotype represents association of type class-directional between aspect and class.

Notations

Table 4. Notation of Aspect-Class Association for Static Figure

ELEMENT	NOTATIONS	EXPLANATIONS
Aspect-Class Association		This notation represents association link of aspect with base class.
Aspect		In a program an abstraction to crosscutting concern.
Class		In UML the depiction of core class.

Association		This represents association between two base class
-------------	---	--

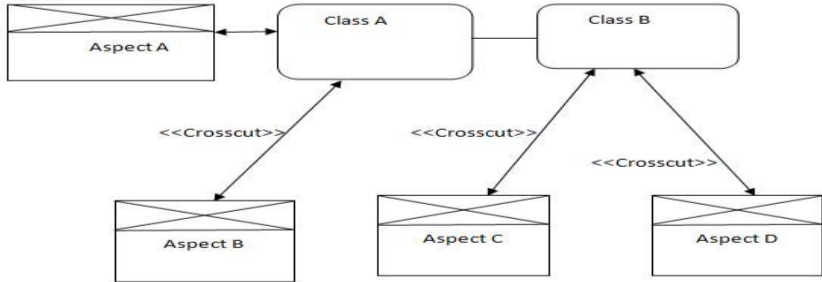


Fig.4. In the above figure shows general purpose Aspect-Class Static Diagram which represents a static relationship among Classes and Aspects. Classes here depict as core class having core functionality. Aspects depict the crosscutting concerns which are in some way crosscut into classes.

Aspect-Class Dynamic Diagram

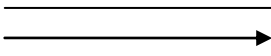
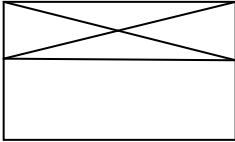
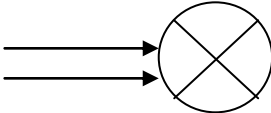
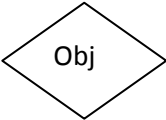
Description

This diagram is known as dynamic diagram as weaving taken place at run time which represents a dynamic process. This process is very important in Aspect-Oriented Software Development. This figure depicts that at the time of dynamic flow of system advices are appended to the base object. By using UML communication diagram, the aspect-class dynamic figure represent the weaving process to simulate.

Notation

Table 2. Notation of Aspect-Class Association for Dynamic Figure

ELEMENT	NOTATION	EXPLANATIONS
---------	----------	--------------

<p>Message</p>	<p>1: method name() </p>	<p>These arrows represent message based information with name of method along occurrence sequence. The head of arrow depicts flow direction.</p>
<p>Aspect</p>		<p>An abstraction of crosscutting concern.</p>
<p>CodeWeaving</p>		<p>The process incorporates aspect's behaviour into base program.</p>
<p>Object</p>		<p>An instance of a class.</p>

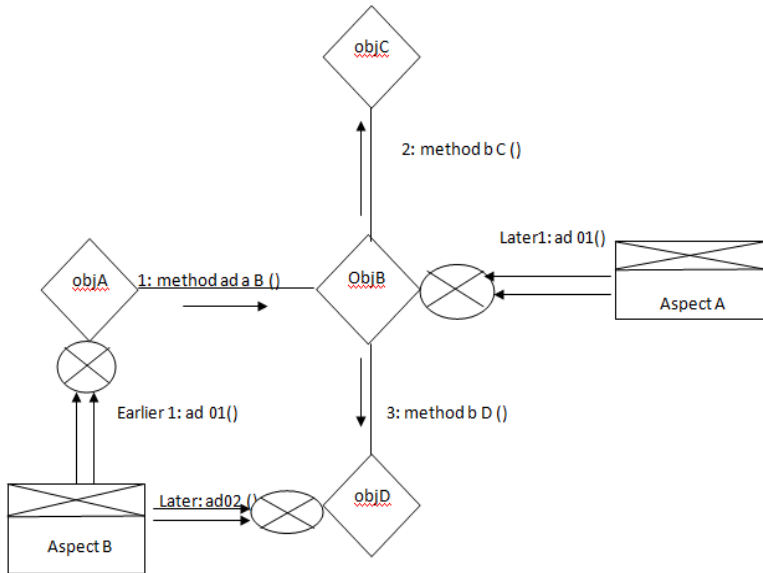


Fig.5. In the above figure shows general purpose Aspect-Class Dynamic Diagram at the time of execution of the system the aspects which have crosscutting concern code weave with core class object. This diagram shows dynamic relationship as it involves a weaving process.

4 Conclusion

Aspect-Oriented Software Development intends to implement a modular approach in the programming phase having more attention given to crosscutting concerns. This paper investigates the concept of crosscutting concerns along with software development principles and suggest a way to deal with these concerns which come across in multiple modules of software development life cycle of complex software.

Addressing the concerns is an important issue for developing a secure application. Also specify the concerns in separate way and applied it to throughout application with proper implementation is an achievable goal to accomplish. For developing a complex application, a framework is needed which is present in this paper in a modeling language form known as ADML. The primary focus of developing this modeling language is to tackle the problem of complicate software development in modular way through aspects and associate it with base constructs through weaving process, so a new application which is easy to use and secure to handle will be developed. The cost of application with development efforts along with its evolution can be achieved by improved modularity.

5 References

1. Stein, D., Hanenberg, S., Unland, R.: Designing aspect-oriented crosscutting UML. In: Workshop on Aspect-Oriented Modeling with UML, in conjunction with the 1st International Conference on Aspect-Oriented Software Development. Enschede, The Netherlands, p. 6 (2002) (accessed on: September 8, 2008).
2. Grassi, V., Sindico, A.: Uml modeling of static and dynamic aspects. In: International Workshop on Aspect-Oriented Modeling, Bonn, Germany, p. 6 (2006) (accessed on: October 16, 2007).
3. OMG. UML 2.0 Infrastructure Specification (2008), <http://www.omg.org> (accessed on: October 20, 2008).
4. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., Griswold, W.G.: An overview of aspectJ. In: Knudsen, J.L. (ed.) ECOOP 2001. LNCS, vol. 2072, p. 327. Springer, Heidelberg (2001),
5. Stein, D., Hanenberg, S., Unland, R.: An UML-based aspect-oriented design notation for aspectj. In: 1st International Conference on Aspect-Oriented Software Development, Enschede, The Netherlands (2002),
6. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., Griswold, W.G.: An Overview of AspectJ. In: Lindskov Knudsen, J. (ed.) ECOOP 2001. LNCS, vol. 2072, pp. 327–353. Springer, Heidelberg (2001).
7. Basch, M., Sanchez, A.: Incorporating aspects into the UML. In: International Workshop on Aspect-Oriented Modeling, p. 5 (2003).
8. Stein, D., Hanenberg, S., Unland, R.: An UML-based aspect-oriented design notation for aspectj. In: 1st International Conference on Aspect-Oriented Software Development, Enschede, The Netherlands (2002),
9. Fowler, M.: UML Distilled: A Brief Guide to the Standard Object Modeling Language, 3rd edn., pp. 53–63. Addison-Wesley, Boston (2004).
10. Laddad, R.: AspectJ In Action: Practical Aspect-Oriented Programming, 513 p. Manning Publications Co., Greenwich (2003).
11. Baniassad, E., Clements, P.C., Ara´ujo, J., Moreira, A., Rashid, A., Tekinerdogan, B.: Discovering early aspects. IEEE Software 23(1), 61–70 (2006).
12. VyasVaibhav, Vishwakarma R., Jha C.K., 2016. Modelling Aspects with AODML: Extended UML approach for AOD. International Journal of Information Technology and Computer Science, MECS Publisher (Accepted).

13. Katz, S., Rashid, A.: From aspectual requirements to proof obligations for aspectoriented systems. In: Intl. RE Conf., pp. 48–57 (2004).
14. Lesiecki, N.: Unit test your aspects – eight new patterns for verifying crosscutting behavior. IBM Developer Works (2005).
15. Niu, N., Easterbrook, S., Yu, Y.: A taxonomy of asymmetric requirements aspects. In: Moreira, A., Grundy, J. (eds.) Early Aspects Workshop 2007 and EACSL 2007. LNCS, vol. 4765, pp. 1–18. Springer, Heidelberg (2007).
16. VyasVaibhav, Vishwakarma R., 2016. Aspect Oriented Approach for Securing Web Application. In International Conference for ICCMTC, Allahbad, IEEE.
17. VyasVaibhav, Vishwakarma R., 2016. Integrate Aspects with UML: Aspect Oriented Use case Model. In 4th International Conference on PDGC, JP Institute, Solan, IEEE.

Analysis of Stability and Convergence on Perceptron Convergence Algorithm

Vaibhav Kant Singh

Assistant Professor, Department of Computer Science and Engineering ,
Institute of Technology,
Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur,
Chhattisgarh, India
vibhu200427@gmail.com

Abstract. The Paper deals with the implementation of Perceptron Algorithm on a given data set and its graph plotted for two different values of Learning Rate Constants. The paper shows the convergence behavior portrayed by the problem addressed in the Training set. The paper deals with some of the aspects related to the selection of value for the Learning rate parameter. The two cases are based on the different values of Learning Rate Parameter applied on Perceptron model. The Lippmann algorithm [6,(1987)] is applied on the model and the graphs are plotted on which analysis is made.

Keywords: Activation Dynamics, Convergence, Stability, Synaptic Dynamics.

1 Introduction

Soft-Computing is an emerging Computing field. Soft-Computing as the name implies deals with the flexible nature portrayed by the human beings. There are situations where the computer system fails to give result. As in general computer is a machine that can give answers for what it is coded for. Computers are having programmer's intelligence.

Soft-Computing is not something that is going to create an exact clone of the human being. But is something that is going to make computers to give answers that are going to mimic human's flexible behavior.

1.1 Branches of Soft Computing

Soft-Computing is different from the term Software Computing. Soft-Computing is a term that encompasses several fields like Artificial Neural

Networks, Fuzzy Logic, Genetic Algorithm, Particle Swarm Optimization (PSO) etc.

The fields specified above are having their relationship with the lifestyle, behavior, Origin, Adaptation etc. pertaining with the living organisms.

1.2 Artificial Neural Network

Artificial Neural Network as the name itself reveals is the network of neurons. The neurons present in the network are manmade and that is why the neurons are termed as artificial. The neurons being artificial also reveal that they are mimicking some functioning of the original neuron. The term neuron comes from biology. Neurons are the basic computational unit of human brain. Brain is capable to process and give response through the neurons. In Artificial Neural Network we try to model something that is analogous to the working of human brain. There are so many Artificial Neural Network models that perform a wide range of functionality. There are a number of applications constructed on the Artificial Neural Network paradigm. This paper deals with two important terms coming in Artificial Neural Network i.e. Stability and Convergence.

1.3 Fuzzy Logic, Genetic Algorithm, PSO etc.

Fuzzy Logic is a mathematical formulation which helps us to consider all possible elements with a membership order. Genetic Algorithm deals with the Darwin's theory of survival of the fittest. PSO is something that also is related to nature and is basically utilized as an Optimization technique.

2 Stability

The structure of Artificial Neural Network and the rules that dictate the change in the synaptic weight to learn a pattern are very vital in the study of Artificial Neural Network. The change in weight is governed by Synaptic Dynamics and change in activation value is governed by Activation Dynamics.

Stability is something that is related to the Equilibrium behavior of the activation state of the Artificial Neural Network. In this paper we will see and analyze the behavior portrayed for a problem.

3 Convergence

The ultimate objective of any Learning algorithm in Artificial Neural Network is to gain the pattern information present in the input vector in the form of numeric values contained in the synaptic weights present in the topology. The word Convergence in Artificial Neural Network refers to the adaptive behavior portrayed by the synaptic weights during the course of learning. Learning type either supervised or unsupervised in terms of convergence refers to the same notion of adjustment of weight. In this paper we will see the behavior portrayed by a topology and learning algorithm called Perceptron convergence algorithm.

4 Literature Survey

W.S. McCulloch and W. Pitts [1, (1943)] gave the mathematical formulation mimicking the functioning of the nerve cells. Robbins and Monro [2, (1951)] made a conscious effort in formulating principle for Learning Rate Parameter using stochastic methodology. The major part of the current paper deals with the proposed Artificial Neural Network model by F. Rosenblatt [3, (1958)]. Ljung [4,(1977)] and Kushner and Clark [5,(1978)] deals with stochastic approach for finding Learning rate parameter. The Perceptron Convergence Algorithm by Lippmann [6,(1987)] is the base used in the paper for making analysis of the Perceptron model. Darken and Moody [7,(1992)] proposed search then converge schedule for identification of Learning Rate Parameter value. The author of the paper Vaibhav Kant Singh [8,9,10,11,12,13,14,15,16, (2015-2016)] made several ANN models for solving various problems and did survey and analysis of the various ANN topologies and learning algorithms. The survey and analysis made are used for interpretation of several concepts in the paper.

5 Critical Observation

Different methodologies are given for identification of the value for Learning Rate Parameter. Some of the unique points identified on the context of Learning Rate Parameter are:-

- a) The Learning Rate Parameter is Time Varying
- b) Stochastic Approximation method is proposed for Evaluation
- c) Choice of Constant is possible for Guaranteed Convergence in Stochastic approach
- d) There is a Chance of getting into the danger of Parameter Blowup in case the size of n is small.
- e) Search then converge schedule used user selected constants in the formula for calculating Learning Rate Parameter. The algorithm

proposed by Darken and Moody operates almost equal to the LMS algorithm when the number of iteration is small as compared to the search time constant.

6 The Perceptron Convergence Algorithm

In theory Perceptron Convergence algorithm says that if the patterns i.e. the input patterns are drawn from linearly separable classes that the algorithm will easily converge into set of values that will solve the classification problem. To look into the notion addressed above let us apply the Perceptron Convergence Algorithm on the Training Set present in the Table 1.

6.1 Perceptron Convergence Algorithm

Variables and Arguments Present in Algorithm

$I(T)=(K+1)$ -BY-1 INPUT VECTOR

$= [+1, I_1(T), I_2(T), \dots, I_K(T)]^{TRANS}$

$W(T)=(K+1)$ -BY-1 WEIGHT VECTOR

$= [B(T), W_1(T), W_2(T), \dots, W_K(T)]^{TRANS}$

$B(T)$ =BIAS

$O(T)$ =ACTUAL OUTPUT

$D(T)$ =DESIRED OUTPUT

ϖ = LEARNING RATE CONSTANT

NOTE:

$$SIGNUM(ACTIVATION\ VALUE) = \begin{cases} = +1 & \text{IF ACTIVATION VALUE} > 0 \\ -1 & \text{IF ACTIVATION VALUE} \leq 0 \end{cases}$$

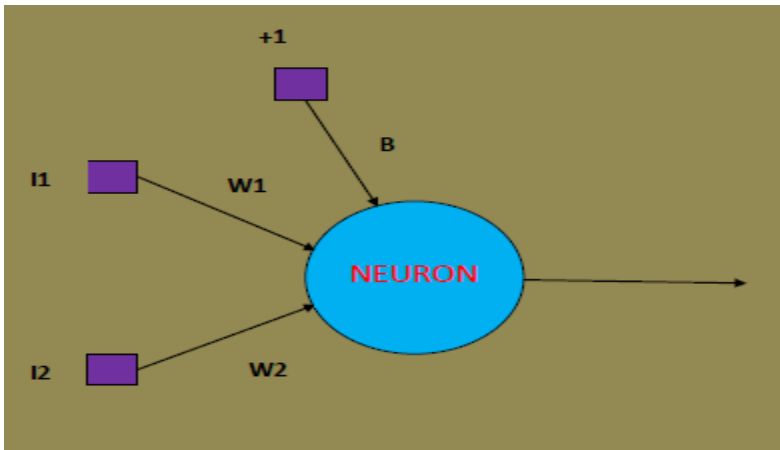


Figure1:Architectural Graph of Perceptron

1. **STEP1[INITIALIZATION]** : SET $W(0)=0$. AFTERWARDS PERFORM THE FOLLOWING CALCULATIONS FOR THE TIME STEPS $I=1,2,3,\dots$
2. **STEP2 [ACTIVATION]** : AT TIME STEP T , ACTIVATE THE PERCEPTRON BY APPLYING CONTINUOUS-VALUED INPUT VECTOR $I(T)$ AND DESIRED OUTPUT $D(T)$.
3. **STEP3 [COMPUTATION OF ACTUAL RESPONSE]** : COMPUTE THE ACTUAL RESPONSE OF THE PERCEPTRON.

$$O(T)=\text{SIGNUM}[(W(T)^{\text{TRANS}}I(T))]$$

WHERE SIGNUM(.) IS THE ACTIVATION FUNCTION

4. **STEP4 [ADAPTATION OF WEIGHT VECTOR]** : UPDATE THE WEIGHT VECTOR OF THE PERCEPTRON:

$$W(T+1)=W(T)+\varpi[D(T)-O(T)]I(T)$$

5. **STEP5 [CONTINUATION]** : INCREMENT TIME STEP T BY 1 AND GO TO STEP2

6.2 MATHEMATICAL IMPLEMENTATION ON TRAINING SET OF Table1

CASE1: [WHEN $\varpi=1$]

Table 1 Table Representing the Training Set [Input Vector] and [Target Vector] .

INPUT VECTOR			DESIRED OUTPUT [TARGET VECTOR]
I0	I1	I2	D
1	0	0	-1
1	0	1	-1
1	1	0	-1
1	1	1	+1

Table 2 Table Representing the SOLUTION for CASE1

INPUT VECTOR		WEIGHT VECTOR		ACTIVATION VALUE		ACTUAL OUTPUT	DESIRED OUTPUT	ITERATION NUMBER	
I0	I1	I2	B	W1	W2	A	O	D	
1	0	0	0	0	0	0	-1	-1	1
1	0	1	0	0	0	0	-1	-1	2
1	1	0	0	0	0	0	-1	-1	3
1	1	1	0	0	0	0	-1	+1	4
1	0	0	2	2	2	2	+1	-1	5
1	0	1	0	2	2	2	+1	-1	6
1	1	0	$\frac{1}{2}$	2	0	0	-1	-1	7
1	1	1	-	2	0	0	-1	+1	8

						2				
1	0	0	0	4	2	0	-1	-1	9	
1	0	1	0	4	2	2	+1	-1	10	
1	1	0	$\frac{-}{2}$	4	0	2	+1	-1	11	
1	1	1	$\frac{-}{4}$	2	0	-2	-1	+1	12	
1	0	0	$\frac{-}{2}$	4	2	-2	-1	-1	13	
1	0	1	$\frac{-}{2}$	4	2	0	-1	-1	14	
1	1	0	$\frac{-}{2}$	4	2	2	+1	-1	15	
1	1	1	$\frac{-}{4}$	2	2	0	-1	+1	16	
1	0	0	$\frac{-}{2}$	4	4	-2	-1	-1	17	
1	0	1	$\frac{-}{2}$	4	4	2	+1	-1	18	
1	1	0	$\frac{-}{4}$	4	2	0	-1	-1	19	
1	1	1	$\frac{-}{4}$	4	2	2	+1	+1	20	
1	0	0	$\frac{-}{4}$	4	2	-4	-1	-1	21	
1	0	1	$\frac{-}{4}$	4	2	-2	-1	-1	22	

1	1	0	$\frac{-}{4}$	4	2	0	-1	-1	23
1	1	1	$\frac{-}{4}$	4	2	2	+1	+1	24

CASE2:[WHEN $\omega=0.5$]

Table 3Table Representing the SOLUTION for CASE2

INPUT VECTOR		WEIGHT VECTOR		ACTIVATION VALUE		ACTUAL OUTPUT	DESIRED OUTPUT		ITERATION NUMBER
I0	I1	I2	B	W1	W2	A	O	D	
1	0	0	0	0	0	0	-1	-1	1
1	0	1	0	0	0	0	-1	-1	2
1	1	0	0	0	0	0	-1	-1	3
1	1	1	0	0	0	0	-1	+1	4
1	0	0	1	1	1	1	+1	-1	5
1	0	1	0	1	1	1	+1	-1	6
1	1	0	$\frac{-}{1}$	1	0	0	-1	-1	7
1	1	1	$\frac{-}{1}$	1	0	0	-1	+1	8
1	0	0	0	2	1	0	-1	-1	9
1	0	1	0	2	1	1	+1	-1	10
1	1	0	$\frac{-}{1}$	2	0	1	+1	-1	11
1	1	1	$\frac{-}{1}$	1	0	-1	-1	+1	12

			2							
1	0	0	$\frac{-}{1}$	2	1	-1	-1	-1	13	
1	0	1	$\frac{-}{1}$	2	1	0	-1	-1	14	
1	1	0	$\frac{-}{1}$	2	1	1	+1	-1	15	
1	1	1	$\frac{-}{2}$	1	1	0	-1	+1	16	
1	0	0	$\frac{-}{1}$	2	2	-1	-1	-1	17	
1	0	1	$\frac{-}{1}$	2	2	1	+1	-1	18	
1	1	0	$\frac{-}{2}$	2	1	0	-1	-1	19	
1	1	1	$\frac{-}{2}$	2	1	1	+1	+1	20	
1	0	0	$\frac{-}{2}$	2	1	-2	-1	-1	21	
1	0	1	$\frac{-}{2}$	2	1	-1	-1	-1	22	
1	1	0	$\frac{-}{2}$	2	1	0	-1	-1	23	
1	1	1	$\frac{-}{2}$	2	1	1	+1	+1	24	

NOTE:

I0=INPUT ATTACHED TO BIAS WEIGHT

I1=FIRST EXTERNAL INPUT

I2=SECOND EXTERNAL INPUT

B=BIAS WEIGHT

W1=WEIGHT ATTACHED TO FIRST EXTERNAL INPUT

W2=WEIGHT ATTACHED TO SECOND EXTERNAL INPUT

A=ACTIVATION VALUE

O=OUTPUT VALUE

D=DESIRED OUTPUT

T=INPUT NUMBER TO THE ANN DURING TRAINING

7 Graph Representing the Stability and Convergence

This section deals with the graphs plotted on behalf of the data obtained during the training process for the two cases discussed in the section above.

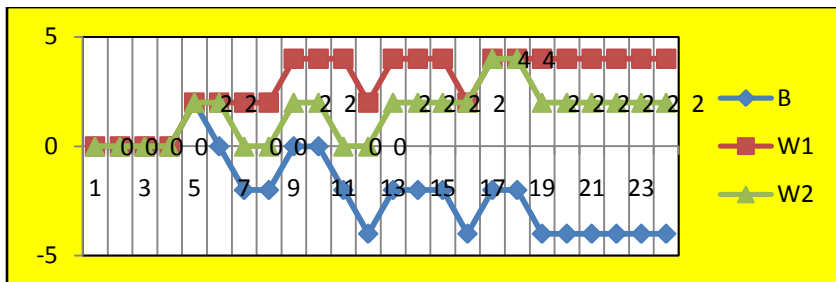


Figure2:-Graph Representing Synaptic Dynamics for Case1 [$\varpi = 1$]

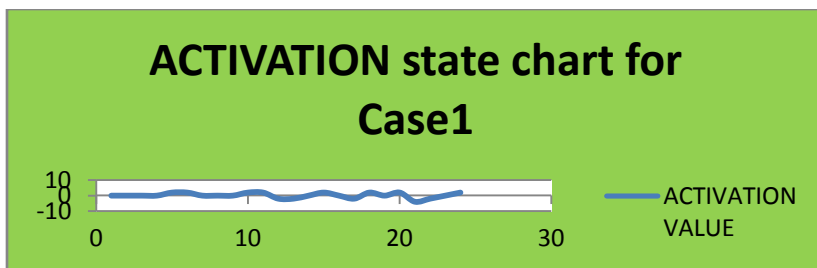


Figure3:Graph representing the Activation Dynamics for Case1 [$\varpi = 1$]

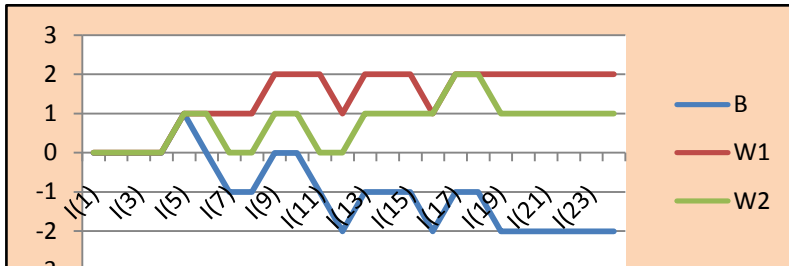


Figure4:Graph representing the Synaptic Dynamics for Case1 [$\bar{\omega} = 0.5$]

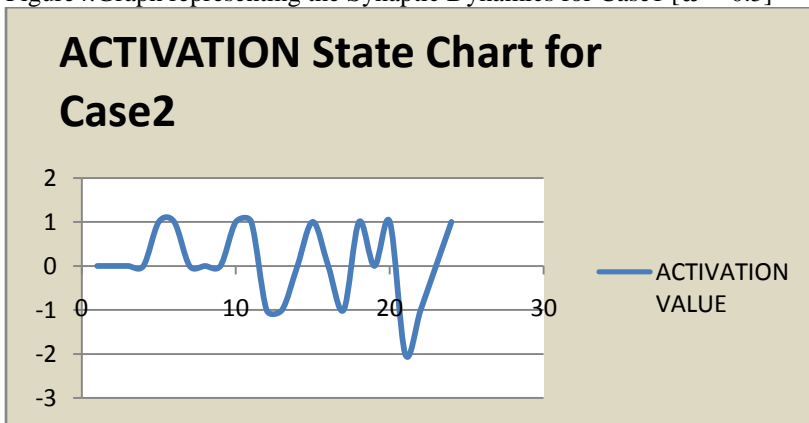


Figure5: Graph representing the Activation Dynamics for Case2 [$\bar{\omega} = 0.5$]

8 Conclusion

From the observations made in the work it is concluded that the Perceptron convergence algorithm as stated in various papers and books is capable to do classification. In other words Perceptron Convergence Algorithm when applied on problem i.e. linearly separable is capable of solving the classification problem. The paper shows the Activation and Synaptic state transition by means of graph. From the graph it is clear that the Learning rate constant does not make very deep impact in convergence. Although there is a slight difference in the upper limit and lower limit values obtained for the free parameters i.e. Bias, Weight attached to external input. The Activation dynamics graph and Synaptic Dynamics Graph for both the cases show similar characteristics. The paper gives the proof of the Perceptron Convergence Algorithm and also graphically shows the

activation and synaptic behavior. The number of iteration after which the network got trained is same for both the cases.

Acknowledgments.

Author :

Mr. Vaibhav Kant Singh would like to thank his Mother Smt. Sushma Singh, Wife Smt. Shubhra Singh, Brothers Mr. Abhinav Kant Singh and Mr. Abhoday Kant Singh, Sister Miss Priyamvada Singh, Daughter Miss Akshi Singh and Son Master Aishwaryat Kant Singh for their support in the completion of the work done. The Author would also like to place his thanks to his Father Shaheed Late Shri Triveni Kant Singh for his blessings.

References

1. McCulloch, W.S., Pitts, W.: A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115—133. (1943)
2. Robbins, H., Monro, S.: A Stochastic Approximation Method. *Annals of Mathematical Statistics*, vol. 22, pp. 400-407. (1951)
3. Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, vol. 65, pp. 386-408. (1958)
4. Ljung, L.: Analysis of Recursive Stochastic Algorithms. *IEEE Transaction on Automatic Control*, vol. AC-22, pp. 551-575. (1977)
5. Kushner, H.J., Clark, D.S.: *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer –Verlag, New York. (1978)
6. Lippmann, R.P.: An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine* vol. 4, pp. 4--22 (1987)
7. Darken, C., Moody, J.: Towards faster stochastic gradient search. *Advances in Neural Information Processing Systems*, vol. 4, pp. 1009--1016. CA : Morgan Kaufmann, San Mateo. (1992)
8. Singh, V.K.: One Solution to XOR Problem using Multilayer Perceptron having Minimum Configuration. *International Journal of Science and Engineering*, vol. 3, no. 2, pp. 32-41, (2015)
9. Singh, V.K.: Two Solutions to the XOR Problem using Minimum Configuration MLP. *International Journal of Advanced Engineering Science and Technological Research*, vol. 3, pp. 16--20. (2015)
10. Singh, V.K.: Proposing Solution to XOR Problem using Minimum configuration MLP. In: *International Conference on Computational Modeling and Security (CMS 2016)*, *Procedia Computer Science*, Elsevier, pp. 255-262, Bangalore, India. (2016)

11. Singh, V.K.: Mathematical Explanation to Solution for Ex-NOR Problem using MLFFN. *International Journal of Information Sciences and Techniques*, vol. 6, pp. 105—122. (2016)
12. Singh, V.K.: ANN Implementation of Constructing Logic Gates Focusing on Ex-NOR. *Research Journal of Computer and Information Technology Sciences*, vol. 4, no. 6, pp. 1--11. (2016)
13. Singh, V.K.: Mathematical Analysis for Training ANNs using Basic Learning Algorithms. *Research Journal of Computer and Information Technology Sciences*, vol. 4, no. 7, pp. 6--13. (2016)
14. Singh, V.K., Pandey, S.: Minimum Configuration MLP for Solving XOR Problem. In: 10th IEEE International Conference on Computing for Sustainable Global Development, IEEE Conference ID:37465, pp. 168--173. IEEE Explore, INDIACom-2016, BVICAM, New Delhi, India (2016)
15. Singh, V.K., Pandey, S.: Proposing an Ex-NOR Solutions using ANN. In: International Conference on Information, Communication and Computing Technology, IIC, Jagan Institute of Management Studies and CSI, pp. 277—284, New Delhi, India (2016)
16. Singh, V.K.: Proposing a New ANN Model for Solving XNOR Problem. In: 5th International Conference on System Modeling and Advancement in Research Trends, IEEE Conference ID:39669, TMU Moradabad, India. (2016)

A Review of Cyber bullying Detection in Social Networking

Prankit Namdeo¹, R.K Pateriya², Sonika Shrivastava³,

¹prankitn@gmail.com, ²pateriyark@gmail.com, ³ms271104@gmail.com

Department of Computer Science and Engineering, MANIT Bhopal -
462003, INDIA

Abstract- With the advancement of Web 2.0, number of social networking platforms is increasing. This has led to the growth of cyber bullying which has become an epidemic. Cyber bullying refers to the use of electronic gadgets to bully a person by sending harmful messages using digital messages. It has now become a platform for insulting and humiliating a person which can affect the person physically, mentally, emotionally and sometimes leading to suicidal attempts in the worst case. This paper is an analysis of cyber bullying and its impact on teens and adults. The main aim of this paper is to cover all the approaches which have been used in cyber bullying detection. It also contains a comparative study of various cyber bullying detection techniques with and without machine learning which will be useful in the future research.

Keywords: Cyber bullying, Machine Learning, Social Media

I Introduction

With the ease of and ubiquitous online access, cyber security is an important concern. The modern day technology is a boon and social media sites cannot be blamed for criminal acts. Cyber bullying is a type of bullying that takes place using electronic technology including devices such as cell phones, computers through social media, text messages, chats etc [1]. Cyber bullying is also defined as “willful and repeated harm inflicted through the medium of electronic text” [2]. It mainly targets children and adolescents as they are most active on social networks. Online and offline bullying are both similar.

Some of the most common forms of cyber bullying are [3]:

Flaming refers to online fights that take place on Internet using vulgar and abusive language. Sometimes a group of people may involve in a heated communication on a particular topic which can lead to cyber bullying.

Harassment is sending offensive, vulgar and threatening messages to someone in order to harm the person.

Denigration is exposing secrets of a person or posting gossips about a person in order to damage the reputation of person which can harm the person socially as well as mentally.

Impersonation involves a false identity of the person to break into the victim's account and posting embarrassing messages on behalf of the victim.

Trickery involves tricking the victim to reveal sensitive details and using those details to cause inconvenience by passing the information to others.

Trolling refers to making fun of a person by making funny posts or comments on his public profile over a social media platform. These funny posts are called as trolls.

An article from The Times of India [4] entitled “\$188,776 Facebook grant for cyber bullying expert Sameer Hinduja” clearly states the increasing cyber bullying from the fact that Sameer Hinduja has received \$188,776 grant from social networking site Facebook to study cyber bullying. The goal of the study is to study the scope of cyber bullying. An article from The Indian Express stated that 50% Indian youths have experienced cyber bullying and found that most of the Indian parents don't find it important to talk to their children about online safety. Although there is age restriction on joining various social networking sites but 10-12 years teens report a very high access to these sites. The Global Youth Online Behavior Survey ranked India third in cyber bullying. A rising number of such cases are being reported, underlining the trend.

National Crime Prevention Council defines cyber bullying as sending text or images to hurt or embarrass another person by using Internet, mobile phones or other devices [5]. According to a research conducted by Symantec, only 25% of the parents were aware that their child was involved in cyber bullying incident [6]. According to a survey [7], majority of cyber bullying is done through Facebook and around 55% of the youth exposed to cyber bullying committed suicide.

The following diagram illustrates the percentage of people where they are bullied most [8]:

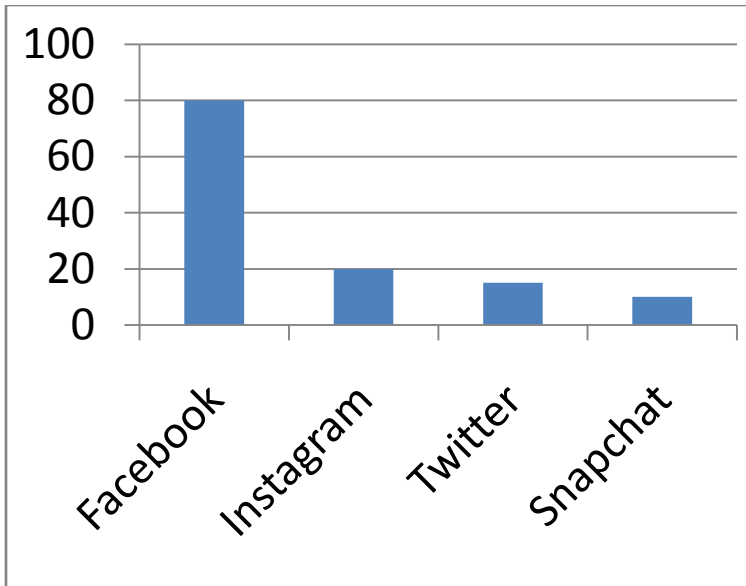


Fig.1 Percentage of people where they are bullied most

A very few research teams are working on the detection of cyber bullying. Many incidents have been reported from all over the world through many new stories. So there is a need of detecting online harassment so that necessary actions can be taken against it. The second section deals with the approaches that are used in cyber bullying detection. The third section contains the datasets used for the previous research and in the next section we offer our proposed work and suggestions for the future research.

II Cyber bullying Detection

According to a study on computer technology to detect cyber bullying [10], a Cyber Bullying Reporting Platform (C.B.R.P.) is adopted to counter cyber bullying by reporting cases of cyber bullying. It is a portal where the victims report the incident. They are required to register on the website and submit the incident. The incident gets submitted to the admin of the website and action is taken by third party as per seriousness of matter.

According to the online available applications and tools [22], eBlaster, Net Nanny, cloud9 and IamBigBrother are some available commercial tools. They work on the basis of Packet sniffing. These tools scan the outgoing and ingoing traffic in a network. But the problem with these tools is that they are based on a simple keyword matching and hence its accuracy is questioned. According to a computer software [13] Bully Tracer which was

designed to detect cyber bullying had a dictionary of bad words to detect offensive text. This approach detected bullying content 85% of the time and innocent content 52% of the time. Improvements to the existing tools are to be made to improve text mining and the bad word set needs to be updated on a daily basis.

According to a study on Psychological Perspective by Feinberg and Robey [11], worked on the psychological aspects of cyber bullying by preventing them using careful observation, monitoring, setting up school campaigns against cyber bullying as most of the affected individuals are teens, by hosting anti-bullying programs, counseling of individuals and by monitoring Internet traffic. Their paper mainly discussed the psychological aspects of preventing cyber bullying. According to The Delete Cyber bullying Project [12], cyber bullying is best detected by simply observing the child. If a child's behavior changes, for example the kid stops using his/her cell phone or computer or any other communication device, if he/she gets upset after taking a call or receiving a text, it indicates that the child is being cyber bullied.

According to a study on Text Mining Approach, the relevant documents are first collected to identify patterns in multiple documents. The data is then pre-processed which involves breaking up of a stream of text into tokens called as tokenization. Subsequently, cleaning up of the text, determination of the relationship of the words with adjacent words to find their meaning [14, 15]. The next step deals with attribute generation where the text document is represented by words. Words and their occurrences are counted and a weight is assigned to each label using an in-built classifier. Then attribute removal is performed and data mining algorithms are applied to this data. The dataset used is MySpace. The text mining cannot detect bullying if it is done in non-curse words which when put together make up an offensive statement.

An approach proposed by Nahar, Li and Pang [16] by using a Graph Model to detect and identify victim and perpetrator worked on detection of the victim and the perpetrator by using 2 phases. The first phase detects harmful messages by employing semantic and weighted features in the feature selection process using L.D.A. (Latent Dirichlet Allocation) algorithm [17]. In the second phase, the predators and victims are identified using HITS algorithm [18]. The person sending the highest bullying content is considered as the bully and the person who receives at least one such message is considered as the victim.

Using Semantic-Enhanced Marginalized Denoising Auto-Encoder [23], in this problem, a new representation learning method is developed via semantic extension of Denoising encoder. Experiments have been conducted over Twitter and MySpace which gave an accuracy of 70.53 % and 65.71 % respectively using different methods such as Bag of Words model etc.

Using Normative Agents to detect cyber bullying before it happens. It focuses on detecting cyber bullying before it happens unlike detecting after it happens. This approach employs normative agents which are physically present in the virtual network and support the victim against attacks [19]. The technique is based on BDI model [20] which detects the violation of predefined norms such as insulting or detection of bad words etc.

According to a study on individual topic sensitive classifiers comments are collected under a label based on sexuality, race and culture, mental capabilities of a person etc [21]. The datasets are subjected to binary and multi-class classifiers to detect comments referring to sensitive topics.

Table 1: A comparison of techniques used to detect cyber bullying

S.No	Technique	Tool	Algorithm	Result
1.	Computer Software	Bully Tracer	-	True Positive Rate (correctly identified instances): 0.85
2.	Semantic analysis	Rapid Miner	Supervised learning algorithm	Confidence (Positive) : 0.615 Confidence (Negative):0.385
3.	Text mining	Weka	Manual labelling	Used Static word set (fixed set of vocabulary) which needs to be updated with time.
4.	Graph model	--	L.D.A and HITS algorithm	--

5.	Semantic Enhanced Marginalized Denoising Auto-Encoder	Stacked denoising encoder	--	Accuracy on Twitter dataset- 70.53 % Accuracy on MySpace- 65.71%
6.	Normative agents	--	BDI Model	Static dictionary of bad words approach which needs to be updated.
7.	Individual topic sensitive classifiers	--	Detects sensitive topics	--
8.	Online available applications, tools	Tools such as eBlaster, cloud9	Simple keyword matching	Accuracy not up to the mark.
9.	Computer Technology	Cyber Bullying Reporting Platform (C.B.R.P.)	--	--
10.	Psychology	--	Campaigns, Anti bullying programs etc	--

In a study by **K. Reynolds, a Kontostathis, and L. Edwards** that used machine learning to detect cyber bullying [6, 22, 26], various machine learning algorithms were applied such as decision tree using J48, a rule based algorithm using JRIP, SVM (support vector machine), Naive Bayes algorithm, Neural Network classifier and k-nearest neighbor approach using IBK (Instance Based Algorithm) using a Weka tool kit giving accuracy of on an average 70 %. Out of all these algorithms, decision tree using J48 algorithm provided the best results. The dataset was downloaded from FormSpring.me, a social networking site, which contains a very high bullying content. The data was labeled using an Amazon Web Service called Turk. A list of bad words downloaded from www.noswearing.com

was used to assign severity levels. The number of bad words was normalized and then used as feature to develop the model.

Table 2: Comparison of machine learning techniques to detect cyber bullying

S.No	Machine Learning Algorithm Used	Accuracy Obtained
1.	Decision Tree Algorithm	78.5%
2.	K nearest neighbour	78.5%
3.	Rule based algorithm	73.77%
4.	Probabilistic Classifier	72.30%
5.	Supervised learning algorithm	60.49%

III Dataset used in cyber bullying detection

Most of the research conducted for cyber bullying detection mainly concentrated on the dataset available on CAW 2.0 (Content Analysis for Web 2.0) consisting of datasets of Kongregate, MySpace, and Slashdot. The data is not labeled too. It was labeled using Amazon Turk Web service. The dataset used in the research of predator and victim conversation is chat logs transcripts from Perverted Justice [25] or by downloading chat logs. A large dataset from Form spring, a question and answer based format social networking has also been used till now in the research. At present, apart from above datasets, there are so many datasets available online which contains bullying data which can be used further.

IV Conclusion

In this paper we illustrated the detection techniques adopted so far to address cyber bullying. Various tools like Weka have been used till now to identify the presence or absence of cyber bullying either by using a set of

static words approach or by applying machine learning classifiers. But there is a need to improve the learning ability of the classifier and to update the list of bad words since cyber bullies use different vocabulary of bad words in some form or the other. So appropriate actions need to be taken to eradicate this evil by adopting techniques that provide accurate results. These techniques can be improved by collecting new datasets and then applying Sentiment analysis using machine learning approach on the social media data to obtain the desired accuracy. Future research can be carried out in detecting bad word which can be updated to improve the learning ability of the classifier and also on collecting new datasets for the further study on cyber bullying detection.

References

- [1] <https://www.stopbullying.gov/cyberbullying/>
- [2] J.Patchin, & S. Hinduja, "Bullies move beyond the schoolyard; a preliminary look at cyber bullying." Youth violence and juvenile justice.4:2 (2006). 148-169.
- [3] Sourabh Parime, Vaibhav Suri "Cyber bullying Detection and Prevention: Data Mining and Psychological Perspective", 2014 International Conference on Circuit, Power and Computing Technologies [ICCPCT]
- [4] <http://www.TIMESOFINDIA.com>
- [5] <http://www.ncpc.org/cyberbullying>
- [6] K. Reynolds, A Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyber bullying," In Proceedings of the 2011 10th international Conference on Machine Learning and Applications Workshops (ICMLA 2011), vol. 2, December 2011. pp. 241-244.
- [7] <http://www.statisticbrain.com/cyber-bullying-statistics/>
- [8] <http://en.wikipedia.org/wiki/Cyber-bullying/>
- [9] A. M. Chandrashekhar, Muktha G S& Anjana D K, "Cyberstalking and Cyber bullying: Effects and prevention measures" Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-3, 2016 ISSN: 2454-1362
- [10] R. Cohen, D. Y. Lam, N. Agarwal, M. Cormier, J. Jagdev, T. Jin, M. Kukreti, J. Liu, K. Rahim, R. Rawat, W. Sun, D. Wang, M. Wexler, "Using Computer Technology to Address the Problem of Cyber bullying", SIGCAS Computers & Society | July 2014 | Vol. 44 | No. 2
- [11] Ted Feinberg, Nicole Robey, "Cyber bullying: Intervention and prevention strategies", Helping children at Home and School volume 3, pp. S4H15-1-4, National association of school psychologists, 2010.
- [12] www.deletecyberbullying.org
- [13] Jennifer Bayzick, April Kontostathis and Lynne Edwards, "Detecting the presence of cyber bullying using Computer Software", WebSci '11,

Koblenz, Germany, National Science Foundation, Grant No. 0916152, pp.1-4, June 2011.

[14] Jim Stern, "Text analytics for social media", S.A.S. Whitepaper, pp.1-13, 2010.

[15] Naveen Kumar, Saumesh Kumar and Padam Kumar, "Parallel Implementation of parts of speech tags for Text mining using grid computing", Advances in Computing and Communications in Computer and Information Science Volume 190, Springer Publications, pp. 461-470, 2011.

[16] Vinita Nahar, Xue Li and Chaoyi Pang. "An effective approach for cyber bullying detection", Volume 3, Issue 5, Communications in Information Science and Management Engineering, pp 238-247, May 2013.

[17] David M. Blei, Andrew Y. Ng and Michael I. Jordan. "Latent Dirichlet allocation", Volume 3, Journal of Machine Learning Research, pp. 993-1022, 2003.

[18] Ramesh Prajapati, "A Survey Paper on HyperlinkInduced Topic Search (HITS)Algorithms for Web Mining", Volume1, Issue 2, International Journal of Engineering Research and Technology, pp.13-20, 2012.

[19] Tibor Bosse and Sven Stam, "A Normative Agent System to Prevent Cyberbullying", in IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011 © IEEE. Doi: 10.1109

[20] www.cs.drexel.edu/~greenie/cs510/bdilologic.pdf.

[21] Mohsen Arab and Mohsen Afsharchi, "A Modularity Maximization Algorithm for Community Detection in Social Networks with Low Time Complexity", presented at The IEEE/WIC/ACM International Conference on Web Intelligence, WI, 2012.

[22] Samaneh Nadali, Masrah Azrifah Azmi Murad, Nurfadhlina Mohammad Sharif, Aida Mustapha, Somayeh Shojae, A Review of Cyber bullying Detection . An Overview. 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)

[23] Rui Zhao and Kezhi Mao, "Cyber bullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder", IEEE Transactions On Affective Computing

[24] Nektaria Potha, Manolis Maragoudakis, "Cyber bullying Detection using Time Series Modeling", 2014 IEEE International Conference on Data Mining Workshop

[25] <http://www.Perverted-Justice.com> .2008.

[26] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. San Francisco, CA: Morgan Kauffman, 2005.

[27] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kauffman, 1993.

[28] D. W. Aha and D. Kibler, "Instance-based Learning Algorithms, "Machine Learning, vol. 6, pp. 37-66, 1991.

Study on threats and improvements in LTE Authentication and Key Agreement Protocol

Ritu Rani¹, Sandeep Kumar², Hitesh Sharma³, Dr. Munish Mehta⁴, Ms. Poonam Saini⁵

²skkansal1993@gmail.com

Department of Computer and Application, National Institute of Technology,
Kurukshetra-136119, Haryana, India

Abstract. This paper presents a study of various authentication models used for implementing security in trending fourth generation cellular networks, which is the result of evolution in 3GPP release 8 known as Long Term Evolution (LTE). 4G/LTE provides high data rate and enhanced capacity due to high bandwidth and high spectral efficiency, which invites many hackers and crackers, who always try to exploit the vulnerabilities in the existing network. In this paper LTE/AKA architecture and its vulnerabilities are analyzed. In the performance evaluation of these models, it is evaluated that there are two ways to secure data and voice transferring between two nodes; these are Symmetric Crypto Key Management and Asymmetric Crypto Key Management. To secure LTE AKA protocols from various attacks, it is required to secure all the parameters used in it, but it will create additional overhead so proper techniques need to be used. This paper gives a brief overview of some of these techniques.

Keywords: AKA, Asymmetric Crypto Key Management, LTE, Symmetric Crypto Key management, 3GPP.

1. Introduction

The recent expansion of Wireless network technology and emergence of novel applications [1] such as multimedia services, high speed internet streaming (telemedicine, tele geo processing and virtual navigation(VoIP), mobile TV etc. with good quality of service and coverage for seamless global roaming requires standardized implementation of 3GPP's release 8 and onwards i.e. LTE and LTE- Advanced in 4G Network. LTE-A promises to deliver high data rate, good QoS, low latency and good coverage with the use of many technologies such as MIMO, OFDMA for downlink, SC-FDMA for uplink and Femtocell to improve indoor coverage and capacity. According to Ericsson's estimate [2], half the world's population will have

LTE coverage by 2017 and many consumer devices—including medical monitors, cameras, and even vehicles—may adopt LTE technology for a new wave of applications. This is to confirm that the mobile industry is reacting swiftly to this new technology as more and more mobile phone producers are also producing LTE compatible phone. For any telecommunication system like 4G network trust and privacy depends upon its security mechanism. LTE/SAE [3] differ from previous generation of cellular networks which uses hybrid packet switched and circuit switched network, but still it is backward compatible. LTE/SAE new architecture is All-IP based and thus carries both data and voice over network. It therefore comes with many security threats such as user privacy concerns, threats to UE/USIM tracking, base stations and handovers, broadcast or multicast signaling, denial of service (DOS), manipulation of control plane, unauthorized access to network, compromise of eNB credential and physical attack on an eNB protocol attack on eNB and attack on the core network and eNB location based attacks [4]. Furthermore, the wireless nature in the LTE networks leads to exponential increase of vulnerabilities, Man in Middle attacks, DoS attacks, eavesdropping etc. Hence the LTE security management becomes the point of interest for various researchers. The Authentication and Key agreement protocol is the main constituent in the LTE Network and play an important role in providing security in 4G network. The 3GPP project releases various AKA protocols as a part of 3GPP's security system starting from 2G-AKA [5], 3G-AKA [6] or UMTS-AKA and finally EPS-AKA [7] for 4G network. UMTS-AKA protocol is designed for 3G network but it has various limitations like tapping user identity and difficulty of sequence numbers etc. EPSAKA protocol due to some additional features solves all limitations of UMTS-AKA protocol but it has also several limitations [4] like user identity attacks, communication cost, bandwidth consumption etc. The user identity attack is due to sending International Mobile Subscriber Identity (IMSI) as clear text from Mobility management entity (MME) to User Equipment (UE) [8].

1.1 EPS-AKA Protocol

LTE System architecture evolution (LTE/SAE) [3], also known as Evolved Packet System (EPS) is all-IP system. EPS as shown in Fig.1 [8] has two elements: 1) Access network which is a network of eNBs also termed as EUTRAN (evolved universal terrestrial radio access network); 2) Core networks of the LTE termed as Evolved Packet Core (EPC) it houses SGW, P-GW, MME and HSS. Serving gateway (SGW) forwards and route packets and handle inter and intra-handover connectivity across eNBs and other technologies of 3GPP. Packet Data Network Gateway (P-GW) provides connectivity between subscriber and the external packet data networks. Home Subscriber server (HSS) is a static catalogue of the subscriber information it is integrated with Authentication Center (AuC) which authenticates the subscriber and encrypts user traffic.

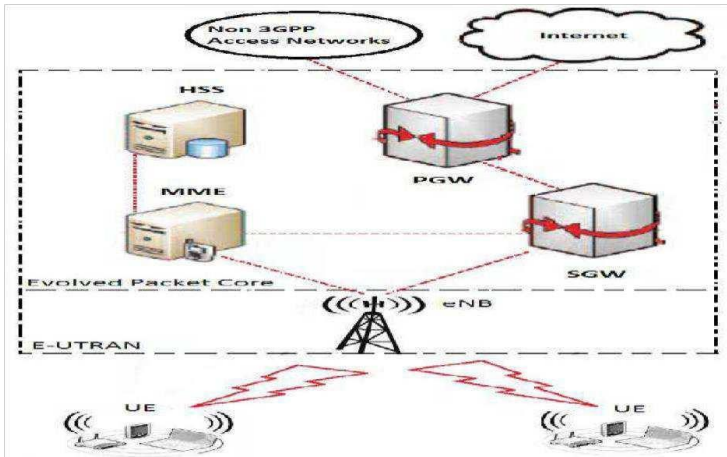


Fig.1. LTE Architecture [8]

For providing mutual authentication between user and the core network LTE uses EPS-AKA protocol which is the last version of the UMTS-AKA. This protocol generates all cipher and integrity keys needed for encrypting the data sent over the network, using a key derivation function. The main components of this protocol are UE or subscriber identity number, eNBs, MME and HSS as shown in Fig.2 [8]. This protocol comes into play when MME sends a user identification request to the UE or subscriber for establishing a connection. Steps of EPS-AKA protocol are as follows [8] [10] [1]:

- 1) On receiving used id request from MME, the UE/ USIM sends a response which contains its IMSI (International Mobile Subscriber Identity), UE security capability and a key set identifier (KSI_{asme}) to MME.
- 2) When MME got the UE identity details it forwards its identity (SN Id), UE IMSI number and Network type (wired or wireless) to HSS for verifying the UE and this request is known as Authentication Data Request.
- 3) Received request is then authenticated by HSS, and find out UE security key (K) from the HSS database. Then it generates a Random number RAND and a self-generating SQNH number. HSS uses five functions to calculate Authentication vector, two of them f₁, f₂ are used for generating MAC (Message Access Code) and Function f₃, f₄, f₅ are used to calculate cipher key, integrity key and authentication vector. Authentication vector is computed as follows: -

- a) Generate RAND
- b) Calculate $MAC=f1(K, SQNH, RAND)$
- c) Calculate expected response($XRES$)= $f2(K, RAND)$
- d) After that several keys are calculated
Cipher Key (CK) = $f3(K, RAND)$

Integrity Key (IK) = $f3(K, RAND)$
 Authentication Token ($AUTN$) = $SQNH || MAC$

Session Key ($Kasme$) = $KDF(CK, IK, SNID)$, where KDF is a key derivation function.

Then AV is computed using (RAND, AUTN, XRES, Kasme)

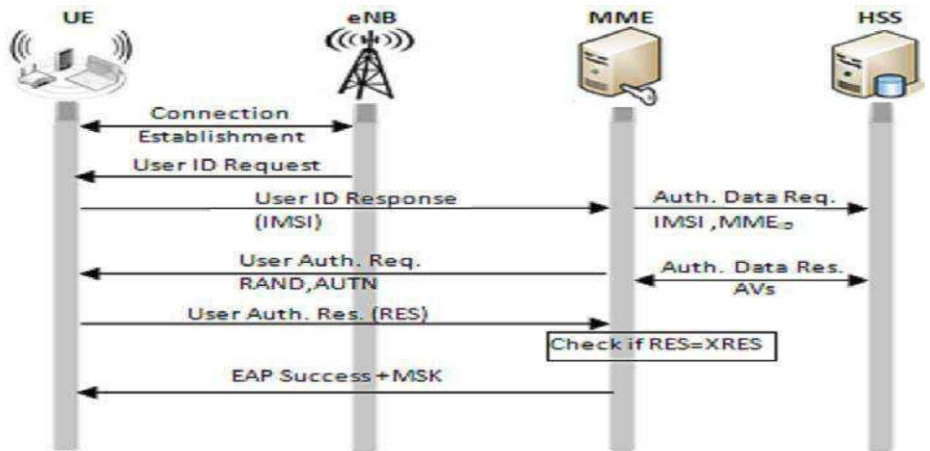


Fig.2. EPS-AKA Authentication protocol [8]

- 4) HSS response to MME by sending AVs which are computed in above step.
- 5) Now MME sends a User Authentication Request which contains RAND, AUTN and KSI_{asme} and it is used to identify $Kasme$.
- 6) On receiving the message UE obtain $SQNH$ and generates the $XMAC$ and for freshness reasons match it with the MAC received in AUTN, then it checks the $SQNH$ received from HSS with its SQN if any of

the two checks fails it sends errors otherwise it computes the response RES and sends it to MME and computes the K_{asme} [8] [10].

7) UE sends RES to MME where it matches it with XRES received from HSS, if both are equal then it sends a success message to UE. EPS-AKA protocol is successfully terminated.

2 Related Works

Many research work are going on to improve the performance of EPS-AKA protocol some of them includes protection from certain type of attacks such as subscriber identity disclosure, Man in the middle attack, Replay, DoS, DDoS, Brute force attack etc. and others include performance enhancement such as minimizing authentication delay, authentication cost, bandwidth and energy consumption. Some of these solutions have been presented in the following papers:

Alezabi, Kamal Ali et. Al [8] proposed an Efficient EPS-AKA (EEPS-AKA) protocol. This protocol is based upon a Simple Password Exponential Key Exchange (SPEKE) protocol; one of the strongest EAP protocol. The proposed protocol improves the SPEKE protocol to protect IMSI and generate stronger keys which is shared among UE, MME and HSS. This method uses symmetric cryptographic method so authentication delay and storage overhead is reduced. This protocol is easier to set up then certificate based authentication methods and therefore is faster than the previous one. The author validates this protocol using Automated Validation of Internet Security Protocol and Applications (AVISPA) tool [9].

Leu, Fang-Yie et. Al [10] analyzed the EPS-AKA protocol and security threats in it. Authors proposed a new authentication model to deal with the security threats of EPS-AKA. The authors presented an improved approached based on Diffe-Hellmen algorithm to secure the RAND which is used for generating a number of keys to secure data or user identity information. The authors also presented a pair key mechanism to generate all parameters i.e. two inputs are required to generate parameters and two dimensional operations which makes this protocol secure from brute force and other attacks. Characteristics of SQN is also enhanced by changing it from sequence number to a random number and now SQN act as a parameter and can be input to various functions to generate other keys for securing data. Parameters previously secured by only one secure key are now protected by more than one key thus provides more safety from various types of attacks.

Abdo, Jacques Bou [1] et. Al paper presented an authentication model which removes the weakness of EPS-AKA inherited from UMTS-AKA. Authors in this paper crypt-analyze Self-Certified Public-key based

Authentication (SPAKA) and Public-Key Broadcast Protocol (PKBP) to solve EPS AKA's privacy and mutual authentication weaknesses and then authors compare this proposed protocol with EPS-AKA and EC-AKA to prove that this proposed protocol is better than other two. Authors use Automated Validation of Internet Security Protocol and Applications (AVISPA) tool [9] to compare protocols which validates that this protocol has less overhead than other two.

Essam Abd El-Wanis [11] et. Al presented a Modified EPS-AKA protocol which uses Symmetric Key Cryptography and SPEKE protocol to solve the vulnerabilities of EPS-AKA protocol by providing stronger mutual authentication between the user and network. Every time a user wants to connect to network a dynamic key is generated and all sent and received messages are encrypted for more security [11]. To verify the results Scyther tool is used along with C Programming. It is more secure but unfortunately its execution time is more than EPS-AKA protocol [11].

3 Conclusions

Despite its security and effectiveness EPS-AKA does not provide full security from various threats like user identity threats, DoS, DDoS, signaling overhead etc. User identity can be revealed by decrypting the IMSI which is sent in clear text in the connection establishment phase. Additionally, forward confidentiality is not fully provided as the primary key can be accessed by getting RAND number sent between UE and HSS. Many solutions have been proposed but none of them is perfect in solving these issues. Some of those solutions use symmetric cryptography which is easy to guess and causes additional overhead. The prevailing security threats can be removed in future by securing IMSI by using public key cryptography method like RSA which prevents the user identity attacks. Several other attacks such as MITM, Replay, DoS and DDoS etc. can be prevented by securing RAND number sent between UE and HSS using RSA algorithm or by using pair key method.

References

1. Abdo, J.B., Demerjian, J., Ahmad, K., Chaouchi, H., Pujolle, G.: EPS mutual authentication and Crypt-analyzing SPAKA 978-1-4673-2088,IEEE (2013)
2. www.technologyreview.com/news/427344/verizon-envisions-4g-wireless-in-just-aboutanything .
3. Aoude, M., Chaouchi, H., Abdo, J.B.: 3rd Generation Partnership Project, 3GPP TS 33.401 V11.2.0 (2011-12), 3GPP System Architecture Evolution (SAE); Security architecture (Release 11),(2012)
4. 3gpp site, <http://www.3gpp.org/>

5. Alezabi, K.A., Hashim, F., Hashim, S.J., Ali, B.M.:G.V6.1.0.:Digital cellular telecommunications system (phase 2+);security aspects (gsm 02.09 version 6.1.0 release) (1997)
6. Alezabi, K.A., Hashim, F., Hashim, S.J., Ali, B.M.: T. G. P. P. (3GPP):3g security; security architecture (release 8), in 3GPP TS 33.102 v8.2.0.(2009)
7. Alezabi, K.A., Hashim, F., Hashim, S.J., Ali, B.M.: T. G. P. P. (3GPP): 3g system architecture evolution (sae); security architecture (release 8), in 3GPP TS 33.401 v8.2.1. (2009)
8. Alezabi, K.A., Hashim, F., Hashim, S.J., Ali, B.M.: An efficient authentication and key agreement protocol for 4G (LTE) networks.Region 10 Symposium, 2014 IEEE. (2014)
9. AVISPA Project, <http://www.avispa-project.org/>
10. Leu, F.Y., You,I., Huang, Y.L., Yim, K., Dai, C.R.: Improving security level of LTE authentication and key agreement procedure. GC'12 Workshop: The 4th IEEE International Workshop on Mobility Management in the Networks of the Future World, 978-1-4673-4941, IEEE (2012)
11. Abdrabou, M.A., Elbayoumy, A.D.E., and Essam Abd EI-Wanis: LTE Authentication Protocol (EPS-AKA) Weaknesses Solution. In: Seventh International Conference on Intelligent Computing and Information Systems (ICICIS'15), IEEE (2015)

A study on Self Healing Functionalities of Self Organizing Networks

Anshita Singh¹, Srishti Shukla², Poonam³

³saini24poonam@gmail.com

Department of Computer Applications, National Institute of Technology,
Kurukshetra, Haryana, India

Abstract. The 3rd Generation Partnership Project (3GPP) initiated successor of UMTS called Long Term Evolution in its 8th release. In its first (8 th) and subsequent releases, it included Self Organizing Network (SON) that promises improvements for future wireless network. Fault Management is an important aspect of SON aiming to automatically diminish the fault by triggering appropriate recovery needed and hence satisfy the operator-specified performance requirements as much as possible. This paper aims to present a Self-healing framework and analyzes the various use-cases and control parameters including Physical Channel Transmit Power, antenna tilt and uplink target received power level P0.

Keywords: Cell Outage Compensation, Cell Outage Detection, Long Term Evolution (LTE), Self-healing, Self-Organizing Network (SON).

1 Introduction

In recent years, the world has experienced tremendous growth in mobile network leading to more complex cellular network which requires huge human effort. Moreover, emerging 5G in few years will increase cost and complexities drastically[1]. These requirements lead to concept of Self Organizing Network (SON).

SON is the key component in LTE network developed by 3rd Generation Partnership Project. It can be defined as the concept of continuously monitoring and making intelligent moves to reduce undesired results[2]. Automation, configuration and optimization processes done automatically instead going for manual work in SON can help to reduce OPEX and CAPEX[1].

There are 3 architectures according to the residing of SON algorithm-Centralized, Distributed and Hybrid. SON functionalities are broadly

categorized in 3 types: Self Configuration, Self Optimization and Self Healing.

Self Configuration reduces the amount of human intervention in installation of eNBs(base stations) by providing plug-and-play function. It provide various features such as Automatic Neighbor Relation Configuration, Automatic Software management and Self Test. Self Optimization functions such as Mobility load balancing (MLB) and RACH optimization[3] adapts the network according to the changing environment to improve the network efficiency. Self Healing function does troubleshooting tasks of the performance failures that affect network without any human interference.

2 Self Healing

Self healing is the most critical aspect of SON that deals with automatic detection, diagnosis and correction of the defects occurred. There are four main stages [4] under this-

∑ Cell Outage Detection- It may be generated by any software or hardware failure or even external failures. Software failures may include channel processing error or radio board failure etc. Power failure, connectivity issues or even misconfiguration are other concerns.

∑ Cell Outage Compensation- Various corrective steps are taken to compensate the outage occurred like changing the neighboring operational cell's parameters.

∑ Cell Outage Recovery- If the changed environmental setup recovers the outage; the things go back-on-track. Else, roll-back to their initial settings.

3 Cell Outage Detection

Cell Detection and management forms the major part of Self Healing Framework[5]. Several Key Performance Indicators (KPIs) can be used for detection purposes. To check KPI, neighboring cell communicates with the Network Management System (NMS) through S1 interface. Cell Outage is announced if the observed values are less than threshold values which is then tried to resolve using some Compensation algorithm. The threshold values used are determined by the network operators. There can be other indications too. Decreased cell efficiency, handover abnormality and forced call drops are some of them.

A novel cell outage detection algorithm that is based on the neighbor cell list reporting of mobile terminals[6]. A degraded cell still works but less-

efficiently carrying much lesser traffic that lowers feasibility. As the result, cell is no more visible to neighboring eNBs or user but from network point of view, it only appears empty and still operational. Using statistical classification techniques and already available measurement data heuristic, the algorithm is able to detect most of the outage situations in less time.

3.1 Cell Outage Detection Algorithm

Step 1: Compute long term throughput value and build Key Performance Indicator (KPI) profile in OAM (Operation and Maintenance) Centre individually.

Step 2: INPUT: Various measurements from base station and OAM Centre.

Step 3: OAM keep monitoring e-NODEs and keep exchanging the information and other measurable parameters such as Radio Link Failure (RLF) etc.

Step 4: Compare current value of KPI of enodeB with profile built in OAM

IF(any deviation from profile is encountered) Trigger alarm, declared cell outaged and start cell outage compensation

ELSE

no action required

ENDIF

4 Cell Outage Compensation

Once the cell outage is acknowledged, cell outage compensation algorithms are used to adjust the parameters of the neighboring cells and to meet the operator's performance requirements. These algorithms are typically iterative processes of parameter alteration and evaluation.

Antenna tilt plays an eminent role in cell compensation. By adjusting antenna tilt with respect to its axis, elucidation is directed further down reducing coverage in more remote location and concentrating energy to the direction of out-aged cell[7]. . This paper presents a heuristic approach for autonomous re-optimization of antenna of the nodes having mutual inference with tilt vector Θ_t and cluster C_j of sector j_t . This increases system's efficiency by 10% and average cell edge efficiency by 5% quantile. This algorithm can be implemented to both centralized and distributed architecture that improves efficiency even up to 100%.

Thus applying antenna tilt changes signal propagation that is determined by the type of electrical and mechanical tilt. Typically, range for tilt angles of

these antennas vary from 0 to 12 based on the vertical beam width and this selection is depended on antenna and site configuration.

The downtilt angle is the balance between other cell's interference reduction and coverage threshold. An optimum downtilt depends on the few factors in which the geometrical factor (Θ_{geo}) has the most significance expressed in [6] as :-

$$\Theta_{geo} = \tan^{-1} \left[\frac{h_{bs} - h_{ms}}{d} \right] \quad (1)$$

where h_{bs} stand for height of base station; h_{ms} stands for height of mobile station antenna and d stands for sector dominance area.

Beside the antenna tilt which is the dominant control parameter in cell compensation, there are other control parameters too that play significant role.

Physical Channel Transmit Power is another control parameter that is used to overcome cell outage [7]. Neighboring cells can extend their service area by increasing their respective Physical Channel Transmit Power. Some neighboring cells also decrease the same to reduce their service area so that interference is diminished.

PUSCH (Physical Uplink Shared Channel) is yet another parameter used in compensation algorithms. Uplink transmit power can be described from target received Power density (P0) [9]. P0 is selected as the adjustment parameter because it provides efficient trade-off between coverage and quality for different scenarios. Increased P0 intensifies the coverage probability in case of outage. On other hand, P0 allows more remote terminals to connect to a given base station by lowering inter-cell interference level leading to better throughput.

A compensation algorithm is based on HandOver(HO) margin modifications, including the faulty cell and its neighbors[9]. It uses HO which has been earlier used as a part of cell optimization for load balancing whenever congestion occurs.

4.1 Cell Outage Compensation Algorithm

Step I: Initialize

Step II: INPUT:

List of all outaged cell and the control parameters Step III: DATA STRUCTURE:

Make list of cell to be optimized using neighboring cell
Step IV: do

Select any cell from list and change parameter of that neighboring cell to reduce outaged area IF(outaged cell is not compensated)

WHILE(new parameter are lesser than threshold values) set the parameters again to the new values
IF(cell is not compensated)
IF(new parameters exceed threshold limit)

rollback to initial state goto LABEL1

ELSE

goto LABEL2 ENDIF
ELSE
goto LABEL1

ENDIF

LABEL2: set the parameters again to the new values ENDLOOP
ENDIF

LABEL1: selected next cell WHILE(until are cell are optimized)
Step V:IF(and cell is not optimized) cell is to compensated manually
ENDIF

5 Conclusion

This paper presents SON's healing functionalities and has summarized how different researches used different control parameters for developing their compensation algorithms. An automatic adjustment of antenna tilt [8] is used in the case of outage that substantially reduce network operating cost so the little communication is needed between nodes. To enhance coverage, the split between the reference signal(RS) and physical downlink shared channel(PDSCH) can be adjusted that raises the RS power(PRS)[7], at the cost of reduced PDSCH power and hence it reduce traffic handling capacity. Also, an adjustment combining all these parameters can be used to generate a new algorithm that may get better results enhancing the performance of the system.

6 Future Approach

For outage detection, false alarm minimization is the big issue; future works will evaluate these false alarms by differentiating them with the actual ones. All the different control parameters discussed in this paper for cell

compensation motivates further work, generating a new algorithm by collaborating and tuning all such parameters and simulating it on different scenarios.

References

1. Soldani, D., and Manzalini, A.,: Horizon 2020 and Beyond- On the 5G Operating System for a True Digital Society. In: IEEE Vehicular Technology Magazine, vol. 10, no. 1, pp. 32--42 (2015).
2. 3GPP TS 32.521.: Self-Optimization OAM- Concepts and Requirements (2009).
3. Marwangi, M. M. S., Faisal, N., Yusof, S.K.S., Rashid, R. A., Ghafar, A. S. A., Saparudin, F. A., and Katiran, N.,: Challenges and practical implementation of self-organizing networks in LTE/LTE-Advanced systems. In: Proceedings of the 5th International Conference on Information Technology & Multimedia, pp. 1--5, Kuala Lumpur (2011).
4. Asghar, M. Z., Hämäläinen, S., and Ristaniemi, T.,: Self-healing framework for LTE networks. In: 2012 IEEE 17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 159--161, Barcelona (2012).
5. Bramah, E. H. B., and Ali, H. A.,: Self healing in Long Term Evolution (LTE) (Development of a cell outage compensation algorithm). In: 2016 Conference of Basic Sciences and Engineering Studies (SGCAC), pp. 75--79, Khartoum(2016).
6. Niemelä, J., Isotalo, T., and Lempiäinen, J.,: Optimum Antenna Downtilt Angles for Macrocellular WCDMA Network. In: EURASIP Journal on Wireless Communications and Networking Vol. 5, pp. 816--827 (2005).
7. Amirijoo, M., Jorguseski, L., Litjens, R., and Nascimento, R.,: Effectiveness of cell outage compensation in LTE networks. In: 2011 IEEE Consumer Communications and Networking Conference (CCNC), pp. 642--647, Las Vegas, NV (2011).
8. Eckhardt, H., Klein, S., and Gruber, M.,: Vertical Antenna Tilt Optimization for LTE Base Stations. In: 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1--5, Yokohama (2011).
9. de-la-Bandera,I., Barco, R., Muñoz, P., Gómez-Andrades,A., Serrano,I.,: Fault compensation algorithm based on handover margins in LTE networks. In: EURASIP Journal on Wireless Communications and Networking (2016).