Online ISSN: 0975-4172
Print ISSN: 0975-4350

Global Journal of Computer Science and Technology

DISCOVERING THOUGHTS AND INVENTING FUTURE

Technology Reforming Ideas

December 2011

Pinnacles

- Association of Data Mining
- Machine Learning Algorithm
- Group-Testing approach
- Magic Square Intrication

The Volume 11
Issue 21
VERSION 1.0

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 11 Issue 21 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

(A Delaware USA Incorporation with "Good Standing"; Reg. Number: 0423089)

Sponsors:Global Association of Research

Open Scientific Standards

Publisher's Headquarters office

Global Journals Inc., Headquarters Corporate Office, Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141 United States
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Association of Research, Marsh Road, Rainham, Essex, London RM13 8EU United Kingdom.

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investers@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Editorial Board Members (HON.)

Dr. Bart Lambrecht
Director of Research in Accounting and FinanceProfessor of Finance
Lancaster University Management School
BA (Antwerp); MPhil, MA, PhD (Cambridge)

Dr. Carlos García Pont
Associate Professor of Marketing
IESE Business School, University of Navarra
Doctor of Philosophy (Management), Massachusetts Institute of Technology (MIT)
Master in Business Administration, IESE, University of Navarra
Degree in Industrial Engineering, Universitat Politècnica de Catalunya

Dr. Fotini Labropulu
Mathematics - Luther College
University of ReginaPh.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windso

Dr. Lynn Lim
Reader in Business and Marketing
Roehampton University, London
BCom, PGDip, MBA (Distinction), PhD, FHEA

Dr. Mihaly Mezei
ASSOCIATE PROFESSOR
Department of Structural and Chemical Biology, Mount Sinai School of Medical Center
Ph.D., Etvs Lornd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram
Department of Accounting and FinanceLancaster University Management SchoolPh.D. (WHU Koblenz)
MBA/BBA (University of Saarbrücken)

Dr. Miguel Angel Ariño
Professor of Decision Sciences
IESE Business School
Barcelona, Spain (Universidad de Navarra)
CEIBS (China Europe International Business School).
Beijing, Shanghai and Shenzhen
Ph.D. in Mathematics
University of Barcelona
BA in Mathematics (Licenciatura)
University of Barcelona

Philip G. Moscoso
Technology and Operations Management
IESE Business School, University of Navarra
Ph.D in Industrial Engineering and Management, ETH Zurich
M.Sc. in Chemical Engineering, ETH Zurich

Dr. Sanjay Dixit, M.D.
Director, EP Laboratories, Philadelphia VA Medical Center
Cardiovascular Medicine - Cardiac Arrhythmia
Univ of Penn School of Medicine

Dr. Han-Xiang Deng
MD., Ph.D
Associate Professor and Research
Department Division of Neuromuscular Medicine
Davee Department of Neurology and Clinical NeuroscienceNorthwestern University
Feinberg School of Medicine

Dr. Pina C. Sanelli
Associate Professor of Public Health
Weill Cornell Medical College
Associate Attending Radiologist
NewYork-Presbyterian Hospital
MRI, MRA, CT, and CTA
Neuroradiology and Diagnostic Radiology
M.D., State University of New York at Buffalo,School of Medicine and Biomedical Sciences

Dr. Roberto Sanchez
Associate Professor
Department of Structural and Chemical Biology
Mount Sinai School of Medicine
Ph.D., The Rockefeller University

Dr. Wen-Yih Sun
Professor of Earth and Atmospheric SciencesPurdue University Director
National Center for Typhoon and Flooding Research, Taiwan
University Chair Professor
Department of Atmospheric Sciences, National Central University, Chung-Li, TaiwanUniversity Chair Professor
Institute of Environmental Engineering, National Chiao Tung University, Hsin-chu, Taiwan.Ph.D., MS The University of Chicago, Geophysical Sciences
BS National Taiwan University, Atmospheric Sciences
Associate Professor of Radiology

Dr. Michael R. Rudnick
M.D., FACP
Associate Professor of Medicine
Chief, Renal Electrolyte and Hypertension Division (PMC)
Penn Medicine, University of Pennsylvania
Presbyterian Medical Center, Philadelphia
Nephrology and Internal Medicine
Certified by the American Board of Internal Medicine

Dr. Bassey Benjamin Esu
B.Sc. Marketing; MBA Marketing; Ph.D Marketing
Lecturer, Department of Marketing, University of Calabar
Tourism Consultant, Cross River State Tourism Development Department
Co-ordinator , Sustainable Tourism Initiative, Calabar, Nigeria

Dr. Aziz M. Barbar, Ph.D.
IEEE Senior Member
Chairperson, Department of Computer Science
AUST - American University of Science & Technology
Alfred Naccash Avenue – Ashrafieh

Contents of the Volume

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Association of Data Mining and Healthcare Domain: Issues and Current State of the Art

By Fawzi Elias Bekri, Dr. A. Govardhan

Jigjiga University, Jigjiga, Ethiopia

Abstract - Data mining has been used prosperously in the favorably perceived areas such as e-business, marketing and retail because of which it is now applicable in knowledge discovery in databases (KDD) in many industrial areas and economy. Data mining is mainly gaining its importance and usage in the areas of medicine and public health. In this paper the investigation of present methods of KDD, applying data mining methods for healthcare and public health has been discussed. The problems and difficulties related to data mining and healthcare in practice are also mentioned. In survey, the use of data mining has increased, along with examination of healthcare institutions so that the health policy prepared is the best, perceive disease causes and protect deaths in hospital and discover the dishonest insurance declaration.

Keywords : *health problems; recognition; data mining; elderly; motion capture; mining methods and algorithms; medicine and science.*

GJCST Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Association of Data Mining and Healthcare Domain: Issues and Current State of the Art

Fawzi Elias Bekri[α] , Dr. A. Govardhan[Ω]

Abstract - Data mining has been used prosperously in the favorably perceived areas such as e-business, marketing and retail because of which it is now applicable in knowledge discovery in databases (KDD) in many industrial areas and economy. Data mining is mainly gaining its importance and usage in the areas of medicine and public health. In this paper the investigation of present methods of KDD, applying data mining methods for healthcare and public health has been discussed. The problems and difficulties related to data mining and healthcare in practice are also mentioned. In survey, the use of data mining has increased, along with examination of healthcare institutions so that the health policy prepared is the best, perceive disease causes and protect deaths in hospital and discover the dishonest insurance declaration.

Keywords : health problems; recognition; data mining; elderly; motion capture; mining methods and algorithms; medicine and science.

I. Introduction

Data mining has been used prosperously in the favorably perceived areas such as e-business, marketing and retail because of which it is now applicable in knowledge discovery in databases (KDD) in many industrial areas and economy. Data mining is mainly used in the field of healthcare.

The rest of the paper organized as fallow. Section II explores the frequently quoted research work in recent literature that reveals the association of data mining and healthcare sector. The section III illustrates an example that concludes the impact of data mining in healthcare sector. Section IV explores the taxonomy of Healthcare issues that demands the data mining as critical requirement. Section V reveals the reckoned obstacles in Knowledge discovery from healthcare databases. In Section VI we discuss the current state of the art in "Data mining in healthcare sector". Section VII explores the conclusion of this paper that fallowed by references.

II. Data Mining In Health Sector

From past decades there occurs the use of definite information and proofs that encourage medical judgment (evidence based medicine or EBM). The father of modern epidemiology, John Snow utilized 1854 maps which consisted bar graphs to find out the flow of cholera to show that it passes through the water that flows beneath [15]. Snow numbered the deaths occurred and then as black bars he marked the sufferers address on the map. He noticed that most of the deaths occurred were surrounded around particular water well in London.

In 1855 polar-area diagrams was introduced by Florence Nightingale to explain that the deaths in army could be reduced to some extent by using hygiene clinical methods. To decrease the death rate, she utilized the diagrams so that the policy-makers can utilize it in their application reforms [1] [9].

During the occurrence of this problem Snow and Nightingale on their own they gathered the information, organized it and then examined the data since it was controllable. Though in present scenario the population is large, gadgets are used to analyze the victims of any disease but still they cannot succeed the result obtained by the investigator in the past. Ere data mining is considered useful as it is helpful in solving different issues that occur while gaining information related to the field of healthcare.

In recent studies the most considered topic is data mining and its uses in the field of medicine and public health. In 2003, Wilson et al studied all the past researches where KDD and data mining were used in the area of healthcare, which he felt to be confusing. Data mining was utilized by few authors for gaining information and others use them in statistical methods within the directory knowledge process [17].

Since the data mining definition is misunderstood in the medical field so the most preferred definition of data mining is a cluster of processes and methods for determining and illustrating models and developments in information [18].

III. Impact of Data Mining in Health Sector

The event that happened at Rizal Medical Center in Pasig City of Philippines in October 2006 can be described to realize the contact of data mining in the area of healthcare. Due to lack of proper hygiene and cleanliness the hospital faced problem of deaths of new-born children because of neonatal sepsis (bacterial infection). Till the increase in deaths no one noticed about the cause and then hospital data was analyzed

Author[α] : Department of Computer Science & Engineering Jigjiga University, Jigjiga, Ethiopia. E-mail : fozbekri@gmail.com

Author[Ω] : Professor of Computer Science & Engineering Principal JNTUH of Engineering College, Jagityal, Karimnagar (Dt), A.P. India . E-mail : govardhan_cse@yahoo.co.in

where the Department of Health (DOH) originated that because of sepsis for instance, 12 out of 28 children born on October 4, died [12]. The DOH could find out the reason behind the cause and restrict them before things went out of control with the proper utilization of data mining to the past records.

IV. The Importance and Uses of Data Mining in Medicine and Public Health

The demand and want for data mining is more in field of healthcare, regardless of variations and conflicts in processes. Various discussions led to the demand of data mining in the field of healthcare which includes both public health as well private health. Many facts can be achieved from the past data stored in computers. Since the data is huge in quantity so it's a disadvantage for persons to examine the entire data and gain awareness [5]. Specialists consider that the improvement in medicals has reduced leading to complication of recent medical data. To overcome this drawback computers and data mining can be utilized.

Evidence-based medicine and prevention of hospital errors: On utilizing data mining on the available data much new informative and possibly life-rescuing information is achieved or else which would have left unutilized. For example, in recent research on hospitals and wellbeing it was originated that almost 87% of the deaths in the United States could have been reduced if the errors would have been lowered by the hospital staff [6].The preventive measurements could have been adopted by the hospitals and government supervisors using data mining to hospital data.

Policy-making in public health : To examine the relatedness among society health centers in Slovenia, Lavrac et al. [19] merged GIS and data mining via Weka with J48. The utilization of data mining in healthcare data helped health centers to determine methods that would lead to policy suggestions to the Public Health Institute. Decision making can be improved by proper utilization of data mining and decision support techniques.

More value for money and cost savings : Extra information can be obtained at less additional price by the firms and institutions using data mining. To know the information about scam in credit cards and insurance claims KDD and data mining is utilized [17].

Early detection and/or prevention of diseases : To obtain the before time recognition of heart problem which is the main public health issue through the world, Cheng et al mentioned the application of arrangement strategy.

Early detection and management of pandemic diseases and public health policy formulation: For before time identification and supervision of pandemics health specialists have preferred to utilize data mining. To obtain the reasons behind occurrence of diseases Kellog et al [3] discussed various methods which is a mixture of spatial modeling, simulation and spatial data mining. The output of examined data mining in the simulated situation can be utilized further to discover and organize disease causes.

V. The Reckoned Obstacles to Apply Data Mining in Medicine and Public Health Sector

Due to the unconventional behavior of the medical work utilization of data mining in the area of medicine is a difficult task. Different natural arguments among the traditional styles of data mining strategies and medicine were seen in the work of Shillabeer et al[11]. Data mining in the field of medical research begins with an assumption and then the outputs are altered so that the assumptions are achieved, which is different in case of standard data mining where data set begins without the assumptions in the beginning. Since traditional data mining is all worried about the development and models in data groups, data mining is more attracted in the alternative that do not match to the development and models. The main difference in methods is that most standard data mining neglects clarifying development and models and mainly focuses on relating. In the field of medicine explanation is a must for the reason that a distinction can lead to variation in the life or death of a human.

Let us consider an example of anthrax and influenza contributes to similar signs of respiratory troubles. When flu spreads the data mining test results in anthrax problem. The critical situation arises when the supposed flu problems is really anthrax epidemic [16]. In all the researches of data mining on disease and cure it is observed that the results were unclear and alerts, which included promoted outcomes but resulting in additional revise. This drawback points towards the present lack of credibility of data mining in the field of healthcare.

The definition of data mining is like a puzzle which results in further problems. The use of data mining in few researches has just explained the use of graphs which is not the correct meaning. According to Shillabeer [10] the confusion of data mining with inappropriate definition is widespread in the healthcare area.

Though the outcomes of data mining are useful the main disadvantage is to influence the health performers to modify their practices. Ayres [2] shows few incidents where the hospital staff when provided with proofs rejected to modify their hospital procedures. In few areas it was noticed that doctors ignored cleansing their hands after autopsy and cured patients with the same hands which lead to an increase in deaths of the patients.

As per Shillabeer [10] research he concluded that the hospital staff desires to accept the view of the head of the medical institution instead of the data mining outcome. His conclusion is suitable because in the medical institutions management's opinion is considered to be the best. One of the major disadvantages of data mining in healthcare is that the data of patients cannot be used for ethical and personal purpose. The amount of data should be less so that exact results can be obtained in case of data mining. The past data consists of private information which helps in knowing the disease and deaths can be prevented.

VI. Current State of The Art

WSARE was proposed by Wong et al [16], it is an approach to find out the before time cure for diseases. WSARE stands for "What's Strange About Recent Events" depends on the organizations policies and Bayesian methods. On simulation models WSARE is utilized which resulted in the exact guess for cure of simulated diseases. Before utilization of WSARE in actual life situations, safety measures must be considered.

Non-invasive diagnosis and decision support[13]: Every patient cannot afford for expensive, persistent and aching diagnostic and laboratory methods. For instance to discover cervical cancer using biopsy in women is a difficult task. K-means gathering algorithm was utilized by Thangavel et al [13] to detect the cervical cancer in women and he originated that the results of gathering data is more accurate then those of medical results. They even originated few factors which can help doctor in taking decision whether the patients suffering from cervical cancer must be suggested with biopsy or not.

Gorunescu et al[20] explained the use of data mining in the improvement of computer-aided diagnosis (CAD) and endoscopic ultrasonographic elastography (EUSE) to generate fresh non-invasive cancer identification. Ultrasound video was utilized by doctors to choose whether a patient must be suggested with biopsy or not in the traditional method.

Depending on the analysis of ultrasound video, the doctor's decision is biased. By utilizing data mining Gorunescu tried to solve this issue in other manner, he and his panel members paid attention on ultrasound videos rather than demographics. In the cases of malignant and benign tumors an organization algorithm was trained using a multi-layer perception (MLP). To differentiate between malignant and benign tumors, the pixels and the content of RGB was examined by the model. The outcome of the model was then utilized in other cases and it was noticed that the models output was appropriate and exact with very few variations in diagnosis.

Adverse drug events (ADEs): Few medicines and vaccines which were declared to be safe on humans are now realized to be injurious to humans when used for long periods. According to Wilson et al [17] he disclosed that data mining is used to investigate the drugs side effects in their data by US Food and Drug Administration. Around 67% of ADEs was discovered five years ago by Multi-item Gamma Poisson Shrinker or MGPS.

Medically Driven Data Mining Application : Recognition of Health Problems from Gait Patterns of Elderly: Bogdan Pogorelc et al [21] discovered a medically driven data mining application system for analyzing of walk models associated to the health problems of old people so that their independent living can be sustained.

The data gathered in this model is done using body antenna and RFID labels. Using motion capture gadgets the walking style of old people was captured, which included labels affixed to their body and antennas fixed in the building. To identify a particular health issue, a labels location is achieved by the antenna and then the responding time series of location directs are examined. To categorize the walking style of old people certain characteristics for training decision tree classifier and KNN classifier were introduced by the authors into:

a. normal
b. with hemiplegia,
c. with parkinson's disease,
d. with pain in the back and
e. with pain in the leg.

Bogdan Pogorelc et al [21] designed an automatic health-state identification, he introduced and examined 13 characteristics that were supported on the 12 labels which were affixed on the shoulders, elbows, wrists, hips, knees, and ankles of the old people. The introduced characteristics that are applied for modeling utilizing the machine learning processes are as shown:

i. Total variation among a) average space among right elbow and right hip and b) average space among right wrist and left hip.
ii. Right elbows normal angle.
iii. Proportion among maximum angle of the left knee and the right knee.
iv. Variation among the maximum and minimum angle of right knee.
v. Variation among a) maximum and minimum height of the left shoulder and b) maximum and minimum height of the right shoulder.
vi. Proportion among a) variation among maximum and minimum height of left ankle and b) maximum and minimum height of right ankle.
vii. Total variation among a) maximum and minimum speed of the left ankle and b) maximum and minimum speed of the right ankle.
viii. Total variation among a) average space among right shoulder and right elbow and b) average space among left shoulder and right wrist.
ix. Average speed of the right wrist.

x. Average angle among a) vector among right shoulder and right hip and b) vector among right shoulder and right wrist.
xi. Regularity of angle of the right elbow passing average angle of the right elbow.
xii. Variation among average height of the right shoulder and average height of the left shoulder.

By applying decision tree classifier and k-nearest neighbor classifier the tests were conducted. The main aim of test was to examine the categorization exactness of models, built applying the machine learning processes. By utilizing stratified 10-fold cross justification the tests appropriateness were achieved. The information for decision tree classifier was obtained from 7 labels and 5mm standard is the variation of sound. The information for KNN classifier was obtained from 8 labels and 0-20mm standard range is the variation of sound. Bogdan Pogorelc et al [21] reported that KNN results are exact when compared to decision tree results. The decision tree obtained 95% of exactness where as KNN obtained more than 99% of exactness.

Observation : In this paper the examining of the walking style of old people is done in connection to the health related issues so that they are maintained with their independent living. At the beginning stage (no sound, all labels) it was found that the decision tree obtained 90.1% exactness and k-nearest neighbor obtained 100% exactness. The elder people were provided by protection and assurance which resulted in less needless ambulance prices. Based on the results we conclude that k-nearest neighbor has achieved exactness of 99% with only 8labels and sound of 0-20mm standard when compared to decision tree which has achieved only 95% of exactness.

The implication of the characteristics applied to model the machine learning is not calculated and the elements picked to instruct the classifier is also not acceptable. So from the results we can say that the researches were just conducted to evaluate the presentation of decision tree and KNN tree considering the training constraints and opted elements. The study must validate the implication of the training characteristics and elements linked to healthcare field. Data mining is utilized in the area of before time identification of diseases, rescuing patients from deaths, enhancement of diagnosis and identification of dishonest health declarations. Data mining can be utilized in healthcare with few warnings.

Signaling Potential Adverse Drug Reactions from Administrative Health Databases: Huidong Jin et al [22] introduced an ADR indicating method that indicates sudden and irregular models feature of ADRs. To carry on the method Huidong Jin et al[22] discussed that all the present post market ADR indicating methods depend on unplanned ADR case results, which undergo grave underreporting and latency data. Because of ADRs there is an increase in hospitalization and deaths universally. On the other hand the administrative health data is gathered universally and regularly. This method consists of a domain-driven facts illustration Unexpected Temporal Association Rule (UTAR), its attractiveness measure, unexlev, and a mining strategy MUTARA (Mining UTARs given the Antecedent). It also proposed HUNT to emphasize the irregular and sudden models by evaluating their grades based on unexlev with those based on established influence.

Observation : Huidong Jin et al[22] proposed two interestingness measures, unexlev and rankratio, in the circumstance of indicating irregular and sudden models features of ADRs from organizational health information. He also introduced two easy but successful mining strategies MUTARA and HUNT to detect pair wise UTARs from the connected organizational health information QLDS. On evaluation to MUTARA the HUNT indicated a small number of strange ADR models.

Predictive Data Mining to Learn Health Vitals of a Resident in a smart Home: Vikramaditya Jakkula[23] mentioned about the practice of analytical data mining to discover health importance of a person staying in a smart home. He introduced a process where tools of smart home are discovering and acquiring their guess capabilities in the course of adjusting to smart home tenants.

During the method of analyzing with introduced model the gathering of data was done using a set of array of motion antennas that gathers information with the help of Argus antenna network. The gathered information is then improved with significant health information gathered by applying digital gadgets. The information gathering process continued for one hundred and fifty days period, which was done on a single tenant of the building.

The information obtained from the motion antennas were applied directly as the information can be applied directly. Data samples can find in the fig 1.

WEKA [24] conducted the tests. The information of 150days was divided and grouped into training and testing data. Among the two tests carried out, the former test forecast health importance sign values.

The first test aimed on forecast examinations. The forecast enhancement is reliant of the classifier to educate. The presentations of different classifiers were analyzed beside the time series health information gathered from the occupants in smart home.

The test conducted concluded that KNN performed very well when compared to other practices like SMO regression, LazyL WL and Multi Layer observation.

The second test aimed on forecasting whether the next given time structure is odd or not?.. in the first test it was concluded that KNN is the best with 51% exactness and in the second test it is concluded that KNN is doing well in forecasting with 85% exactness.

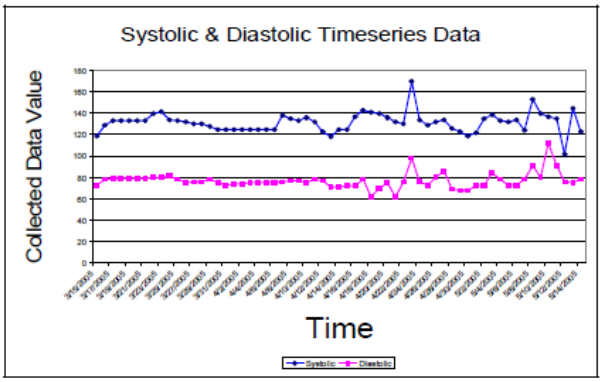
Raw Sensor Data

Timestamp	Sensor State	Sensor ID
3/3/2003 11:18:00 AM	OFF	E16
3/3/2003 11:23:00 AM	ON	G12
3/3/2003 11:23:00 AM	ON	G11
3/3/2003 11:24:00 AM	OFF	G12

Raw Health Data

Timestamp,weight,temperature,Systolic,Diastolic,Pulse
2005-10-17 22:02:38,168,98.6,130,80,82
2005-10-18 13:08:36,168,98.3,124,77,89
2005-10-19 12:41:36,168,97.6,127,78,75
2005-10-20 01:18:00,168,97.6,129,78,74

(a) Sensor readings as collected in a smart home.



(b) Systolic and Diastolic time series data plot.

Figure 1 : Data samples collected from sensors.

Observation : The tests indicate the prospective of data mining utilization in smart home study. In this process various learning strategies were evaluated and the conclusion was KNN performed well with 51% exactness for forecasting important health sign values and also has 85% exactness of forecasting of odd periods. Main emphasis was laid on forecasting the exactness of categories and the drawback is the validation of constraints opted as information features.

Patient Histories derived from Electronic Health Records: Jeremy Rogers et al[26] proposed data mining model called CLEF Chronicle to derive Patient Histories from Electronic Health Records, which is a representation of how a patient's illness and treatments unfold through time. Its primary goal is efficient querying of aggregated patient data for clinical research, but it also supports summarization of individual patients and resolution of co-references amongst clinical documents.

Properties of CLEF Chronicle: The CLEF Chronicle for an individual patient seeks to represent their clinical story entirely as a network of typed instances and their interrelations. Figure 1 illustrates the general flavor of what we are trying to represent: a patient detects a painful mass in their breast, as a result of which a clinic appointment occurs, where drug treatment for the pain is arranged and also a biopsy of the (same) mass in the (same) breast. The first clinic arranges a follow-up appointment, to review the (same) biopsy, which finds cancer in the (same) mass. This (same) finding is in turn the indication for radiotherapy to the (same) breast.

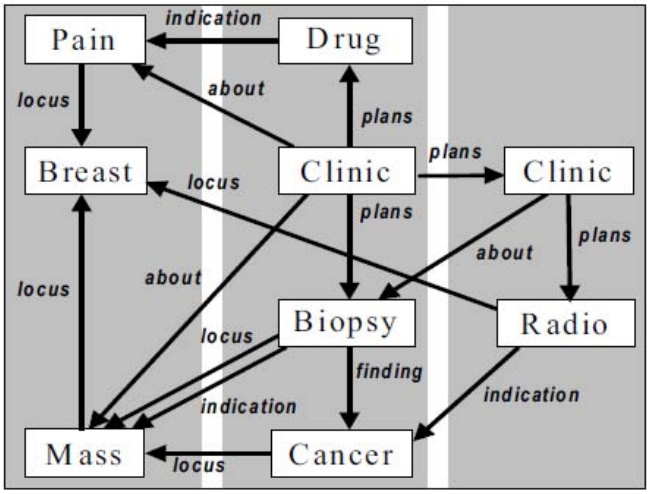


Figure 1[26]: Informal View of Patient-History Fragment
(NOTE: Time-flow is roughly left ⍰ right)

In addition to the obvious structural difference between this representation and that of traditional electronic records, any clinical content represented as a CLEF Chronicle should have two central properties:
Parsimony – traditional patient records contain multiple discrete mentions of relevant instances in the real world (the tumor, the breast etc). A CLEF Chronicle should have only one occurrence of each.

Explicitness – traditional patient records may only imply clinically important information (e.g. the fact of relapse). This must be explicit within a CLEF Chronicle Representation.

Functions of CLEF Chronicle :

- The CLEF Chronicle is intended to support more detailed, and more expressive, querying of aggregations of patient stories than is currently possible whilst at the same time improving the efficiency of complex queries. More detailed, because the Chronicle is more explicit: we can now ask e.g. how many patients 'relapsed' within a set period of treatment. More expressive, because the typing information associated with each Chronicle instance is drawn from a rich clinical ontology, such that queries may be framed in terms of arbitrarily abstract concepts: we can ask how many cancers of the lower limb were recorded, and expect to retrieve those of all parts of the lower limb. This is more efficient, because the traditional organization of patient records tends to require much serial or nested processing of records.
- An individual Chronicle can serve as an important knowledge resource during its own reconstruction from available electronic sources of traditional clinical records. In particular, a Chronicle can help resolve the frequent co-references and repeated

references to real-world instances such as characterize traditional records. For example, heuristic and other knowledge linked to an ontology of Chronicle data types can be used to reject any request to instantiate more than one identifier for a [Brain], or any attempt to merge two mentions of [Pregnancy] separated by more than 10 months.

- The Chronicle is intended to serve as a knowledge resource from which summarizing information or abstractions may be inferred. For example, deducing 'anaemia' from a run of discrete low blood counts obtained from the traditional record, or 'remission' from several years of clinical inactivity and 'relapse' when this is followed by a flurry of tests and a new course of chemotherapy. Similarly, where the record does not explicitly say why drug X was given, reasoners browsing a Chronicle may identify condition Y because the drug has no other plausible context of use.
- The CLEF Chronicle is intended to support automatic summarization of patient records. Given the sometimes chaotic nature of real patient records, manual case summarization is recognized as good clinical practice. Manual derivation of such summaries from the content of the record is, however, notoriously time consuming whilst the result of such labours is notoriously out of date whenever it would be most clinically valuable.

Observation: The idea of representing clinical information as some form of semantic net, particularly focusing on why things were done, is not new: echoes of it can be found in Weed's work on the problem oriented record [27]. Ceusters and Smith more recently advocated the resolution of co-references in clinical records to instance unique identifiers (IUIs) [28]. The semantic web initiative offers new possibilities for implementing such an approach, but the lack of any suitable clinical data severely constrains any practical experimentation. The model CLEF discussed here effective to provide a useful means to explore some of the computational and representational issues that arise.

Detecting Non-compliant Consumers in Spatio-Temporal Health Data [25]: K.S. Ng et al[25] attempt to describe their experience with applying data mining techniques to the problem of fraud detection in spatio-temporal health data in Medicare Australia. A modular framework that brings together disparate data mining techniques was adopted by authors. Several generally applicable techniques for extracting features from spatial and temporal data are also discussed. The system was evaluated with input from domain experts and observed high hit rates. Finally they concluded some conventions drawn from the experience.

Experimental Objectives that the authors considered was

i. Is the characterization of prescription shoppers given is accurate?
ii. Can Consumer RAS be used to identify prescription shoppers that do not rigidly fit the strong criteria. The false identification of genuinely ill patients that exhibit certain characteristics of prescription shopping as prescription shoppers have in the past been an issue for Medicare Australia. Can Consumer RAS avoid making such errors?

To conduct the experimental study, the data was picked randomly from a populous postcode in a major capital city of Australia in which they try to identify possible fraudulent activities.

The software that was used is LOF implementation in the dprep package in R for our analysis. The modified Huff model and the temporal feature extraction scheme was implemented in-house using C++.

Experiment I The first experiment seeks to verify the accuracy of our quantitative characterization of prescription shoppers. Only 12 people in the chosen postcode satisfy our criteria. These are passed on to domain experts for evaluation. We are interested in the percentage of these people that are true prescription shoppers.

Experiment II The second experiment seeks to verify whether it is feasible to use outlier detection technique to identify, with low false positive rate, prescription shoppers that do not fit the criteria identified. To do that, we remove all consumers identified in Experiment I from the data, perform a LOF analysis on the rest, and then pick out consumers that have high volumes of drugs of concern for evaluation by the experts. Fourteen consumers were picked this way. Some of these consumers have genuine needs for their drugs. In the aim to know, whether prescription shoppers tend to exhibit higher LOF scores compared to patients with genuine needs? In other words, do prescription shoppers show up as statistical outliers in the data? We excluded older consumers in this experiment because they are more likely to have genuine medical conditions and we do not want to spend time analyzing such patients. For the LOF analysis, the data are normalized and we compute the minimum LOF value with $k \in [20, 50]$. People identified in Experiment I were removed from the LOF analysis because we do not want to miss people who look normal with respect to that suspicious group.

Observation: Further work is required to assess the potential application of this work within the Medicare Australia compliance framework. Though the authors concluding a high degree of confidence in the validity of methodology for detecting prescription shoppers, it is presently unclear the extent to which the system could be used as a standalone and only method for identifying

all prescription shoppers within a population. It is likely that the value of approach lies in targeting higher risk prescription shoppers. The true extent of our false negative rate with respect to the entire population, not just the targeted subset, needs to be quantified.

The main limitation for the system is not being able to see the MBS side of the story to augment what it can infer from a consumer's PBS record. Such restrictions on linking MBS and PBS claims data are due to legislative requirements and will continue to pose difficulties for us.

A second limitation is that the system was designed from the beginning to look at individual consumers. From a cost-benefit perspective, detection of a colluding group of consumers is clearly more useful.

VII. Conclusion

The research of data mining utilization in medicine and public health offered only synopsis of present observations and disputes. Healthcare institutions and firms would utilize this process of data mining to gain more facts and knowledge from the information that is already present in their institution records. An institution must mention all rules and strategies on the safety and confidentiality of victim's data, before boarding on data mining. The same rule must be said to its partners and implied to other institutions and branches. Public health related issues like pandemic occurrence, the want to identify the commencement of disease in a non-invasive, easy manner and the want to be more reactive towards the patients. All these factors must be considered important and the desire for health institutions to combine information and data mining must be utilized to examine this information.

References References Referencias

1. Audain, C. 2007. Florence Nightingale. Online: http://www.scottlan.edu/lriddle/women/nitegale.htm. Accessed 30 July 2009.
2. Ayres, I 2008. Super Crunchers. New York: Bantam Books.
3. Bailey-Kellog, C. Ramakrishnan, N. and Marathe, M. Spatial Data Mining to Support Pandemic Preparedness. SIGKDD Explorations (8) 1, 80-82.
4. Cao, X., Maloney, K.B. and Brusic, V. 2008. Data mining of cancer vaccine trials: a bird's-eye view. Immunome Research, 4:7. DOI:10.1186/1745-7580-4-7
5. Cheng, T.H., Wei, C.P., Tseng, V.S. 2006 Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).
6. Health Grades, Inc. 2007. The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.
7. Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. 2004. Survey of fraud detection techniques. In Networking, Sensing and [8]Control, 2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.
8. Nightingale, F 1858. Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army.
9. Shillabeer, A 29 July 2009. Lecture on Data Mining in the Health Care Industry. Carnegie Mellon University Australia.
10. Shillabeer, A. and Roddick, J 2007. Establishing a Lineage for Medical Knowledge Discovery. ACM International Conference Proceeding Series. (311) 70, 29-37.
11. Tandoc, E.S 14 October2006.http://services.inquirer.net/print/print.php?article_id=26612.
12. Thangavel, K., Jaganathan, P.P. and Easmi, P.O. Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering [14]Technique. Asian Journal of Information Technology (5) 4, 413-417.
13. Tufte, E. 1997. Visual Explanations. Images and Quantities, Evidence and Narrative. Connecticut: Graphics Press.
14. Wong, W.K., Moore, A., Cooper, G. and Wagner, M . What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks, 2005. Journal of Machine Learning Research. 6, 1961- 1998.
15. Wilson A., Thabane L., Holbrook A 2003). "Application of data mining techniques in pharmacovigilance". British Journal of Clinical Pharmacology. (57) 2, 127-134.
16. Witten, I. H. and Frank, E. Data mining : practical machine learning tools and techniques. Morgan Kaufmann series in data management systems. Morgan Kaufman 2005.
17. Nada Lavrac, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, Andrej Kobler: Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics 40(4): 438-447 2007.
18. Kenneth Revett, Florin Gorunescu, Abdel-Badeeh M. Salem: Feature selection in Parkinson's disease: A rough sets approach. IMCSIT 2009: 425-428.
19. Pogorelc, B.; Gams, M.; , "Medically Driven Data Mining Application: Recognition of Health Problems from Gait Patterns of Elderly," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.976-980, 13-13 Dec. 2010.
20. Huidong Jin; Jie Chen; Hongxing He; Kelman, C.; McAullay, D.; O'Keefe, C.M.; , "Signaling Potential Adverse Drug Reactions from Administrative Health Databases," Knowledge and Data Engineering, IEEE Transactions on , vol.22, no.6, pp.839-853, June 2010.

21. Jian Xu; Maynard-Zhang, P.; Jianhua Chen; , "Predictive Data Mining to Learn Health Vitals of a Resident in a Smart Home," Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on , vol., no., pp.163-168, 28-31 Oct. 2007.
22. I.H. Witten and Eibe Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
23. Ng, K.S.; Shan, Y.; Murray, D.W.; Sutinen, A.; Schwarz, B.; Jeacocke, D.; Farrugia, J.; , "Detecting Non-compliant Consumers in Spatio-Temporal Health Data: A Case Study from Medicare Australia," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.613-622, 13-13 Dec. 2010.
24. J. Rogers; C. Puleston; A. Rector; , "The CLEF Chronicle: Patient Histories Derived from Electronic Health Records," Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on , vol., no., pp.x109, 2006
25. Weed LI (1969) Medical records medical education, and patient care. The problem-oriented record as a basic tool. Cleveland, OH: Case Western Reserve University.
26. Ceusters W, Smith B. (2005) Strategies for referent tracking in Electronic Health Records. Journal of Biomedical Informatics (in press).

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Modeling and Counter Measures of Flooding Attacks to Internet Threat Monitors (ITM): Using Botnet and Group-Testing approach

By K Munivara Prasad, A Rama Mohan Reddy, V Jyothsna

Sree Vidyanikethan Engg.College, Tirupati

Abstract - The Internet Threat Monitoring (ITM),is a globally scoped Internet monitoring system whose goal is to measure, detect, characterize, and track threats such as distribute denial of service(DDoS) attacks and worms. To block the monitoring system in the internet the attackers are targeted the ITM system. In this paper we address flooding attack against ITM system in which the attacker attempt to exhaust the network and ITM's resources, such as network bandwidth, computing power, or operating system data structures by sending the malicious traffic. We propose an information-theoretic frame work that models the flooding attacks using Botnet on ITM.we propose a novel group testing (GT)-based approach deployed on back-end servers, which not only offers a theoretical method to obtain short detection delay and low false positive/negative rate, but also provides an underlying framework against general network attacks.

General Terms : *Computer networks, network security, Attacks and Internet.*

Keywords : *Internet Threat Monitors (ITM), DDoS, Flooding attack, Botnet and Honeypot, Group testing.*

GJCST Classification : *K.6.5*

Strictly as per the compliance and regulations of:

Modeling and Counter Measures of Flooding Attacks to Internet Threat Monitors (ITM): Using Botnet and Group-Testing approach

K Munivara Prasad [α], A Rama Mohan Reddy [Ω], V Jyothsna [β]

Abstract - The Internet Threat Monitoring (ITM),is a globally scoped Internet monitoring system whose goal is to measure, detect, characterize, and track threats such as distribute denial of service(DDoS) attacks and worms. To block the monitoring system in the internet the attackers are targeted the ITM system. In this paper we address flooding attack against ITM system in which the attacker attempt to exhaust the network and ITM's resources, such as network bandwidth, computing power, or operating system data structures by sending the malicious traffic. We propose an information-theoretic frame work that models the flooding attacks using Botnet on ITM.we propose a novel group testing (GT)-based approach deployed on back-end servers, which not only offers a theoretical method to obtain short detection delay and low false positive/negative rate, but also provides an underlying framework against general network attacks.

General Terms : *Computer networks, network security, Attacks and Internet.*

Keywords : *Internet Threat Monitors (ITM), DDoS, Flooding attack, Botnet and Honeypot, Group testing.*

I. Introduction

Internet security is increasing in importance. Yet, despite decades of research, we are still unable to make secure computer networks. Further, more sophisticated and new attacks are expected to continue posing a greater degree of threat to Internet services. As a result, a more fundamental model, in terms of theoretical and system perspectives, regardless of attack types must be investigated. An essential problem to overcome for any defense mechanism is the fact that malicious traffic/packets can be similar to legitimate ones.

Denial-of-Service (DoS) is a major security problem in computer systems and networks. In a DoS attack, a group of attackers try to make a service unavailable to legitimate clients for unacceptably long periods of time. Service-level DoS attacks target server resources by issuing legitimate-like service requests at a high rate to overwhelm the victim servers. These attacks, attempt to exhaust the victim's resources, such as network bandwidth, computing power, or operating system data structures. Flood attack, Ping of Death attack, SYN attack, Teardrop attack, DDoS , and Smurf attack are the most common types of DoS attacks. The hackers who launch DDoS attacks typically target sites or services provided by high-profile organizations, such as government agencies, banks, credit-card payment gateways, and even root name servers.

A flooding-based Distributed Denial of Service (DDoS) attack is a very common way to attack a victim machine by sending a large amount of unwanted traffic. Network level congestion control can throttle peak traffic to protect the network. Network monitors are used to monitor the traffic in the networks to classify them as genuine or attack traffic and also these monitors gives the traffic as an input to several DDoS detection algorithms for detection of DDoS attacks. However, it cannot stop the quality of service (QoS) for legitimate traffic from going down because of attacks. Two features of DDoS attacks hinder the advancement of defense techniques. First, it is hard to distinguish between DDoS attack traffic and normal traffic. There is a lack of an effective differentiation mechanism that results in minimal collateral damage for legitimate traffic. Second, the sources of DDoS attacks are also difficult to find in a distributed environment. Therefore, it is difficult to stop a DDoS attack effectively.

The Internet Threat Monitoring (ITM) System basically has two main components one is centralized data center and another is the number of monitors which are distributed across the Internet. Each monitor covers the range of IP addresses and monitors the traffic to send the traffic logs to data center. The data center now collects the traffic logs from monitors and analyzes the collected traffic logs to publish reports to ITM system users.

The collected logs, as a random sample of the Internet traffic, can still provide critical insights for the public to measure, characterize, and track/detect Internet security threats. The idea of ITM systems dates back to DShield and CAIDA network telescope [4], [5], which have been successfully used to analyze the activities of worms and DDoS attacks [3], [6].The reason is that if an attacker discovers the monitor locations, it can easily avoid detection (by ITM systems) by bypassing the monitored IP addresses and directing the attack to the much larger space of unmonitored IP addresses. Furthermore, such an attacker may even mislead the reports published by an ITM system by

Author [α]: Assistant Professor(SL) Sree Vidyanikethan Engg.College, Tirupati. E-mail : prasadkmv27@gmail.com

Author [Ω] : Professor and Head Dept.of CSE Sri Venkateswara university college of Engg.S V University,Tirupati.

Author [β] : Assistant Professor, Department of Information technology, Sreevidyanikethan Engineering College, Tirupati

manipulating traffic to the identified monitors, generating highly skewed samples. Since ITM reports are trusted by the public as a random (unbiased) sample of Internet traffic, the confidentiality of monitor locations is vital for the usability of ITM systems.

The monitor locations of an ITM system can be compromised by introducing several attacks by the attackers which includes Localization attacks [1] and DDoS Attacks which exploits some vulnerability or implementation bug in the software implementation of a service to bring that down or that use up all the available resources at the target machine or that consume all the bandwidth available to the victim machine, this is called as Bandwidth attacks.

The main goal of our work is to perform the identification of attackers much faster by testing them in group instead of one by one. This will help to detect the flooding attacks, So that if any ITM is found under attack, we can immediately identify and filter the attackers out of its client set. Apparently, this problem resembles the group testing (GT) theory [14] which aims to discover defective items in a large population with the minimum number of tests where each test is applied to a subset of items, called pools, instead of testing them one by one. Therefore, we apply GT theory to this network security issue and propose specific algorithms and protocols to achieve high detection performance in terms of short detection latency and low false positive/negative rate.

In this paper we introduce an information theoretic frame work model to existing flooding attacks on ITM system monitors. In the flooding attack the attacker sends the large volume of unwanted traffic to the targeted monitor by using the botnet or huge number of compromised systems. Based on the Information-theoretic model we propose a Group Testing based approach to detect flooding attacks.

II. Related Work

Probing traffic based Localization attack [7][8] in which an attacker sends high rate short length port scan messages to the targeted network to compromise the monitor locations in ITM system. Then, attacker queries the data center to determine whether a short spike of high-rate traffic appears in the queried time-series data, for confirmation of the attack.

A steganographic localization attack [9] an attacker launches a stream of low-rate port-scan probing traffic which is marginally modulated by a secret Pseudonoise (PN) code. While the low-rate property prevents the exhibition of obvious regularity of the published traffic data at the data center, based on the carefully synchronized PN code, the attacker can still accurately identify the PN-code-modulated traffic in the retrieved published traffic data from the data center. Thereby, the existence of monitors in the targeted network can be compromised. To this end, the PN-code-based steganographic attack presented in our paper can be understood as a covert channel problem [10], because the attack traffic encoded by a signal blends into the background traffic and is only recognizable by the attacker which knows the secret pattern of the PN code.

In [1] introduced the information theoretic framework to evaluate the effectiveness of the localization attacks by using the minimum time length required by an attacker to achieve a predefined detection rate as the metric. But this frame work is defined in specific to the localization attacks only; they are not given any solution for other DDoS attacks. The frame work allows the ITMs which are registered within the data center given, and the access is restricted to that private region only. But public access of the ITMs and data center allows more scope to provide security against different attacks.

Group Testing (GT) based approach [17] is used to detect the application denial of service attacks, there the efficiency of the attack detection is improved by testing the traffic by group instead of one by one. The GT approach also minimizes the false positives and false negatives comparatively to the input traffic. But the GT approach can be applied to the DDoS detection whenever the number of attackers known in advance. This assumption will not be suitable for all DDoS attacks.

III. Proposed Work

In [1] the authors define a model in which the ITMs in the networks sends the traffic logs periodically to the data center and the data center collects the traffic logs and publishes the reports to ITM system users which are registered, that means it creates the private environment or region .In the private region the scope for DDoS attacks are very less, and they are restricted this model only for Localization attacks. In this section we have defined a model which will provide the following extensions.

Public accessing : Public accessing of the data center increases the network usage and provides better communication with the outside world rather than private environment. In this any user from outside the private region can get the communication with the private network, if the user is genuine he can get the status of the monitor before sending the data to internal monitors, to avoid the attacks. If the user is an attacker, then this status information can be misused to perform the attacks on the monitor. The data center sends the status information to any users (public or private) based on the request query, but the private (internal) users can get the highest priority.

Usage of Botnets for Flooding Attack : A denial-of-service (DoS) attack is an explicit attempt by attackers to prevent an information service's legitimate users from using that service. In a DDoS attack, these

attempts come from a large number of distributed hosts that coordinate to flood the victim with an abundance of attack packets simultaneously. The attacker may use the botnets [11], [12] and other alternatives to launch the attack.

a) Flooding

Launching a flooding attack : Once the DDoS network has been set up and the infrastructure for communication between the agents and the handlers established, all that an attacker needs to do is to issue commands to the agents to start sending packets to the victim host. The agents try to send unusual data packets (all TCP flags set, repeated TCP SYN packets, Large ICMP packets) to maximize the possibility of causing disruption at the victim and the intermediate nodes. There are certain basic packet attack types which are favorites of the attack tool designers. All the attack tools use a combination of these packet attack types to launch a DDoS attack. The basic attack types are

i. TCP floods : A stream of packets with various flags (SYN,RST, ACK) are sent to the victim machine. The TCP SYN flood works by exhausting the TCP connection queue of the host and thus denying legitimate connection requests. TCP ACK floods can cause disruption at the nodes corresponding to the host addresses of the floods as well. Also the one known tool that uses TCP ACK flooding (mstream [13]) has been known to cause disruptions in a router even with a moderate packet rate. Both TCP SYN flooding and the mstream attack constitute a group of attacks known as asymmetric attacks (Attacks where a less powerful system can render a much more powerful system useless).

ii. ICMP floods (e.g ping floods) : A stream of ICMP packets is sent to the victim host. A variant of the ICMP floods is the Smurf attack in which a spoofed IP packet consisting of an ICMP ECHO_REQUEST is sent to a directed broadcast address. The RFC for ICMP specifies that no ECHO_REPLY packets should be generated for broadcast addresses, but unfortunately many operating systems and router vendors have failed to incorporate this into their implementations. As a result, the victim host (in this case the machine whose IP address was spoofed by the attacker) receives ICMP ECHO_REPLY packets from all the hosts on the network and can easily crash under such loads. Such networks are known as amplifier networks and thousands of such networks have been documented.

iii. UDP floods : A huge amount of UDP packets are sent to the victim host. Trinoo is a popular DDoS tool that uses UDP floods as one of its attack payloads.

b) Bots

Studying the evolution of bots and botnets provides insight into their current capabilities. One of the original uses of computer bots was to assist in Internet Relay Chat (IRC) channel management [14]. IRC is a chat system that provides one-to-one and one-to-many instant messaging over the Internet. Users can join a named channel on an IRC network and communicate with groups of other users. Administering busy chat channels can be time consuming, and so channel operators created bots to help manage the operation of popular channels. One of the first bots was Eggdrop, which was written in 1993 to assist channel operators [1].

In time, IRC bots with more nefarious purposes emerged. The goal of these bots was to attack other IRC users and IRC servers. These attacks often involved flooding the target with packets (i.e., DoS attacks). The use of bots helped to hide the attacker because the attack packets were sent from the bot rather than directly from the attacker (assuming a non-spoofed attack). This new level of indirection also allowed multiple computers to be grouped together to perform distributed attacks (DDoS) and bring down bigger targets. Larger targets required more bots, and so attackers looked for methods to recruit new members. Since very few users would agree to have their computers utilized for conducting packet floods, attackers used trojaned files and other surreptitious methods to infect other computers.

c) IRC- based Command and Control

A bot must communicate with a controller to receive commands or send back information. One method for establishing a communication channel is to connect directly to the controller. The problem is that this connection could compromise the controller's location. Instead, the bot controller can use a proxy such as public message drop point (e.g., a well-known message board). However, because websites and other drop points can introduce significant communication latency, a more active approach is desirable. A well-known public exchange point that enables virtually instant communication is IRC.

IRC provides a common protocol that is widely deployed across the Internet and has simple text-based command syntax. There is also a large number of existing IRC networks that can be used as public exchange points. In addition, most IRC networks lack any strong authentication, and a number of tools to provide anonymity on IRC networks are available. Thus, IRC provides a simple, low-latency, widely available, and anonymous command and control channel for botnet communication. An IRC network is composed of one or more IRC servers as depicted in Figure 1.

In a typical botnet, each bot connects to a public IRC network or a hidden IRC server on another compromised system. The bot then enters a named channel and can receive commands directly from a controller or even from sequences encoded into the title of the channel.

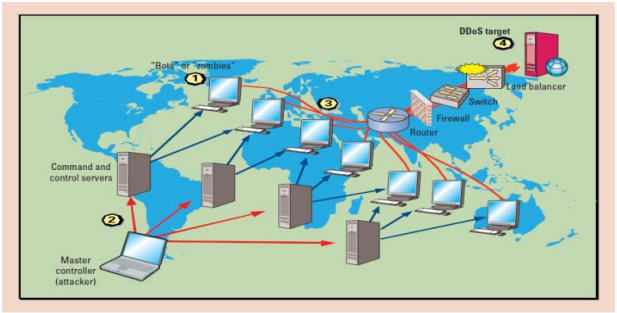


Figure 1 : Compromised computers. In a distributed denial-of-service attack (DDoS), these computers serve three major roles: master controller, command and control server, and bot.

d) Group Testing

The first application of group testing was during WWII; instead of testing every blood sample individually, groups of samples were pooled together and tested collectively. If the outcome of the group test is negative, all samples in the group are good (disease-free). Although group testing has been used since then in many security and networking applications, such as data forensics, cryptography, multiple-access channels, and broadcast security against jamming, our work is the first to apply this powerful theory to the DoS attack problem.

Group testing aims mainly at identifying the defective (special) members of a population with few tests. There are two classes of group-testing mechanisms. "Non-adaptive", or single-stage, specifies all tests simultaneously without the benefit of using the outcomes of previous tests to determine the present test. Adaptive (multi-stage) group testing uses feedback from previous test results to determine subsequent tests.

i. *Basic Idea*

The basic group testing is modeled as a T × N matrix, where N is the total number of members, and T is the number of tests. Matrix rows represent tests and columns represent group members. When a matrix element (i, j) is set to 1, this means that member j participates in test i. An example of a group testing matrix for a population of 10 members with 4 tests is shown in Figure 1. Test results are represented as a vector with an element for each test. For simplicity we assume that the test results are binary. So, a test result is set to 1 if the corresponding test returns a positive result, that is, if the test was applied to a group with at least one defective member. In the example shown in Figure 1 the 2nd, 6th, and 8th members are defective.

	Members (defectives underlined)										Test Results
	1	2	3	4	5	6	7	8	9	10	
Tests	0	1	0	0	0	0	1	1	1	0	1
	1	0	1	0	0	0	1	0	1	1	0
	0	0	0	1	0	1	0	1	0	1	1
	1	0	0	1	1	0	1	0	0	0	0

ii. *Detection of defective members*

The detection algorithm discovers the defective members using the result vector and Members Test (defectives underlined) Results. An example of a group-testing matrix. the matrix. The algorithm we use in this paper works by excluding a negative (non-defective) member if it participates in a "large enough" number of tests with a negative result. For instance, if we assume that a member has to participate in only one negative test to be excluded, then in the above example; members 1, 3, 4, 5, 7, 9, and 10 will be excluded. This leaves us with the defective members 2, 6, and 8 as suspects. In this specific example, all defective elements are detected, and all non-defective members are cleared.

iii. *Apply to Attack Detection*

A detection model based on GT can be assumed that there are T virtual servers and N clients, among which d clients are attackers shown in fig. Consider the matrix M_{txn} ,the clients can be mapped into the columns and virtual servers into rows in M, where M [i,j]=1 if and only if the requests from client j are distributed to virtual server i. With regard to the test outcome column V, we have V [i]=1 if and only if virtual server i has received malicious requests from at least one attacker, but we cannot identify the attackers at once unless this virtual server is handling only one client. Otherwise, if V [i]=0, all the clients assigned to server i are legitimate. The d attackers can then be captured by decoding the test outcome vector V and the matrix M.

iv. *False Positive and False Negative Probabilities*

A false positive is when a non-defective member gets falsely identified as defective, while a false negative is when a defective member ends up being not detected. In the example above, both the false positive probability and the false negative probability are 0. In general, simple detection algorithm discussed above detects all defective members with the false positive probability

$$FP = [1-p(1-p)d]T,$$

Where d is the number of defective members in the group and T is the number of tests used to detect defective members. By differentiating the above equation with respect to p, the optimal value of p, the value that yields the minimum false positive probability, is 1/d+1. Thus, the minimum false positive probability for a given number of tests T is:

$$FP = [1 - \frac{1}{d+1}(1 - \frac{1}{d+1})^d]^T$$

Finally, from Eqn. 1 we derive the number of tests, Tfp, required to achieve a target false positive probability fp:

$$T_{fp} = \frac{log(fp)}{log(1 - \frac{1}{d+1}(1 - \frac{1}{d+1})^d)}$$

As we will show later, Tfp is O(d), that is, the number of tests is in the order of number of defective members attackers) not the total number of members (N).

The false negative probability is

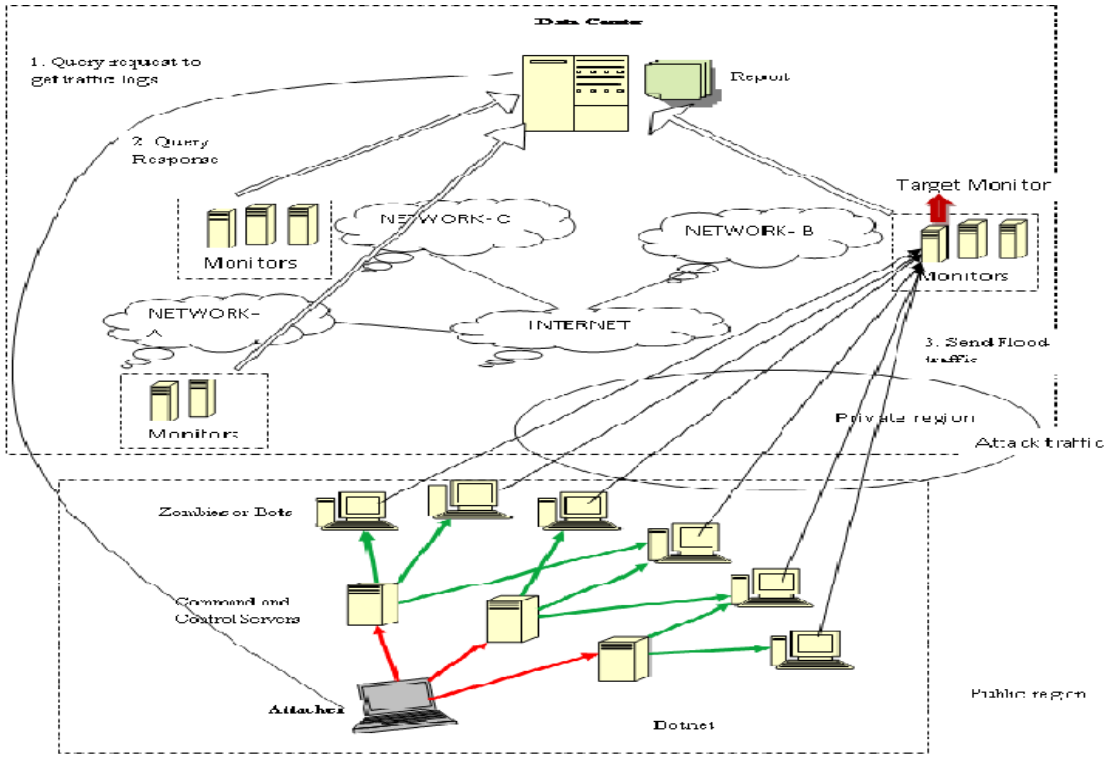


Figure 2 : Work flow of flooding attacks using botnet.

$$FN = 1 - [1 - \frac{1}{d+1}(1.0 - \rho_{attack})(1 - \frac{1}{d+1} \cdot \rho_{attack})^{d-1}]^T$$

Where ṗattack refers the the probability of attack and d is the expected number of attackers.

IV. Proposed Model

In this paper we divided the entire model into two regions namely private region and public region. The Internet Threat Monitors (ITM) are distributed across the Internet and each monitor records the traffic addressed to range of IP addresses and send the traffic logs periodically to the data center. The data center then analyzes the traffic logs collected from the monitors and publishes the reports to ITM system users. The collection of monitors under the data center forms the private region because the ITMs are registered before sending the logs to the data center. Any user can get the reports of the requested ITM by sending the query request to the data center and the data center is answerable to all the ITMs which are registered.

The public region of our model specifies the unregistered users of the data center who does not have any permission to access the data center, but they can get the traffic reports related to any ITM by sending the query request to the data center. The data center scope is extended to the public domain but it can only give the traffic reports to the public users. Allowing the public users or network accessing to the data center and monitors, causes decrease in the performance because of the overload of the data center. These can be balanced by introducing the priorities to the users; the internal or private region users have the highest priority than the public users .This priorities does not disturb the existing scenario but this can enhance the service to the public domain ,this will not be a over burden to the data center.

In This section we are constructing the botnet as the public user network without having any registration with data center and performing the flooding attack on the ITM which is local to the data center.

i. *Generation of flooding attack with Botnet*

A DDoS (Flooding) attack mechanism typically includes a network of several compromised computers [15]. These compromised computers serve three major role -master controller, command and control (C&C) server, and bot. An attacker prepares a DDoS attack by exploiting vulnerabilities in one computer system and making it the DDoS "master controller." From here, the attacker identifies and communicates with other

compromised systems. A C&C server is a compromised host with a special program running on it, this server distributes instructions from the attacker to the rest of the bots, which form a botnet[11]. (A bot is a compromised host that runs a special program.) Each C&C server is capable of controlling multiple bots, each of which is responsible for generating a stream of packets to the intended victim. Often, the bots employed to send the flood of requests are infected with a virus that lets attackers use them anonymously.

A Flooding attack happens in several phases :

- Discover vulnerable hosts. To launch a DDoS attack, attackers first build a network of computers that they can use to produce the volume of traffic needed to deny services to legitimate users. To create this network, they first scan and identify vulnerable sites or hosts. Vulnerable hosts are usually those that run either no antivirus software or an out-of-date version, or those that aren't properly patched. Attackers use these compromised hosts for further scanning and compromises.
- Establish a botnet. After gaining access, attacker must then install attack tools on the compromised hosts to form a botnet.
- Launch an attack. In the next phase, attackers send commands to C&C servers for their bots to attack by sending hundreds of thousands of requests to the target simultaneously.
- Flood a target. In the final phase, monitor receives a flood of requests to the point where they can't operate effectively.

ii. *Conformation of attack*

The attacker queries the data center for the traffic reports. Such traffic reflects both flooding requests traffic and other traffic collected from all monitors. Then the attacker confirms the attack by checking the status of the ITM in the traffic reports published by the data center.

V. Prevention

Preventive mechanisms attempt either to reduce the possibility of DDoS attacks or enable potential victims to endure the attack without denying services to legitimate users.

- System security mechanisms increase a host's overall security posture and prevent it from becoming part of a botnet or a DDoS victim. Examples of system security mechanisms include reliable firewall filtering, proper system configuration, effective vulnerability management, timely patch installation, robust antivirus programs, controlled and monitored system access, and solid instruction detection.
- Resource multiplication mechanisms provide an abundance of resources to counter DDoS threats, such as increasing the capacity of network bandwidth, routers, firewalls, and servers. Additional examples include deploying information services at diverse network locations and establishing clusters of servers with load-balancing capabilities. Resource multiplication essentially raises the bar on how many bots must participate in an attack to be effective. While not providing perfect protection, this last approach has often proved sufficient for small- to mid-range DDoS attacks.

Preventing Flooding Attacks

In this section we introduce a general methodology to prevent flooding attacks. It is based on the following line of reasoning:

1) To mount a successful Flooding attack, a large number of compromised machines are necessary.
2) To coordinate a large number of machines, the attacker needs a remote control mechanism.
3) If the remote control mechanism is disabled, the Flooding attack is prevented.

Our methodology to mitigate flooding attacks aims at manipulating the root-cause of the attacks, i.e., influencing the remote control network. Our approach is based on three steps:

1. Infiltrating the remote control network.
2. Analyzing the network in detail.
3. Shutting down the remote control network.

In the first step, we have to find a way to smuggle an agent into the control network. In this context, the term agent describes a general procedure to mask as a valid member of the control network. This agent must thus be customized to the type of network we want to plant it in. The level of adaptation to a real member of the network depends on the target we want to infiltrate. For instance, to infiltrate a botnet we would try to simulate a valid bot, maybe even emulating some bot commands.

Once we are able to sneak an agent into the remote control network, it enables us to perform the second step, i.e., to observe the network in detail. So we can start to monitor all activity and analyze all information we have collected.

In the last step, we use the collected information to shut down the remote control network. Once this is done, we have deprived the attacker's control over the other machines and thus efficiently stopped the threat of a flooding attack with this network. Again, the particular way in which the network is shut down depends on the type of network.

VI. Detection of Flooding Attacks

In this section we present efficient way of detecting the attacks on the ITMs in the given information theoretic frame work. We divide the attack

detection process into two phases, Firstly the primary detection of DDoS attacks on the ITMs and the later is the detection of flooding attacks on the ITMs.

In the primary detection phases the system detects the attacks based on traffic information aggregated from all monitors in the ITM system. If the overall traffic rate (e.g., volume in a given time interval) exceeds a predetermined threshold, the defender issues an alarm. The threshold value can be maintained either at data center or the individual ITMs based on the type of schemes used [1] in the network.In the primary detection phase the system detects some attack was happened in the network. If the detection scheme is centralized, then whenever the aggregate traffic exceeds the threshold maintained at the data center then the data center finds the attack and that attacked monitor can be identified by verifying the individual traffic logs of each ITM from the report. Otherwise if the detection strategy is distributed then each monitor maintained an individual threshold and checked the aggregate traffic regularly. If the traffic exceeds the threshold then it find the attack was happened and sends the status as attacked to the data center. After getting the attacked status from the ITM the data center blocks the corresponding ITM and displays the status of the ITM as blocked in the status reports, which will avoids the further traffic to or from the attacked ITM with the rest of the networks.

The second stage of detection specifies the detection of the flooding attacks. Once the attack is conformed then the data center identifies the attacked monitor and the traffic logs will be handover to the flooding detection phase. The flooding detection phase then performs the group testing (GT) on the traffic, then identifies the attackers from the client traffic set.

In this section we define group testing (GT) based DDoS detection methods for flooding detection on ITMs in information frame work defined, and also detection of false positives and false negatives in the network for large flow size.

a) BOTNET Detection

Botnets are a very real and quickly evolving problem that is still not well understood. In this paper, we outline the problem and investigate methods of stopping bots. We identify three approaches for handling botnets:

1) Prevent systems from being infected,
2) Directly detect command and control communication among bots and between bots and controllers, and,
3) Detect the secondary features of a bot infection such as propagation or attacks.

The first approach is to prevent systems from being infected. There are a range of existing techniques, including anti-virus software, firewalls, and automatic patching.

The second approach is to directly detect botnet command and control traffic. Botnets today are often controlled using Internet Relay Chat (IRC) and one possible method of detecting IRC-based botnets is to monitor TCP port 6667 which is the standard port used for IRC traffic. One could also look for non-human behavioral characteristics in traffic, or even build IRC server scanners to identify potential botnets.

We argue there is also a third approach that detects botnets by identifying secondary features of a bot infection such as propagation or attack behavior. Rather than directly attempting to find command and control traffic, the key to this approach is the correlation of data from different sources to locate bots and discover command and control connections.

In this paper we investigate the second approach for stopping botnets. The problem with the first approach is that preventing all systems on the Internet from being infected is nearly an impossible challenge. As a result, there will be large pools of vulnerable systems connected to the Internet for many years to come.

b) Detecting Command and Control

To combat the growing problem of bots, we identified two approaches for detecting botnets: detect the command and control communication, or detect the secondary features of a bot infection. In this section we study methods of detecting botnets by directly locating command and control traffic.

i. *IRC-based Botnet Detection*

Today, most known bots use IRC as a communication protocol, and there are several characteristics of IRC that can be leveraged to detect bots. One of the simplest methods of detecting IRC-based botnets is to offramp traffic from a live network on known IRC ports (e.g., TCP port 6667) and then inspects the payloads for strings that match known botnet commands. Unfortunately, botnets can run on non-standard ports. Another method is to look for behavioral characteristics of bots. One study found that bots on IRC were idle most of the time and would respond faster than a human upon receiving a command. The system they designed looked for these characteristics in Netflow traffic and attempted to tag certain connections as potential bots [15].

The approach was successful in detecting idle IRC activity but suffered from a high false positive rate. Given problems such as false positives on live networks, another approach is to use a non-productive resource or honeypot.

One group set up a vulnerable system and waited for it to be infected with a bot. They then located outgoing connections to IRC networks and used their own bot to connect back and profile the IRC server [16]. However, they did not take the next step and develop a detection system based on the technique.

Rather than connecting to the IRC server directly, another approach is to use a honeypot to catch the bot and then look for characteristics of command and control traffic in the outgoing connections. We located all successful outgoing TCP connections and verified that they were all directly related to command and control activity by inspecting the payloads. There were a wide range of interesting behaviors, including connections from the bot to search engines to locate and use bandwidth testers, downloading posts from popular message boards to get server addresses, and the transmission of comprehensive host profiles to other servers.

These profiles included detailed information on the operating system, host bandwidth, users, passwords, file shares, filenames and permissions for all files, and a number of other minute details about the infected host.

We then analyzed all successful outgoing connections and looked for specific characteristics that could be used to identify botnet command and control traffic. The results suggested that there are no simple characteristics of the communication channels themselves that can be used for detection. For example, the length of the outgoing connections varied widely, with certain connections lasting more then 9 hours and others less than a second.

The number of bytes transferred per connection also varied widely even when we separated out IRC communication from other command and control activity. The results from our analysis nor the results from previous bot detection efforts has revealed any simple connection-based invariants useful for network detection. One might inspect every payload of every packet however this is currently very costly on high throughput networks. More importantly, attackers can make small modifications that make detection nearly impossible.

ii. *Limitations of Honeypot detection*

Efficiency : The efficiency of the detection system is depends on the number of honeypots placed in the network. If one honeypot is used to perform the detection in centralized approach, then more than one ITM is attacked automatically the honeypot will be overloaded and it takes more time to detect the attackers. Otherwise if the detection system is distributed, then the efficiency of the detection system is improved, but it is very much cost effective, practically not possible for the large networks.

In GT approach the detection process can be carried out only at data center by collecting the traffic logs from the attacked ITM through data center same as the centralized detection of the honeypot approach. Unlike honeypot detection, the efficiency of the detection process does not depend on the traffic because huge amount of traffic also processed in terms of pools in the GT approach and handled successfully. The pools or groups can be as inputs for multiple rounds of different tests in GT approach to check for different anomalies in the input malicious traffic.

Reliability : One honeypot for each ITM approach is reliable but it is practically very difficult to manage and maintained. If the centralized honeypot compromises then total detection process will be vanished.

The reliability of the GT approach depends on the groups or pools of the malicious traffic considered as the input for the GT approach. The number of tests performed on the traffic improves the detection efficiency and covers wide range of possible attacks of DDoS attacks on ITM.

Scalability : If the detection approach is centralized then no need to use additional honeypots to the network except the honeypot placed at the data center when ever new ITM entered into the private region. In the distributed detection approach new honeypot is attached whenever the new ITM entered into the network.

When the network increases or new ITMs entered into the private region of the network, it does not create additional load on the existing data center. The GT approach handles the input traffic without depends on the number of ITMs or the data centers in the network.

Load Sharing : Every honeypot has its own capacity of handling the load or traffic in the network. In the centralized detection approach if the traffic exceeds its capacity, then the total detection system is vanished. The same problem occurs in case of distributed detection approach also.

If the load in the network increases then the GT approach forms more number of input pools and the tests are applied on the pools repeatedly to perform the attack detection.

False positives and false negatives: In the honeypot based flooding detection false positives and false negatives are not explicitly considered. The detection process finds the root of the attack, by blocking the IRC server.

In GT approach the false positives and false negatives calculated explicitly by conducting specific tests on the input pools of the traffic. False positives and false negatives improve the detection process in terms of considering the attack traffic as genuine and vice-versa.

c) Attack detection using Group testing approach

In the detection model[17], each testing pool is mapped to a virtual server within a back-end server machine. Although the maximum number of virtual servers can be extremely huge, since each virtual server requires enough service resources to manage client requests, it is practical to have the virtual server quantity

(maximum number of servers) and capacity (maximum number of clients that can be handled in parallel) constrained by two input parameters K and w, respectively.

The maximum number of attackers d is assumed known beforehand. Scenarios with nondeterministic d are out of the scope of this paper. In fact, these scenarios can be readily handled by first testing with an estimated d, then increasing d if exactly d positive items are found.

Each back-end server works as an independent testing domain, where all virtual servers within it serve as testing pools. In the following sections, we only discuss the operations within one backend server, and it is similar in any other servers. The detection consists of multiple testing rounds, and each round can be sketched in four stages.

First, generate and update matrix M for testing.

Second, "assign" clients to virtual servers based on M. The back-end server maps each client into one distinct column in M and distributes an encrypted token queue to it. Each token in the token queue corresponds to a 1-entry in the mapped column, i.e., client j receives a token with destination virtual server i iff M[i,j]= 1. Being piggybacked with one token, each request is forwarded to a virtual server by the virtual switch. In addition, requests are validated on arriving at the physical servers for faked tokens or identified malice ID. This procedure ensures that all the client requests are distributed exactly as how the matrix M regulates and prevents any attackers from accessing the virtual servers other than the ones assigned to them.

Third, all the servers are monitored for their service resource usage periodically, specifically, the arriving request aggregate (the total number of incoming requests) and average response time of each virtual server are recorded and compared with some dynamic thresholds to be shown later. All virtual servers are associated with positive or negative outcomes accordingly.

Fourth, decode these outcomes and identify legitimate or malicious IDs. By following the detection algorithms all the attackers can be identified within several testing rounds. To lower the overhead and delay introduced by the mapping and piggybacking for each request, the system is exempted from this procedure in normal service state. As shown in Fig. 3, the back-end server cycles between two states, which we refer as NORMAL mode and DANGER mode. Once the estimated response time (ERT) of any virtual server exceeds some profile-based threshold, the whole backend server will transfer to the DANGER mode and execute the detection scheme. Whenever the average response time (ART) of each virtual server falls below the threshold, the physical server returns to NORMAL mode.

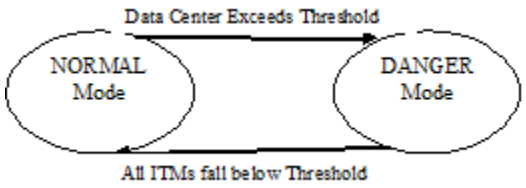


Fig.3 : Two state Diagram if the system.

Based on the system framework above, we propose three detection algorithms SDP, SDoP, and PND in this section. Note that the length of each testing round is a predefined constant P; hence, we analyze the algorithm complexity in terms of the number of testing rounds for simplicity.

i. *Sequential Detection with Packing*

This algorithm investigates the benefit of classic sequential group testing, i.e., optimizing the grouping of the subsequent tests by analyzing existing outcomes. Similar to traditional sequential testing, each client (column) only appears in one testing pool (server) at a time. However, to make full use of the available K servers, we have all servers conduct test in parallel.

ii. *Sequential Detection without Packing*

Considering the potential overload problem arises from the "packing" scheme adopted in SDP, we propose another ralgorithm where legitimate clients do not shift to other servers after they are identified. This emerges from the observation that legitimate clients cannot affect the test outcomes since they are negative.

The basic idea of the SDoP algorithm can be sketched below. Given a suspect IDs set S with initial size n, evenly assign them to the K server machines, similar to SDP in the first round. For the following rounds, assign suspect IDs to the K servers instead of |A| available ones. For the identified legitimate IDs, never move them until their servers are to be overloaded. In this case, reassign all legitimate IDs over the K machines to balance the load. For server with positive outcome, the IDs active on this server but not included by the set of identified legitimate ones, i.e., suspect IDs, will still be identified as suspect. However, if there is only one suspect IDs of this kind in a positive server, this ID is certainly an attacker.

iii. *Partial Non adaptive Detection*

Considering the fact that in the two sequential algorithms mentioned, we cannot identify any attackers until we isolate each of them to a virtual server with negative outcome, which may bring up the detection latency. In this scenario, the requests from the same client will be received and responded by different servers in a round-robin manner. Different from SDP and SDoP, a d-disjunct matrix is used as the testing matrix in this scheme and attackers can be identified without the need of isolating them into servers.

The attack traffic can be identified by using any of the three methods defined and QoS of the system is

completely depends on the number of tests performed, number of rounds conducted on the pools .Once the attack traffic is identified ,then it is easy to find the attack source using the traffic entities. In our paper we focused on the flooding based attacks and these are generated using the botnet. The attacker floods the target ITM by sending the commands through C & C server. Here the attack traffic contains the C & C server address, and then it is very easy to block or point out that server to avoid the flooding attacks on the system.

While identifying the attack traffic with GT approach one can remember that the data should not be lost; these can be effectively done with false positives and negatives.

Despite the number of needed testing rounds differs for these three algorithms above, the time complexity of calculating each testing round for these algorithms is approximate in practice. It is trivial to see that the costs for SDP and SDoP are negligible, but not for PND algorithm which involves polynomial computation on Galois Field. However, considering that the upper bound of both the number of clients n and attackers d is estimated, the detection system can pre compute the d-disjunct matrices for all possible (n, d) pairs offline, and fetch the results in real time. Therefore, the overhead can be decreased to O(1) and the client requests can be smoothly distributed at the turn of testing rounds without suffering from long delays of matrix update.

VII. Conclusion and Future Work

The frame work integrates active real time flooding attack flow identification from botnet with GT approach. The GT approach has been used at Data center to detect the attack traffic that interns helpful, while identifying the C&C server to block the flooding attack against the ITM. The false positive and false negatives can be effectively minimized to improve the QoS factors of the system.

Some of the avenues for further extensions are with larger and heterogeneous networks. Back tracking can be applied on attack flows to reach the attack source. Both of them hold promise for evaluating and improving our DDoS detection and defense method and data center information protection. The data center load can be still minimized by used some distributed load sharing algorithms.

References Références Referencias

1. wei yu, nan zhang, xinwen fu, Riccardo bettati, and wei zhao, "localization attacks to internet threat monitors:Modeling and countermeasures"on ieee transactions on computers, vol. 59, no. 12, december 2010.
2. J. Mirkovic and P. Reiher, "A Taxonomy of DDOS Attack and DDOS Defense Mechanisms," ACM SIGCOMM Computer Comm.Rev., vol. 34, no. 2, pp. 39-53, Apr. 2004.
3. SANS, Internet Storm Center, http://isc.sans.org/, 2010.
4. D. Moore, G.M. Voelker, and S. Savage, "Inferring Internet Deny-of-Service Activity," Proc. 10th USNIX Security Symp. (SEC), Aug. 2001.
5. Yegneswaran, P. Barford, and S. Jha, "Global Intrusion Detection in the Domino Overlay System," Proc. 11th IEEE Network and Distributed System Security Symp. (NDSS), Feb. 2004.
6. M. Bailey, E. Cooke, F. Jahanian, J. Nazario, and D. Watson, "The Internet Motion Sensor: A Distributed Blackhole Monitoring System," Proc. 12th Ann. Network and Distributed System Security Symp. (NDSS), Feb. 2005.
7. J. Bethencourt, J. Frankin, and M. Vernon, "Mapping Internet Sensors with Probe Response Attacks," Proc. 14th USNIX Security Symp. (SEC), July/Aug. 2005.
8. Y. Shinoda, K. Ikai, and M. Itoh, "Vulnerabilities of Passive Internet Threat Monitors," Proc. 14th USNIX Security Symp. (SEC), July/Aug. 2005.
9. X. Wang, W. Yu, X. Fu, D. Xuan, and W. Zhao, "Iloc: An Invisible Localization Attack to Internet Threat Monitoring Systems," Proc.IEEE INFOCOM (Mini-Conf.), Apr. 2008.
10. S. Cabuk, C. Brodley, and C. Shields, "Ip Covert Timing Channels:Design and Detection," Proc. 2004 ACM Conf. Computer and Comm.Security (CCS), Oct. 2004.
11. E. Cooke, F. Jahanian, and D. McPherson, "The Zombie Roundup: Understanding, Detecting, and Disrupting Botnets," Proc. Steps to Reducing Unwanted Traffic on the Internet Workshop (SRUTI), July 2005.
12. F.C. Freiling, T. Holz, and G. Wicherski, "Botnet Tracking: Exploring a Root-Cause Methodology to Prevent Distributed Denial-of-Service Attacks," Proc. 10th European Symp. Research in Computer Security (ESORICS), Sept. 2005.
13. The mstream distributed denial of service attack tool.http://staff.washington.edu/dittrich/misc/mstream.analysis.txt.
14. J. Oikarinen and D. Reed. RFC 1459: Internet Relay Chat Protocol, 1993.
15. St´ephane Racine. Analysis of Internet Relay Chat Usage by DDoS Zombies. Master's thesis, Swiss Federal Institute of Technology Zurich, April 2004.
16. The Honeynet Project. Know your enemy: rackingbotnets.http://www.honeynet.org/papers/bots/,March 2005.
17. Ying Xuan, Incheol Shin, My T. Thai, Member, and Taieb Znati, Member, Detecting Application Denial-of-Service Attacks: A Group-Testing-Based Approach IEEE transactions on parallel and distributed systems, vol. 21, no. 8, august 2010.

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Novel Advent for Add-On Security by Magic Square Intrication

By S. Praveen Kumar, K. Naveen Kumar, S. Sreenadh, B. Aravind, K. Hemnath Kumar

GITAM University

Abstract - The efficiency of a cryptographic algorithm is based on its time taken for encryption / decryption and the way it produces diverse cipher text from a clear text. The RSA, the extensively used public key algorithm and other public key algorithms may not guarantee that the cipher text is copiously secured. As an alternative approach to handling ASCII characters in the cryptosystems, a magic square implementation is deliberated of in this work. It attempts to augment the efficiency by providing add-on security to the cryptosystem. This approach will boost the security due to its complexity in encryption because it deals with the magic square. Here, encryption / decryption is based on numerals generated by magic square rather than ASCII values. In this, we add the ASCII value and the numeral in the consequent magic square. Because to encrypt the plaintext characters, their ASCII values are taken and if a character occurs in numerous places in a plaintext there is a possibility of same cipher text is produced. To surmount the problem, this paper attempts to develop a technique in which a constant is added for the recurring character.

Keywords : *ICryptography, Encryption, Decryption, Cipher text, Secret key, Public key, Steganography, Magic Square, ASCII value.*

GJCST Classification : *K.6.5*

Strictly as per the compliance and regulations of:

Novel Advent for Add-On Security by Magic Square Intrication

S. Praveen Kumar [α], K. Naveen Kumar [Ω], S. Sreenadh[β], B. Aravind [Ψ], K. Hemanth Kumar[¥]

Abstract - The efficiency of a cryptographic algorithm is based on its time taken for encryption / decryption and the way it produces diverse cipher text from a clear text. The RSA, the extensively used public key algorithm and other public key algorithms may not guarantee that the cipher text is copiously secured. As an alternative approach to handling ASCII characters in the cryptosystems, a magic square implementation is deliberated of in this work. It attempts to augment the efficiency by providing add-on security to the cryptosystem. This approach will boost the security due to its complexity in encryption because it deals with the magic square.

Here, encryption / decryption is based on numerals generated by magic square rather than ASCII values. In this, we add the ASCII value and the numeral in the consequent magic square. Because to encrypt the plaintext characters, their ASCII values are taken and if a character occurs in numerous places in a plaintext there is a possibility of same cipher text is produced. To surmount the problem, this paper attempts to develop a technique in which a constant is added for the recurring character.

Thus, instead of taking ASCII values for the characters to encrypt, preferably dissimilar numerals representing the position of ASCII values are taken from magic square. This proposed work provides another layer of security to any public key algorithms such as RSA, Elgamal etc., since, this model is acting as a wrapper to a public key algorithm, it ensures that the security is enhanced.

Keywords : Cryptography, Encryption, Decryption, Cipher text, Secret key, Public key, Steganography, Magic Square, ASCII value.

I. Introduction

Cryptography : Cryptography is the science of securing data. Classical cryptanalysis involves an appealing amalgamation of analytical interpretation, application of mathematical tools and pattern finding. The method of disguising plaintext in such a way as to hide its substance is called encryption. Encrypting plaintext results in unreadable gibberish called cipher text. The process of reverting cipher text to its original plaintext is called decryption.

The benefit of cryptography is that it works in a combination with a key ,a word number or phrase to encrypt the plain text. The same plain text encrypts to different cipher text with different keys. A key is a value that works with a cryptographic algorithm to fabricate a precise cipher text. The bigger the Key the more secure the cipher text.

Algorithms : There are two modules of key-based algorithms, symmetric (or secret-key) and asymmetric (or public-key) algorithms.

There is a problem of key distribution in the case of symmetric algorithms is solved with help of public key in asymmetric algorithms. The concept of Public key is using a pair of keys , a public key which encrypts data ,and a corresponding private key for decryption.

II. Literature Review

Steganography is the knack of writing concealed messages in such a way that no one, apart from the sender and intended recipient, suspects the existence of the message, a form of security through obscurity. DES is the archetypal block cipher — an algorithm that takes a fixed-length string of plaintext bits and transforms it through a series of complicated operations into another cipher text bit string of the same length.PGP is often used for signing, encrypting and decrypting texts, E-mails, files, directories and whole disk partitions to increase the security of e-mail communications. It was created by Philip Zimmermann in 1991.

Gopinanath Ganapathy, and K. Mani (2009) developed "Add-on Security Model for Public-key Cryptosystem based on Magic Square Implementation". This paper discusses how encryption and decryption is done by using RSA cryptosystem along with the elements of doubly even magic square.

R. L. Rivest (1996) developed "Hand book of Applied Cryptography". In this, emphasis is on those that are both (believed to be) secure and practically useful.

Objective : The objective of the method is to provide a secure way of sending a message using some standards and to overcome from man in the middle attack, spoofing etc.

Using Magic square for add-on security : In our approach , add on security is provided by dealing with magic square. We encrypt a file of any length using a magic square. We append each element of magic square to each character in the file until the end of the

Author [α] : IT department, GIT, GITAM University 09989590690.
E-mail : spkmtech@gmail.com
Author [Ω] : IT department, GIT, GITAM University 09963778239.
E-mail : nkumarkuppili@gmail.com
Author [β] : IT department, GIT, GITAM University 09491552723.
E-mail : Sreenadh.sadasivuni@gmail.com
Author [Ψ]: IT department, GIT, GITAM University 09676067261.
E-mail : aravind.barla22@gmail.com
Author [¥] : IT Dept, GIT, GITAM University 9494329294,
E-mail : hemanth903@gmail.com

file. The size of the file may surpass the number of elements in the magic square. If number of characters is auxiliary than the number of elements then we start totaling from the first element of the magic square. . We browse the file, which is to be encrypted and a password which is known only to sender and receiver. Encryption is done, and a copy of encrypted file is saved automatically. For decryption we browse the file and same password is given. A copy of decrypted (original) file is also saved. It is a reverse process of encryption in which instead of adding elements we will subtract elements. In this way, add-on security is provided by dealing with magic square.

Magic Square : A magic square of order a is an arrangement of a2 numbers, usually distinct integers, in a square, such that the a numbers in all rows, all columns, and both diagonals sum to the same constant. A normal magic square contains the integers from 1 to a2. The term "magic square" is also sometimes used to refer to any of various types of word square.

Given an $a \times a$ normal magic square, suppose S is the number that each row, column and diagonal must add up to. Then since there are a rows the sum of all the numbers in the magic square must be . But the numbers being added are 1, 2, 3, ... a^2 , and so 1 + 2 + 3 + ...+ a^2 = a x S. In summation notation, $\sum_{i=1}^{n^2} i = a.S$. Using the formula for this sum, we have $a.S = \frac{a^2(a^2+1)}{2}$ and then solving for S gives $S = \frac{a(a^2+1)}{2}$. Thus, a 3×3 normal magic square must have its rows, columns and diagonals adding to

$$S = \frac{3\,(3^2+1)}{2} = \frac{30}{2} = 15$$

In the same way for a 4×4 square S= 34, Benjamin Franklin's 8×8 to S = 260, and so on.

The magic sum for an $a \times a$ normal magic square can be found by filling the square with the numbers 1, 2, 3, ... a^2 . first going across the top row, then the second row, and so on -- and then adding the numbers along either of the diagonals. For instance, to find the magic sum of a4×4normal magic square, we form the following square:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

and then compute 1 + 6 + 11 + 16 = 34. See if you can figure out why this works for any $a \times a$ normal magic square. As defined above, a normal magic square uses the numbers 1, 2, 3 , . .. a^2 . Some people relax this restriction to permit any positive integers, calling the resulting square simply a magic square -- without the adjective "normal." It is easy obtain magic squares of this generalized type from an existing magic square. One way is simply to multiply every number used by some positive constant, and/or add a positive constant to every number used. It should be easy for you to see why this will always work.A more interesting problem, however, concerns the process of creating normal magic squares. For odd values of a, there is a simple procedure for constructing a normal magic square.

In our approach, we add each element of magic square to each character in the file until the end of the file. The size of the file may exceed the number of elements in the magic square. If number of characters is more than the number of elements then we start adding from the first element of the magic square.

Code snippet:

Encryption:

In the beneath code we can see the processing of the magic square and the encryption of the data by using it. First the magic square serial (password) is processed and then using the serial given the data is encrypted by totaling each element of the magic square to each character in the file until the end of the file.

```
for (int i = 2; i <= n*n; i++)
 if (magic[(row + 1) % n][(col + 1) % n] == 0)
row = (row + 1) % n; col = (col + 1) % n;
 else
 row = (row - 1 + n) % n;   // don't change col
 magic[row][col] = i;
for(int i=0;i<n;i++)
for(int j=0;j<n;j++)
c=fis.read();
if(c!=-1)
 s=(char)(c+magic[i][j]);
fos.write(s); l++;
if(l==(n*n))
i=-1; l=0;
fos.close(); fis.close();
 i=n; j=n;
```

Decryption:

The following code decrypts the cipher into Plain text.

During the process of decryption the magic square serial (password) is to be entered ,if the password matches with the password given during encryption then the cipher is converted into plain text.

```
for (int i = 2; i <= n*n; i++)
 if (magic[(row + 1) % n][(col + 1) % n] == 0)
row = (row + 1) % n; col = (col + 1) % n;
 else
 row = (row - 1 + n) % n;   // don't change col
 magic[row][col] = i;
for(int i=0;i<n;i++)
for(int j=0;j<n;j++)
c=fis.read()
if(c!=-1)
s=(char)(c-magic[i][j]);
fos.write(s) l++;
if(l==(n*n))
i=-1; l=0;
else
fos.close(); fis.close();
i=n; j=n;
```

III. Result

Choose encrypt or decrypt

Browse the file for Encryption and enter password (magic square sequence)

Encryption from plain text to cipher

Browse file for Decryption and enter the password specified during encyption

Decryption from cipher to plain text

IV. Conclusion & Future Enhancement

An alternative advance to existing ASCII based cryptosystem a number based approach is thought of an implemented. This methodology will add on one more layer of security, by adding numerals of the magic square for the text. It provides security to the files in PCs and also can be transferred through pen drives. We can apply this algorithm for any type of files like .txt, .doc, .jpeg etc. The size of the file is very small to carry and is simply a jar file.

Further we can feed into any public key algorithms like RSA, ElGamal etc. Thus it provides add on security to existing algorithms.

References References Referencias

1. Gopinanadh Ganapathi, and K.Mani, "Add on security model for public key Cryptosystem based on magic square implementation", India, 2009.
2. R. L. Rivest, "Hand book of Applied Cryptography", 1996.
3. Herbert Schildt, "The Complete Reference", Seventh Edition, Tata McGraw-Hill Publishing Company Limited.
4. Patrick Naughton, Michael Morrison, :The Java Handbook", Publisher: Osborne/ McGraw-Hill.
5. Benson, W.H. & Jacoby, O. ; New Recreations with Magic Squares. Dover Publications Inc. New York.
6. Gauthier, N. 'Singular matrices applied to 3x3 Magic Squares'. Math. Gazette.81, (1997).
7. Thompson, A.C. ; 'Odd Magic Powers'. American Math. Monthly, 101(4),(April 1994).
8. Ollerenshaw, K. & Bondi, H. ; 'Magic Squares of Order 4'. Phil. Trans. Royal Soc. London.
9. Dudeney, H.E. ; The Canterbury Puzzles (and other curious problems). T. nelson & sons Publishers. 2nd Edition. 1927.
10. Gardner, M. 'Mathematical games: a breakthrough in magic squares and the first perfect magic cube.' Scientific American.
11. Fletcher, T.J.; Linear Algebra through its Applications. Van Nostrand Reinhold,New York. 1972.
12. Ward, James E. III; 'Vector spaces of Magic Squares.', Math. Magazine .
13. van den Essen, A. ; 'Magic Squares and Linear Algebra', American Math.Monthly.
14. Heinrich, C.J.; 'Magic Squares and Linear Algebra', American Math. Monthly.
15. Leinbach C. & Pountney D.C. 'Appropriate use of Computer Algebra Systems in Teaching Mathematics' Pennsylvania Council of Teachers of Mathematics.

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Multi-Sensor Image Fusion for Impulse Noise Reduction in Digital Images

By M.Premakumar, N.Sowjanya, Dr.P.RajeshKumar

Andhra University

Abstract - This paper introduces the concept of Multi-sensor image fusion technique for impulse noise reduction in digital images. Image fusion is the process of combining two or more images into a single image while retaining the important features of each image. Multiple sensor image fusion is an important technique used in military, remote sensing and medical applications. The images captured by five different sensors undergo filtering using five different vector median filtering algorithms and the filtered images are fused into a single image, which combines the uncorrupted pixels from each one of the filtered image. The fusion algorithm is based on quality assessment of the spatial domain from the individual de-noised images. The performance evaluation of our algorithm is evaluated using PSNR between original image and individually filtered and the fused image. Experimental results show that this fusion algorithm produce a high quality image compared to individually de-noised images.

Keywords : *Image Fusion, Image Restoration, Image Processing, Impulse Noise.*

GJCST Classification : *I.4.3*

Strictly as per the compliance and regulations of:

Multi-Sensor Image Fusion for Impulse Noise Reduction in Digital Images

M.Premakumar[α], N.Sowjanya[Ω], Dr.P.RajeshKumar[β]

Abstract - This paper introduces the concept of Multi-sensor image fusion technique for impulse noise reduction in digital images. Image fusion is the process of combining two or more images into a single image while retaining the important features of each image. Multiple sensor image fusion is an important technique used in military, remote sensing and medical applications. The images captured by five different sensors undergo filtering using five different vector median filtering algorithms and the filtered images are fused into a single image, which combines the uncorrupted pixels from each one of the filtered image. The fusion algorithm is based on quality assessment of the spatial domain from the individual de-noised images. The performance evaluation of our algorithm is evaluated using PSNR between original image and individually filtered and the fused image. Experimental results show that this fusion algorithm produce a high quality image compared to individually de-noised images.

Keywords : Image Fusion, Image Restoration, Image Processing, Impulse Noise.

I. Introduction

Digital images are often corrupted during acquisition, transmission or due to faulty memory locations in hardware [1]. The impulse noise can be caused by a camera due to the faulty nature of the sensor or during transmission of coded images in a noisy communication channel [2]. Consequently, some pixel intensities are altered while others remain noise free. The noise density (severity of the noise) varies depending on various factors namely reflective surfaces, atmospheric variations, noisy communication channels and so on. In most image processing applications the images captured by different sensors are combined into a single image, which retains the important features of the images from the individual sensors, this process is known as image fusion[3][4]. In this paper, the images captured by multiple (five) sensors are differently noised depending on the proximity to the object, environmental disturbances and sensor features. These noise images are filtered using five different vector median filtering algorithms such as Vector Median Filter, Rank Conditioned Vector Median Filter, Rank Conditioning and Threshold Vector Median Filter, Center Weighted Vector Median Filter and Absolute Deviation Vector Median Filter. The filtered images are fused into a single image using the quality assessment of spatial domain from the de-noised images, thus producing a high quality image. The performance evaluation of the image fusion is evaluated using PSNR between the original and fused image. This paper is organized as follows: Section II present the impulse noise in images, Section III present different filtering algorithms, Section IV present the image fusion algorithm, Section VI present experimental results and finally Section VII reports conclusion.

Author α Ω : Department of ECE, Sri Vishnu Engineering College for Women, Bhimavaram.
Author β : Associate professor, Department of ECE A.U. College of Engineering, Andhra University.

II. Impulse Noise in Images

Impulse noise [5] corruption is very common in digital images. Impulse noise is always independent and uncorrelated to the image pixels and is randomly distributed over the image. Hence unlike Gaussian noise, for an impulse noise corrupted image all the image pixels are not noisy, a number of image pixels will be noisy and the rest of pixels will be noise free. There are different types of impulse noise namely salt and pepper type of noise and random valued impulse noise.

In salt and pepper type of noise the noisy pixels takes either salt value (gray level -225) or pepper value (grey level -0) and it appears as black and white spots on the images. If p is the total noise density then salt noise and pepper noise will have a noise density of $p/2$. This can be mathematically represented by (1)

$$y_{ij} = \begin{cases} \text{zero or 255 with probability } p \\ x_{ij} \text{ with probability } 1\text{-}p \end{cases} \quad (1)$$

Where y_{ij} represents the noisy image pixel, p is the total noise density of impulse noise and xij is the uncorrupted image pixel.

In case of random valued impulse noise, noise can take any gray level value from zero to 225. In this case also noise is randomly distributed over the entire image and probability of occurrence of any gray level value as noise will be same. We can mathematically represent random valued impulse noise as in (2).

$$y_{ij} = \begin{cases} n_{ij} \text{ with probability } p \\ x_{ij} \text{ with probability } 1\text{-}p \end{cases} \quad (2)$$

Where n_{ij} is the gray level value of the noisy pixel.

III. Fltering Algorithms

In the **Vector median filter** (VMF) [6] for the ordering of the vectors in a particular kernel or mask a

suitable distance measure is chosen. The vector pixels in the window are ordered on the basis of the sum of the distances between each vector pixel and the other vector pixels in the window.

The sum of the distances is arranged in the ascending order and then the same ordering is associated with the vector pixels. The vector pixel with the smallest sum of distances is the vector median pixel. The vector median filter is represented as

$$X_{VMF} = \textit{vectormedian}\ (\text{window}) \tag{3}$$

If δ_i is the sum of the distances of the i^{th} vector pixel with all the other vectors in the kernel, then

$$\delta_i = \sum_{j=1}^{N} \Delta(X_i, X_j \tag{4}$$

where $(1 \leq i \leq N)$ and X_i and X_j are the vectors, N=9.

$\Delta(X_i, X_j)$ is the distance measure given by the L_1 norm or the city block distance which is more suited to non correlated noise. The ordering may be illustrated as

$$\delta_1 \leq \delta_2 \leq \delta_3 \leq ...,\leq \delta \tag{5}$$

and this implies the same ordering to the corresponding vector pixels i.e.

$$X_{(1)} \leq X_{(2)} \leq ...,\leq X_{(9)} \tag{6}$$

where the subscripts are the ranks. Since the vector pixel with the smallest sum of distances is the vector med ian pixel, it will correspond to rank 1 of the ordered pixels, i.e.,

$$X_{VMF} = X_{(1)} \tag{7}$$

The *Rank Conditioned Vector Median Filter* [7] improves the performance of the vector median filter. The vector median of the kernel replaces the central pixel when the rank of the central pixel is greater than a predefined rank of a healthy vector pixel inside the window. The rank of the healthy vector pixel is obtained by simulating RCVMF codeon a noiseless image. Then, the mean value of the obtained ranks of the central vector pixel is calculated. This value is rounded off to a whole number, and it is considered to be rank of the healthy vector pixel of a kernel.The rank conditioned vector median filter can be expressed as:

$$\begin{aligned} X_{RCVMF} &= X_{VMF}, & \text{if } r_c > r_k \\ &= X_C & \text{otherwise} \end{aligned} \tag{8}$$

Where r_c is a rank of the central vector pixel and c=5.and r_k is the predefined healthy vector pixel rank inside the window.

The **Rank-Conditioned and Threshold Vector Median Filter** [8] aims to further enhance the RCVMF by incorporating an additional test – a distance threshold for the detection of impulses. In RCVMF, a central vector having greater than the predefined rank implies a corrupt vector. However, it may not be true always, because the vectors may be close as per the distance measure. Hence another criterion Θ, is taken into account. It is the distance between the central vector pixel and the vector pixel corresponding to the predefined rank. To find out the value of this pre-determined distance threshold Θ, the code simulating RCTVMF is executed on a noiseless image. Then the mean of the obtained θ values is calculated and used for the simulations at various noise percentages. The distance is calculated as follows :

$$D = \Delta(X_c, X_{(k)}) \tag{9}$$

Where Xc is the central vector and $X_{(k)}$ $(1 < k < 9)$ is a rank ordered and healthy vector pixel inside the window. On the basis of the above information, the filter has the following form :

$$\begin{aligned} X_{RCTVMF} &= X_{VMF}, & \text{if } r_c > r_k \text{ and } D > \Theta \\ &= X_c & \text{otherwise} \end{aligned}$$

In **Center Weighted Median Filter** [8] the kernel vector pixels are assigned some non negative values called weights. The central vector pixel is assigned a non negative weight while the weight of the neighboring pixels is kept unity. The weights denote the number of copies is obtained. The output Y (say), of a weighted median filter of span N (where N generally denotes the kernel size, N=9) associated with N integer weights,

$$W=[W_1.W_2.....W_N] \tag{10}$$

is given by ,

$$Y = \text{vectormedian}\ [W_1*X_1, W_2*X_2, ...,W_9*X_9]$$

Where the vectormedian [.] denotes the vector median operation.

In **Absolute Deviation Vector Median Filter** [9], the impulse noise detection mechanism does not require the distance calculation and subsequent ordering of the vectors of a kernel. The algorithm deals with the difference values of the red (R) and the green (G) intensities denoted by Ω_{RGi} (say), and the difference values of the green (G) and blue (B) intensities denoted by Ω_{GBi} (say), (where $1 \leq i \leq N$, N=9).

In a 3×3 kernel, it has been observed empirically that Ω_{RGi} and Ω_{GBi} values closely correspond to each other. Thus the mean absolute deviation $\mathbf{D'}_{RG}$ and $\mathbf{D'}_{GB}$ i.e. the mean of $\mathbf{D}_{RGj}$ and $\mathbf{D}_{GBj}$ (where $1 \leq j \leq N$, and $j \neq c$, c = (N+1)/2, N = 9) has small values. $\mathbf{D}_{RGj}$ and $\mathbf{D}_{GBj}$ are the absolute deviation values of Ω_{RGj} and Ω_{GBj} from Ω'_{RG} and Ω'_{GB} respectively. Ω'_{RG} and Ω'_{GB} denote the mean of Ω_{RGj} and Ω_{GBj} (where $1 \leq j \leq N$, and $j \neq c$, c = (N+1)/2, N = 9).

The absolute deviation of the central vector Ω_{RGc} and Ω_{GBc} values from Ω'_{RG} and Ω'_{GB} is obtained respectively as $\mathbf{D}_{RGc}$ and $\mathbf{D}_{GBc}$.

If the absolute deviation $\mathbf{D}_{RGc}$ and $\mathbf{D}_{GBc}$ of the central vector pixel exceeds the value of $\mathbf{D'}_{RG}$ or $\mathbf{D'}_{GB}$ respectively for a 3X3 kernal, the central vector pixel is to be replaced by the vector median of the kernel. The algorithm is represented as follows.

In a 3X3 kernel,

Step 1: Find the difference values of red(R) and the green(G) intensities denoted by Δ_{RGi} and the difference values of the green(G) and blue(B) intensities denoted by Δ_{GBi} .

$$\Delta_{RGi} = X(i, R) - X(i,G)$$
$$\Delta_{GBi} = X(i,G) - X(i,B) \qquad \text{Where } i= 1\ldots9 \tag{11}$$

Step 2: Calculate the mean of Δ_{RGi} and Δ_{GBi} denoted by Δ^1_{RGi} and Δ^1_{GBi}. Where i=1,2…9 and i≠5.

Step 3: Calculate the absolute deviation between Δ and Δ^1.

$$D_{RGi} = | \Delta_{RGi} - \Delta^1_{RGi} |$$
$$D_{GBi} = | \Delta_{GBi} - \Delta^1_{GBi} | \quad \text{where } i= 1,2,\ldots9 \tag{12}$$

Step 4 : Calculate the mean of D_{RGi} and D_{GBi} denoted as D^1_{RGi} and D^1_{GBi}. where i=1,2….9 and i≠5.

Step 5 : Now

$$D_{RGc} = | \Delta_{RGc} - \Delta^1_{RGi} |$$
$$D_{GBc} = | \Delta_{GBc} - \Delta^1_{GBi} | \quad \text{where } i= 1,2,\ldots9,\ i\neq 5 \text{ and c is the central vector .} \tag{13}$$

Step 6 : If $D_{RGc} > D^1_{RGi}$ or $D_{GBc} > D^1_{GBi}$ where i=1,2…9 and i≠5. central vector is corrupted, hence central vector is replaced by Vector Median of kernel.

IV. Image Fusion Algorithm

The block diagram for multi-sensor image fusion is shown in figure 1. The algorithm for the multi-sensor image fusion using quality assessment of spatial domain is as follows:

The images captured by different sensors are filter using five different filtering algorithms. These five filtered images are fused into a single image having all objects in focus without producing details that are non-existent in the given images. The algorithm consists of the following steps:

1. Let $I^1, I^2, \ldots\ldots, I^5$ be the noisy images of an object or scene captured by sensors $S^1, S^2, \ldots\ldots, S^5$ respectively. Let I^i be of size NXN where $i = 1,2,\ldots,5$.
2. Filter the noisy images using five different filtering algorithms. The filtered images are denoted as R^i.
3. The recovered images R^i for $i=1,2,..,5$ are divided into non-overlapping rectangular blocks (or regions) with size of mxn . The j^{th} image blocks of R^i are referred by R^i_j
4. Quality assessment value (λ) of R^i_j is calculated and the results of R^i_j are denoted by λ^i_j. Quality Assessment value λ is given by λ^i_j.

Quality Assessment value λ is given by

$$\lambda=\lambda_1-\lambda_2 \tag{14}$$

where

$$\lambda_1=\frac{1}{2}\left(trace(J)+\sqrt{trace(J)-4\det(J)}\right)$$
$$\lambda_2=\frac{1}{2}\left(trace(J)-\sqrt{trace(J)-4\det(J)}\right) \tag{15}$$

The covariance matrix of the gradient vectors for all b^2 sites in this block is given by

$$J=\frac{1}{b^2}\sum_{s\in B} g_s^T g_s = \begin{bmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{bmatrix} \tag{16}$$

In order to determine the sharper image block, the quality assessment value of image blocks from 5 recovered images are sorted in descending order and the same ordering is associated with image blocks. The block with the maximum quality assessment is kept in the fused image. The fusion mechanism is represented as follows:

If λ^i_j is the quality assessment value of block R^i_j, the ordering of assessment values is given by

$$\lambda_{(1)} > \lambda_{(2)} > \ldots\ldots > \lambda_{(5)} \tag{17}$$

and this implies the same ordering to the corresponding blocks

$$R_{(1)} > R_{(2)} > \ldots > R_{(5)}$$

Where the subscripts are the ranks of the image blocks. Since the block with the largest quality assessment value is in the fused image, it will correspond to rank 1 of the ordered blocks ie;

$$Fused\ Block = R_{(1)} \tag{18}$$

V. Experimental Results

The proposed method of image fusion for impulse noise reduction in images was tested on the true color remote sensing image with 290x290 pixels. The images are captured by five different sensors with different noise densities. The noisy images are filtered using five different vector median filtering algorithms. The filtered images are fused into a single image using the Image fusion method based on the quality assessment in spatial domain.. The experimental results are shown in Figure 2. Table (1) shows the PSNR value of fused image with different noise densities of input images with respect to original image.

VI. Conclusion

This paper presents the multi-sensor image fusion method for impulse noise reduction in digital images. This technique can be used in military, remote sensing and medical applications. The experimental results show that our fusion algorithm works better in removal of impulse noise in digital images. The proposed method is simple and can be used for realtime imaging applications.

References Références Referencias

1. Tao Chen, Kai-Kaung Ma and Li-Hui Chen, "Tri-state median filter for image Denoising", IEEE Transactions on Image Processing, Vol 8, no.12, pp.1834-1838, December 1999.
2. Reihard Berstein," Adaptive nonlinear filters for simultaneous removal of different kinds of noise in images," IEEE Trans on circuits and systems , Vol.cas-34, no 11,pp.127-1291, Nivember 1987.
3. O.Rockinger," Image sequence fusion using ashift invariant wavelet transform," IEEE transactions on image processing, 3:288- 291, 1997.
4. Chaveli Ramesh and T.ranjith, "Fusion performance measures and a lifting wavelet transform based algorithm for image fusion", In Proc. 5th International conference on image fusion, july 2002,pp 317-320.
5. S.Indu, Chaveli Ramesh, " A noise fading technique for images corrupted with impulse noise", Proceedings of ICCTA07, IEEE.
6. J.Astola, P.Haavisto and Y.Neuro," Vector Median Filters", Proc. of IEEE, vol. 78. No.4. pp. 687-689, April 1990.
7. LL. Alparone, S. Baronti and R. Carla, "Express Letters - Two- Dimensional Rank-Conditioned Median Filter," IEEE Trans. On Circuits and Systems-II: Analog and Digital Signal Processing, vol. 42, no. 2, pp. 130- 132 , February 1995.
8. K. M. Singh and P. K. Bora, "Adaptive Vector Median Filter for Removal Impulses from Color Images," Proc. of Int.Symposium on Circuits and Systems, vol. 2, pp. II-396 - II-399, May 2003.
9. R.H.Laskar, B.Bhoowmick, R.Biswas and S.Kar, " Removal of Impulse Noise from Color Image", TENCON IEEE 2009.

Block Diagram of Multi Sensor Image Fusion :

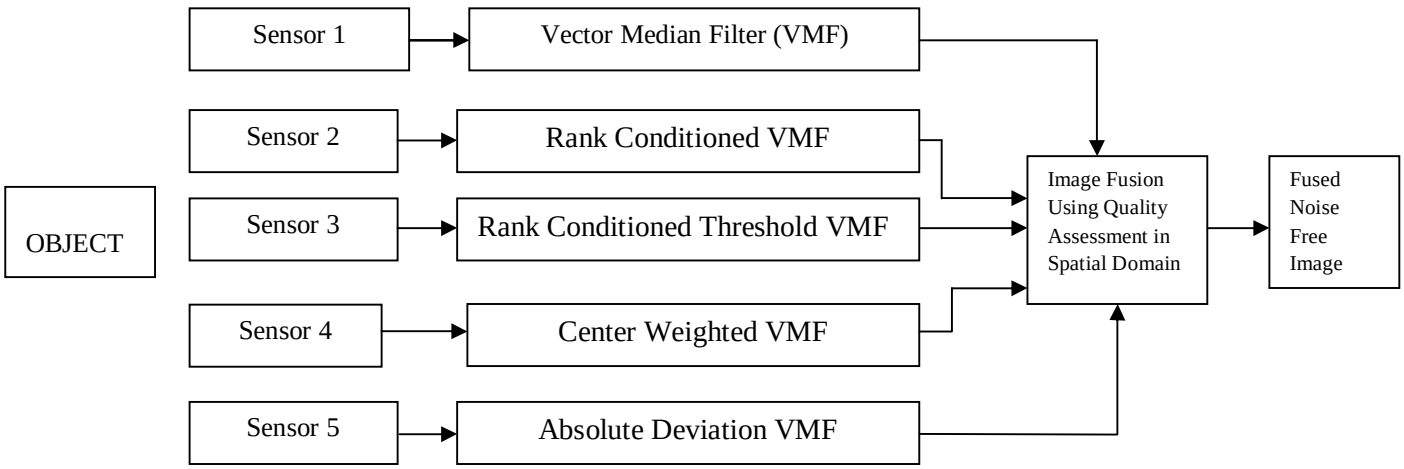


Figure 1 : Block Diagram of multisensor Image fusion

Multi sensor Image Fusion:

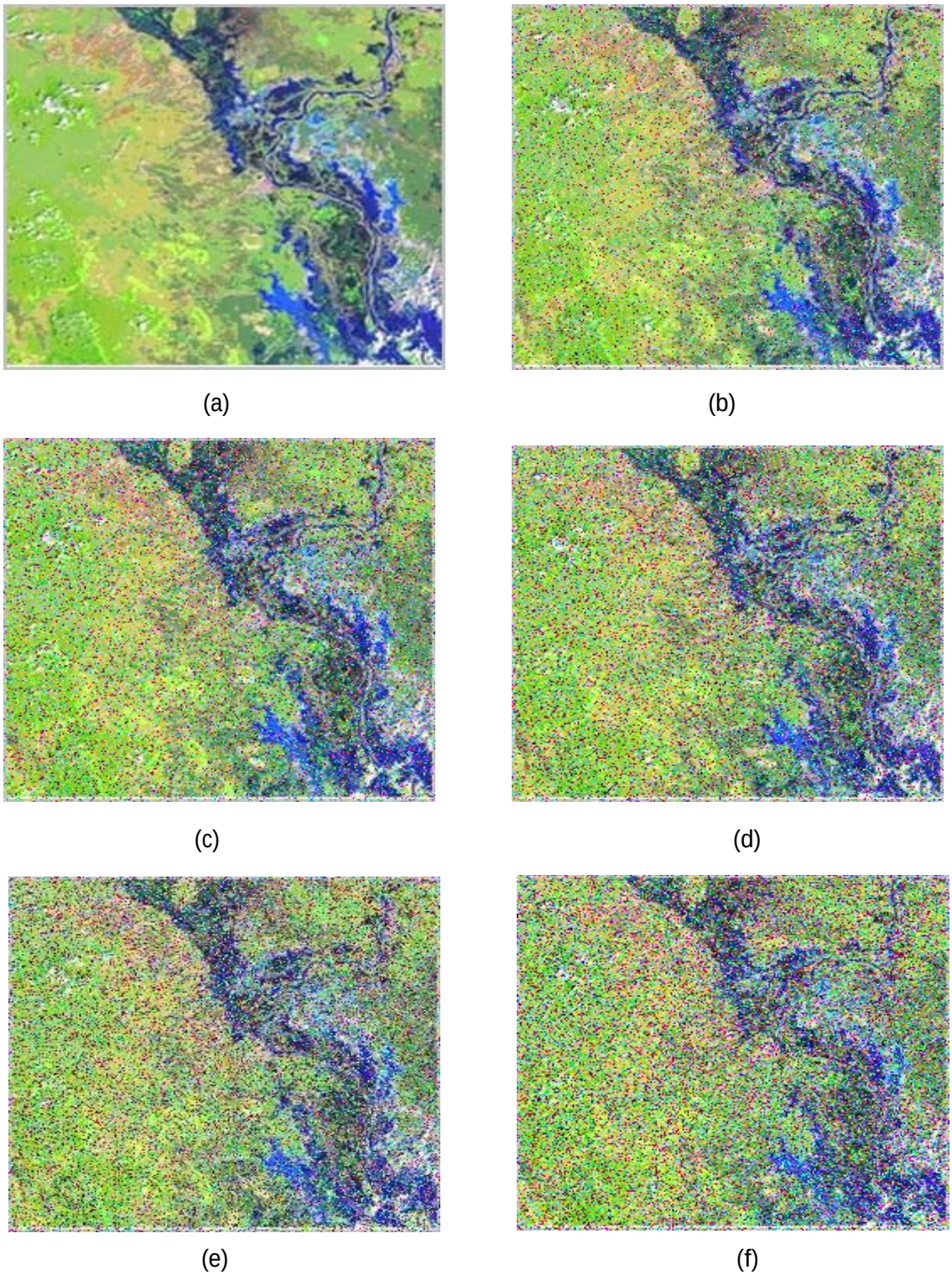
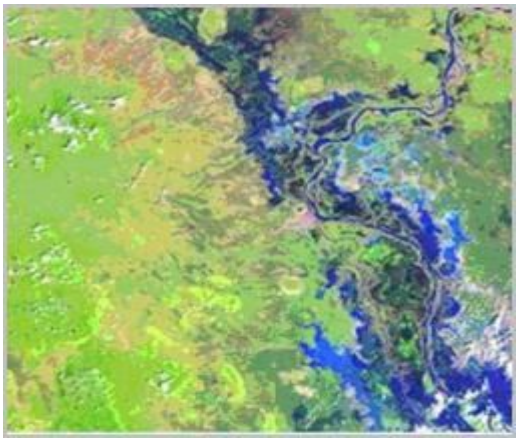
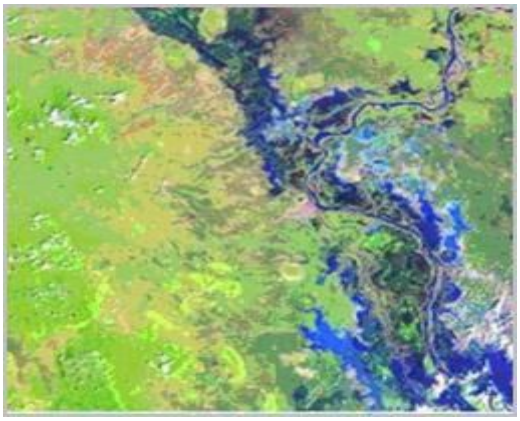


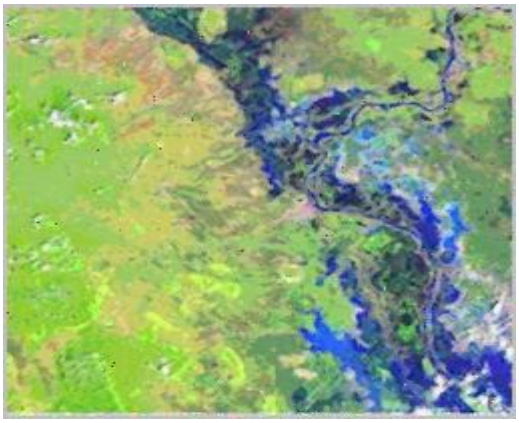
Figure 2(a) : River Image captured by five sensors with impulse noise. (a) Original Image (b) 10% Noise (c) 15% Noise (d) 20% Noise (e) 25% Noise (f)30% Noise



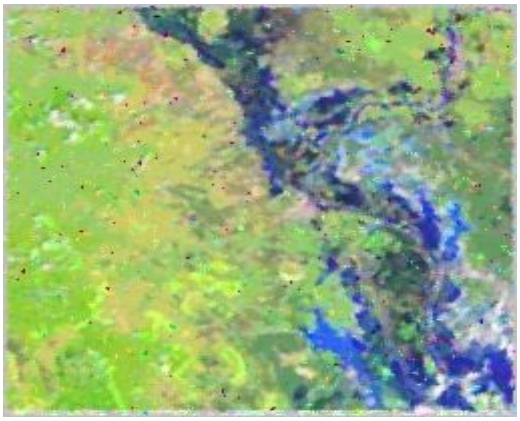
(a)



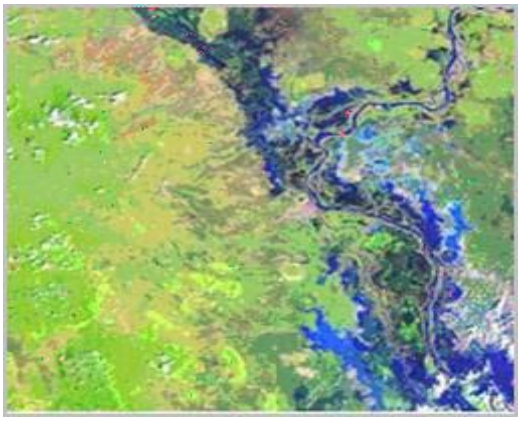
(b)



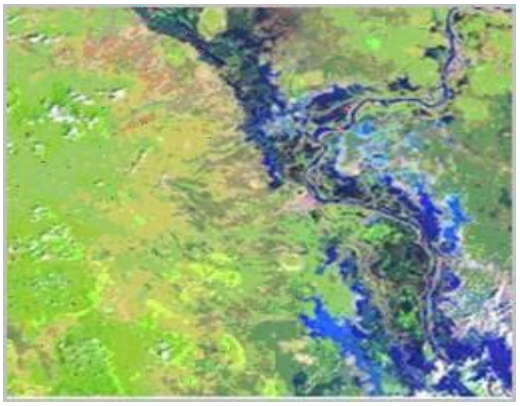
(c)



(d)



(e)



(f)

Figure 2(b) : Images filtered by different Filters. (a) Vector Median Filter (b) Rank conditioned Vector Median Filter (c) Rank conditioned & threshold Vector Median Filter (d) Center weighted Vector Median Filter (e) Absolute Deviation Vector Median Filter (f) Fused Image

Table 1: comparison of performance of the different filters and fused image with respect to original image.

Noise %	Filter	PSNR
10	Vector median filter(VMF)	31.25
15	Rank conditioned VMF	30.522
20	Rank conditioned & threshold VMF	27.961
25	Center weighted VMF	22.73
30	Absolute deviation VMF	29.46
	Fused Image	33.6

This page is intentionally left blank

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Safety Critical Systems Analysis

By K. Amarendra, A. Vasudeva Rao
Dadi Institute of Engineering & Technology Visakhapatnam Dt, India

Abstract - A brief overview of the fields that must be considered when designing, implementing safety-critical systems is presented. The notion of safety is most likely to come to mind when we drive a car, fly on an airliner, or take an elevator ride. In each case, we are concerned with the threat of a mishap, which defined as an unplanned event or series of events that result in death, injury, occupational illness, damage to or loss of equipment or property or damage to the environment.

Keywords : *Safety, Design, Implementation, Applications.*

GJCST Classification : *K.6.5*

Strictly as per the compliance and regulations of:

Safety Critical Systems Analysis

K. Amarendra [α], A. Vasudeva Rao [Ω]

Abstract- A brief overview of the fields that must be considered when designing, implementing safety-critical systems is presented. The notion of safety is most likely to come to mind when we drive a car, fly on an airliner, or take an elevator ride. In each case, we are concerned with the threat of a *mishap*, which defined as an unplanned event or series of events that result in death, injury, occupational illness, damage to or loss of equipment or property or damage to the environment.

Keywords: *Safety, Design, Implementation, Applications,*

I. Introduction

A safety critical system is a system where human safety is dependent upon the correct operation of system. Safety is considered not only for software elements but also for hardware, electrical hardware, operators or users etc. If the failure of a system could lead to consequences that are determined to be unacceptable then the system is safety critical.

Safety-critical systems, a term whose customary meaning is systems whose failure might danger human life, lead to substantial economic loss, or cause extensive environmental damage. Many modern systems depend on computers for their correct operation. The future is likely to increase dramatically the number of computer systems that we consider to be safety-critical. The dropping cost of hardware, the improvement in hardware quality, and other technological developments ensure that new applications will be sought in many domains.

Traditional Systems

Traditional areas that have been considered the home of safety-critical systems include medical care, commercial aircraft, nuclear power, and weapons. Failure in these areas can quickly lead to human life being put in danger, loss of equipment, and so on. Computerized equipment is making inroads in procedures such as hip replacement, spinal surgery, and ophthalmic surgery. In all three of these cases, computer controlled robotic devices are replacing the surgeons traditional tools, and providing substantial benefits to patients.

Non Traditional Systems

The scope of the safety-critical system concept is broad, and that breadth has to be taken into account when practitioners and researchers deal with specific systems. Some of the examples of Non Traditional systems are transportation control, banking and financial systems, electricity generation and distribution, telecommunications, and the management of water systems. All of these applications are extensively computerized, and computer failure can and does lead to extensive loss of service with consequent disruption of normal activities.

Separating safety-critical and safety-related systems from systems where safety integrity is unable to be established or maintained is an important aspect of system safety design. When implementing a system safety program it is important to ***suspect all*** components as being unsafe unless assured otherwise and then ***target the few*** areas where safety requirements are allocated. Coupling between components of complex systems can be subtle and interaction with non-safety related systems have led to harmful outcomes in safety related systems.

Traditional Areas

Traditional areas that have been considered the home of safety-critical systems include medical care, commercial aircraft, nuclear power, and weapons. Failure in these areas can quickly lead to human life being put in danger, loss of equipment and so on. Computers are used in medicine far more widely than most people realize. The idea of using a microprocessor to control an insulin pump is quite well known. The fact that a pacemaker is largely a computer is less well known. The extensive use of computers in surgical procedures is almost unknown except by specialists. Computerized equipment is making inroads in procedures such as hip replacement, spinal surgery, and ophthalmic surgery. In all three of these cases, computer controlled robotic devices are replacing the surgeons traditional tools, and providing substantial benefits to patients.

Non Traditional Areas

The scope of the safety-critical system concept is broad, and that breadth has to be taken into account when practitioners and researchers deal with specific systems. A closer examination of the topic reveals that many new types of system have the potential for very high consequences of failure, and these systems should probably be considered safety-critical also. It is obvious that the loss of a commercial aircraft will probably kill people. It is not obvious that loss of a telephone system could kill people.

Author α : Associate Professor & Head, Department of Computer Science & Engineering,Dadi Institute of Engineering & Technology, Anakapalle – 531002, Visakhapatnam Dt, India.
E-mail : hodcse@dietakp.com
Author Ω : Associate Professor, Department of Computer Science & Engineering,Dadi Institute of Engineering & Technology, Anakapalle – 531002, Visakhapatnam Dt, India .
E-mail : vasudevarao@dietakp.com

Emergency service is an example of a critical infrastructure application. Other examples are transportation control, banking and financial systems, electricity generation and distribution, telecommunica - tions, and the management of water systems. All of these applications are extensively computerized, and computer failure can and does lead to extensive loss of service with consequent disruption of normal activities. In some cases, the disruption can be very serious. Widespread loss of water or electricity supply has obvious implications for health and safety. Similarly, widespread loss of transportation services, such as rail and trucking, would affect food and energy distribution. It is prudent to put the computer systems upon which critical infrastructures depend into the safety-critical category.

II. Safety Boundaries

Functional Safety Boundaries

a) The need for having boundaries

Taking the extreme position, very few systems are fully independent in their operation and to be completely assured of the absence of interaction or common cause failure between the safety-related and other systems would take an inordinate amount of time and effort. This could cause the opposite effect to delay introducing the safety benefits of the deployment of a safety-related system. At some point a determination must be made that all possible influences are controlled or risks sufficiently known so the safety analysis can be bounded.

b) Objectives of functional safety boundaries

Minimise the interfaces across the safety boundary to direct focus on the safety separation implemented in these;

Minimise likelihood of common-cause failures across the boundary;

Exclude non safety related functions where these are volatile or subject to undefined or non-safety related controls;

Allow a Safety Integrity Level (SIL) to be achieved within the boundary.

c) Identifying safety functions

A useful method to establish the functional safety boundary between systems or subsystems is to undertake a Fault Tree Analysis (FTA) of the contributing factors to failure of the system, which may lead to hazardous events identified in the preliminary hazard analysis. The first attempt at a boundary would be around the systems that are implicated in the FTA. This FTA needs to be extensive and complete from all initiating situations to the system failure that is a casual factor for the hazardous event. Then flowing down the tree, mark off those functions that are related to systems that should be excluded due to:

- The possibility of common-cause failure;
- High levels of complexity and non-deterministic Failure rate; or
- Components that may not always be present or enabled.

d) The problem with software

At a system level, this process looks reasonably straightforward but the problem comes with setting boundaries with distributed software architectures. In this situation it is very difficult to identify boundaries that don't involve the possibility of common-cause failures.

Common-cause failures and dependencies extending over the distributed communication networks must also be considered and the functional safety boundary set accordingly. These may include:

- Global variables accessed by network
- Security attack and security blocking issues
- Affects of network lock-up on functional safety

The separation requirements over the functional safety boundary must take these failures into account.

III. Disciplines

The criterion used is that these disciplines are at the heart of the safety-critical electronic and information technology components of modern vehicles. Safety-critical systems have many requirements that stem from several engineering disciplines. The main disciplines having a direct bearing on designing safety critical systems are: domain engineering, embedded systems engineering, protocol and network engineering, safety engineering, reliability engineering, real-time systems engineering, and systems engineering.Currently, several design and implementation options are available to a researcher, developer, or designer. In terms of protocols, one can choose among CAN, TTCAN, Switched Ethernet, TTP/C, Flex Ray and others. Because of cost, flexibility, the intended application theoretical, advances implementation technology, and other issues, it is not straightforward to decide what protocol or network technology is the best.

a. *Domain Engineering :* Safety-critical systems exist in a certain application context. Certainly the details of safety-critical aerospace systems are different from those of the space shuttle, process control, or automotive. It is important that it can be used to tune or optimize certain mechanisms (e.g., communications, fault tolerance, fail status, etc)
b. *Embedded System Engineering : S*afety-critical systems are embedded systems such as micro-controllers; real-time operating systems, memory configurations, and I/O are relevant.
c. *Protocol and Network Engineering :* Protocols and networking are at the heart of distributed safety-

critical systems. The degree of flexibility offered by the protocol is in order to experiment with and provide higher layer protocols (HLP). Still another issue is the application of inter-networking (using bridges, switches, and routers) at the vehicle level.

d. *Safety Engineering :* system (availability) reliability deals with the problem of ensuring that a system performs a required task or mission (at) for a specified time. *System safety* is concerned with ensuring that a *mishap* does not occur in the process. Usually, there are some failures exits like *benign failures and catastrophic failures.*
e. *Reliability Engineering :* It deals with the available operation of a system even under the failure of system components. The primary mechanism is the use of redundant components to design fault tolerant systems. There are two schemes to handle the replacement of failed components. They are *static and dynamic redundancy.*
f. *Real time Engineering :* Techniques for ensuring that a system meet timeliness requirements are important for safety-critical applications. A distinction is made between hard real-time and soft real-time systems. Safety-critical systems certainly belong to the category of hard real-time systems. To see if a system meets real-time requirements, schedulability analysis is used and this methodology is well known for single-processor or multiprocessor operating systems.
g. *Systems Engineering :* System engineering emphasizes formal processes that start with a system's requirements and specification, and includes an iterative design, test, and verification cycle.

IV. Applications

Safety critical systems are whose failure results in loss of life, property damage or damage to the environment. There are many well known examples in application areas such as medical devices, aircraft flight control weapons and nuclear systems.

Example of a safety critical system is an aircraft fly by wire control system, where the pilot inputs commands to the control computer using a joystick, and the computer manipulates the actual aircraft controls. The lives of hundreds of passengers are totally dependent upon the continued correct operation of such a system.

Moving down to earth, railway signalling systems must enable controllers to direct trains, while preventing trains from colliding. Like an aircraft fly by wire, lives are dependent upon the correct operation of the system. However, there is always the option of stopping all trains if the integrity of the system becomes suspect. You can't just stop an aircraft while the fly by wire system is fixed!

Software in medical systems may be directly responsible for human life, such as metering safe amounts of X-rays. Software may also be involved in providing humans with information, such as information which a doctor uses to decide on medication. Both types of system can impact the safety of the patient.

Big civil engineering structures are designed on computers and tested using mathematical models. An error in the software could conceivably result in a bridge collapsing. Aircraft, trains, ships and cars are also designed and modelled using computers.

Even something a simple as traffic lights can be viewed as safety critical. An error giving green lights to both directions at a cross road could result in a car accident. Within cars, software involved in functions such as engine management, anti-lock brakes, traction control, and a host of other functions, could potentially fail in a way which increases the likelihood of a road accident.

V. Challenges

In one way or another, many people in the software business are working on safety-critical systems technology. Many more systems than one might expect have to be viewed as safety-critical and the number is increasing all the time. So what are the major challenges that we face?

In some cases, what amount to completely new technologies is required? The number of interacting safety-critical systems present in a single application will force the sharing of resources between systems. This will eliminate a major architectural element that gives confidence in correct operation—physical separation. Knowing that the failure of one system cannot affect another greatly facilitates current analysis techniques. This will be lost as multiple functions are hosted on a single platform to simplify construction and to reduce power and weight requirements. Techniques that provide high levels of assurance of non-interference will be required.

Breakdowns in the interplay between software engineering and systems engineering remains a significant cause of failures. It is essential that comprehensive approaches to total system modelling be developed so that properties of entire systems can be analysed. Such approaches must accommodate software properly and provide high fidelity models of critical software characteristics. They must also deal with the issue of assured non-interference.

Defective software specifications are implicated in many serious failures, and it is clear that we have difficulty stating exactly what software is required to do. There are many aspects of specification that are not supported by any current technique, and, even where specification techniques do exist, there remains a lack of integration to permit whole specification analysis.

VI. Designing

The design of any safety critical system must be as simple as possible taking no unnecessary risks.

Software point of view, this usually involves minimizing the use of interrupts and minimizing the use of concurrency within the software.

Ideally, a safety critical system requiring a high integrity level would have no interrupts and only one task. However, this is not achievable in practice.

There are two distinct philosophies for the specification and design of safety critical systems.

- To specify and design a "perfect" system, which cannot go wrong because there are no faults in it, and to prove that there are no faults in it.
- To aim for the first philosophy is to accept that mistakes may have been made, and to include error detection and recovery capabilities to prevent errors from actually causing a hazard to safety.

The first of these approaches can work well for small systems, which are sufficiently compact for formal mathematical methods to be used in the specification and design, and for formal mathematical proof of design correctness to be established.

The second philosophy, of accepting that no matter how careful we are in developing a system, that it could still contain errors, is the approach more generally adopted. This philosophy can be applied at a number of levels:

- Within a routine, to check that inputs are valid, to trap errors within the routine, and to ensure that outputs are safe;
- Within the software, to check that system inputs are valid, to trap errors within the Software, and to ensure that system outputs are safe;
- Within the system, as independent verification that the rest of the system is behaving correctly, and to prevent it from causing the system to become unsafe;

The safety enforcing part is usually referred to as an interlock or protection subsystem. Designing safety-critical systems is a complex endeavour particularly if extensive use of advanced electronics and information technology is used. The increased use of microcontrollers in modern automotive systems has brought many benefits such as the merging of chassis control systems for active safety with passive-safety systems. Unfortunately, it has also brought the potential for catastrophic failures. Thus, the widespread application of microcontrollers requires extreme care in order to produce a dependable system. Dependability involves reliability, safety, availability, and security but in this paper we are only concerned with safety, reliability, and availability.

The area of system safety is well-established and procedures exist to identify and analyse electromechanical hazards along with techniques to eliminate or limit hazards in a final product. Unfortunately, much more is known about how to engineer safe mechanical systems than safe computing systems, particularly when software is a major component of the engineered system. With the increased used of software in safety-critical components of complex systems, governments agencies and other institutions are increasingly including requirements for software hazard analysis and verification of software safety.

Security : It has become clear that security attacks against information systems are a large and growing problem. Attacks against both public and private networks can have devastating effects. The Internet is being used increasingly to provide communication service to business, and security attacks against the Internet are a troubling problem for network users.

Although Internet attacks are important, private networks are a bigger concern. Money is moved locally and around the World on private networks owned by financial institutions. Transportation systems are monitored and controlled using mostly private networks. A successful attack against certain private networks could permit funds or valuable information such as credit card numbers to be stolen, transportation to be disrupted, and so on. The potential for loss is considerable, and, although no physical damage would be involved in security failures, the consequences of failure are such that many systems that only carry information should be regarded as safety-critical.

VII. Implementation

Some programming language features prone to problems than others. This is because of number of reasons. Those are

1) Programmers do errors while using the feature.
2) Poor compilation or poor implementation.
3) Programs written may be difficult to analyze and test.

Few programming language features that cause problems:

1) *Usage of pointers :* It is very difficult to use the pointers in programming language .In order to use pointers; the developers' need great understanding of memory address and management. Programs which use pointers can be difficult to understand or analyze.
2) *Memory Management :* The memory allocation and de-allocation is related to pointers. every programmer allocate memory but sometimes they forget to de-allocate .Compilers and operating systems frequently fail to fully recover de-allocated memory. The result is errors which are dependent on execution time, with a system mysteriously failing after a period of continuous operation.
3) *Multiple Entry and Exits :* More number of exit and

entry points to loops, blocks, procedures and functions, is really just a variation of unstructured programming. However, controlled use of more than one exit can simplify code and reduces the risk.

4) *Type of Data :* where the type of data in a variable changes, or the structure of a record changes, is difficult to analyze, and can easily confuse a programmer leading to programming errors.
5) *Declaration & Initialization :* A simple spelling mistake can result in software which compiles, but does not execute correctly. In the worst case individual units may appear to execute correctly, with the error only being detectable at a system level. Declaration must be perfect.
6) *Parameter Passing :* passing one procedure or function as a parameter to another procedure or function, is difficult to analyze and test thoroughly.
7) *Recursion :* Recursion is calling a function itself. It is difficult to analyze and test thoroughly. Recursion can also lead to unpredictable real time behavior.
8) *Concurrency and Interrupts :* These features are supported directly by some programming languages only. Use of concurrency and interrupts is some what produce ambiguity.

The use of such programming language features in safety critical software is discouraged.

Most modern programming languages encourage the use of block structure and modular programming, such that programmers take good structure for granted. Well structured software is easier to analyze and test, and consequently less likely to contain errors.

The features of few programming languages which can be used to increase reliability are:

1) *Perfect data usage :* The data is only used and assigned where it is of a compatible type.
2) *Constraint checking :* Ensure that arrays bounds are not violated, that data does not overflow, that zero division does not occur.
3) *Parameter checking : T*o ensure that parameters passed to or from procedures and functions are of the correct type, are passed in the right direction (in or out) and contain valid data.

There are no commonly available programming languages which provide all of the good language features. The solution is to use a language subset, where a language with as many good features as possible is chosen, and the bad features are simply not used. Use of a subset requires discipline on behalf of the programmers and ideally a subset checking tool to catch the occasional mistake. An advantage of a subset approach is that the bounds of the subset can be flexible, to allow the use of some features in a limited and controlled way.

Ada is the preferred language for the implementation of safety critical software because it can be used effectively within the above constraints. The most popular Ada subset for safety critical software is the SPARKAda subset.

SPARKAda is a subset of the Ada Programming Language that restricts several features of Ada such as unrestricted tasking. SPARKAda includes a built-in toolset called the "Examimator" which tests the entire source code for conditional and unconditional data flow errors which in theory would deem the source code exception free. The disadvantage to SPARKAda is that is closed and proprietary which increases the cost of implementation. Since it is a closed format, outside community support is restricted and there is a higher risk of implementation with only one vendor to rely on for technical support and language updates.

VIII. Testing & Verification

Safety critical testing : Testing of safety-critical systems follows two important strategies which are systematic rigorous testing and static analysis. While there is no substitute for rigorous testing at many levels: Unit, regression, functionality and integration testing, testing effectiveness depends on the quality of the test cases used. The best test suites are those that have good code coverage. Statement coverage and condition Coverage are the most commonly used metrics. Full Condition coverage is considered essential for safety-critical code, such as flight control software. Achieving full coverage can be exceedingly time-consuming and expensive.

Safety critical software functions provide the source of requirements to be tested. Testing shall be performed to verify correct incorporation of software safety requirements. Testing must show that hazards have been eliminated or controlled to an acceptable level of risk. Additional hazardous states identified during testing shall undergo complete analysis prior to software delivery or use. Software safety testing of Safety-Critical Computer Software Components (SCCSC) shall be included in the integration and integration and acceptance tests. Acceptance testing shall verify correct operation of the SCCSCs in conjunction with system hardware and operators[36]. It shall verify correct operation during stress conditions and in the presence of system faults. It is important to tailor the safety-critical testing effort to emphasize the parts of the software that need additional analysis and testing. The greatest effort must be placed on the hazards posing the highest risk. We consider it adequate to divide the software into two risk groups for test purposes.

Verification is the most important and most expensive group of activities in the development of safety critical systems, with verification activities being associated with each stage of the development lifecycle. An added complication is that *independent verification* is usually required. The means by which this is achieved depends upon the integrity level. Independent

verification can vary from independent witnessing of tests, participation at reviews and audit of the developer's verification, to fully independent execution of all verification activities. Independent verification is an addition to verification conducted by developers, not a substitute for it.

According to ISO 9001 activity, reviews will be conducted as a part of verification. Reviews become more formal, including techniques such as detailed walkthroughs of even the lowest level of design. The scope of reviews is extended to include safety criteria.

IX. Conclusion

The choice of a language can have a significant impact on the success or failure of a safety-critical system. The language can impact the ease of validation, the number of defects, and many important parts of the development process. Few languages are inherently "safe" as well as having good tool support, good documentation and wide usage.

A general-purpose language, which is made "safe" by use of a subset and good tool support, is the best choice for a safety-critical system. Modelling languages show excellent promise as implementation languages for all types of software development, not just safety critical.

Safety critical software is a complex subject. This paper will give an analysis of safety critical system means about design, implementation, verification, Applications etc.

Although safety critical systems have been in use for many years, the development of safety critical software is still a relatively new and immature subject. New techniques and methodologies for safety critical software are a popular research topic with universities, and are now becoming available to industry.

Tools supporting the development of safety critical software are now available, making the implementation of safety critical standards a practical prospect.

References References Referencias

1. Robyn R. Lutz, "Software Engineering for Safety: a Roadmap", *Proceedings of the Conference on The Future of Software Engineering*, June 04-11, 2000, Limerick, Ireland, pp. 213-226.
2. Alan C. Tribble et al. "Software Safety Analysis of a Flight Guidance System", *Proceedings of the 21st Digital Avionics Systems Conference* (DASC'02), Irvine, California, Oct. 27-31, 2002.
3. Debra S. Herman, "*Software Safety and Reliability Basics*", (ch.2), Software Safety and Reliability: Techniques, Approaches, and Standards of Key Industrial Sectors Wiley-IEEE Computer Society Press, 2000.
4. Dale M. Gray. *Frontier Status Report* #203, 19 May 2000, www.asi.org
5. John C. Knight. "Safety Critical Systems: Challenges and Directions" *Proceedings of the 24th International Conference on Software Engineering* (ICSE), Orlando, Florida, 2002.
6. N.Leveson, Safeware: *System Safety and Computers*, Addison Wesley, 1995.
7. L.Pullum, *Software Fault Tolerance: Techniques and Implementation*, Artech House, 2001.
8. W.R. Dunn, *Practical Design of Safety-Critical Computer Systems*, Reliability Press, 2002.
9. Kopetz, H.,*Real-Time Systems, Design Principles for Distributed Embedded Applications*, Kluwer Academic Publishers, 1997.
10. Conmy, P., Nicholson, M., Purwantoro, Y.,M., and McDermid, J. (2002) *Safety Analysis and Certification of Open Distributed Systems*.
11. J. A. McDermid, The cost of COTS, *IEE Colloquium - COTS and Safety critical systems* London,1998.
12. IEC 61508 *Functional Safety of electrical / electronic / programmable electronic safety-related systems* Geneva: International Electrotechnical Commission, 1998.
13. Tindell, K., "*Analysis of Hard Real-Time communications*", Real-Time Systems,vol 9,pp,147-171,1995.
14. Jesty, P.H., Hobley, K.M., Evans, R., and Kendall, I.,"Safety Analysis of Vehicle-Based Systems," *Proceedings of the 8th Safety-critical Systems Symposium*, 2000.
15. Raghu Singh. "A Systematic Approach to Software Safety". *Proceedings of Sixth Asia Pacific Software Engineering Conference* (APSEC), Takamatsu, Japan, 1999.
16. N. G. Leveson "Software Safety: Why, what, and how". *ACM Computing Surveys*, 18(2):125-163, June 1986.
17. The University of York, *Safety critical systems engineering, system safety engineering,* Modular MSc, diploma, certificate, short courses1999.
18. The University of York, Heslington, U.K.; www.cs.york.ac.uk/MSc/SCSE.
19. *The Hazards Forum, Safety-related systems: Guidance for engineers,* The Hazards Forum (1995). London,U.K.;www.iee.org.uk/PAB/SCS/hazpub.htm.

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Mining Frequent Item Sets from incremental database: A single pass approach

By Sandhya Rani Jetti, Sujatha D

Aurora's Technological and Research Institute, Hyderabad, India

Abstract - Apriori based Association Rule Mining (ARM) is one of the data mining techniques used to extract hidden knowledge from datasets that can be used by an organization's decision makers to improve overall profit. Performing Existing association mining algorithms requires repeated passes over the entire database. Obviously, for large database, the role of input/output overhead in scanning the database is very significant. We propose a new algorithm, which would mine frequent item sets with vertical format. The new algorithm would need to scan database one time. And in the follow-up data mining process, it can get new frequent item sets through 'and operation' between item sets. The new algorithm needs less storage space, and can improve the efficiency of data mining.

Keywords : *Apriori; vertical format; Association Rule ; data mining.*

GJCST Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Mining Frequent Item Sets from incremental database: A single pass approach

Sandhya Rani Jetti [α], Sujatha D [Ω]

Abstract - Apriori based Association Rule Mining (ARM) is one of the data mining techniques used to extract hidden knowledge from datasets that can be used by an organization's decision makers to improve overall profit. Performing Existing association mining algorithms requires repeated passes over the entire database. Obviously, for large database, the role of input/output overhead in scanning the database is very significant. We propose a new algorithm, which would mine frequent item sets with vertical format. The new algorithm would need to scan database one time. And in the follow-up data mining process, it can get new frequent item sets through 'and operation' between item sets. The new algorithm needs less storage space, and can improve the efficiency of data mining.

Index Terms : Apriori; vertical format; Association Rule; data mining.

I. Introduction

Data mining technology is mainly used to process massive amounts of data and information. It can help us find the potential and meaningful knowledge. Association rule mining is one of the most active research methods. It was proposed by Agrawal etc. to analysis market basket question. The purpose is to discover the association rules among different commodities included in trading database..

II. Apriori Algorithm

a) Brief overview

Apriori[1] algorithm is a classical algorithm to find association rules which was presented by Agrawal and Srikant in 1993. It is the boolean association rules algorithm to mining frequent item sets. The basic idea of the algorithm is to recursively generate frequent item sets. The basic idea of the algorithm is that it starts with the only *1*-item sets, recursively generates a frequent *2*-item set, and then produces a frequent *3*-item set, and continues like this, until all frequent item sets are produced. The algorithm will stop till generate all the frequent item sets.

b) Properties

The properties of Apriori algorithm is that an item set is frequent and its all non-empty subsets are frequent. In other words, if an item set is not frequent, all of its super-sets will not be frequent.

c) Pseudo code

```
Input D min_sup
Output Lk
L1=find_frequent_1-itemsets (D);
for(k=2;Lk-1≠Φ;k++)
{
Ck=Apriori_gen(Lk-1);
For each item.Tid D
{
Ct=subset(Ck,t);
for each candidate c Ct
c.count++;
}
Lk={c ,Ck,c.count>=min_sup}
}
return L=kLk
procedure Apriori_gen(Lk-1: frequent (k-1)-item sets)
for each item set p1,Lk-1
for each item set p2,Lk-1
if(p1[1]=p2[1]),(p1[2]=p2[2]),...,(p1[k-
1]=p2[k-1]) then{
c=p1 connect p2;
if has_infrequent_subset(c,Lk-1) then
delete c;
else add c to Ck ;
}
return Ck;
procedure has_infrequent_subset(c:candidate k-item set;
Lk-1:frequent (k-1)-item sets)
for each (k-1)-item set s of c
If s not in Lk-1 then
return TURE;
return FALSE;
```

d) Algorithm weaknesses

a). This algorithm may generate many candidate item sets in the calculation process. When the number of frequent *1*-item set is very big or the frequent patterns are very long, the number of the generated candidate item sets will increase sharply. So the algorithm's efficiency will fall dramatically. For example, if the number of frequent *1*-item set is 104, the number of candidate *2*-item set we need to generate, will be 107. If the length of frequent mode is 100, we will need to generate 2100 candidate sets. If we search so many candidate sets, the efficiency of the algorithm should be fairly low.

Author α : Student, CSE, Aurora's Technological and Research Institute, Hyderabad, India. E-mail : Sandya.jetti@gmail.com
Author Ω : Head Of the Dept, CSE, Aurora's Technological and Research Institute , Hyderabad, India.
E-mail : sujatha.dandu@gmail.com

b). This algorithm needs to scan the database many times and check candidate item sets by matching pattern. If the database is very large and the patterns needed to be matched are also very long, the efficiency of the algorithm will be greatly reduced.

III. Related Work

a) Brief overview

Apriori is a horizontal format algorithm. It is for mining frequent item sets. It needs to scan the database repeatedly to get support degree of the candidate sets. The time consumption in this process is the key of the algorithm. A variety of improved algorithms proposed to reduce the Comparison times between candidate sets and the Transaction records. Such as the DHP[2], FP-growth[3] and so on. They all have played a certain role in this regard. Here, we think that if there is a new algorithm can eliminate this comparison process, it will make the performance improved greatly.

This paper presents a new algorithm, which mine frequent item sets with vertical format. The new algorithm only needs to scan database one time to get frequent *1*-item set. In the follow-up of the mining process we statistic support degree of candidate sets which is used to generate table. We needn't to scan database again.

b) Algorithm Benefits

Benefits of vertical algorithm for mining frequent item sets: It will be able to judge whether the non-frequent item sets before generating candidate item sets. This can save time. Since each TID set of *k*-item set to carry the complete information that can calculate the support degree, so we needn't to scan the database to calculate the support degree of (*k+1*)-item set.

c) Description of the algorithm

First, scan the database to generate frequent *1*-item set.Second, transform horizontal format of frequent *1*-item set into vertical format. Then do 'and operation' among each element of frequent item set Lk and record the result. If the result is more than the min_sup, we'll obtain a candidate set Ck+1, else we will do the next 'and operation'. We will stop doing the 'and operation' till the following situations come: there is a request item set left and we have no way to do the 'and operation' or all the results of 'and operation' is less than min_sup.

IV. Algorithm Examples

a) Example Process

a. Define min_sup=2, scan the database and generate item sets which are vertical data format. They are listed in Table 1.

Table I : To Simplify the Database

Item set	TID set,
I1	1,4,5,7,8,9 (b1)
I2	1,2,3,4,6,8,9 (b2)
I3	3,5,6,7,8,9 (b3)
I4	2,4 (b4)
I5	1,8 (b5)

b. b1∩b2={1,4,8,9} FLAG=4
b1∩b3={5,7,8,9} FLAG=4;
b1∩b4={4} FLAG=1<2 delete
b1∩b5={1,8} FLAG=2;
b2∩b3={3,6,8,9} FLAG=4
b2∩b4={2,4} FLAG=2;
b2∩b5={1,8} FLAG=2
b3∩b4=Φ delete;
b3∩b5={8} FLAG=1<2 delete
b4∩b5=Φ delete

c. Arrange the above datas to generate *2*-frequent set shown in Table 2.

Table II : Frequent 2-Item Set

Item set	*TID set*
I1I2	1,4,8,9 (b1)
I1I3	5,7,8,9 (b2)
I1I5	1,8 (b3)
I2I3	3,6,8,9 (b4)
I2I4	2,4 (b5)
I2I5	1,8 (b6)

d. If the first item in the item set is the same, connect and intersect them.

b1∩b2={8,9} FLAG=2
b1∩b3={1,8} FLAG=2
b2∩b3={8} FLAG=1<2 delete
b4∩b5=Φ delete
b4∩b6={8} FLAG=1<2 delete
b5∩b6=Φ delete

e. Arrange the above data is to generate frequent *3*-item set shown in Table 3

Item set	**TID set**
I1I2I3	8,9 (b1)
I1I2I5	1,8 (b2)

f. b1∩b2=8 FLAG=1<2 delete

So the maximal frequent item set is {I1,I2,I3} and {I1,I2,I5}.

V. Related Work

The literature "A vertical format algorithm for mining frequent item sets" that considered as motivation for this proposal discussed an effective Apriori approach that avoids the multiple passes to the database. This model named by the author as vertical approach.

VI. My Contribution

The solution discussed in the literature "A vertical format algorithm for mining frequent item sets" given single pass approach to perform Apriori. This solution limited to the stable dataset. I would like to extend this work to handle the multi pass problem in incremental dataset also known as streaming data.

VII. Acknowledgment

I would like to thank Prof D.Sujatha(HOD), Sr. Asst.Prof A. Poongodai and Prof D. Sujatha (Aurora's Technological and Research Institute, Hyderabad, India) for proposing the concept of **Mining Frequent Item Sets from incremental database: A single pass approach** as well as providing their careful reading and valuable suggestions. I would also like to thank the anonymous referees for their helpful comments, correction and suggestions to improve this work.

References References Referencias

1. Margaret H.Dunham, Data Mining Itroductory and Advanced Topics,Tsinghua University Press,2005 pp.145–155.
2. Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, 2nd ed. China Machine Press, 2006, pp.155–160.
3. Wang Cuiru, Wang Shaohua, An Improved Apriori Algorithm for Association Rules. Computer Technology and Applications, February 2008.
4. Song Jingjing Mining Maximal Frequent Patterns in a Unidirectional FP-tree,Henan University,Henan Zhengzhou, May2007.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Generation of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model

By Momotaz Begum, Sumaya Kazary, Md. Jakir Hossain, Sohag Kumar Bhadra, Md. Rokon Uddin

Dept. of CSE, DUETGazipur, Bangladesh

Abstract - There are lots of work for inferring genetic network architectures from state transition tables which correspond to time series of gene expression patterns, using the Boolean network model. Results of those computational experiments suggested that a small number of state transition **(INPUT/OUTPUT)** pairs are sufficient in order to infer the original Boolean network correctly. Tatsuya **AKUTSU**, Satoru **MIYANO** and Satoru **KUHARA** gave a mathematical proof for this. So there is possibility to devise an algorithm to generate all consistent genetic networks from a small number of gene expression patterns under the Boolean network model.

General Terms : *Gene network, gene regulatory network, Genetic network, gene expression pattern, Boolean network, microarray data, Sum of product, Boolean algebra.*

Keywords : *Generation, consistent network.*

GJCST Classification : *C.2.1*

Strictly as per the compliance and regulations of:

Generation of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model

Momotaz Begum[α], Sumaya Kazary[Ω], Md. Jakir Hossain[β], Sohag Kumar Bhadra[ψ], Md. Rokon Uddin[¥]

Abstract - **There are lots of work for inferring genetic network architectures from state transition tables which correspond to time series of gene expression patterns, using the Boolean network model. Results of those computational experiments suggested that a small number of state transition (INPUT/OUTPUT) pairs are sufficient in order to infer the original Boolean network correctly. Tatsuya AKUTSU, Satoru MIYANO and Satoru KUHARA gave a mathematical proof for this. So there is possibility to devise an algorithm to generate all consistent genetic networks from a small number of gene expression patterns under the Boolean network model.**

General Terms : Gene network, gene regulatory network, Genetic network, gene expression pattern, Boolean network, microarray data, Sum of product, Boolean algebra.

Keywords : Generation, consistent network.

I. Introduction

Inference of gene regulation mechanism from time series of gene expression patterns are getting more important especially due to the invention of DNA microarray technology. Expression profiles of several thousands of genes are now being produced for further analyses. Some methods have been proposed for the inference of gene regulation mechanism from time series of gene expression patterns.

A statistical method is proposed by Arkin, Shen and Ross. They used correlation matrices to infer chemical reaction networks from time series of measured concentration of species. Though they treated chemical reaction networks, they also suggested that their method might be applied to genetic networks. However, it seems difficult to apply their method to the inference of large scale networks.

A metabolic pathway is suggested to be inferred by DeRisi, Iyer and Brown from gene expression patterns of Saccharomyces cerevisiae obtained by using DNA microarrays. A network model similar to the Boolean network model is constructed by Yuh, Bolouri and Davidson from time series of expression patterns relating to a sea urchin gene. But their inference methods are not systematic or automatic. Besides, some studies have been done on the inference of genetic networks from state transition data using the Boolean network .

On the other hannd, Liang, Fuhrman and Somogyi proposed an algorithm named REVEAL for inference of Boolean networks (corresponding to genetic networks) from state transition tables (corresponding to time series of gene expression patterns). REVEAL used information theoretic principles in order to reduce the search space. They made some computational experiments on REVEAL. The results suggested that only a small number of state transition pairs (100 pairs from 1015) were sufficient for inferring Boolean networks with 50 nodes (genes) whose in degree (the number of input nodes to a node) was bounded by 3.

Tatsuya AKUTSU, Satoru MIYANO and Satoru KUHARA gave a mathematical proof for their observation. We will extend their algorithm to identify genetic networks from gene expression patterns derived by gene disruptions and gene over expressions using a Boolean network-like model. They proved mathematically a lower bound and an upper bound of the number of expression patterns required to identify the network correctly. In this paper we will try to extend try to provide an algorithm to generate all the consistent genetic networks from a small number of gene expression patterns under the Boolean network model.

II. Generation Approach

a) Genetic network

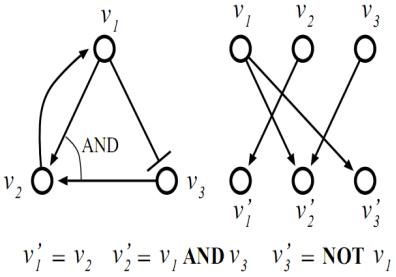
A **gene regulatory network** or **genetic regulatory network (GRN)** is a collection of DNA segments in a cell which interact with each other (indirectly through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA.

b) Boolean Network

A Boolean network G (V;F) consists of a set V = $\{v_1,v_2,\ldots\ldots,v_n\}$ of nodes representing genes and a list F = $(f_1,f_2,\ldots\ldots,f_n)$ of Boolean functions, where a Boolean function $f_i(v_{i1},v_{i2},\ldots,v_{ik})$ with inputs from specified nodes $v_{i1},v_{i2}\ldots,v_{ik}$ is assigned to each node v_i. For a subset U

Author α : Lecturer Dept. of CSE, DUETGazipur, Bangladesh.
E-mail : momotaz03_duet@yahoo.com
Author Ω : Assitant professor Dept. of CSE, DUET Gazipur, Bangladesh.
E-mail : kazal_duet@yahoo.com
Author β : CSE, DUET Gazipur,Bangladesh.
E-mail : newjakir@gmail.com
Author ψ : CSE, DUET Gazipur,Bangladesh.
E-mail : bhadra035@gmail.com
Author ¥ : CSE, DUET, Gazipur, Bangladesh.
E-mail : rokon.duet@gmail.com

of V, an expression pattern of U is a function from U to {0,1}. An expression pattern of V is also called a state of a Boolean network. That is, represents the states of nodes (genes), where each node is assumed to take either 0 (not-express) or 1 (express) as its state value.



v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

$v_1' = v_2 \quad v_2' = v_1 \text{ AND } v_3 \quad v_3' = \text{NOT } v_1$

Figure 1(a) : A genetic network represented by Boolean network

c) Definition of Problem

Let $(I_j;O_j)$ be a pair of expression patterns of $\{v_1,v_2,\ldots,v_n\}$, where I_j corresponds to the INPUT and O_j corresponds to the OUTPUT. We call the pair (Ij;Oj) an example. We say that a node vi in a Boolean network G(V;F) is consistent with an example (Ij;Oj), if Oj(vi)= $f_i(I_j(v_{i1},\ldots I_j(v_{ik}))$ holds. We say that a Boolean network G(V;F) is consistent with $(I_j;O_j)$ if all nodes are consistent with (Ij;Oj). For a set of examples EX = $\{(I_1;O_1),(I_2;O_2),\ldots\ldots..(I_m;O_m)\}$, we say that G(V;F) (resp. node vi) is consistent with EX if G(V;F) (resp. node vi) is consistent with all (Ij;Oj) for 1 <= j <= m. Then, the problems are defined as:-

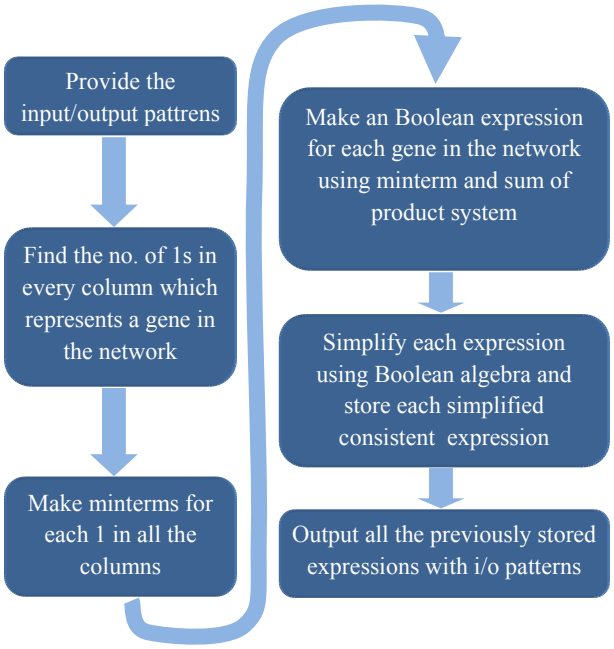
Consistency : Given n (the number of nodes) and EX, decide whether or not there exists a Boolean network consistent with EX and output one if it exists;

Generation : Given n (the number of nodes) and EX, generate all the number of Boolean networks consistent with EX.

	v_1 v_2 v_3	v_1' v_2' v_3'	
I_1	1 0 0	0 0 1	O_1
I_2	0 1 0	0 1 1	O_2
I_3	0 1 1	1 0 0	O_3
G_1	$v_1' = v_2$, $v_2' = v_2$ AND (NOT v_3) $v_3' =$ NOT v_3		

Table 1(a) : A genetic network with its input/output patterns and a consistent Boolean network

d) Process Development



e) Formulation of minterms

For a boolean function of n variables $x_1, \ldots, x_n$, a product term in which each of the n variables appears once (in either its complemented or uncomplemented form) is called a minterm. Thus, a minterm is a logical expression of n variables that employs only the complement operator and the conjunction operator.

For example, abc, ab'c and abc' are 3 examples of the 8 minterms for a Boolean function of the three variables a, b and c. The customary reading of the last of these is a AND b AND NOT-c.

There are 2^n minterms of n variables, since a variable in the minterm expression can be in either its direct or its complemented form--two choices per n variables.

It is apparent that minterm *n* gives a true value (i.e., 1) for just one combination of the input variables. For example, minterm 5, *a b' c*, is true only when *a* and *c* both are true and *b* is false—the input arrangement where a = 1, b = 0, c = 1 results in 1.

If one is given a truth table of a logical function, it is possible to write the function as a "sum of products". This is a special form of disjunctive normal form. For example, if given the truth table for the arithmetic sum bit *u* of one bit position's logic of an adder circuit, as a function of *x* and *y* from the addends and the carry in, *ci*:

Table 1(a) : Minterms formulation from a Boolean network

ci	x	y	u(ci,x,y)
0	0	0	0
0	0	1	1

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

Observing that the rows that have an output of 1 are the 2nd, 3rd, 5th, and 8th, we can write u as a sum of minterms m_1,m_2,m_4, and m_7. If we wish to verify this: $u(ci, x, y) = m_1 + m_2 + m_4 + m_7 = (ci'\, x'\, y) + (ci'\, x\, y') + (ci\, x'\, y') + (ci\, x\, y)$ evaluated for all 8 combinations of the three variables will match the table.

On the other hand if we assign serial numbers to each row of the input expression patterns from *0* to 2^n*-1* then some of products can be expressed with summation notation ($\sum$).

For example, for the above table as in output column the row 1,2,4 and 7 contain 1, so the output ex[ression can be formulated as $u(ci, x, y) = \sum(1,2,4,7)$.

It is apparent that this notation can be simplified using the help of computer more easily and efficiently.

f) Algorithm

Suppose we have n number of nodes (v_1, v_2,……..v_n) and a pair of expression pattern (I,O), where I correspond to input pattern and O corresponds to output pattern. V_1 to v_n are assumed to take either 0 (not-express) or 1 (express) as its state value. First we generate a Boolean network consistent with (I,O) and then eliminate nodes which aren't essential to generate more networks. The idea is as follows:-

1. Repeat step 2 for each node $v_i \in V$.
2. Generate Boolean equation and assign in $\widehat{v_i}$ using Sum of Product (SOP) notation based on input and output expression patterns (I and O).
3. Repeat step 4 and 5 for each $\widehat{v_i}$.
4. Eliminate variables v_i from $\widehat{v_i}$ using Boolean algebra until elimination is impossible.
5. This is a new Boolean network consistent with input and output expression patterns (I and O), store it.

g) Complexity

Consider the following expression pattern:-

Input			Output		
v_1	v_2	v_3	$\widehat{v_1}$	$\widehat{v_2}$	$\widehat{v_3}$
0	0	0	0	1	1
0	0	1	0	0	0
0	1	0	1	0	0
0	1	1	0	0	1
1	0	0	0	0	0
1	0	1	1	1	0
1	1	0	0	0	1
1	1	1	0	0	1

Table 2(a) : Generation of Boolean network from input /output expression patterns

As here are 3 genes we have got 3 Boolean expressions for them. They can be formulated according to previous discussion .

They are listed below:-

$$\widehat{v_1} = \sum(3,5),\ \widehat{v_2} = \sum(0,5),\ \widehat{v_3} = \sum(0,3,6,7).$$

Using the above example first we'll calculate the running time of the propose 4d algorithm.

To generate an Boolean equation for any state first we have to check how many 1s it has in output pattern. It can have maximum 2^n 1s. So $O(n*2^n)$ is the running time for generation of each equation. As there are n nodes and one variable is associated with each node so running time of step two of the algorithm is $O(n^2 * 2^n)$.

Again for each node, there can be maximum 2^n minterms and step 4 eliminates one minterm in each iteration. Now two minterms can be chosen in $\binom{m}{2}$ ways (let $m = 2^n$). If we reduce one minterm at a time then complexity of step 4 can be calculated as follows:-

In first iteration there are m minterms. So we can choose two of them to reduce into one in $\binom{m}{2}$ ways. In second iteration there are m-1 minterms. So we can choose two of them to reduce into one in $\binom{m-1}{2}$ ways and so on. Similarly at the end we should have two minterms and we can choose them in one way. So the complexity will be

$$f(m) = \binom{m}{2} + \binom{m-1}{2} + \binom{m-2}{2} + \ldots\ldots\ldots\ldots + \binom{m-(m-3)}{2} + \binom{m-(m-2)}{2}$$

$$= \binom{m+1}{3}$$

$$= \frac{(m+1)!}{3!(m-2)!}$$

$$= \frac{m^3-m}{6}$$

$$= \frac{(2^n)^3-2^n}{6}$$

$$= \frac{8^n-2^n}{6}$$

This function expresses the complexity of one simplifying one equation. There are at most n equations. So the overall complexity is given by $\frac{8^n-2^n}{6}*n$.

So the running time of the algorithm is $O(\frac{8^n-2^n}{6}*n + n^2 * 2^n) = O(n*8^n)$.

III. Experimental Outcomes

a) Performance Analysis

This paper improves the previous woks done by the authors worldwide.

The improvements are listed below:-

1) It generates only the consistent network. So overhead of checking inconsistent networks is eliminated.
2) Huge improvement is achieved in running time.
3) As all consistent networks are generated, so counting problem is solved simultaneously.
4) The in degree/out degree have no limit. A node (gene) can be expressed by any number of nodes(genes).

IV. Conclusion

We have proved mathematically that to generate all consistent Boolean networks it requires $O(n*8^n)$ times. For that purpose, we proposed a simple algorithm. Of course, real biological systems are different from Boolean networks nodes in a Boolean network take binary values which are updated synchronously, whereas quantities of gene expressions in real cells are not binary and are changing continuously in time. However, owing to its simplicity, the proposed algorithm can be extended in various way. This algorithm can be extended to identify, count and enumerate all the consistent networks also.

Finally, we believe that our theoretical results, along with the mathematical computations encourage the attempts to discover the gene regulation mechanism from time series of gene expression patterns.

References Références Referencias

1. Tatsuya AKUTSU, Satoru MIYANO and Satoru KUHARA,Identification Of Genetic networks From A Small Number Of Gene Expression Patterns Under The Boolean Network Model.
2. S. Liang, S. Fuhrman and R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, Paci_c Symposium on Biocomputing 3, 18 (1998).
3. Chia-Chin Wu, Hsuan-Cheng Huang, Hsueh-Fen Juan and Shui-Tein Chen, GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data
4. D.A. Kightley, N. Chandra, and K. Elliston, Inferring Gene Regulatory Networks fromRaw Data: A Molecular Epistemics Approach, Pacific Symposium on Biocomputing 9:510-520(2004).
5. Rui Xu, Donald C. Wunsch II, and Ronald L. Frank, Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization
6. T. Akutsu, S. Kuhara, O. Maruyama and S. Miyano, Identification of gene regulatory networks by strategic gene disruptions and gene over expressions, Proc. 9th ACM-IAM Symp. Discrete Algorithms, 695(1998).
7. A.Arkin,P. Shen and J. Ross, A test case of correlation metric construction of a reaction pathway from measurements, Science 277, 1275(1997).
8. J.L. DeRisi, V.R. Lyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278, 680(1997).
9. M.J. Kearns and U.V. Vazirani, An Introduction to Computational Learning Theory, The MIT Press (1994).
10. H.H. McAdams and L. Shapiro, Circuit simulation of genetic networks,Science 269, 650 (1995).
11. R. Somogyi and C.A. Sniegoski, Modeling the complexity of genetic networks: Understanding multigene and pleiotropic regulation, Complexity 1, 45 (1996).
12. R. Thomas, D. Thie_ry and M. Kaufman, Dynamical behaviour of biological regulatory networks -I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state, Bulletin of Mathematical Biology 57, 247 (1995).
13. Wikipedia (http://en.wikipedia.org/wiki/Genetic_net work), *the free encyclopedia.*
14. Donald E. Knuth, Concrete Mathematics
15. X. Wen et al**,** Large-scale temporal gene expression mapping of central nervous system development, Proc. Natl. Acad. Sci. USA 95, 334 (1998)
16. A.Wuensche**,** Genomic regulation modeled as a network with basins of attraction, Paci_c Symposium on Biocomputing 3, 89 (1998). 11. C-H. Yuh, H.
17. Bolouri and E.H. Davidson, Genomic Cis-regulatory logic: experimental and computational analysis of a sea urchin gene, Science 279, 1896 (1998).

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Prototype centric (PC) software development process model: A machine learning based Hybrid Software Development Model

By Nabil Mohammed Ali Munassar, Dr. A. Govardhan

Jawaharlal Nehru Technological University Hyderabad

Abstract - Here in this paper we propose a Machine learning technique based Hybrid software development process model called prototype centric, in short can refer as PC. The proposed hybrid model works by considering any one or more traditional models as source models. We also conduct empirical study to analyze the performance of the PC over other traditional models that are most frequently quoted in literature.

Keywords : *Hybrid Software Development Method, Conventional Software Development Methods, Agile Software Development Methods, Empirical Studies, Software Engineering.*

GJCST Classification : *D.2.9*

Strictly as per the compliance and regulations of:

Prototype centric (PC) software development process model: A machine learning based Hybrid Software Development Model

Nabil Mohammed Ali Munassar [α], Dr. A. Govardhan [Ω]

Abstract - Here in this paper we propose a Machine learning technique based Hybrid software development process model called prototype centric, in short can refer as PC. The proposed hybrid model works by considering any one or more traditional models as source models. We also conduct empirical study to analyze the performance of the PC over other traditional models that are most frequently quoted in literature.

Keywords : Hybrid Software Development Method, Conventional Software Development Methods, Agile Software Development Methods, Empirical Studies, Software Engineering.

I. Introduction

All software, especially large pieces of software produced by many people, should be produced using some kind of methodology. Even small pieces of software developed by one person can be improved by keeping a methodology in mind. A methodology is a systematic way of doing things. It is a repeatable process that we can follow from the earliest stages of software development through to the maintenance of an installed system. As well as the process, a methodology should specify what we're expected to produce as we follow the process. A methodology will also include recommendation or techniques for resource management, planning, scheduling and other management tasks. Good, widely available methodologies are essential for a mature software industry. A good methodology addresses the following issues: Planning, Scheduling, Resourcing, Workflows, Activities, Roles, Artifacts, Education. There are a number of phases common to every development, regardless of methodology, starting with requirements capture and ending with maintenance. During the last few decades a number of software development models have been proposed and discussed within the Software Engineering community. With the traditional approach, you're expected to move forward gracefully from one phase to the other. With the modern approach, on the other hand, you're allowed to perform each phase more than once and in any order. [1,10].

Author α : PhD Scholar in Computer Science & Engineering Jawaharlal Nehru Technological University Hyderabad Kuktapally, Hyderabad-500085, Andhra Pradesh, India. E-mail : Nabil_monaser@hotmail.com
Author Ω : Professor of CSE & School of Information Technology Jawaharlal Nehru Technological University Hyderabad Kuktapally, Hyderabad- 500 085, Andhra Pradesh, India.
E-mail : govardhan_cse@yahoo.co.in

II. Related Work

Conventional heavyweight, document-driven software development methods can be characterized as extensive planning, codified process, rigorous reuse, heavy documentation and big design up front [3]. The conventional methods were predominant in the software industry up until the mid 1990s. Since then, the conventional methods have been replaced by lightweight agile software development methods mostly in small-scale and relatively simple projects. This phenomenon is mainly due to the conventional methods' shortcomings, including a slow adaptation to rapidly changing business requirements, and a tendency to be over budget and behind schedule [3, 6, 9, 15]. The conventional methods also have failed to provide dramatic improvements in productivity, reliability, and simplicity [9].

Some researchers reported that during their project development experience, requirements often changed by 25% or more [5]. An interesting research mentioned that the conventional methods were not initially designed to respond to requirements change occurring in the middle of the development process, and the ability to take action appropriate to the change often determines the success or failure of a software product. According to the Standish Group report , numerous projects with the conventional methods in various industry and government sectors were completed with fewer features and functionalities than specified in the user requirements. It is also a challenge for the conventional methods to create a complete set of requirements up front due to constant changes in the technology and business environments.

Despite the existing shortcomings, the conventional methods are still widely used in industry, particularly, for large-scale projects. The driving force of this broad utilization of the conventional methods comes from their straightforward, methodical, and structured nature [12], as well as their capability to provide predictability, stability, and high assurance [6].

Agile software development methods focus on iterative and incremental development, customer

collaboration, and frequent delivery through a light and fast development life cycle. There are many positive benefits of the agile approaches. Shorter development cycles, higher customer satisfaction, lower bug rates, and quicker adaptation to rapidly changing business requirements have been reported [6].

III. Hybrid Software Development Process Model

The proposed hybrid software development process model works as prototype centric with one or more traditional models as source. In short we there after refer as PC. The fig. 1 describes the proposed risk analysis process that mingles with each stage of the SDLC. Here in PC the risk analysis is strategic and supports to predict the risk that influence the cost and targeted outcomes. This prediction can help the experts involved to change the current action to decrease the severity of the risk predicted. Fig. 2 describe the risk analysis strategy proposed as key aspect of the PC. Here in risk analysis process we opt to machine learning technique called support vector machines in short SVM. The Risk analysis stage of the PC targets the SDLC logs available as input to train the SVM for better predictions. The feature extraction process that is part of SVM training process can be done with support of mathematical model called Quantum particle swarm optimization. The usage of these technologies described in fallowing section.

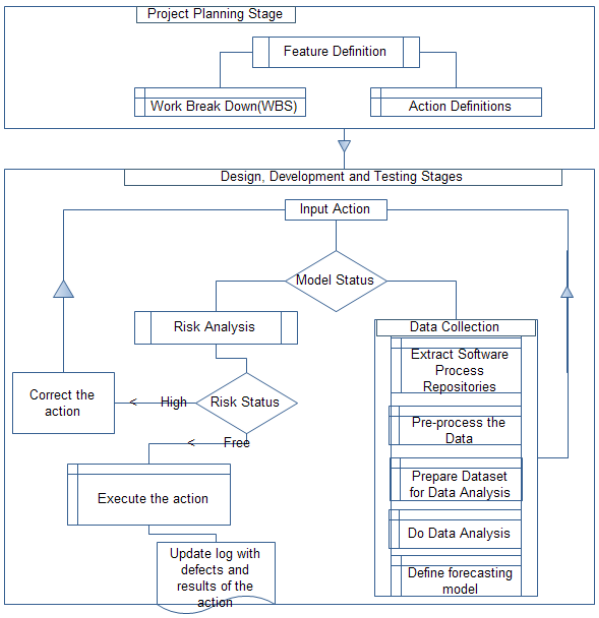


Fig.1 : Hybrid Software development process model

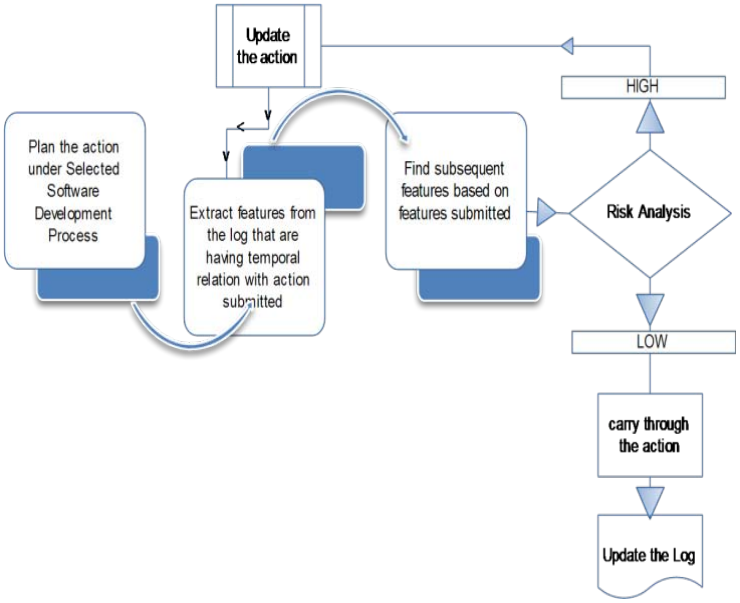


Fig. 2 : Risk Analysis Process

IV. Risk Analysis Using LS-SVM and Q-QPSO

a) LS-SVM

Support vector machine (SVM) introduced by Vapnik[5, 6] is a valuable tool for solving pattern recognition and classification problem. SVMs can be applied to regression problems by the introduction of an alternative loss function. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive application[5]. SVM shows outstanding performances because it can lead to global models that are often unique by embodies the structural risk minimization principle[7], which has been shown to be superior to the traditional empirical risk minimization principle. Furthermore, due to their specific formulation, sparse solutions can be found, and both linear and nonlinear regression can be performed. However, finding the final SVM model can be computationally very difficult because it requires the solution of a set of nonlinear equations (quadratic programming problem). As a simplification, Suykens and Vandewalle[8] proposed a modified version of SVM called least-squares SVM (LS-SVM), which resulted in a set of linear equations instead of a quadratic programming problem, which can extend the applications of the SVM. There exist a number of excellent introductions of SVM [8, 9] and the theory of LS-SVM has also been described clearly by Suykens et al[7, 8] and application of LS-SVM in quantification and classification reported by some of the works[10, 11].

In principle, LS-SVM always fits a linear relation ($y = w\ x + b$) between the regression (x) and the dependent variable (y). The best relation is the one that minimizes the cost function (Q) containing a penalized regression error term:

$$Q = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 \qquad (1)$$

Subject to

$$y_{i=w^T\phi(x_i)+b+e_i} \quad i=1,\ldots,N \qquad (2)$$

The first part of this cost function is a weight decay which is used to regularize weight sizes and penalize large weights. Due to this regularization, the weights converge to similar value. Large weights deteriorate the generalization ability of the LS-SVM because they can cause excessive variance. The second part of cost function is the regression error for all training data. The relative weight of the current part compared to the first part can be indicated by the parameter 'g', which has to be optimized by the user.

Similar to other multivariate statistical models, the performances of LS-SVMs depends on the combination of several parameters. The attainment of the kernel function is cumbersome and it will depend on each case. However, the kernel function more used is the radial basis function (RBF), a simple Gaussian function, and polynomial functions where width of the Gaussian function and the polynomial degree will be used, which should be optimized by the user, to obtain the support vector. For the RBF kernel and the polynomial kernel it should be stressed that it is very important to do a careful model selection of the tuning parameters, in combination with the regularization constant g, in order to achieve a good generalization model.

b) Q- QPSO

Millie Pant et al[12] attempt to optimize the QPSO by replacing least good swarm particle with new swarm particle. An interpolate equation will be traced out by applying a quadratic polynomial model on existing best fit swarm particles. Based on emerged interpellant, new particle will be identified. If the new swarm particle emerged as better one when compared with least good swarm particle then replace occurs. This process iteratively invoked at end of each search lap.

The computational steps of optimized QPSO algorithm are given by :

Step 1 : Initialize the swarm.
Step 2 : Calculate mbest
Step 3 : Update particles position
Step 4 : Evaluate the fitness value of each particle
Step 5 : If the current fitness value is better than the best fitness value (Pbest) in history Then Update Pbest by the current fitness value.
Step 6 : Update Pgbest (global best)
Step 7 : Find a new particle
Step 8 : If the new particle is better than the worst particle in the swarm, then replace the worst particle by the new particle.
Step 9 : Go to step 2 until maximum iterations reached.

The swarm particle can be found using the following.

$t_i = \sum_{k=1}^{3} p_i^2 - q_i^2) * f(r)$	$p=a,q=b,r=c \ for\ k=1;$ $p=b,q=c,r=a \ for\ k=2;$ $p=c,q=a,r=b \ for\ k=3$
$t1_i = \sum_{k=1}^{3} p_i - q_i) * f(r)$	$p=a,q=b,r=c \ for\ k=1;$ $p=b,q=c,r=a \ for\ k=2;$ $p=c,q=a,r=b \ for\ k=3$

$$x_i = 0.5 * (\frac{t_i}{t1_i})$$

In the above math notations 'a' is best fit swarm particle, 'b' and 'c' are randomly selected swarm particles x_i is new swarm particle.

c) LS-SVM Regression and QPSO based hyper parameter selection

Consider a given training set of N data points $\{x_t, y_t\}_{t=1}^{N}$ with input data $x_t \in R^d$ and output $y_t \in R$. In feature space LS-SVM regression model take the form

$$y(x) = w^T \varphi(x) + b \qquad (1)$$

Where the input data is mapped $\varphi(.)$.

The solution of LS-SVM for function estimation is given by the following set of linear equations:

$$\begin{bmatrix} 0 & 1 & .... & 1 \\ 1 & K(x_1,x_1)+1/C & .... & K(x_1,x_1) \\ . & . & . & . \\ . & . & . & . \\ 1 & K(x_1,x_1) & & K(x_1,x_1)+1/C \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ . \\ . \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ . \\ . \\ y_1 \end{bmatrix} \qquad (2)$$

$$\text{Where } K(xi, xj) = \phi(x_i)^T \phi(x_j)^T \ \text{ for } i, j = 1 \ldots L$$

And the Mercer's condition has been applied.

This finally results into the following LS-SVM model for function estimation:

$$f(x) = \sum_{i=1}^{L} \alpha_i K(x, x_i) + b \qquad (3)$$

Where α , b are the solution of the linear system, K(.,.) represents the high dimensional feature spaces that is nonlinearly mapped from the input space x. The LS-SVM approximates the function using the Eq. (3).

In this work, the radial basis function (RBF) is used as the kernel function:

$$k(x_i, x_j) = \exp(-\| x - x_t \|^2 / \sigma^2)$$

In the training LS-SVM problem, there are hyper-parameters, such as kernel width parameter σ and regularization parameter C, which may affect LS-SVM generalization performance. So these parameters

need to be properly tuned to minimize the generalization error. We attempt to tune these parameters automatically by using QPSO.

d) Hyper-Parameters Selection Based on Q-QPSO

To surpass the usual L2 loss results in least-square SVR, we attempt to optimize hype parameter selection.

There are two key factors to determine the optimized hyper-parameters using QPSO: one is how to represent the hyper-parameters as the particle's position, namely how to encode [13,14]. Another is how to define the fitness function, which evaluates the goodness of a particle. The following will give the two key factors.

i. *Encoding Hyper-parameters*

The optimized hyper-parameters for LS-SVM include kernel parameter and regularization parameter. To solve hyper-parameters selection by the proposed Q-QPSO, each particle is requested to represent a potential solution, namely hyper-parameters combination. A hyper-parameters combination of dimension m is represented in a vector of dimension m, such as $x_i = (\sigma, C)$. The resultant Hyper-parameter optimization under Q-QPSO can found in fallowing the Eq. (4).

a. Fitness function

The fitness function is the generalization performance measure. For the generation performance measure, there are some different descriptions. In this paper, the fitness function is defined as:

$$fitness = \frac{1}{RMSE(\sigma, \gamma)} \qquad (4)$$

Where RMSE (σ ,γ) is the root-mean-square error of predicted results, which varies with the LS-SVM parameters (σ ,γ) . When the termination criterion is met, the individual with the biggest fitness corresponds to the optimal parameters of the LS-SVM.

There are two alternatives for stop criterion of the algorithm. One method is that the algorithm stops when the objective function value is less than a given threshold ε; the other is that it is terminated after executing a pre-specified number of iterations. The following steps describe the Q-QPSO-Trained LS-SVM algorithm:

1) Initialize the population by randomly generating the position vector iX of each particle and set iP = iX;
2) Structure LS-SVM by treating the position vector of each particle as a group of hyper-parameters;
3) Train LS-SVM on the training set;
4) Evaluate the fitness value of each particle by Eq.(4), update the personal best position iP and obtain the global best position gP across the population;
5) If the stop criterion is met, go to step (7); or else go to step (6);
6) Update the position vector of each particle according to Eq.(7), Go to step (3);
7) Output the gP as a group of optimized parameters.

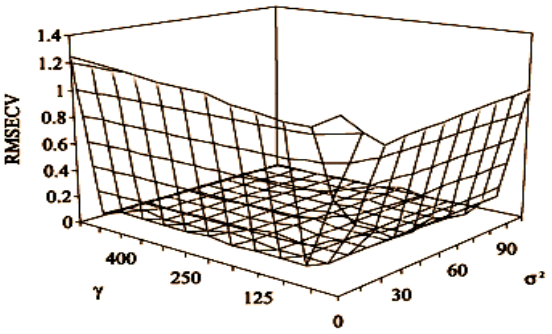


Fig.3 : Hyper-Parameter optimization response surface under Q-QPSO for LS-SVM

V. Risk Analysis Method Proposed

This section explains the algorithm for proposed risk analysis in various stages of SDLC, where the feature extraction can be done under LS-SVM regression and Q-QPSO.

- The SDLC log considered into multitude blocks of SDLC stages.
- Collect the resultant approximate and details features of each block.
- Apply LS-SVM regression under Q-QPSO on each feature matrix that generalizes the training data by producing minimum support vectors required.
- Estimate the features determined levels.
- Apply the risk analysis process by comparing the features of the assigned action and subsequent related actions of the current SDLC stage.
- Identify the risk status

VI. Empirical Study and Results Discussion

The performance analysis of the proposed Software development process model is carried by conducting empirical study on various projects development process logs. We opted to different logs that belong to applications of different sizes from low to high and enterprise level.

a) Empirical analysis of the small size software development process logs

We opted to a small size off the shelf application development process log to analyze the performance of the proposed hybrid software development process model that can referred as prototype centric in short PC. This selected off the shelf product actually developed under waterfall model. We conducted some empirical analysis for waterfall prototyping.

Empirical analysis has been conducted by considering the features of each individual action of each SDLC stage and applied risk analysis process as discussed in section IV. And then we conducted a comparative study between risk status identified and actual impact available in the log. The results that we observed are interesting and concluded that this model is having much influence in SDLC stages

1. Development
2. Testing

Table 1 represents the actual deviation ratio of waterfall model and predicted possible deviation ratio for PC with waterfall model as source. The Fig. 4 indicate the accuracy in risk analysis approach proposed in PC with waterfall as source model. We can observe that the proposed model is impressive at prediction particularly in development testing stages. Therefore we can conclude that PC with Waterfall model as source can minimize the cost and involvement of the high risk.

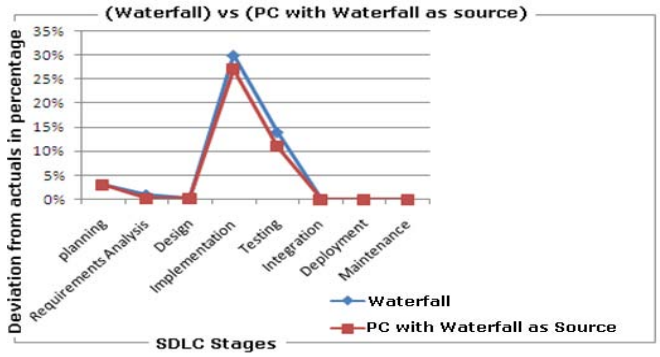
Table 1: performance and deviation analysis of PC with Waterfall as source

(a) deviation analysis

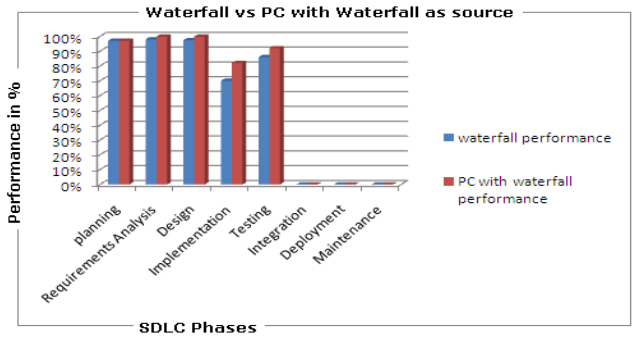
Waterfall vs PC with Waterfall as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Actual deviation	3%	1%	0.30%	30%	14%	0%	0%	0%
Predicted deviation	3%	0.20%	0.15%	27%	11%	0%	0%	0%

(b) Performance analysis

Waterfall vs PC with Waterfall as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
waterfall performance	97%	98%	97.40%	70%	86%	0%	0%	0%
PC with waterfall performance	97%	99.80%	99.85%	82%	92%	0%	0%	0%



(a) Deviation Comparison chart



(b) Performance Comparison chart

Fig. 4 : Deviation ratio and performance ratio of waterfall and PC with Waterfall as source

b) Empirical analysis of the mid size software development process logs

We opted to a mid size work flow engine application development process log to analyze the performance of the PC. This selected product actually developed under spiral model with less expertise resources. We conducted some empirical analysis for Spiral prototyping as described in Section IV. And then we conducted a comparative study between risk status identified and actual impact available in the log. The results that we observed are interesting and concluded that this model is having much influence in SDLC stages

1. Design
2. Development
3. Testing

Table 2 represents the actual deviation ratio of spiral model and predicted possible deviation ratio for PC with spiral model as source. The Fig. 5 indicate the accuracy in risk analysis approach proposed in PC with spiral as source model. We can observe that the proposed model is impressive at prediction particularly in design, development and testing stages. Therefore we can conclude that PC with spiral model as source

can minimize the cost and involvement of the high risk even under less expertise resource availability. Where in the case of spiral model high expert resources are must to minimize the cost and risk involvement.

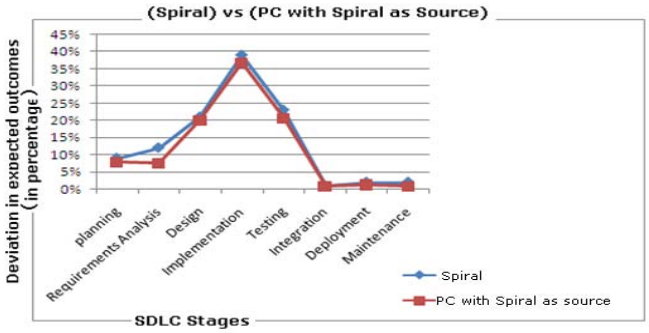
Table 2 : Deviation ratio and performance ration of PC with spiral as source

(a) Deviation ratio

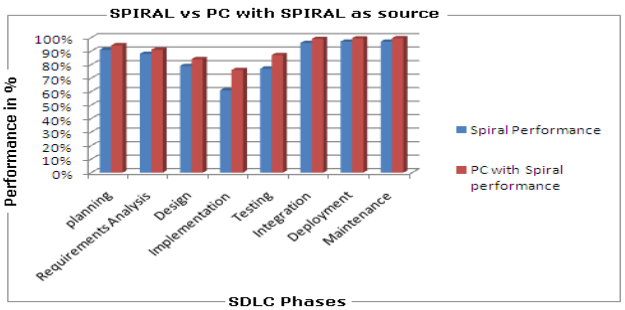
Spiral Vs PC with spiral as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Actual deviation	9%	12%	21%	39%	23%	1%	2%	2%
Predicted deviation	7.90%	7.60%	20.10%	36.78%	20.67%	0.90%	1%	0.90%

(b) Performance ratio

Spiral Vs PC with spiral as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Spiral Performance	91%	88%	79%	61%	77%	96%	97%	97%
PC with Spiral performance	94.30%	91.00%	84.00%	76.00%	87%	99.00%	99%	99.60%



(a) Deviation ratio chart



(b) Performance ratio chart

Fig. 5 : Deviation ratio and Performance ratio of Spiral and PC with spiral as source

c) Empirical analysis of the big size software development process logs

We opted to a big size tailor made java bean framework development process log to analyze the performance of the PC. This selected product actually developed under Incremental model. We conducted some empirical analysis for incremental prototyping as described in section IV. And then we conducted a comparative study between risk status identified and actual impact available in the log. The results that we observed are interesting and concluded that this model is having much influence in SDLC stages

1. Design
2. Development
3. Testing
4. Integration

Table 3 represents the actual deviation ratio of incremental model and predicted possible deviation ratio for PC with incremental model as source. The Fig. 6 indicate the accuracy in risk analysis approach proposed in PC with incremental model as source. We can observe that the proposed model is impressive at prediction particularly in design, development, testing and integration stages. Therefore we can conclude that PC with incremental model as source can minimize the cost and involvement of the high risk. This becomes practical because the proposed model prediction ability of deviations in requirement analysis. Where in the case of incremental model, risk involvement is high since requirement analysis not done in beginning that reflects as high cost and risk involvement.

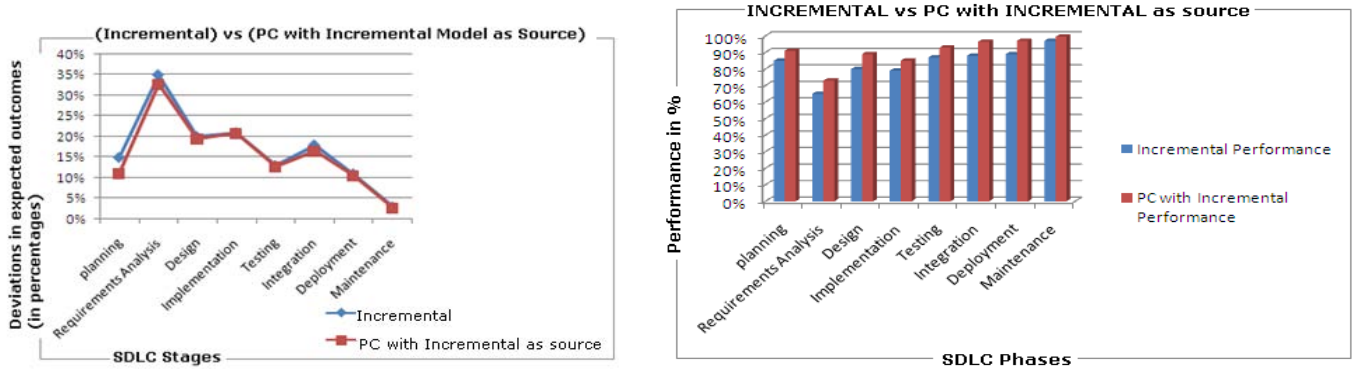
Table 3 : Deviation Ratio and performance ratio of Incremental model and PC with incremental model as source

(a) Deviation Ratio

Incremental Vs PC with incremental as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Actual deviation	15%	35%	20%	21%	13%	18%	11%	3%
Predicted deviation	11%	32.70%	19.40%	20.80%	12.70%	16.50%	10.60%	2.70%

(b) Performance Ratio

Incremental Vs PC with incremental as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Incremental Performance	85%	65%	80%	79%	87%	88%	89%	97%
PC with Incremental Performance	91%	73.00%	89.00%	85.00%	93%	96.50%	97.00%	99.50%



(a) Deviation ratio chart

(b) Performance Ratio Chart

Fig. 6 : Deviation ratio and Performance Ratio of Incremental and PC with Incremental model as source

d) Empirical analysis of the big enterprise software development process logs

We opted to a big enterprise MVC based media sharing web application development process log to analyze the performance of the PC. This selected product actually developed under Agile. We conducted some empirical analysis for agile prototyping as described in section IV. And then we conducted a comparative study between risk status identified and actual impact available in the log. The results that we observed are interesting and concluded that this model is having much influence in SDLC stages.

1. Design
2. Development
3. Testing
4. Integration
5. Maintenance

Table 4 represents the actual deviation ratio of incremental model and predicted possible deviation ratio for PC with incremental model as source. The Fig. 7 indicates the accuracy in risk analysis approach proposed in PC with agile model as source. Since agile is combination of iterative and incremental models, so that he advantages of PC with incremental model as source those we observed in earlier section are applicable as it is. We can observe that the proposed model is impressive at prediction particularly in design, development, integration, testing and maintenance stages. Therefore we can conclude that PC with incremental model as source can minimize the cost and involvement of the high risk. This becomes practical because the proposed model prediction ability of deviations in requirement analysis. Where in the case of agile model, risk involvement is high since requirement analysis not done in beginning and that reflects as high cost and risk involvement. It is obvious in agile model that expert resources are must to avoid the project to deviate from the expected outcome. Because of risk analysis and prediction strategy introduced in PC, an interesting issue about PC with agile as source model is that risk involvement can be minimized even under resources with moderated expertise.

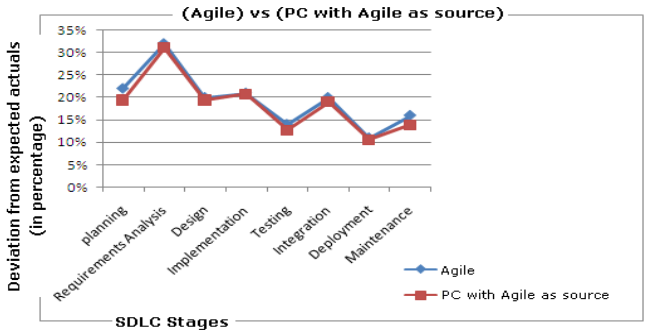
Table 4 : Deviation and performance analysis of PC with Agile model as source

(a) Deviation Ratio

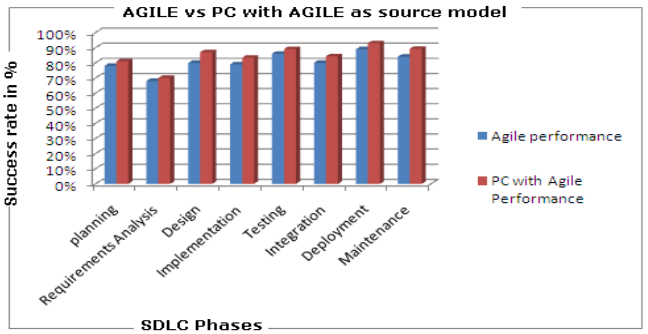
Agile vs PC with Agile as source	Planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenance
Actual deviation	22%	32%	20%	21%	14%	20%	11%	16%
Predicted deviation	19.40%	31.10%	19.40%	20.80%	12.70%	19.04%	10.60%	13.90%

(b) Performance Ratio

Agile vs PC with Agile as source	planning	Requirements Analysis	Design	Implementation	Testing	Integration	Deployment	Maintenane
Agile Performance	78%	68%	80%	79%	86%	80%	89%	86%
PC with agile performance	80.60%	69.90%	81.40%	80.80%	88.40%	19.04%	91.10%	88.00%



(a) Risk prediction ratio between Agile and PC with Agile as source



(b) SDLC phases level Success ratio between Agile and PC with Agile as source

Fig. 7 : Risk prediction ratio and SDLC phase level success ratio of PC with Agile model as source

e) Feature wise performance analysis of existing and proposed software development process models

Table 5 : Comparison report of the existing and proposed Software development process Models

Feature	Waterfall Model	Prototype Model	Spiral Model	Iterative Model	Agile Model	Prototype Centric(PC)
Requirement Specifications	Beginning	Frequently Changed	Beginning	Beginning	Frequently Changed	Dependent of Risk Analysis report
Understanding Requirements	Well Understood	Not Well understood	Well Understood	Not Well understood	Well understood	Well understood
Cost	Low	High	Intermediate	Low	Very high	Moderate
Guarantee of Success	Low	Good	High	High	Very high	Very high
Resource Control	Yes	No	Yes	Yes	No	Yes
Cost Control	Yes	No	Yes	No	Yes	Sure
Simplicity	Simple	Simple	Intermediate	Intermediate	Complex	Moderate
Risk Involvement	High	High	Low	Intermediate	Moderate	Dependent of Source Model
Expertise Required	High	Medium	High	High	Very high	Dependant of source model
Changes Incorporated	Difficult	Easy	Easy	Easy	difficult	Moderate
Risk Analysis	Only at beginning	No Risk Analysis	Yes	No	yes	On each Stage of source model

User Involvement	Only at beginning	High	High	Intermediate	high	Dependent of Risk Analysis report
Overlapping Phases	No	Yes	Yes	No	yes	Dependant of source model
Flexibility	Rigid	Highly Flexible	Flexible	Less Flexible	highly flexible	Highly Flexible

i. *Simplicity*

Data was obtained for a cost driver value of 'multi-skilled and experienced'. The data indicates that the waterfall and prototype models are most suitable for projects in which simplicity is the main factor. The spiral and iterative models have limited impact because they have intermediate with regard to simplicity factor and agile is not feasible[15], while the Prototype Centric model is most optimal because of its ability to minimize the complex nature of the source model. But due to modular evaluation, more time and money is required to complete a software project.

ii. *Risk Involved*

The data indicates that the Spiral model is most suitable for projects because software projects using this model involve low risk, where as waterfall model is unsuitable because high risk is involved in software projects. But Prototype centric can be optimal regardless of the source model to minimize the risk.

iii. *Expertise Required*

Data was obtained for a cost driver value of 'range of development experience' The Prototyping models are most appropriate where only developers with a range of experience are available. The waterfall, spiral and iterative models are slightly less suitable because they require personnel with high level of expertise, whereas the agile process model is inappropriate because it requires personnel with very high expertise and experience. The strong positive value for the Prototyping model may suggest the developers, instead of managers, are performing objective setting and evaluation. The proposed Prototype centric PC can improvise the other models performance even under resources with less expertise.

iv. *Changes Incorporated*

From the analysis of data, it is observed that the prototype, spiral and iterative models are most suitable of all as they requires less changes to be incorporated after the project is complete. Because if model needs more changes during usage, software projects takes more cost and also time for its updating etc. While the Waterfall model and agile models are totally inappropriate because if it requires the changes to be incorporated, then many difficulties do arise while incorporating changes in the software project [16].

v. *Risk Analysis*

Data was obtained for a cost driver value of 'risk involvement (expressed as 'complex, difficult or challenging to implement' or 'very complex or novel algorithm'). Data shows waterfall model have risk involved only at beginning, while the prototype model and iterative model don't involves any risk analysis while being used in any software projects. While on the contrary the spiral model and agile process model have risk analysis being used in any software project.

vi. *User Involvement*

Data was obtained and it is observed that waterfall model has very less involvement of the users because it requires user involvement only at the beginning of project. Iterative model needs intermediate user involvement, whereas spiral model and agile process models require high user involvement as a requirement of these models [17].

M. Overlapping Phases

From the research it was seen that Waterfall model and iterative model have no overlapping phases while the prototype model, spiral model process models requires overlapping phases. In the point of prototype centric it is obvious that the behavior of source model need to be considered.

vii. *Flexibility*

Data was obtained for a cost driver value of 'range of flexibility'. Data shows that PC process model and prototype models are highly flexible and are most appropriate, spiral and waterfall models also performs much better when those considered as source process models for PC. As an individual Waterfall model is rigid but as a source model of PC performs better.

VII. Conclusion

Based on the results of the empirical analysis conducted in section VI, we can conclude that regardless of the source model the Prototype Centric is modest in all desired features, particularly in terms of cost, resource utilization and balanced SDLC. It helps to work with any one or more traditional models as source under any circumstances such as resource availability with less expertise. As the methodology we allowed to perform risk analysis, it is stable regardless of the software application size.

References References Referencias

1. Molokken-Ostvold et.al, "A comparison of software project overruns - flexible versus sequential development models", Volume 31, Issue 9, Page(s): 754 – 766, IEEE CNF, Sept. 2005.
2. Boehm, B. W. "A spiral model of software development and enhancement", ISSN: 0018-9162, Volume: 21, Issue: 5, on page(s): 61-72, May 1988.

3. Abrahamsson P. et.al, “Agile Software Development Methods: Review and Analysis”, ESPOO, VTT Publications 478, VTT Technical Research Centre of Finland. http:/www.fi/pdf/publications/2002/ P478 .pdf, 2002.
4. Vapnik, V.; Statistical Learning Theory, John Wiley: New York, 1998.
5. Cortes, C.; Vapnik, V.; Mach. Learn. 1995, 20, 273.
6. Sun J, Xu W, Feng B, A Global Search Strategy of Quantum- Behaved Particle Swarm Optimization. In Proc. of the 2004 IEEE Conf. on Cybernetics and Intelligent Systems, Singapore: 291 – 294, 2004.
7. Suykens, J. A. K.; Vandewalle, J.; Neural Process. Lett. 1999, 9, 293.
8. Suykens, J. A. K.; van Gestel, T.; de Brabanter, J.; de Moor, B.; Vandewalle, J.; Least-Squares Support Vector Machines, World Scientifics: Singapore, 2002.
9. Zou, T.; Dou, Y.; Mi, H.; Zou, J.; Ren, Y.; Anal. Biochem. 2006, 355, 1.
10. Ke, Y.; Yiyu, C.; Chinese J. Anal. Chem. 2006, 34, 561.
11. Niazi, A.; Ghasemi, J.; Yazdanipour, A.; Spectrochim. Acta Part A 2007, 68, 523.
12. Millie Pant, Radha Thangaraj, and Ajith Abraham. 2008. A new quantum behaved particle swarm optimization. In Proceedings of the 10th annual conference on Genetic and evolutionary computation (GECCO '08), Maarten Keijzer (Ed.). ACM, New York, NY, USA, 87-94. DOI=10.1145/1389095.1389108 http://doi.acm .org/ 10.1145/1389095.1389108.
13. Liu J, Sun J, Xu W, Quantum-Behaved Particle Swarm Optimization with Adaptive Mutation Operator. ICNC 2006, Part I, Springer-Verlag: 959 – 967, 2006.
14. M. Barni, F. Bartolini, and A. Piva, "Improved Wavelet- Based Watermarking Through Pixel-Wise Masking," IEEE Transactions on Image Processing, Vol. 10, No. 5, IEEE, pp. 783-791, May 2001.
15. Dennis, A., Wixom, B. H. and Tegarden, D. (2002), Systems Analysis and Design: An Object-Oriented Approach, John Wiley & sons, New York.
16. Roger S. Pressman, “Software Engineering a practitioner’s approach”, McGraw-Hill, 5th edition, 200.
17. M M Lehman,”Process Models, Process Programs, Programming Support”, ACM, 1987.

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Energy Efficient Routing Protocols and Algorithms for Wireless Sensor Networks – A Survey

By R.A.Roseline, Dr.P.Sumathi
Chikkanna Government Arts College, Tirupur, India

Abstract - Wireless Sensor Networks (WSNs) are an emerging technology for monitoring physical world. The sensor nodes are capable of sensing various types of environmental conditions, have some processing capabilities and ability to communicate the sensed data through wireless communication. Routing algorithms for WSNs are responsible for selecting and maintaining the routes in the network and ensure reliable and effective communication in limited periods. The energy constraint of WSNs make energy saving become the most important objective of various routing algorithms. In this paper, a survey of routing protocols and algorithms used in WSNs is presented with energy efficiency as the main goal.

Keywords : *Wireless Sensor Networks, Routing Protocols, energy efficiency.*

GJCST Classification : *C.2.1*

Strictly as per the compliance and regulations of:

Energy Efficient Routing Protocols and Algorithms for Wireless Sensor Networks – A Survey

R.A.Roseline[α], Dr.P.Sumathi[Ω]

Abstract - Wireless Sensor Networks (WSNs) are an emerging technology for monitoring physical world. The sensor nodes are capable of sensing various types of environmental conditions, have some processing capabilities and ability to communicate the sensed data through wireless communication.

Routing algorithms for WSNs are responsible for selecting and maintaining the routes in the network and ensure reliable and effective communication in limited periods. The energy constraint of WSNs make energy saving become the most important objective of various routing algorithms. In this paper, a survey of routing protocols and algorithms used in WSNs is presented with energy efficiency as the main goal.

Keywords : Wireless Sensor Networks, Routing Protocols, energy efficiency.

I. Introduction

Wireless Sensor Networks (WSN) are found in many applications including environmental monitoring, health applications, military surveillance, habitat monitoring and smart homes. A Wireless Sensor Network consists of many sensor nodes deployed in environment and connected to a base station that processes the sensed data from the sensors. One of the key characteristics of sensor nodes is that they are energy constrained[9]. Typically sensor nodes rely on finite energy sources like battery for power in unmanned positions.

Due to massive number of deployment and remote, unattended positions, replacements of batteries are quite impossible. Harvesting energy from the environment is currently a promising but under developed research area and therefore, energy has to be used judiciously. The expectancy of longer lifetime of sensor nodes has put researchers to work on every possible aspect of sensor nodes in gaining energy efficiency.

II. Classification of Sensor Networks and Design Objectives

Sensor Networks can be classified on the basis of their mode of functioning and the type of target application into two major types. They are

Author α : Post Graduate and Research Department of Computer Science, Government Arts College, Coimbatore, India. E-mail : roselinera@yahoo.com

Author Ω: Department of Computer Science, Chikkanna Government Arts College, Tirupur, India. E-mail : sumathirajes@hotmail.com

a) Proactive Networks

The nodes in this network switch on their sensors and transmitters periodically, sense the data and transmit the sensed data. They provide a snapshot of the environment and its sensed data at regular intervals. They are suitable for applications that require periodic data monitoring like moisture content of a land in agriculture.

b) Reactive Networks

The nodes in this network react immediately to sudden and drastic changes in the value of the sensed attribute. They are therefore suited for time critical applications like military surveillance or temperature sensing.

Most sensor networks are application specific and have different application requirements. Thus, all or part of the following main design objectives is considered in the design of sensor networks[11,13]:

(i) ***Small node size:*** Since sensor nodes are usually deployed in a harsh or hostile environment in large numbers, reducing node size can facilitate node deployment. It will also reduce the power consumption and cost of sensor nodes.

(ii) ***Low node cost:*** Since sensor nodes are usually deployed in a harsh or hostile environment in large numbers and cannot be reused, reducing cost of sensor nodes is important and will result into the cost reduction of whole network.

(iii) ***Low power consumption:*** Since sensor nodes are powered by battery and it is often very difficult or even impossible to charge or recharge their batteries, it is crucial to reduce the power consumption of sensor nodes so that the lifetime of the sensor nodes, as well as the whole network is prolonged.

(iv) ***Reliability:*** Network protocols designed for sensor networks must provide error control and correction mechanisms to ensure reliable data delivery over noisy, error-prone, and time-varying wireless channels.

(v) ***Scalability:*** Since the number sensor nodes in sensor networks are in the order of tens, hundreds, or thousands, network protocols designed for sensor networks should be scalable to different network sizes.

(vi) ***Self-configurability:*** In sensor networks, once deployed, sensor nodes should be able to autonomously organize themselves into a communication network and reconfigure their in the event of topology changes and node failures.

(vii) ***Channel utilization:*** Since sensor networks have limited bandwidth resources, communication protocols designed for sensor networks should efficiently make use of the bandwidth to improve channel utilization.

(viii) ***Fault tolerance:*** Sensor nodes are prone to failures due to harsh deployment environments and unattended operations. Thus, sensor nodes should be fault tolerant and have the abilities of self-testing, self-calibrating, self-repairing, and self-recovering.

(ix) ***Adaptability:*** In sensor networks, a node may fail, join, or move, which would result in changes in node density and network topology. Thus, network protocols designed for sensor networks should be adaptive to such density and topology changes.

(x) ***Security:*** A sensor network should introduce effective security mechanisms to prevent the data information in the network or a sensor node from unauthorized access or malicious attacks.

III. Energy Efficient Wireless Sensor Network Protocols

Protocols for Sensor networks must be designed in such a way that the limited power available at the sensor nodes is efficiently used. Routing in WSN is quite challenging due to its inherent constraints and basic characteristics that distinguish WSN from other wireless networks. They are

- There is no global addressing scheme in WSN. Therefore, routing protocols of IP based networks cannot be used in WSN.
- Characteristic of data flow in WSN is a bit different. Data from multiple nodes actually go to a single point that is a sink or base station.
- Data from multiple sources can create significant redundancy in the data traffic.
- Nodes are tightly constrained about resources.

There are a handful number of routing protocols have been proposed for WSN. These protocols can be broadly categorized into six different types, namely, data - centric, hierarchical, location-aware, mobility based, heterogeneity – based and Quality of Service (QoS) based.

a) Data Centric Protocols

Data-centric protocols aim at aggregating the data by the intermediate sensors on the data originating from the source sensors and send the aggregated data toward the sink. This results in energy savings due to lesser transmission required to send the data from the sources to the sink. In this section, some the data-centric, energy efficient routing protocols for WSNs are discussed.

i. *Directed Diffusion*

Directed diffusion [7, 8] is a data-centric routing protocol for sensor query dissemination and processing. It is energy-efficient, scalable and robust.

A sensing task is described by a list of attribute-value pairs. The sink specifies a low data rate for incoming events at the beginning of the directed diffusion process. The sink can thereafter reinforce one particular sensor to send events with a higher data rate by resending the original interest message with a smaller interval.

ii. *Sensor Protocols for Information via Negotiation (SPIN)*

SPIN [10, 23] protocol was developed to overcome the problems like implosion and overlap caused by flooding protocols. The SPIN protocols are able to compute the energy consumption required to compute, send, and receive data over the network.

SPIN uses meta-data as the descriptors of the data that the sensors want to disseminate. The notion of meta-data avoids the occurrence of overlap given the sensors can name the interesting portion of the data they want to get. The size of the meta data should be less than that of the corresponding sensor data.

SPIN-1(SPIN_PP) uses negotiation mechanism to reduce the consumption of the sensors. SPIN-2(or SPIN-EC) uses a resource –aware mechanism for energy savings.

iii. *Energy-Aware Data-Centric Routing(EAD)*

EAD[1] is energy aware and helps extend network lifetime.EAD is a distributed routing protocol, which builds a virtual backbone composed of active sensors that are responsible for in-network data processing and traffic relaying.

The network is represented by a broadcast tree spanning all sensors in the network and routed at the gateway, in which all leaf nodes' radios are turned off while all other nodes correspond to active sensors forming the backbone and thus their radios are turned on.

b) Hierarchical Protocols

Hierarchical clustering in WSN is an energy efficient protocol with three main elements: sensor nodes (SN), Base station (BS) and Cluster Heads (CH). The SNs are sensors deployed in the environment to collect data. The main task of a SN in a sensor field is to detect events, perform quick local data processing, and transmit the data. The BS is the data processing point for the data received from the sensor nodes, and from where the data is accessed by the end-user. The CH acts as a gateway between the SNs and BS. The CH is

the sink for the cluster nodes, and the BS is the sink for the cluster heads. This structure formed between the sensor nodes, the sink and the base station can be replicated many times, creating the different layers of the hierarchical WSN.

i. *Low Energy Adaptive Clustering Hierarchy(LEACH):*

LEACH [24, 25] was the first dynamic energy efficient cluster head protocol proposed for WSN using homogeneous stationary nodes .

In LEACH all nodes have a chance CH and therefore energy spent is balanced for every node. The CH for the Clusters are selected based on their energy load. After its election, the CH broadcasts a message to other nodes, which decide which cluster they want to belong to, based on the signal strength of the CH. The clusters are formed dynamically in each round and the data collection is centralised. A TDMA schedule created by the CH is used to gather data from the sensors. The operation of LEACH is divided into rounds having two phases each namely

c) a setup phase to organize the network into clusters, CH advertisement, and transmission schedule creation and

d) a steady phase to for data aggregation, compression and transmission to the sink.

LEACH reduces energy consumption by

a. minimizing the communication cost between sensors and their CH.

b. Turning off non-head nodes when not required.

ii. *Power-Efficient Gathering in Sensor Information Systems (PEGASIS):*

PEGASIS [20] is an extension of the LEACH protocol, and simulation results show that PEGASIS is able to increase the lifetime of the network twice as much as the LEACH protocol.

PEGASIS forms chains from sensor nodes , each node transmits the data to neighbour or receives data from a neighbour and only one node is selected from that chain to transmit data to the BS. The data is finally aggregated and sent to the BS. PEGASIS avoids cluster formation, and assumes that all the nodes have knowledge about the network , particularly their positions using a greedy algorithm. Although clustering overhead is avoided, PEGASIS requires dynamic topology adjustment since the energy status of its neighbour is necessary to know where to route its data. This involves significant overhead particularly in highly utilised networks.

iii. *Threshold Sensitive Energy Efficient Sensor Network Protocol (TEEN):*

TEEN [3] is a energy efficient hierarchical clustering protocol which is suitable for time critical applications TEEN has SNs reporting data to CHs. The CH sends aggregated data to the next higher level CH until data reaches the sink. TEEN is designed for reactive networks, where the sensor nodes react immediately to sudden changes in the value of the sensed attribute. Sensor nodes sense the environment continuously, but data transmission is done occasionally and this helps in energy efficiency. This protocol sends data if the attribute of the sensor reaches a Hard Threshold and a small change -the Soft Threshold. The drawback of this protocol is that if the threshold is not reached, the nodes may not communicate and. we do not know if a node is dead.

iv. *Adaptive Periodic Threshold Sensitive Energy Efficient Sensor Network Protocol (APTEEN):*

APTEEN [2] is an improvement to TEEN and aims at periodic data collection and reacting to time critical events. It is a hybrid clustering based protocol and supports different types of queries like (i) historical query, to get results on past data (ii) one-time query that gives a snapshot of the environment and (iii) persistent queries, to monitor an event for a time period. The cluster head selection in APTEEN is based on the mechanism used in LEACH-C. The cluster exists for an interval called the cluster period, and then the BS re-groups clusters, at the cluster change time.

APTEEN used modified TDMS, where each node in the cluster is assigned a transmission slot, to avoid collisions. For query responses it uses node pairs. If adjacent nodes sense similar data, only one of them responds to a query, the other one goes to sleep mode and thereby saves energy.

v. *Hybrid, Energy-Efficient Distributed Clustering (HEED)*

HEED [16, 17] is an extension of LEACH and uses residual energy and node degree or density asymmetric for cluster selection to achieve power balancing. HEED has the following features.

(i) prolongs network lifetime by distributing energy consumption,
(ii) terminates clustering process within a constant number of iterations,
(iii) minimizes control overhead and
(iv) produces well distributed CHs and compact clusters.

HEED selects CHs based on the residual energy of the SNs and intra-cluster communication cost as a function of cluster density or node degree. HEED clustering improves network lifetime over LEACH clustering randomly selects CHs and cluster size and therefore nodes die faster.

vi. *Clustered Aggregation Technique (CAG)*

CAG[29] is a protocol for reactive networks and the first in-network aggregation algorithm exploiting spatial correlation, which trades a negligible quality of result (precision) for a significant energy saving. CAG forms clusters of nodes sensing similar values .The CAG algorithm operates in two phases: query and

response. During the response phase, CAG transmits the value of aggregated data within the cluster to the BS. CAG achieves efficient in-network storage and processing by allowing a unified mechanism between query routing (networking)and query processing (application). CAG generates synopsis by filtering out insignificant elements in data streams to minimize response time, storage, computation, and communication cost. CAG uses only sensor values from the cluster heads to compute the aggregates and so it is a lossy clustering algorithm.

vii. *Updated CAG Algorithm*

Updated CAG Algorithm[30] is an improvement of CAG algorithm, where the clusters are still formed from nodes sensing similar values within a given threshold, but in this case, the clusters remain as long as the sensor values stay within a given threshold over time(temporal correlation). This ensures that the performance of CAG become independent of the magnitude of sensor readings and network topology. When used in the interactive mode, the protocol alternates query and response phases. This algorithm builds a new forwarding tree each time a query is sent out. This rebuilding of trees frequently is a waste if the sensed data is almost the same over time.

viii. *Energy Efficient Homogeneous Clustering Algorithm for Wireless Sensor Networks*

Energy Efficient Homogeneous Clustering Algorithm for Wireless Sensor Networks [21] is a algorithm that proposes homogeneous clustering for WSNs that save power and prolongs network life. The life span of the network is increased by homogeneous distribution of nodes in the clusters. A new CH is selected based on the residual energy of existing cluster heads, holdback value and nearest hop distance of the node. The cluster members are uniformly distributed , and thus , the life of the network is extended.

c) Location-Based Protocols

Sensor nodes are addressed by means of their locations in location based protocols. Energy consumption is estimated by the distance between two sensor nodes and so location information is essential. Some queries from sensor nodes are also location specific and so location-based sensors find a wide number of applications. We present some location-based protocols in this section.

i. *Geographic and Energy-Aware Routing (GEAR)*

GEAR[27] is an energy-efficient routing protocol for routing queries to target regions in a sensor field. Sensors have localization hardware like GPS so that they know their current positions. The sensors are aware of their locations and their residual energy and also the locations and residual energy of their neighbours. GEAR uses a recursive geographic forwarding algorithm to disseminate the packet inside the target regions for data communication. GEAR also uses energy aware heuristics that are based on geographical information to select sensors to route a packet towards its destination.

ii. *Geographic Adaptive Fidelity (GAF)*

GAF [28] is an energy aware routing protocol proposed for MANETs but can also be used for WSNs because it aims at energy conservation. GAF turns off unnecessary sensors while keeping a constant level of routing fidelity (or uninterrupted connectivity between communicating sensors). A sensor field is divided into grid squares and every sensor uses its location information, which can be provided by GPS or other location systems. The sensor associates itself with a particular grid and this helps GAF to identify the sensors.

The state transition diagram in GAF has three states:

(i) *Sleeping state:* A sensor turns off its radio in the sleeping state.

(ii) *Discovery state:* A sensor exchanges discovery messages to learn about other sensors in the same grid.

(iii) *Active state:* A sensor periodically broadcasts its discovery message to inform equivalent sensors about its state. GAF aims to maximize the network lifetime by reaching a state where each grid has only one active sensor based on sensor ranking rules. The residual energy levels helps to rank the sensors. A sensor with a higher rank handles routing within their corresponding grids.

iii. *Minimum Energy Communication Network(MECN):*

MECN [22] is a location-based protocol for achieving minimum energy for randomly deployed networks, which uses mobile sensors to maintain a minimum energy network. It computes an optimal spanning tree with sink as root that contains only the minimum power paths from each sensor to the sink. This tree is called minimum power topology. It has two phases:

(i) Enclosure Graph Construction: MECN constructs sparse graph, called a enclosure graph, based on the immediate locality of the sensors. An enclosure graph is a directed graph that includes all the sensors as its vertex set and edge set is the union of all edges between the sensors and its neighbours located in their enclosure regions.

(ii) Cost distribution: In this phase non-optimal links of the enclosure graphs are simply eliminated and the resulting graph is a minimum power topology. This graph has a directed path from each sensor to the sink and consumes the least total power among all graphs having directed paths from each sensor to the sink. Every sensor broadcasts its cost to its neighbours, where the cost of a node is the minimum power required for this sensor to establish a directed path to the sink.

iv. *Small Minimum-Energy Communication Network (SMECN)*

SMECN[14] is a routing protocol that improves MECN by constructing a minimal graph characterised with regard to the minimum energy property. This property ensures that there s minimum energy-efficient path between any pair of sensors in a graph that has the smallest cost in terms of energy consumption over all possible paths. In SMENC protocol every sensor broadcasts a neighbour discovery message using some initial power to discover its neighbours. It then checks whether the theoretical set of neighbours that are computed analytically is a subset of the set of the set of sensors that replied to that neighbour discovery message. The sensor uses a corresponding power p to communicate with its immediate neighbours for this case and else it increments p and rebroadcasts its neighbour discovery message.

v. *Coordination of Power Saving with Routing(SPAN)*

SPAN[4,5] is a routing protocol is applied to WSNs though it was proposed for MANETs since it is energy efficient. This protocol turns off the radio when not in use since the wireless network interface of a device is often the single largest consumer of power. Span helps sensors to join a forwarding backbone topology as coordinators that will forward packets on behalf of other sensors between any source and destination.

d) Heterogeneous-Based Protocols

Heterogeneous-based protocols ate used for heterogeneous networks where there are two types of sensors namely line-powered sensors that have no energy constraint, and battery-powered sensors having limited lifetime. The battery powered sensors have limited energy and so protocols should minimize their data communication and computation. We present some heterogeneous-based protocols in this section.

i. *Cluster-Head Relay Routing (CHR)*

CHR Routing protocol[26] uses two types of sensors to form a heterogeneous network with a single sink:

(i) A large number of low-end sensors denoted by L-sensors and
(ii) A small number of powerful high-end sensors denoted by H-sensors

Both types of sensors are static and location-aware. These sensors are randomly deployed over the environment and CHR partitions the heterogeneous network into clusters or groups of sensors with L-sensors and headed by a H-sensor.

Within the cluster, the L-sensors sense the environment and send the data to H-sensor in multihop routing. The H-sensors are responsible for data fusion within their own clusters and forwards them to the sink. Therefore H-sensors are used for long-range data communication to the sink and other H-sensors and L-sensors are used for short-range data communication between L-sensors and its cluster head.

ii. *Information-Driven Sensor Query (IDSQ)*

IDSQ[15,19] maximises information gain and minimises detection latency and energy consumption for target localization and tracking by dynamic sensor querying and data routing. In order to conserve energy only a subset of sensors are active at times when there are critical events to report in some parts of the sensed network. The choice of this active subset of sensors is balanced by the communication costs needed for communication of all sensors. A leaded is selected in this protocol that decides the optimal subset of sensors necessary for information sensing from the network.

e) Mobility-Based Protocols

Mobility based protocols have mobile sinks that are responsible for data collection from the network. In this section, we discuss mobility-based protocols that aim at energy efficiency.

i. *Data MULES Based Protocols*

Data MULEs(Mobile Ubiquitous LAN extensions) are used collect data from a sparse network while reducing energy consumption of the sensors[18].The MULE architecture has three main layers:

(i) The bottom layer consists of wireless static sensors that are responsible for sensing the environment.
(ii) The middle layer has mobile entities(MULEs) that collect the data from the sensors by moving in proximity to them and deliver them to access points.
(iii) The top layer contains WAN connected devices and access points/central repositories for analysing the sensed data. These access points communicate with a central data warehouse and synchronise the data collected, reduces redundant data and acknowledges the receipt of the of the data sent by the MULEs.

ii. *Scalable Energy - Efficient Asynchronous Dissemination(SEAD):*

SEAD [6] is proposed to trade-off between minimizing the forwarding delay to a mobile sink and energy savings. The source sensor reports sensed data to multiple mobile sinks and the protocol consists of three main components namely:

(i) Dissemination tree(d-tree) construction,
(ii) Data dissemination and
(iii) Maintaining links to mobile sinks.

SEAD assumes that sensors are aware of their own geographic locations. Data dissemination tree is built for every sensor routed at itself and all the dissemination trees for other sensor nodes are constructed separately. SEAD sits on top of a location aware routing protocol and can be viewed as an overlay network.

f) Quality of Service-Based Protocols

Quality of Service (QoS) requirements like delay, reliability and fault tolerance are as important in routing in WSNs as energy efficiency. A routing protocols that support QoS with energy efficiency is discussed in this section.

i. *Energy-Aware QoS Routing Protocol*:

Real-time traffic is generated by imaging sensors in this QoS energy aware routing protocol [12].This protocol finds the least cost and energy efficient path and the link cost is a function that captures the nodes' energy reserve , transmission energy, error rate and some communication parameters. The queuing model allows service sharing for real time and non-real-time traffic. This algorithm performs well with respect to QoS and energy metrics.

IV. Conclusion

The ultimate aim of a routing protocol design is to extend the lifetime of the network by keeping the sensors alive for a maximum time. Since energy spent on transmission is very high compared to that of sensing, the routing algorithm should be designed to reduce energy consumption while transmitting data.

In this paper, different routing protocols and algorithms based on data-centric routing, grouping or clustering of sensors, location information, network heterogeneity and QoS have been discussed. This survey helps in understanding the working of these protocols and the advantages of these algorithms combined together may be a good research direction for future applications.

References References Referencias

1. A. Boukerche, X. Cheng, and J. Linus, "Energy-aware data - centric routing in microsensor networks", Proceedings ACM MSWiM, in conjunction with ACM MobiCom, San Diego, CA, Sept. 2003.
2. A.Manjeshwar and D.P. Agarwal," APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless in Wireless Sensor Networks ",in the Proceedings of the 2nd International Workshop of Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco CA, April 2001.
3. A.Manjeshwar and D.P. Agarwal, "TEEN : A Protocol for Enhanced Efficiency in Wireless Sensor Networks", in the Proceedings of the 1 st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA, April 2001.
4. B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks", Proceedings *ACM MobiCom'01*, Rome, Italy, July 2001.
5. B. Chen, K. Jamieson, H. Balakrishnan, and R. Morris, "Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks", Wireless *Networks,* vol. 8, no.5 Sept 2002.
6. B. Karp and H.T. Kung, "GPSR: Greedy perimeter stateless routing for wireless networks'', *proceedings ACM MobiCom '00,* Boston, MA ,Aug .2000 , pp . 243-254.
7. C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva "Directed diffusion for wireless sensor networking",*IEEE/ACM Transactions on Networking,* vol. 11.,no. 1, Feb. 2003.
8. C.Intanagonwiwat , R.Govindan , and D.Estrin, "Directed diffusion : A scalable and robust communication paradigm for sensor networks", *Proceedings ACM MobiCom'00,* Boston, MA, Aug.2000
9. I.F. Akyildiz , W .Su , Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey" ,*computer networks (Elsevier)Journal,* Vol.38, no.4 ,Mar.2002,pp. 393-422 .
10. J. Kulik, W. Heinzelman, and H. Balakrishnan, "Negotiation - based protocols for disseminating information in wireless sensor networks", *Wireless Networks,* vol. 8, no. 2/3, Mar.-May 2002.
11. Jamal Al-Karaki, and Ahmed E.Kamal, "Routing Technique in Wireless Sensor Networks: A Survey", IEEE Communications Magazine, vol 11,no.6, Dec.2004,pp 6-28.
12. K.Akkaya and M. Younis, "An Energy-Aware QoS Routing Protocol for Wireless Sensor Networks," in the *Proceedings of the IEEE Workshop on Mobile and Wireless Networks(MNV2003)*,Providence, Rhode Island ,May2003.
13. Kemal Akkaya and Mohamed Younis, "A Survey on Routing Protocols for Wireless Sensor Networks", *Ad hoc Networks*, vol. 3, no. 3, May 2005.
14. L. Li and J. Y. Halpern, "Minimum-energy mobile wireless networks revisited", *Proceedings IEEE ICC'01,* Helsinki, Finland, June 2001.
15. M. Chu, H. Haussecker, and F. Zhao, "Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks", *Interternational Journal of High performance computing Applications,* vol.16, no.3, Feb.2002 , pp. 293 - 313.
16. Ossama Younis and Sonia Fahmy "Heed: A Hybrid, energy, Distributred Clustering Approach for Ad-hoc Networks:, *IEEE Transactions on Mobile Computing Issues in Wireless Networks and Mobile Computing*,vol 3,no. 4, Dec 2004.pp.366-369.
17. Ossama Younis and Sonia Fahmy, "Distributed Clustering in Ad-hoc sensor Networks:A Hybrid,

Energy - efficient Approach", September 2002. International Journal of Computer Science.
18. R.C.Shah , S. Roy , S.Jain, and W .Brunette, "Data MULEs: Modeling a three –tier architecture for sparse sensor networks" ,*proceedings SN P A'03,* Anchorage , AK, May 2003 , pp. 30-41.
19. S. Lindsey, C.S. Raghavendra, and K.M. Sivalingam, "Data gathering in algorithms in sensor networks using energy metrics '', *IEEE Transactions on Parallel and Distributed Systems,*vol. 13, no. 9,Sept. 2002,pp.924-95.
20. S. Lindsey and C.S. Raghavendra, "PEGASIS: Power-efficient Gathering in Sensor Information System", *Proceedings IEEE Aerospace Conference*, vol. 3, Big Sky, MT,Mar. 2002, pp. 1125-1130.
21. S.K.Singh, M.P.Singh and D.K.Singh, "Energy–efficient Homogeneous Clustering Algorithm for Wireless Sensor Network", International Journal of Wireless & Mobile Networks(IJWMN), Aug.2010,vol.2.no 3,pp 49-61
22. Rodoplu and T. H. Meng, "Minimum energy mobile wireless networks", *IEEE Journal on Selected Areas in Communications,* vol. 17, no. 8, Aug. 1999.
23. W. R. Heinzelman, J. Kulik, and H. Balakrishnan, "Adaptive protocols for information dissemination in wireless sensor networks", *Proceedings ACM MobiCom* '99, Seattle, *WA,* Aug.1999.
24. W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy- efficient Communication Protocol for Wireless Microsensor Networks", in IEEE Computer Society *Proceedings of the Thirty Third Hawaii International Conference on System Sciences (HICSS '00),* Washington, DC, USA, Jan. 2000, vol. 8, pp. 8020
25. W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks" in *IEEE Tmnsactions on Wireless Communications* (October 2002), vol. 1(4)
26. X. Du and F. Lin, "Improving routing in sensor networks with heterogeneous sensor nodes", Proceedings IEEE VTC '05, Dallas TX, Sept 2005.
27. Y.Xtu, J. Heidemann and D.Estrin, "Geography-informed energy conservation for ad-hoc routing", Proceedings ACM/IEEE MobiCom '01, Rome, Italy, July 2001.
28. Y.Yu, R.Govindan and D.Estrin, "Geographical and energy aware routing: A recursive data dissemination protocol for wireless sensor networks:, *Technical Report,UCLA/CSD-TR-01-0023, UCLA,* Computer Science Department, Mat 2001.
29. Yoon S., Shahabi C.,"Exploiting Spatial Correlation Towards an Energy Efficient Clustered Aggregation Techique (CAG)", IEEE Conference on Communications, 2005.
30. Yoon S., Shahabi C., "An Experimental Study of the Effectiveness of Clustered Aggregation (CAG) Levaraging Spatial and Temporal Correlations in Wireless Sensor Networks" ACM Transactions on Sensor Networks,USC,CS Dept Technical Report 05-869, August 2005.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Volume 11 Issue 21 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Automatic License Plate Recognition (ALPR) for Bangladeshi Vehicles

By Ashim Kumar Ghosh, Swapan K.D. Sharma, Md. Nazrul Islam, Suchitra Biswas, Sameya Akter

haka University of Engineering and Technology

Abstract - This paper presents Automatic License Plate extraction, character segmentation and recognition method for license plate of Bangladeshi vehicles with chain code and neural network. In Bangladesh, license plate models are not followed strictly. Characters on plate are in Bangla and English languages and also are in one or two lines. Due to dissimilarity in the model of license plates, vehicle license plate extraction, character segmentation and recognition are key issue. Different types of algorithm already applied and the performance is examined for English license plate. We describe the license plate extraction, character segmentation and recognition work, with Bangla characters. License plate extraction is performed using Sobel filter, connected component analysis and morphological operations. Character segmentation is performed in different levels by using scanning the binary image horizontally and vertically and connected component analysis. Character recognition is carried out using chain code generation and stored knowledge of the network.

Keywords : *Vehicle license plate, line segmentation, word segmentation, character recognition, chain code, neural network.*

GJCST Classification : *I.5.4*

Strictly as per the compliance and regulations of:

Automatic License Plate Recognition (ALPR) for Bangladeshi Vehicles

Ashim Kumar Ghosh [α], Swapan K.D. Sharma [Ω], Md. Nazrul Islam [β], Suchitra Biswas [Ψ], Sameya Akter [¥]

Abstract - This paper presents Automatic License Plate extraction, character segmentation and recognition method for license plate of Bangladeshi vehicles with chain code and neural network. In Bangladesh, license plate models are not followed strictly. Characters on plate are in Bangla and English languages and also are in one or two lines. Due to dissimilarity in the model of license plates, vehicle license plate extraction, character segmentation and recognition are key issue. Different types of algorithm already applied and the performance is examined for English license plate. We describe the license plate extraction, character segmentation and recognition work, with Bangla characters. License plate extraction is performed using Sobel filter, connected component analysis and morphological operations. Character segmentation is performed in different levels by using scanning the binary image horizontally and vertically and connected component analysis. Character recognition is carried out using chain code generation and stored knowledge of the network.

Keywords : Vehicle license plate, line segmentation, word segmentation, character recognition, chain code, neural network.

I. Introduction

Automatic License Plate Recognition (ALPR) is an important area of research due to its applications. It is a machine vision technique used to identify vehicles by their license plates without direct human involvement. The Intelligent Transportation System provides the data of vehicle information which can be used in follow up, analysis and monitoring. The complexity of automatic license plate recognition work varies throughout the world. For the standard license plate, ALPR system is easier to read and recognize. In Bangladesh this task becomes much difficult due to variation in plate model and specialties of Bangla scripts.

Bangla text can be partitioned into three zones. The upper zone indicate the portion above the headline called 'Marta', the middle zone indicate the portion of basic characters and compound characters under the head-line and lower zone is the portion where some of the modifiers can exist in .The ALPR algorithm consists of three steps: license plate locating as well as true license plate extraction, character segmentation and character recognition. From the input image, the license plate detection, noise removal, invert and skew detection is performed from license plate in license plate extraction phase. In character segmentation phase each and every character is isolated and segmented accordingly followings step line segmentation, word segmentation and character segmentation, based on the selection of good features of characters, each character is recognized using neural network, in the character recognition phase. Extraction of license plate is difficult job; mainly due to license plate occupy a small part of whole image, difference in license plate models and cause of environmental factors. This step affects the accuracy of character segmentation and recognition work.

The rest of this paper is prepared as follows segment II shortly illustrates the applications of the ALPR system in different areas, Segment III described related works, Segment IV described true license plate extraction process, Segment V includes steps of the character segmentation and Segment VI includes steps character recognition process. Segment VII discusses its experimental results and lastly Segment VIII concludes the paper.

II. Application

ALPR system is mostly used in Intelligent Transportation System. ALPR is important in the area of highway toll collection, traffic problems, borders and custom security, premises where high security is needed, like national assembly, V.I.P houses and so on.

III. Related Work

Different techniques are developed for license plate extraction. Hao Chen [1] et al planned the method, several candidates based on texture information similar

Author [α] : B.Sc. in engineering degree in Computer Science and Engineering (CSE) from Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh. Ph.:+88 01920092314. E-mail : ashim.cse06@gmail.com

Author [Ω] : B.Sc. in engineering degree in Computer Science and Engineering (CSE) from Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh. Ph: +88 01913570037. E-mail : Swapan_cse48@yahoo.com

Author [β] : Assistant Professor in Computer Science and Engineering (CSE) with Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh. Ph: +88 01732183690. E-mail : nazrul_ruet@yahoo.com

Author [Ψ] : B.Sc. in engineering degree in Computer Science and Engineering (CSE) from Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh. Ph: +88 01918173506 E-mail : laxmi.chitra@yahoo.com

Author [¥] : B.Sc. in engineering degree in Computer Science and Engineering (CSE) from Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh. Email : sameyaakter@yahoo.com , PH : +88 01726133150.

to license plate are extracted and auto-correlation based binary image and projection algorithm are used to verify the true candidate plate. Gisu Heo [2] developed license plate detection technique using group of lines forming rectangle at the plate boundary, followed by this step is the vertical edge density technique to find out the plate area. Ozbay et al [3] developed smearing algorithm to locate the license plate. Mei Yu et al [4] proposed vertical edge detection followed by size, shape filter for edge area and edge matching technique based on plate model. Farhad Faradji et al [5] first used Sobel edge detection on the image. Next, vertical projection analysis was used to locate plate area. False candidates were removed using compact factor, which estimated the densest vertical edge area declaring true license plate. Every character on detected license plates is segmented in character segmentation step. Segmentation techniques based on projection analysis, Hough transform, region growing are proposed in the text. Xinagjian He et al [6] used horizontal and vertical projection analysis for character segmentation. Yuangang Zhang et al [7] developed character segmentation using Hough Transform. In this, horizontal edges of the plate area were decided initially, using Hough Transform, which helped to segment the characters with the large rotation. Characters were segmented using vertical projection analysis based on the prior knowledge of the plate model. Feng Yang et al [8] developed region growing algorithm for character segmentation. Shen Zheng Wang et al [9] used connected component analysis for character segmentation.

In this paper ALPR work for Bangladeshi car is presented. Images are taken out with different lighting conditions, different background and direction. Histogram equalization, median filter are used which take care of lighting and contrast problem. Sobel vertical edge detection and morphology is employed to locate the license plate. Horizontal and vertical scanning is used to segment the line, word and characters. For character recognition work chain code and neural network is used.

IV. License Plate Extraction

License plate extraction is the key step in ALPR system, which maintains the accuracy of the system significantly. The goal of this phase, given an input image, is to produce a region that contains true license plate.

a) Image capturing and noise removing

In this system a high resolution digital camera is used to capture an image. Images are taken in different background, illumination conditions and at various distances from the camera to vehicle. Images are resized to (1024 X 768). All the processing steps are executed on gray scale image. Preprocessing is mainly used to enhance the processing speed, improve the contrast of the image and to reduce the noise in the image. Therefore, it is necessary to filter this noise before we process the image. Usually the low-pass filter approach is used to reduce the problem of low quality and low contrast in vehicle images. Original image that captured by digital camera and corresponding gray scale image are showed in Fig. 1(a) and (b).

Fig.1 : (a) Original image (b) Gray scale of image

b) Vertical Edge detection

The license plate region contains plentiful edges with respect to background. Sobel edge detection is used to find out the regions which have high pixel variance value. To extract candidate license plate area from the entire image, threshold is used to select rows which are having particular white pixel density. Fig. 2 shows the result of effect of using, Sobel edge detection and threshold [10].

Fig.2 : (a) Sobel vertical edge detection (b) Effect of threshold

c) Candidate Plate Area Detection

Morphological operations aim to remove unrelated objects in the image. Dilation and erosion are used to extract candidate plate areas from the entire image. Sometimes background areas may also get declared as candidate plate. Hence to remove the fake candidates, plate validation is done using the aspect ratio of the plate and horizontal cuts in the license plate [11]. Invert and threshold operation is performed for true license plate extraction.

Fig.3 : (a) Extracted Candidate plate (b) After Inverting and threshold

d) Skew Detection and Correction

The captured image may be skewed, so skew detection and correction necessary to make text lines

horizontal. Skew angle makes the text lines of the document image with the horizontal direction. Skew correction can be accomplished in two steps. Firstly, we will estimate the skew angle and secondly, we will rotate the image by skew angle.

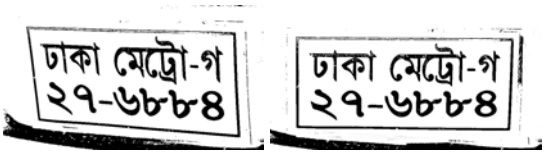


Fig.4 : (a) Skewed image (b) After skew correction

e) True license Plate Extraction

After the detection of candidate license plate area, Bounding Box analysis is used to extract plate area from the image. From the Bounding Box analysis, respective row and column indices of plate area are found out. The result is as shown in Fig. 5.

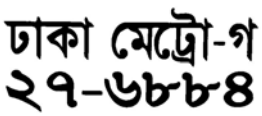


Fig.5 : Extracted true license plate

V. Character Segmentation

Character isolation from the license plate region is very important and crucial step in ALPR system, which influences the accuracy of character recognition significantly. The goal of this phase, given the license plate image, is to segment all the characters, without loosing features of the characters. This phase consists of the sequence of operations as, line segmentation, word segmentation, character segment and connected component analysis.

a) Line segmentation

Line segmentation has been executed by scanning the input image horizontally. Frequency of black pixels in each row is counted in order to construct the row histogram. The position between two consecutive lines, where the number of pixels in a row is zero denotes a boundary between the lines. Line segmentation process shown in figure.

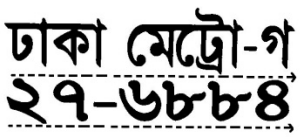


Fig.6 : Line segmentation

b) Word Segmentation

Each line is scanned vertically for word segmentation. Number of black pixels in each column is calculated to construct column histogram. The portion of the line with continuous black pixels is considered to be a word in that line. If no black pixel is found in some vertical scan that is considered as the spacing between words. Thus different words in different lines are separated. So the image file can now be considered as a collection of words. Fig. 7 shows the word segmentation process.

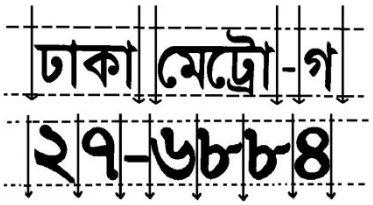


Fig.7 : Word Segmentation

c) Zones of Bangla script

From fig. 8 we see that Bangla text may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic characters below the head-line and lower zone is the portion where some of the modifiers can reside. The imaginary line separating middle and lower zone is called base line.

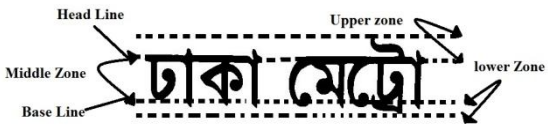


Fig.8 : The zones of bangle script

d) Detection and Deletion of Matra

To segment the character separately from the segmented word, Firstly we find out the headline of the word which is called 'Matra'. From the word, a row histogram is constructed by counting frequency of each row in the word. The row with highest frequency value indicates the headline.

Fig.9 : After Matra elimination

Detection of character between baseline and headline After removing the headline the characters in a word are isolated and can easily be separated. Vertical scan is initiated from the row that is just beneath the 'Matra' row to find the differentiation line between characters. If during scan, one can reach the base line without touching any black pixel then this scan successfully found a differentiation line between characters. Fig. 9 illustrates the character segmentation process.

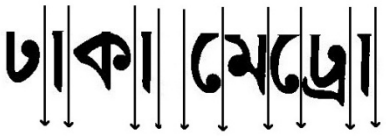


Fig.10 : Character segmentation

e) Detection below the baseline

A greedy search technique is initiated for the presence of black pixels below the baseline, the result will some connected components below the baseline. The components below the baseline contain lowest point called 'Base point'. Baseline is highest frequency row of base points. After determining the baseline, The Depth First Search (DFS) technique easily extracts the characters below the baseline.

Fig.11 : Below the base line detection

VI. Character Recognition

The character recognition phase consists of three steps:

1) Character normalization
2) Feature extraction
3) Character classification and recognition

a) Character Normalization

Segmented characters may have variation in size. In this step, all the characters are normalized to predefine valued in pixel. Characters may have variable width horizontally and vertically, each character image is normalized to a size.

b) Feature Extraction

The objective of feature vector is to define characteristic features of the characters. Selecting the most appropriate feature of each character can facilitate data visualization, data understanding and also reduce the measurement, storage requirements, training and utilization time. Initially, connected component extraction and the centroid of the character image is determined. Chain code was introduced by Freeman [12] as a way to represent lines or boundaries of shapes by a connected sequence of straight line segments of particular length and direction. A chain code has two components .They are: 1) The coordinates of the starting point; 2) A chain of codes; that represents the relative position of the starting pixel and its followers. The chain code is generated by using the changing direction of the connected pixels contained in a boundary. The character has been divided into connected components and boundaries of the connected components are recognized. Freeman chain code works on the observation that each pixel has eight neibourhood pixels is given in fig.12. Transitions are specified for axes with predetermined angles. By monitoring strokes of each character, 13 different angles are decided to count the transitions.

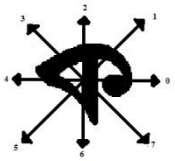


Fig.12 : Slope convention for Freeman chain code

For a closed boundary, its chain code obviously depends on the starting point of the boundary. To make it invariant to the starting point, the chain code can be normalized according to the following method [13]: A chain code can be treated as a circular sequence of direction numbers. Therefore, the starting point can be redefined so that the resulting sequence of numbers forms a minimum integer.

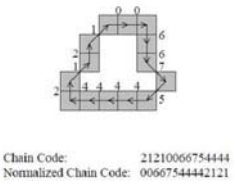


Fig.13 : Bidirectional chain code

c) Classification and Recognition

A neural network [14] is a extremely parallel distributed processor that has a natural propensity for storing experiential knowledge. First knowledge is acquired by the network through a learning process and then storing knowledge is for recognition of character. We use feed forward neural network for the classification and recognition Bangla characters. We trained the neural network by normalized feature vector obtained for each character in the training set. Our layer neural networks have been used with two hidden layers for improving the classification capability. For 32 dimensional feature vectors and 4 layers is used for each Character and recognized using stored knowledge of the network. Use of two hidden layers increases the recognition rate extensively.

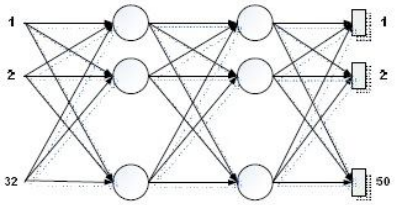


Fig.14 : A neural network with 4 layers

VII. Experimental Results and Performance Analysis

Different sized JPEG colored images are used in our experiment. Total 300 images are used to test the algorithm. The images are taken with different background as well as lighting conditions. Experiments show that the algorithm has good performance on license plate extraction, and character segmentation work. This work is implemented using MATLAB 7.0. Result found 84% for license plate extraction and 80% for character recognition. Deep shadows and reflections have an effect on license plate extraction work. Because of rough lighting, true license plates could not get correctly extracted. Failure in character segmentation phase when two characters are joined together. Good Performance of the ALPR system depends on good feature extraction of character.

VIII. Conclusions

An algorithm for vehicle License plate extraction, character segmentation and recognition is offered. Experiment consists of images with different size, lighting, background, camera angle and distance etc. The experimental results show that, license plates are extracted truly with higher success. Character recognition phase using connected component analysis, Freeman chain code, and Neural network that works well. We suggest to use multilayer feed forward neural network for classify and recognition of character. Recognition performance will be increased if the network trains with distorted character and with good shaped character.

References Références Referencias

1. Hao Chen, Jisheng Ren, Huachun Tan, Jianqun Wang, " A novel method for license plate localization", 4th Proc. of ICIG 2007, pp. 604-609.
2. Gisu Heo, Minwoo Kim, Insook Jung, Duk Ryong Lee,Il Seok Oh, "Extraction of car license plate regions using line grouping and edge density methods", International Symposium on Information Technology Convergence, 2007, pp. 37-42.
3. Serkan Ozbay, Ergun Ercelebi, "Automatic vehicle identification by plate recognition", Proc. of PWASET, vol. 9, no. 4, 2005, pp. 222-225.
4. Mei Yu and Yong Deak Kim, "An approach to Korean license plate recognition based on vertical edge matching", IEEE International Conference on System, Man and Cybernetics, 2000, vol.4, pp. 2975-2980.
5. Farhad Faradji, Amir Hossein Rezaie, Majid Ziaratban,"A Morphological based License Plate Location", ICIP, 2007, pp. I 57-I 60.
6. Xiangjian He et al, "Segmentation of characters on car license plates", 10th Workshop on Multimedia Signal Processing, Oct. 2008, pp. 399-402.
7. Yungang Zhang, Changshui Zhang, "A New algorithm for character segmentation of license plate", Proc. Of IEEE Intelligent Vehicles Symposium, 2003, pp. 106-109.
8. Feng Yang, Zheng Ma, Mei Xie, "A Novel approach for license plate character segmentation", ICIEA, 2006, pp.1-6.
9. Shen-Zheng Wang, His-Jian Lee, "Detection and recognition of license plate characters with different appearances," Proc. of 16th International Conference on Pattern Recognition, vol.3, 2003, pp. 979-983.
10. Rafel C. Gonzalez, Richard E. Woods, "Digital Image Processing," 2nd edition, Prentice Hall, Inc., 2002.
11. Tran Dur Duan, Duong Anh Duc, Tran Le, Hong Du,"Combining Hough transform and contour algorithm for detecting vehicle license plate", Proc. Of International
12. Symposium Intelligent Multimedia, Video and Speech Processing , 2004, pp. 747-750.
13. H. Freeman, "Computer processing of line drawing images," Computer Survey, Vol.6, pp.57-97, 1974.
14. R. Gonzalez and R. E. Woods, Digital Image Processing, Prentice Hall, 2002.
15. S. Haykin, Neural Networks, A Comprehensive Foundation (2nd Edition)
16. http://www.mathworks.com
17. http://en.wikipedia.org

Global Journals Inc. (US) Guidelines Handbook 2011

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

- 'FARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FARSC" can be added to name in the following manner. eg. **Dr. John E. Hall, Ph.D., FARSC or William Walldroff Ph. D., M.S., FARSC**
- Being FARSC is a respectful honor. It authenticates your research activities. After becoming FARSC, you can use 'FARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 60% Discount will be provided to FARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- FARSC will be given a renowned, secure, free professional email address with 100 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- FARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 15% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- Eg. If we had taken 420 USD from author, we can send 63 USD to your account.
- FARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- After you are FARSC. You can send us scanned copy of all of your documents. We will verify, grade and certify them within a month. It will be based on your academic records, quality of research papers published by you, and 50 more criteria. This is beneficial for your job interviews as recruiting organization need not just rely on you for authenticity and your unknown qualities, you would have authentic ranks of all of your documents. Our scale is unique worldwide.
- FARSC member can proceed to get benefits of free research podcasting in Global Research Radio with their research documents, slides and online movies.
- After your publication anywhere in the world, you can upload you research paper with your recorded voice or you can use our professional RJs to record your paper their voice. We can also stream your conference videos and display your slides online.
- FARSC will be eligible for free application of Standardization of their Researches by Open Scientific Standards. Standardization is next step and level after publishing in a journal. A team of research and professional will work with you to take your research to its next level, which is worldwide open standardization.

- FARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), FARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 80% of its earning by Global Journals Inc. (US) will be transferred to FARSC member's bank account after certain threshold balance. There is no time limit for collection. FARSC member can decide its price and we can help in decision.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

- 'MARSC' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'MARSC" can be added to name in the following manner. eg. Dr. John E. Hall, Ph.D., MARSC or William Walldroff Ph. D., M.S., MARSC
- Being MARSC is a respectful honor. It authenticates your research activities. After becoming MARSC, you can use 'MARSC' title as you use your degree in suffix of your name. This will definitely will enhance and add up your name. You can use it on your Career Counseling Materials/CV/Resume/Visiting Card/Name Plate etc.
- 40% Discount will be provided to MARSC members for publishing research papers in Global Journals Inc., if our Editorial Board and Peer Reviewers accept the paper. For the life time, if you are author/co-author of any paper bill sent to you will automatically be discounted one by 60%
- MARSC will be given a renowned, secure, free professional email address with 30 GB of space eg.johnhall@globaljournals.org. You will be facilitated with Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.
- MARSC member is eligible to become paid peer reviewer at Global Journals Inc. to earn up to 10% of realized author charges taken from author of respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account or to your PayPal account.
- MARSC member can apply for free approval, grading and certification of some of their Educational and Institutional Degrees from Global Journals Inc. (US) and Open Association of Research, Society U.S.A.
- MARSC is eligible to earn from their researches: While publishing his paper with Global Journals Inc. (US), MARSC can decide whether he/she would like to publish his/her research in closed manner. When readers will buy that individual research paper for reading, 40% of its earning by Global Journals Inc. (US) will be transferred to MARSC member's bank account after certain threshold balance. There is no time limit for collection. MARSC member can decide its price and we can help in decision.

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJMBR for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5.STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. **Pick a good study spot:** To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. **Think technically:** Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be

sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to

shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic

principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation)
by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350