

GLOBAL JOURNAL

OF COMPUTER SCIENCE AND TECHNOLOGY: C

Software & Data Engineering

Data Leakage Detection

Ipod System's Usability

Highlights

Clone Detection Techniques

Temporal Remote Sensing

Discovering Thoughts, Inventing Future

VOLUME 13

ISSUE 6

VERSION 1.0



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY: C
SOFTWARE & DATA ENGINEERING

GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY: C
SOFTWARE & DATA ENGINEERING

VOLUME 13 ISSUE 6 (VER. 1.0)

OPEN ASSOCIATION OF RESEARCH SOCIETY

© Global Journal of Computer Science and Technology. 2013.

All rights reserved.

This is a special issue published in version 1.0 of "Global Journal of Computer Science and Technology" By Global Journals Inc.

All articles are open access articles distributed under "Global Journal of Computer Science and Technology"

Reading License, which permits restricted use. Entire contents are copyright by of "Global Journal of Computer Science and Technology" unless otherwise noted on specific articles.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without written permission.

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website <http://globaljournals.us/terms-and-condition/menu-id-1463/>

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

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: *Open Association of Research Society*
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.)

John A. Hamilton, "Drew" Jr.,
Ph.D., Professor, Management
Computer Science and Software
Engineering
Director, Information Assurance
Laboratory
Auburn University

Dr. Henry Hexmoor
IEEE senior member since 2004
Ph.D. Computer Science, University at
Buffalo
Department of Computer Science
Southern Illinois University at Carbondale

Dr. Osman Balci, Professor
Department of Computer Science
Virginia Tech, Virginia University
Ph.D. and M.S. Syracuse University,
Syracuse, New York
M.S. and B.S. Bogazici University,
Istanbul, Turkey

Yogita Bajpai
M.Sc. (Computer Science), FICCT
U.S.A. Email:
yogita@computerresearch.org

Dr. T. David A. Forbes
Associate Professor and Range
Nutritionist
Ph.D. Edinburgh University - Animal
Nutrition
M.S. Aberdeen University - Animal
Nutrition
B.A. University of Dublin- Zoology

Dr. Wenying Feng
Professor, Department of Computing &
Information Systems
Department of Mathematics
Trent University, Peterborough,
ON Canada K9J 7B8

Dr. Thomas Wischgoll
Computer Science and Engineering,
Wright State University, Dayton, Ohio
B.S., M.S., Ph.D.
(University of Kaiserslautern)

Dr. Abdurrahman Arslanyilmaz
Computer Science & Information Systems
Department
Youngstown State University
Ph.D., Texas A&M University
University of Missouri, Columbia
Gazi University, Turkey

Dr. Xiaohong He
Professor of International Business
University of Quinipiac
BS, Jilin Institute of Technology; MA, MS,
PhD,. (University of Texas-Dallas)

Burcin Becerik-Gerber
University of Southern California
Ph.D. in Civil Engineering
DDes from Harvard University
M.S. from University of California, Berkeley
& Istanbul University

Dr. Bart Lambrecht

Director of Research in Accounting and Finance
Professor 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 Regina
Ph.D., M.Sc. in Mathematics
B.A. (Honors) in Mathematics
University of Windsor

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., Eötvös Loránd University
Postdoctoral Training,
New York University

Dr. Söhnke M. Bartram

Department of Accounting and Finance
Lancaster University Management School
Ph.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
Neuroscience
Northwestern 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

PRESIDENT EDITOR (HON.)

Dr. George Perry, (Neuroscientist)

Dean and Professor, College of Sciences

Denham Harman Research Award (American Aging Association)

ISI Highly Cited Researcher, Iberoamerican Molecular Biology Organization

AAAS Fellow, Correspondent Member of Spanish Royal Academy of Sciences

University of Texas at San Antonio

Postdoctoral Fellow (Department of Cell Biology)

Baylor College of Medicine

Houston, Texas, United States

CHIEF AUTHOR (HON.)

Dr. R.K. Dixit

M.Sc., Ph.D., FICCT

Chief Author, India

Email: authorind@computerresearch.org

DEAN & EDITOR-IN-CHIEF (HON.)

Vivek Dubey(HON.)

MS (Industrial Engineering),

MS (Mechanical Engineering)

University of Wisconsin, FICCT

Editor-in-Chief, USA

editorusa@computerresearch.org

Sangita Dixit

M.Sc., FICCT

Dean & Chancellor (Asia Pacific)

deanind@computerresearch.org

Suyash Dixit

(B.E., Computer Science Engineering), FICCTT

President, Web Administration and

Development , CEO at IOSRD

COO at GAOR & OSS

Er. Suyog Dixit

(M. Tech), BE (HONS. in CSE), FICCT

SAP Certified Consultant

CEO at IOSRD, GAOR & OSS

Technical Dean, Global Journals Inc. (US)

Website: www.suyogdixit.com

Email: suyog@suyogdixit.com

Pritesh Rajvaidya

(MS) Computer Science Department

California State University

BE (Computer Science), FICCT

Technical Dean, USA

Email: pritesht@computerresearch.org

Luis Galárraga

J!Research Project Leader

Saarbrücken, Germany

CONTENTS OF THE VOLUME

- i. Copyright Notice
 - ii. Editorial Board Members
 - iii. Chief Author and Dean
 - iv. Table of Contents
 - v. From the Chief Editor's Desk
 - vi. Research and Review Papers
-
- 1. Database Autopsy Close Look to Database Auditing for Oracle Database. *1-11*
 - 2. Neuro-Fuzzy based Software Risk Estimation Tool. *13-18*
 - 3. Data Preprocessing in Multi-Temporal Remote Sensing Data for Deforestation Analysis. *19-25*
 - 4. Deriving Association between Student's Comprehension and Facial Expressions using Class Association Rule Mining. *27-33*
 - 5. Data Leakage Detection by using Fake Objects. *35-39*
 - 6. Ipod System's Usability: An Application of the Fuzzy Logic. *41-46*
 - 7. Object Serialization Formats and Techniques a Review. *47-49*
 - 8. A Review of Clone Detection Techniques using Model Semantics. *51-53*
-
- vii. Auxiliary Memberships
 - viii. Process of Submission of Research Paper
 - ix. Preferred Author Guidelines
 - x. Index



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING

Volume 13 Issue 6 Version 1.0 Year 2013

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals Inc. (USA)

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

Database Autopsy Close Look to Database Auditing for Oracle Database

By Elham Iskandarnia

AMA International university, Bahrain

Abstract - Today, that business has different rules and regulation and supplied threats; organizations must go well beyond securing their data, and managing their database. Essentially, Data have to be perpetually monitored to be aware of who what, to all their data. Database auditing involves monitoring database to be aware of what user of proceedings. In this article we will offer a novel procedure for finding auditing records from different locations that DBMS keeps records, further more we will discuss which user, or system activity to keep records to do auditing efficiently and also avoid over use of system resources which will caused on slow transaction time.

GJCST-C Classification : H.2.m



DATABASE AUTOPSY CLOSE LOOK TO DATABASE AUDITING FOR ORACLE DATABASE

Strictly as per the compliance and regulations of:



RESEARCH | DIVERSITY | ETHICS

Database Autopsy Close Look to Database Auditing for Oracle Database

Elham Iskandarnia

Abstract - Today, that business has different rules and regulation and supplied threats; organizations must go well beyond securing their data, and managing their database. Essentially, Data have to be perpetually monitored to be aware of who what, to all their data. Database auditing involves monitoring database to be aware of what user of proceedings. In this article we will offer a novel procedure for finding auditing records from different locations that DBMS keeps records, further more we will discuss which user, or system activity to keep records to do auditing efficiently and also avoid over use of system resources which will caused on slow transaction time.

The Problem and Its Background

1. INTRODUCTION

Today, that business has different rules and regulation and supplied threats, organizations must go well Beyond securing their data, and managing their database. Essentially, Data have to be perpetually monitored to be aware of who what, to all their data. Database auditing involves monitoring database to be aware of what user of proceedings. Database consultant and DBA often set up auditing policy for security purposes, for example, to ensure that everybody get access to what it has permission. This study looks at auditing as a concept, what are threats and how to diagnose that threat when they are happening.

One of challenge regarding auditing that Organizations today are facing is data security, they have to recognize threats and threat them in a more cost effective way. There are some policies which are force by platform the DBA must understand the cost of this policy of database performance and base of need modify or unable these predefine policies as well as create new policies base of their needs.

By understanding the audit concept companies can decrease the increasing cost of database security.

a) Back Ground of Study

Database security is a very broad area that addresses many issues like legal and ethical, policy issues at governmental, or corporate and system-related issues such as system level.

Threats to database results in the loss or degradation of some or all of following commonly accepted security Goals: integrity, availability, and confidentiality.

Grate amount of different errors can leak into organizational databases; these errors may range from data entry errors to violations of accounting standards. It seems that although database systems have radically changed the file system in terms speed and competence, detecting errors and keeping quality did not cope with speed of events.

To protect database against types of threats, it is common to implement four kind of control measurement.

- Access control
- Authentication
- Authorization
- Auditing

Database auditing is directly related to database security. Auditing is one aspect of database security. In practice, if there is need to secure a particular database system, then auditing is essential. Auditing mechanism implemented on a database system facilitates the security implemented in a system.

Three different strategies for database auditing are introduced and compare in term of efficiently (Data Base Auditing, Levant V. Orman, and Cornell University).

In this paper we will discuss Data base Autopsy, that is when a threat happened you use tools to find out what was type of threat, who attack your data base, when that happened. And from this information you can review and change your Authentication, authorization policies.

Regulatory compliance requirements can be met for your data base by carrying out auditing which enforces internal controls, so that unwanted changes can be prevented.

b) Statement of Problem

i. General Objective

The outcomes of database auditing can be used by administrator to watch the activities of Information System Users (ISU). System user can be aware of database and its capabilities (sophistic user) or they can be completely unaware of facilities offered by database (naïve user). The information that they can access is defined as constrained by the explicit privileges which is defined by the Database

Management System (DBMS), or any other tier in multi tier architecture like Application Servers (AS) or in any another tier. However if a naughty user (inside organization) or a malicious user (from outside) attempts to access a piece of information for whom he has no privileges, an audit trail must be made. And his activity must be trace and record.

ii. *Specific Objective*

1. What are different types of threats in Database Management system?
2. What is the importance of database auditing?
3. What in formations must be kept in database auditing report?
4. What are locations you can find auditing records?
5. What are the effects of big auditing file on system performance?

c) *The Scope and Delimitation*

The Information system which are manipulating and storing financial, accounting, or other legally sensitive data can use this study. Legal vulnerabilities can be created by even basic trading operation, in today increasingly litigious world. For example.

The investor who lose large amount in trading may hope to recover his capital, and/or the investor's employee may seek to escape responsibility, by legal action against the broker. If related trade data has been modified by the broker without an acceptable Autopsy, which can be seen as *apparent* indicator from the investor's opposing results which are due to the misconduct of the broker.

Auditing, which means capturing and storing information about what is happening in the system, increase the amount of work the system must do. Auditing must be focused so that only events that are of interest are captured .Well designed auditing policies has minimal impact on system performance .Improperly focused auditing can significantly affect performance.

The scope and depth of the audit should be designed to meet the specific objective of organizations base of its environments and type of threats it is facing.

The coverage of this study is the Oracle database 11g, without SAM, which data renovation period is specified by DBA However, we will discuss FGA auditing in this paper, but will not consider RAID system.

The researcher is limited this research to user of relational database.

d) *Importance & Significance of Study*

Database auditing and database security are directly related to each other. One aspect of database security is auditing. Auditing is essential tasks for DBA of database management system which need to secure a particular database system, then. Auditing mechanism which is deploying on a database system facilitates the security implemented in a system.

Auditing, which means capturing and storing information about what is happening in the system, increase the amount of work the system must do.

Auditing must be focused so that only events that are of interest are captured .Properly focused auditing has minimal impact on system performance .Improperly focused auditing can significantly affect performance.

The security administration will use guides provided by Database auditing to develop and enforce a well defined set of security policies based on the initial set of business rules. At the beginning of the specific information system project' the DBA will decide the business rules and later on, modify it with the passage of time, based on the behavior of users.

A wide range of events that can be tracked and collected in Auditing of database records. This creates an additional processing and disk I/O load on database server machine and hence degrades its performance. In other word auditing, which means capturing and storing information about what is happening in the system, increase the amount of work the system must do. Auditing must be focused so that only events that are of interest are captured Properly focused auditing has minimal impact on system performance .but with gathering extra and not needed auditing record you will scarified the performance of database.

So we divide the auditing to

- *Mandatory Auditing:* All oracle database audit certain action regardless of other audit option or parameter .The reason for mandatory audit logs is that the database needs to record some database activities, such as connection by privileged users.
- *Standard Database Auditing:* Enabled at the system level by using the audit trail initialization parameter. After you enabled auditing select the object and privileges that you want to audit and set auditing properties with audit command.
- *Value base Auditing:* Extends standards database auditing, capturing not only the audited event that occurred but also the actual values that were inserted, updated or deleted, Values base auditing is implemented through database trigger.
- *Fine-Grained:* Auditing(FGA) Extends standard database auditing, capturing the actual SQL statement that was issued rather than only the fact that the event occurred.
- *SYSDBA Auditing:* separates the auditing duties between the DBA and an auditor or security administrator who monitors the DBA activates in an operating system audit trail.

So, it is not recommended to keep the excessive auditing of so as to avoid demeaning of performance of the database server. For example, in case of auditing DML commands, it is recommended to

keep insert, update and delete statements are recorded in audit logs, not all retrieval statements.

Following statics may show the important of security and auditing in data base (Neon Enterprise Software).

- Percentage of companies suffers from increasing of security budget? 54%
- Number of companies that claim their job will get more strategic in 2013? 69%
- Number of companies that claim their job will get more compliance? 75%
- Number of companies claim that their top priority will be related to network/security integration? 52%
- How many companies will be more interested to buy best security they can with their budject? 80%
- Sensitive data is losing important data? 68%

This study can be used by

- All organization using database, especially oracle database.
- Researchers, who are interested about DBMS and its features.
- Database Professionals: Data base administrators, database sysmans who are responsible for database recoveries.
- Teachers and students how are working with database in different aspect.

II. RELATED STUDIES

A database is considered as a core asset of an organization. Numerous approaches dealing with Data base Auditing Interface for operating systems and networks have been developed .Nevertheless, they are not sufficient for protecting databases. Still there are many discussions about the abstract and high-level architecture of a DBMS including an ID component.

However, this work mainly focuses on debating generic solutions rather than proposing solid algorithmic solutions. There are many researches about auditing in work done by LIU the authors compare strategies adapted by selected set of vendor for auditing database.

In study done by Cornell University the author introduce three major strategies of database auditing to maintain integrity [1]. Then these strategies are compared in term of efficiency and effectiveness in eliminating error, to do so they compute the optimum timing under each strategy.

In studies submitted to IEEE conference they demonstrate a way in which deductive database cart be used to address elusive problems in the establishment and maintenance of a database audit trail. However, they discus situation and examples of issues that arise in many application computing environments. The means deductive approach can be applied to more than database issues. The main features of the relational

model, is extended to more wider concerns of application development in Deductive computing which is an example of an emerging paradigm[2].

In study published in ACM proposed an approach for detecting different type of anomalies and anomalous access patterns in DBMS. They have developed three models, of different granularity, to represent the SQL queries appearing in the database log files. We will use their work to find out what information must be extract from log file to show access patterns of the queries[8].

a) Conceptual Frame Work

We fist will discuss the tolls you can use for auditing, then we will look inside the system and search for different location database stores auditing records, then we will consider what is actions that DBMS produces and audit records about it and how we can add actions to keep data about according to our needs and requirements.

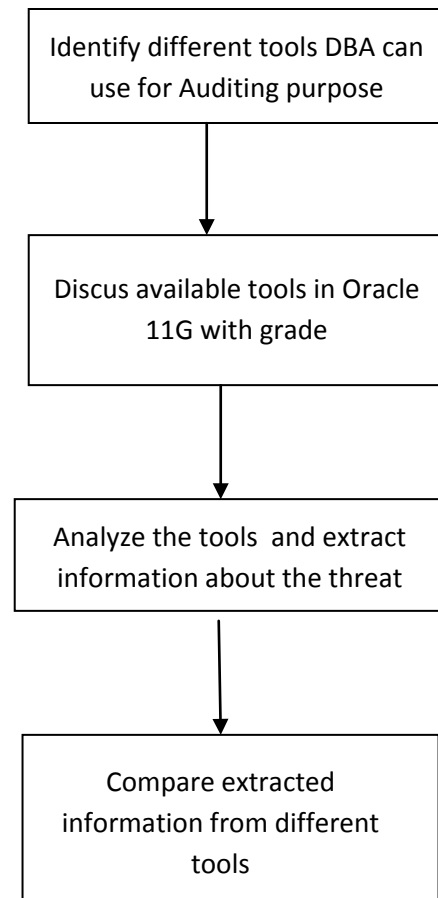


Figure 1 : Conceptual Framework

Research Methodology

III. SPECIFIC RESEARCH PURPOSE AND RESEARCH QUESTIONS

The purpose of this study is to help database Administration to identify threats and configure the database to response by predefine procedure to that trends, also it will help oracle DBA to identify the harm the tread will make to system performance or security of data.

The study will address following questions on specific:

- What are the kinds of threats that can affect database performance?
- What are different threats that will attack database security?
- Which location database will store data's about threats?
- How to setup database management system to do predefine actions against each threat?
- What are important information regarding threats. How to extract information from system log?
- Which statement can create audit log?
- How can you add or edit default commends that make audit logs?

a) Research Methodology

Meta Analysis Research methodology is adopted in accomplishing my research goal. The research intends to study the commonly used auditing method and its drawbacks, also the additional factor which solves the problems in widely used methodology. The research analyzed various auditing records and suggested an efficient procedure to abstract information more effectively and efficiently from recourses. Future more, a novel model that avoids extra lose of system resources is proposed.

b) Research Design

i. Source of Data

Database audit records for statement, privilege, and object auditing are stored in the table SYS.AUD\$. Depending on how extensive you're auditing and retention policies are, you will need to periodically delete old audit records from this table. The database does not provide an interface to assist in deleting rows from the audit table, so you will need to do so yourself.

So we will take a close look at SYS. Audis table and also very important parameter which is audit _trail.

We will look at four level of auditing which oracle 11 G do auditing.

- Statement
- Privilege
- Object
- Fine-grained access

c) Research Plan

A database Administrator knows that when a threat is happened the first priority is backing up and restores the system, but it is also very important to find what the reason for system failure was. To answer this question ,priors to any failure ,you have to analyze your system as database Administrator and decide what are the possible threat for your system, and active or inactive default auditing records ,or modify and extra information to your audit tables.

All DBA know that finding the proper records for your query are one of skills that they will gain it by time and trying, so in this article we will help database administrator to find required information about threat effectively and sufficiently.

Also we will talk about how and when you will purge your unwanted information from Audit table and export them for future used.

Results and Discussions

IV. LOCATION OF AUDIT RECORDS

Oracle 11 record audit records in two locations

- Database
- Operating-system Files

Oracle decided where to keep record by investigating value of initialization parameter **audit trail**. The default is DB, as in AUDIT_TRAIL=DB, you can change this value to DB, EXTENDED to record audit records in the database together with bind variables (SQLBIND) and the SQL.

Statement triggering the audit entry (SQLTEXT). AUDIT_TRAIL=OS tells the database to record audit records in operating-system files.

To change value for audit _trail you have to edit your pile or file. For example, the following statement will change the location of audit records in the spilled.


```

C:\Windows\system32\cmd.exe - sqlplus
C:\Users\adel>sqlplus
SQL*Plus: Release 11.1.0.6.0 - Production on Thu Mar 14 09:02:36 2013
Copyright (c) 1982, 2007, Oracle. All rights reserved.
Enter user-name: sys/sys as sysdba
Connected to:
Oracle Database 11g Enterprise Edition Release 11.1.0.6.0 - Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options
SQL> alter system set audit trail=db scope=spfile_
    
```

Figure 2 : Set Audit Scope

The audit_trail parameter can also have values XML and XML, EXTENDED. With these two options, audit records are written to OS files in XML format. The value of NONE disables auditing.

Keep in mind that you should bounce your database instance for change to take effect. When recorded in the database, most audit entries are recorded in the SYS AUD \$ table. On UNIX systems,

operating-system audit records are written into files in the directory specified by the initialization parameter audit_file_dest (which is set to \$ORACLE_BASE/admin/\$ORACLE_SID/a dump if the database is created using DBCA). On Windows systems.

These audit records are written to the **Event Viewer** log file. So we can find the locations for gathering information as describe in following chart.

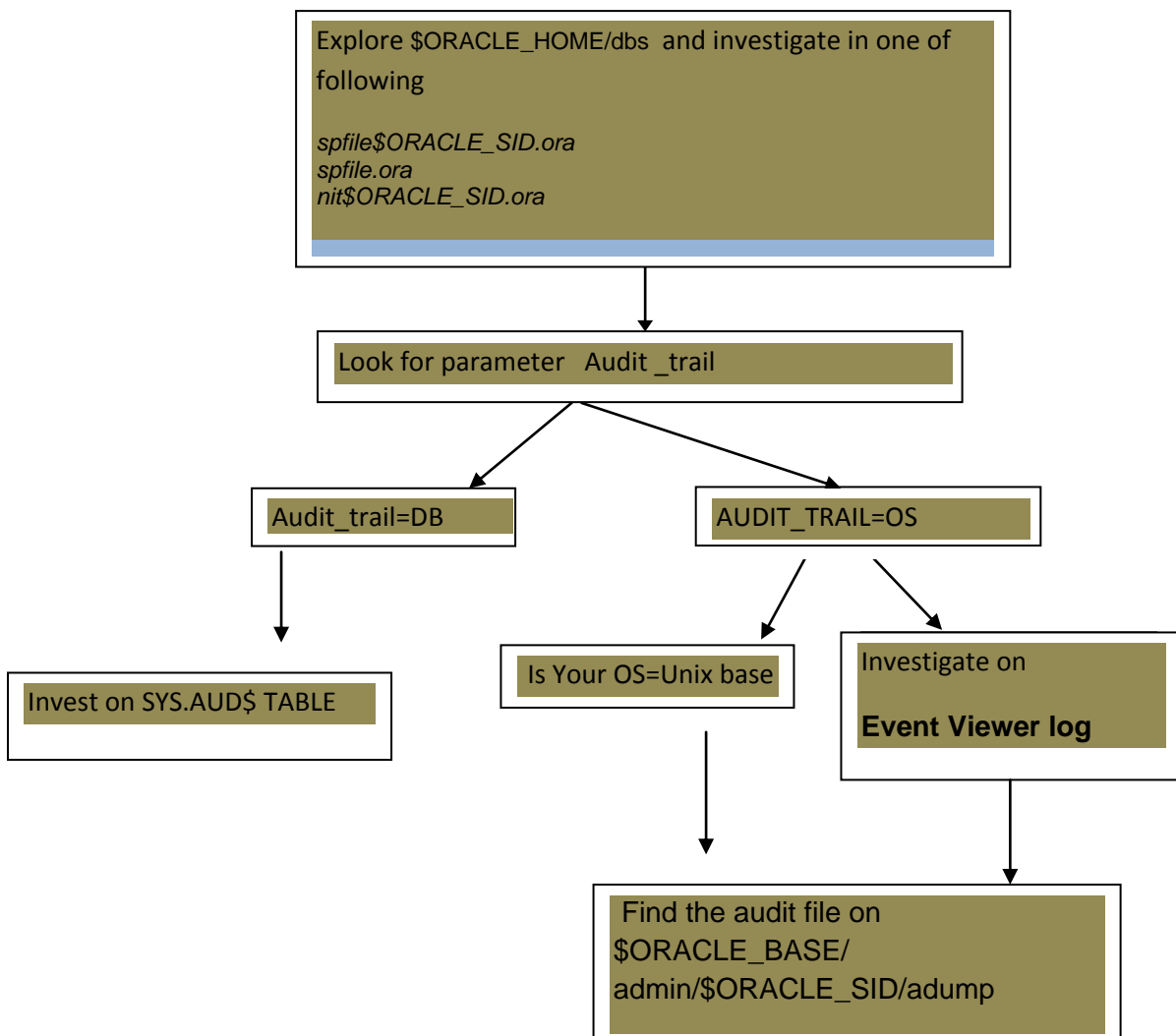


Figure 3 : Procedure for finding audit information

What to audit?

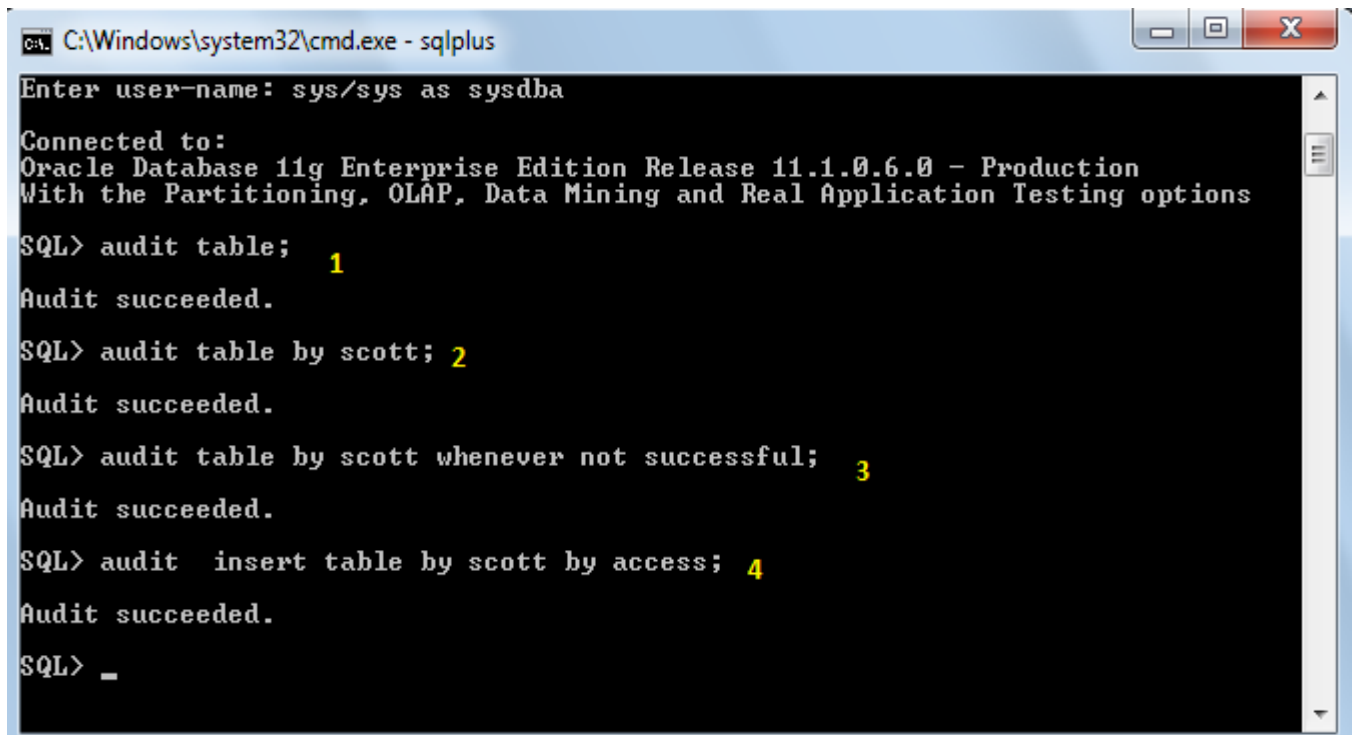
Auditing involves monitoring and recording specific database activity. An Oracle 11g database supports four levels of auditing:

- Statement
- Privilege
- Object
- Fine-grained access

We discuss managing each one of this level in following sections.

a) Management Statement Auditing

Statement auditing involves monitoring and recording the execution of specific types of SQL statements. Executions of some statements are enabling by default, but you can modify this list by adding or deleting some statements to this list as explain in code.



```

C:\Windows\system32\cmd.exe - sqlplus

Enter user-name: sys/sys as sysdba

Connected to:
Oracle Database 11g Enterprise Edition Release 11.1.0.6.0 - Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

SQL> audit table; 1
Audit succeeded.

SQL> audit table by scott; 2
Audit succeeded.

SQL> audit table by scott whenever not successful; 3
Audit succeeded.

SQL> audit insert table by scott by access; 4
Audit succeeded.

SQL> _
    
```

Figure 4 : Modify Audit Record

1. Audit the SQL statements CREATE TABLE, DROP TABLE, and TRUNCATE TABLE, use the TABLE audit option like this. You can add attribute for this command .The attributes are by user, whenever successful by session, whenever not successful and...
2. To record audit entries for specific users only, include a BY USER clause in the AUDIT statement. For example, to audit CREATE, DROP, and TRUNCATE TABLE statements for user scott only ...
3. Frequently, you want to record only attempts that fail—perhaps to look for users who are probing the system to see what they can get away with. To further limit auditing to only these unsuccessful executions, use a WHENEVER clause.
4. You can alternately specify WHENEVER SUCCESSFUL to record only successful statements. If you do not include a WHENEVER clause, both successful and unsuccessful statements trigger audit records. You can further configure non-DDL statement to record one audit

entry for the triggering session or one entry for each auditable action during the session. Specify BY ACCESS or BY SESSION in the AUDIT statement.

There are many auditing options other than TABLE or INSERT TABLE. Table bellow shows all the statement-auditing options.

Statement-Auditing Option	Triggering SQL Statements
ALTER SEQUENCE	ALTER SEQUENCE
ALTER TABLE	ALTER TABLE
COMMENT TABLE	COMMENT ON TABLE COMMENT ON COLUMN
DATABASE LINK	CREATE DATABASE LINK DROP DATABASE LINK
DELETE TABLE	DELETE
EXECUTE PROCEDURE	Execution of any procedure or function or access to any cur-sor or variable in a package
GRANT PROCEDURE	GRANT on a function, package, or procedure
GRANT SEQUENCE	GRANT on a sequence
GRANT TABLE	GRANT on a table or view
INDEX	CREATE INDEX
INSERT TABLE	INSERT into table or view
LOCK TABLE	LOCK
NOT EXISTS	All SQL statements
PROCEDURE	CREATE FUNCTION DROP FUNCTION CREATE PACKAGE CREATE PACKAGE BODY DROP PACKAGE CREATE PROCEDURE DROP PROCEDURE
PROFILE	CREATE PROFILE ALTER PROFILE DROP PROFILE
ROLE	CREATE ROLE ALTER ROLE DROP ROLE SET ROLE
SELECT SEQUENCE	SELECT on a sequence
SELECT TABLE	SELECT from table or view
SEQUENCE	CREATE SEQUENCE DROP SEQUENCE
SESSION	LOGON
SYNONYM	CREATE SYNONYM DROP SYNONYM
SYSTEM AUDIT	AUDIT NOAUDIT
SYSTEM GRANT	GRANT REVOKE
TABLE	CREATE TABLE DROP TABLE TRUNCATE TABLE
TABLESPACE	CREATE TABLESPACE ALTER TABLESPACE DROP TABLESPACE
TRIGGER	CREATE TRIGGER ALTER TRIGGER (to enable or disable) ALTER TABLE (to enable all or disable all)
UPDATE TABLE	UPDATE on a table or view
USER	CREATE USER ALTER USER DROP USER
VIEW	CREATE VIEW DROP VIEW

Table 4.1 : Auditing Option

All needed information about STAEMENT auditing will exist in the DBA_STMT_AUDIT_OPTS data dictionary view. You can query this view to find needed informations, the recorded in formations are

```

C:\Windows\system32\cmd.exe - sqlplus
CREATE EXTERNAL JOB
BY ACCESS

27 rows selected.

SQL> desc dba_stmt_audit_opts
Name                                         Null?    Type
-----
USER_NAME                                   UARCHAR2(30)
PROXY_NAME                                  UARCHAR2(30)
AUDIT_OPTION                                NOT NULL  UARCHAR2(40)
SUCCESS                                     UARCHAR2(10)
FAILURE                                     UARCHAR2(10)
SQL>
    
```

Figure 5 : Structure of Table

You can enable administrator auditing by setting the initialization parameter `AUDIT_SYS_OPERATIONS=TRUE`. All the activities performed connected as SYS or SYSDBA/SYSOPER privileges are recorded in the OS audit trail.

If you enable `AUDIT SESSION`, the database creates one audit record when a user logs on and updates that record when the user logs off successfully.

These session audit records contain some valuable information that can help you narrow the focus

of your tuning efforts. Among the information recorded in the audit records are the username, logon time, logoff time, and the number of physical reads and logical reads performed during the session. By looking for sessions with high counts of logical or physical reads, you can identify high-resource-consuming jobs and narrow the focus of your tuning efforts.

The following are statements will create records by default-

ALTER ANY PROCEDURE	CREATE ANY TABLE	DROP USER
ALTER ANY TABLE	CREATE EXTERNAL JOB	EXEMPT ACCESS POLICY
ALTER DATABASE	CREATE PUBLIC DATABASE LINK	GRANT ANY OBJECT PRIVILEGE
ALTER PROFILE	CREATE SESSION	GRANT ANY PRIVILEGE
ALTER SYSTEM	CREATE USER	GRANT ANY ROLE
ALTER USER	DROP ANY PROCEDURE	ROLE
CREATE ANY LIBRARY	DROP ANY TABLE	SYSTEM AUDIT
CREATE ANY PROCEDURE	DROP PROFILE	

You can restricts and limit this by using command "no audit" in order to not decrease your operation time and also avoid overwrite of audit file which cause lose important information so you need .

b) Examining the Audit Trail

Statement, privilege, and object audit records are written to the `SYS.AUD$` table and made available via the data dictionary views `DBA_AUDIT_TRAIL`

(displays all standard audit trail entries) and `USER_AUDIT_TRAIL` (the standard audit trail entries related to the current user. For example, you can view the user, time, and type of statement audited for user Scott by executing the following:

```

SELECT username, timestamp, action name
FROM dba_audit_trail
WHERE username =scott;
    
```

ORA USER	TIMESTAMP	ACTION NAME
SCOTT	6/15/2004 18:43	LOGON
SCOTT	6/15/2004 18:44	LOGOFF
SCOTT	6/15/2004 18:46	LOGON
SCOTT	6/15/2004 18:46	CREATE TABLE

Table 4.2

c) Managing of Privilege Auditing

Privilege auditing involves monitoring and recording the execution of SQL statements that require a specific system privilege, such as `SELECT ANY TABLE`

or `GRANT ANY PRIVILEGE`. You can audit any system privilege. As we discussed before the command that you will use is `AUDIT` statement, specifying the system

privilege that you want to monitor, or user, or even include DML privilege.

If you want to make report of which privilege is recording in audit records you will query DBA_PRIV_AUDIT_OPTS data dictionary views.

To disable auditing of a system privilege, use a NOAUDIT statement. The NO AUDIT statement.

Allows the same BY options as the AUDIT statement.

d) Managing of Objects Auditing

Object auditing involves monitoring and recording the execution of SQL statements that require a specific object privilege, such as SELECT, INSERT, UPDATE, DELETE, or EXECUTE. Unlike either statement or system privilege auditing, schema object auditing cannot be restrict to specific users, it is enabled for all users or no users.

You enable object auditing with an AUDIT statement, specifying both the object and object privilege that you want to monitor. For example, to audit

SELECT statements on the HR.EMPLOYEES TABLE, execute the following:

AUDIT select ON hr. employee;

You can further configure these audit records to record one audit entry for the triggering session or one for each auditable action during the session by specifying BY ACCESS or BY SESSION in the AUDIT statement. This access/session configuration can be defined differently for successful or unsuccessful executions.

The object-auditing options that are enabled in the database are recorded in the DBA_OBJ_AUDIT_OPTS data dictionary view. Unlike the statement and privilege _AUDIT_OPTS views,

The DBA_OBJ_AUDIT_OPTS data dictionary view always has one row for each auditable object in the database. There are columns for each object privilege that auditing can be enabled on, and in each of these columns, a code is reported that shows the auditing options let's see following codes.

```

C:\Windows\system32\cmd.exe - sqlplus
SQL> desc dba_OBJ_audit_optss
ERROR:
ORA-04043: object dba_OBJ_audit_optss does not exist

SQL> desc dba_OBJ_audit_opts
Name                                         Null?    Type
-----
OWNER                                         VARCHAR2(30)
OBJECT_NAME                                  VARCHAR2(30)
OBJECT_TYPE                                  VARCHAR2(23)
ALT                                           VARCHAR2(3)
AUD                                           VARCHAR2(3)
COM                                           VARCHAR2(3)
DEL                                           VARCHAR2(3)
GRA                                           VARCHAR2(3)
IND                                           VARCHAR2(3)
INS                                           VARCHAR2(3)
LOC                                           VARCHAR2(3)
REN                                           VARCHAR2(3)
SEL                                           VARCHAR2(3)
UPD                                           VARCHAR2(3)
REF                                           CHAR(3)
EXE                                           VARCHAR2(3)
CRE                                           VARCHAR2(3)
REA                                           VARCHAR2(3)
WRI                                           VARCHAR2(3)
FBK                                           VARCHAR2(3)

SQL> SELECT owner,object_name,fbk
2  from dba_obj_audit_opts;

no rows selected

SQL> select * from dba_obj_audit_opts;

no rows selected
    
```

Figure 6 : Empty View

As you can see the view is empty because the audit is not activate ,so we will activate object auditing for employees table of hr schema. And try to access the table.

Figure 7 : Enabling Auditing

```
SQL> audit select on hr.employees;
Audit succeeded.

SQL> conn hr/hr;
Connected.

SQL> select first_name from employees;
```

No we will look at view again

```
SQL> conn sys/sys as sysdba;
Connected.
SQL> select * from dba_obj_audit_opts;
```

OWNER	OBJECT_NAME	OBJECT_TYPE	ALT	AUD	COM	DEL	GRA	IND	INS	LOC	REN	SEL	UPD	REF	EXE	CRE
HR	EMPLOYEES	TABLE	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	S/S	-/-	-/-	-/-	-/-

Figure 8 : View with audit record

To disable object auditing, use a NOAUDIT statement, which allows the same WHENEVER options as the AUDIT statement.

e) Purging the Audit Trail

Database audit records for statement, privilege, and object auditing are stored in the table SYS.AUD\$. Depending on how extensive your auditing and retention policies are, you will need to periodically delete old audit records from this table. The database does not provide. An interface to assist in deleting rows from the audit table, so you will need to do so yourself.

To purge audit records older than 90 days, execute the following as user SYS:

```
DELETE FROM sys.aud$ WHERE timestamp#
< SYSDATE -90>;
```

You might want to copy the audit records into a different table for historical retention or export them to an operating-system file before removing them. It is a good practice to audit changes to the AUD\$ table so that you can identify when changes were made.

f) Using of DBCA to create base line for Auditing

The Oracle Database Configuration Assistant (DBCA) is a Java-based tool used to create Oracle Databases the DBCA provides a flexible and robust environment in which you not only can create databases but also can generate templates containing the definitions of the databases created. This provides you with the ease of using a GUI-based interface with the flexibility of Oracle-generated XML-based templates that you can use to maintain a library of database definitions.

g) Look at FGA Auditing

Oracle auditing can be divided into two basic categories: standard auditing and FGA. Standard

auditing provides the ability to audit based on user, privileges, schemas objects, and statements. For example, it can be based on a specific type of SQL statement (create, alter, update, delete...). FGA provides the ability to audit access to specific application table columns conditionally based on factors such as IP address or the program name used to connect to the database.

Starting with Oracle Database 11g, the Oracle Database Configuration Assistant (DBCA) can automatically configure Oracle recommended minimum audit settings for compliance and internal controls. These audit settings are associated with important security relevant SQL statements and privileges and are listed in the Oracle security documentation. After creating a database with DBCA, the database will audit the following privileges and SQL statements by default:

ALTER ANY
PROCEDURE
CREATE ANY TABLE
GRANT ANY OBJECT
PRIVILEGE
ALTER ANY TABLE
CREATE EXTERNAL
JOB GRANT ANY
PRIVILEGE
ALTER DATABASE
CREATE PUBLIC DATABASE
LINK

LINK
GRANT ANY ROLE
ALTER PROFILE CREATE
SESSION PROFILE
ALTER SYSTEM CREATE USER
PUBLIC SYNONYM
ALTER USER DATABASE LINK
ROLE
AUDIT SYSTEM DROP ANY
PROCEDURE SYSTEM AUDIT

CREATE ANY JOB DROP ANY
TABLE SYSTEM GRANT
CREATE ANY
LIBRARY
DROP PROFILE
CREATE ANY
PROCEDURE
DROP USER

h) *Managing Fine-Grained Auditing*

Fine-grained auditing (FGA) lets you monitor and record data access based on the content of the data. With FGA, you define an audit policy on a table and optionally a column.

When the specified condition evaluates to TRUE, an audit record is created, and an optional event-handler program is called. You use the PL/SQL package DBMS_FGA to configure and manage FGA. The implement of this type of auditing need creating package in Pl/sql which is out of scope of this article.

Conclusions and Recommendations

V. CONCLUSIONS

In this study we discussed the important role of auditing not only for detecting mistrustful behavior also providing proof and reasons to auditors, and also it is recommended to use. Oracle database auditing because of it minimal impact for high audit trail load.

We also discussed different methodology in audits which include trigger or Transactional log but you cannot use these methods for some events which go beyond server events.

VI. RECOMMENDATIONS

1. Use oracle database auditing even if you have large amount of audit trail load.
2. Write audit record to Operating system.
3. Set enough size for OS audit file.
4. Set auditing as part of your defense architecture as follow:
 - i. Set full audit trail of logon and logoff, and record all failed login attempts, as first category
 - ii. Audit Data Control Language (DCL) of the database. For second category.
 - iii. The third category is to audit Data Definition Language (DDL) which changes database schema.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Huang, Liu, "A logging schema for Database Auditing", IEEE conference publication, Computer

Science and Engineering, 2009, Huang, liu page 390-393.

2. Qiang, Liu, Lian-zong, "A Framework for database Auditing" IEEE conference publications, Forth conference on Computer Science and Convergence Information 2009, page 902, 910.
3. Oracle White Paper—Oracle Database Auditing: Performance Guidelines.
4. Oracle Database New Features Guide 11g Release 1 (11.1) B28279-02.
5. Oracle Database 11g New Features for DBAs and Developers, by Sam R. Alapati and Charles Kim, Apress, ISBN: 978-1-59059-910.
6. Ramez Elmasri & Shamkant B. Navathe, Fundamentals of Database Systems, Sixth Edition, Addison-Wesley, 2009.
7. LI Yung, ACM publication, Proceedings of the 40th ACM technical symposium on Computer science education page 241-245, ISBN: 978-1-60558-183.
8. Thomas Connolly & Carolyn Begg, Database Systems: A practical approach to Design, Implementation and Management, Fifth Edition, Addison Wesley, 2010.
9. Oracle Database 11g The Complete Reference (Osborne ORACLE Press Series), Kevin Loney, Publisher: McGraw-Hill Osborne Media; 1 edition (December 17, 2010).





This page is intentionally left blank



Neuro-Fuzzy based Software Risk Estimation Tool

By Pooja Rani & Dalwinder Singh Salaria

Lovely Professional University, Punjab

Abstract - To develop the secure software is one of the major concerns in the software industry. To make the easier task of finding and fixing the security flaws, software developers should integrate the security at all stages of Software Development Life Cycle (SDLC). In this paper, based on Neuro-Fuzzy approach software Risk Prediction tool is created. Firstly Fuzzy Inference system is created and then Neural Network based three different training algorithms: BR (Bayesian Regulation), BP (Back propagation) and LM (Levenberg-Marquardt) are used to train the neural network. From the results it is conclude that for the Software Risk Estimation, BR (Bayesian Regulation) performs better and also achieves the greater accuracy than other algorithms.

Keywords : software security, software threat, neural network, fuzzy logic, neuro-fuzzy.

GJCST-C Classification : D.2.9



Strictly as per the compliance and regulations of:



Neuro-Fuzzy based Software Risk Estimation Tool

Pooja Rani ^α & Dalwinder Singh Salaria ^σ

Abstract - To develop the secure software is one of the major concerns in the software industry. To make the easier task of finding and fixing the security flaws, software developers should integrate the security at all stages of Software Development Life Cycle (SDLC). In this paper, based on Neuro-Fuzzy approach software Risk Prediction tool is created. Firstly Fuzzy Inference system is created and then Neural Network based three different training algorithms: BR (Bayesian Regulation), BP (Back propagation) and LM (Levenberg-Marquardt) are used to train the neural network. From the results it is conclude that for the Software Risk Estimation, BR (Bayesian Regulation) performs better and also achieves the greater accuracy than other algorithms.

General terms : software risk prediction.

Keywords : software security, software threat, neural network, fuzzy logic, neuro-fuzzy.

I. INTRODUCTION

Software systems are being used in every area to perform the different kind of activities all over the world. Due to the rapid growth of internet, technology advancement and the extensively usage of software systems results in security threats that are increasing day by day. So security becomes important concern to be considered. Threat can be any undesired event that is having potential to harm the system. Software Threat Modeling is an approach that deals with the identification, mitigation and prioritization of attacks that have to address. To predict the model for software threats, there are number of techniques like: Statistical techniques, Neural Network, Genetic Algorithm, Support Vector Machine, Fuzzy Logic and hybrid approaches: Neural Network with Genetic Algorithm, Neural Network with Support Vector Machine and Neuro-Fuzzy are being used. As it is fact that each technique has their own pros and cons. It cannot be say that one technique can overcome the limitations of all other techniques. But from the past research work, its find that the hybrid approaches provide more level of accuracy than the individual approaches.

In this paper, to create the prediction model for Software Risk, Hybrid Neuro-Fuzzy approach has been used.

Neural network based three different training algorithms: BR, BPA and LM are used.

Author ^α : Student, Department of CSE Lovely Professional University Phagwara, Punjab – 144411. E-mail : erpoojapuri88@gmail.com

Author ^σ : Assistant Professor, CSE Dept Lovely Professional University Phagwara, Punjab – 144411. E-mail : ds_salaria@yahoo.com

II. REVIEW OF LITERATURE

For Software threat prediction, various statistical approaches as well as advanced approaches are introduced in different areas where Software systems are being used. For Cyber Threat, Cyber threat trend analysis model is proposed using Hidden Markov Model (HMM), to forecast the Cyber threat trend. HMM is a tool in which hidden state is determined. After comparison with existing techniques, the proposed model provides accurate results [1]. MERIT workshop and training programs are conducted for effective training about insider threat awareness. Insider threats are those undesired events that are performed by the legitimate users [2]. Threat Analysis and Modeling (TAM) tool is used to identify the threats and evaluate the risks. This process is useful in business applications [3]. To identify the most critical large system threats, Cyber Threat Tree is implemented as directed graph known as Multiple Valued Decision Diagram (MDD). Cyber Threat Markup Language (Cyma) is used for cyber threat tree representation. Multiple Valued Logic function is used to represent the threat states and their interdependence [4].

In the area of Software Security, to identify the security vulnerabilities in software systems and to show the sequential events that occur during an attack, Regular expression based attack patterns are created. Identification of vulnerabilities is done via matching sequence of components that trigger an event during an attack [5]. Threat Mitigation, Monitoring and Management Plan (TMMMP) approach is discussed to identify the threats, to monitor the remedial measures and to deal with management plans in case of failure of remedial measures. It uses Defense In Depth (DID) strategy for threat mitigation and risk management associated with threats [6]. To identify the security flaws at early stages of software development life cycle, Extended Model Driven Architecture (MDA) approach is introduced with quantitative security assessment model. It will provide the feedback at every stage of software development life cycle [7]. To prioritize the identified threats, Common Vulnerability Scoring system (CVSS) based Risk ranking Tool is used. This tool converts Yes/No values into numerical values and then calculates the risk score using CVSS. It helps to software developer by answering the impact and exploitability of threats [8]. To overcome the limitation i.e. identification of effects by

new security threats and to developing proper countermeasures, two kind of security patterns are introduced i.e. Software Requirement Patterns (SRPs) and Software Design Patterns (SDPs). To identify the threats Software Requirement Patterns (SRPs) are used. Software Design Patterns (SDPs) are used for the identification of remedial measures against identified threats [9].

In the Networked organizations, to enhance the security by prioritizing threats and vulnerabilities, a new methodology is proposed that integrates threat modeling with formal threat analysis. This method is divided into three phases: Threat modeling, asset mapping and mitigation plan that enable the system to identify, quantify the threats and vulnerabilities [10]. For identification of threats in networked organizations, a new approach is introduced that provides reliability statistics to defense analyst to identify the top node in the network. It is useful to identify the top threats in networked organizations [11].

Now a day's modern technique Neural Network is emerged. It is also used to model the software threats. For an intrusion detection system, user behavior modeling approach is introduced that use the neural algorithm and provides better results than existing results [12]. With the use of hybrid approach i.e. Neural network and support vector machine, Intrusion detection system is constructed. It is observed that the performance of this hybrid approach is superior and deliver accurate results [13]. As we know new intrusions are introduced day by day, so there is need to update the new rules to intrusion detection systems. To meet this requirement, a new intrusion detection system is presented with Genetic algorithm approach [14].

To model the real world risk scenarios, risk analysis modeling is introduced that uses fuzzy logic technique. Fuzzy logic model the vagueness in natural way. Thus it provides the accurate recommendations [15]. For electronic commerce development, web based Fuzzy Decision Support System (FDSS) is introduced. This will help to identify electronic ecommerce risk factors [16]. With the use of fuzzy logic secure software system (SSS) approach is introduced. It will help to avert the failed state of the system [17].

For the development of marketing strategy, hybrid intelligent system is developed with the combined approach of Neural Network, Fuzzy Logic and expert system. For the settlement of marketing strategy, this hybrid system is useful to produce intelligent advice [18]. Neural fuzzy scheme is proposed for the development of Direction of arrival (DOA) estimation algorithm by Self-constructing Neural fuzzy Inference Network (SONFIN). The performance of this newly developed algorithm is superior than RBFN [19]. To calibrate the conversion ratios for backfiring technique, calibrated model is generated by using neuro-fuzzy approach. From this model, it is concluded that higher

accuracy is achieved for software size estimation [20]. To make the decision about Distributed Intrusion Prediction and Prevention system (DIPPS) , a model named Hierarchical Neuro-Fuzzy Online Risk Assessment(HiNFRA) using Neuro-fuzzy approach is introduced. This model by using Neuro-fuzzy approach results in more robustness and better performance [21].

III. NEURO-FUZZY RISK PREDICTION MODEL

For the prediction of risk, Neuro-Fuzzy approach is used in this paper. Because the combination of Neural Network and Fuzzy Logic results in such hybrid intelligent system that is having learning ability to optimize its parameters with the use of neural network and to represent the knowledge in an interpretable manner, with the use of Fuzzy System. The hybrid Neuro-Fuzzy technique is well suitable to those areas or applications, where the interpretation and interaction of user is required. Neuro-Fuzzy approach provides more accurate results than other existing hybrid techniques.

a) Fuzzy Inference System

Fuzzy Inference System is based on the concept of Fuzzy set, If Then Rules and Defuzzification. In this paper, MATLAB Fuzzy toolbox that is Graphical User Interface tool used to build the Fuzzy Inference System. To determine how Neuro-Fuzzy approach can be applied to evaluate the Software risk, some of the software factors that affect the security vulnerability are considered. These risk factors are abstracted from [22] [23] [24]. Regarding these input attributes, Corresponding security vulnerability output in the form of Low, Medium, High, Very Low and Very High are obtained from Software industry experts in from of surveys. The total 17 input risk attributes includes the following.

1. Faulty/Changing Requirements.
2. Lack of user Co-operation.
3. Poor Project Planning.
4. Poor Project Management and Resource Estimation.
5. Undefined Project Milestones.
6. Personnel Shortfalls.
7. Insufficiently Trained Team Members.
8. Lack of Specialization.
9. Inexperienced Project Manager.
10. Schedule variation.
11. Budget variation.
12. Deviation From Software Requirements.
13. Shortfalls in Externally Furnished Components.
14. Shortfalls in Externally Performed Tasks.
15. Limitations on Real Time Performance Activities or Tasks.
16. Computer Science Difficulties.
17. Wrong Functions, Properties and UI(User Interface) Development.

i. *Fuzzification*

Fuzzification is the process to describe the input parameters through linguistic variables with meaning like 'Low','High','Medium','Very Low' and 'Very High'. Fuzzy sets are representation of input parameters. These sets are represented by Membership Functions. Input parameters are represents by Zmf (Z- shaped built-in membership function). Similarly, Output parameters are represented by Gauss (Gaussian curve built-in membership function).

ii. *Rule Evaluation*

The total 137 if-then rules are generated after the creation of input output fuzzy sets and Membership functions. In the rules 'T' means "True" and representing value 1 and 'F' means "False" and representing value 0. The rules created in rule base of Fuzzy Inference System (FIS) are represented in the following format:

If(Fault/Changing Requirements is 'T') and (Lack of user Co-operation is 'F') and (Poor Project Planning is 'T') and (Poor Project management and Resource Estimation is 'F') and (Undefined Project Milestones is 'T') and (Personnel Shortfalls is 'F') and (Insufficiently Trained Team Members is 'T') and (Lack of Specialization is 'F') and (Inexperienced Project Manager is 'T') and (Schedule variation is 'F') and (Budget variation is 'T') and (Deviation From Software Requirements is 'F') and (Shortfalls in Externally Furnished Components is 'T') and(Shortfalls in Externally Performed Tasks is 'F') and (Limitations on Real Time Performance Activities or Tasks is 'T') and (Computer Science Difficulties is 'F') and (Wrong Functions, Properties and UI Development is 'T').

iii. *Defuzzification*

Defuzzification is the process to calculate the output, after applying if-then rules. It refers the way in which fuzzy sets are transformed into numerical value. Seventeen Input Parameters and Output parameter named Security Vulnerability are represented in Fig 1. Fuzzy Inference System Editor is used to achieve this representation.

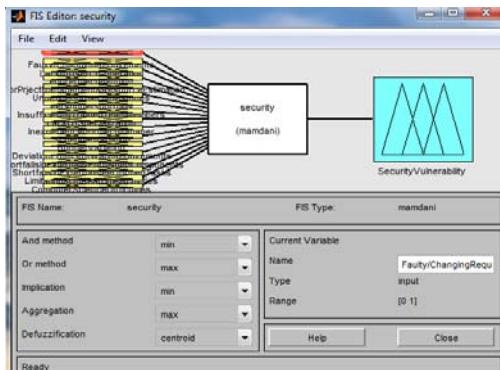


Figure 1 : Using FIS Editor Input and Output Parameters Representation

For a given set of input parameters like [Faulty/Changing Requirements Lack of user Co-

operation Poor Project Planning Poor Project management and Resource Estimation Undefined Project Milestones Personnel Shortfalls Insufficiently Trained Team Members Lack of Specialization Inexperienced Project Manager Schedule variation Budget variation Deviation From Software Requirements Shortfalls in Externally Furnished Components Shortfalls in Externally Performed Tasks Limitations on Real Time Performance Activities or Tasks Computer Science Difficulties Wrong Functions, Properties and UI Development] say [1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1} Rule Viewer is used to see the output of Security Vulnerability i.e. generated 0.5 is specified at the top of graph corresponding to considered set of input variables in Fig 2 shown below.

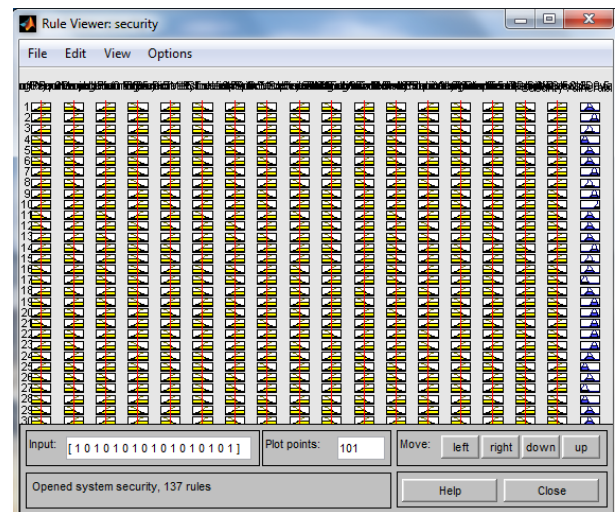


Figure 2 : Security Vulnerability Generation in Rule Viewer

b) *Neural Network Architecture*

After completing the work of Fuzzy System now next step to move on to Neural Network. In this paper Neural Network based three different algorithms are used: Levenberg-Marquardt (trainlm), Back propagation algorithm, and Bayesian Regulation.

Levenberg-Marquardt (trainlm) is a network training function that according to Levenberg-Marquardt optimization updates its weight and bias values. It is fastest algorithm. Limitation of Levenberg-Marquardt algorithm is that it consumes more memory.

Back propagation (triangdx) is a learning algorithm means it learns from many inputs for desired output. It is very simple. It does not require any specialization. But the Limitation of this algorithm is that its having low prediction capabilities. Due to low prediction capabilities, it does not provide accurate results.

Bayesian Regulation (Trainbr) is advanced method. This algorithm is more suitable for those prediction cases where large number of inputs is used to predict the output. Many researchers has used

Liebenberg-Marquardt and Back-propagation algorithm for training phase.

IV. EXPERIMENTAL ANALYSIS

A feed-forward network with three different training algorithms: BR, BPA and LM are used. 12 neurons for input layer, 12 for hidden layer and 1 for output layer are used for the implementation of Neural Network.

a) Source of Training Data

As it above discussed that after generating the fuzzy rules, output is generated corresponding to fuzzy set of input variables. This training data is used to train the neural network.

b) Tool Development

For the prediction of Risk, Risk development tool is generated using MATLAB. As three different algorithms BR, BPA and LM are used so three different Graphical User Interfaces are created. Firstly Using BR algorithm GUI (Graphical User Interface) is created and shown below in fig 3.

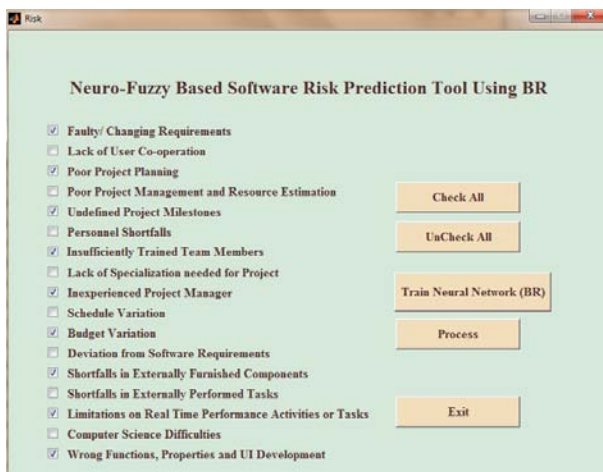


Figure 3 : Neuro- Fuzzy based Software Risk Prediction Tool using BR

Secondly GUI (Graphical User Interface) is created by using BPA as shown below in fig 4.

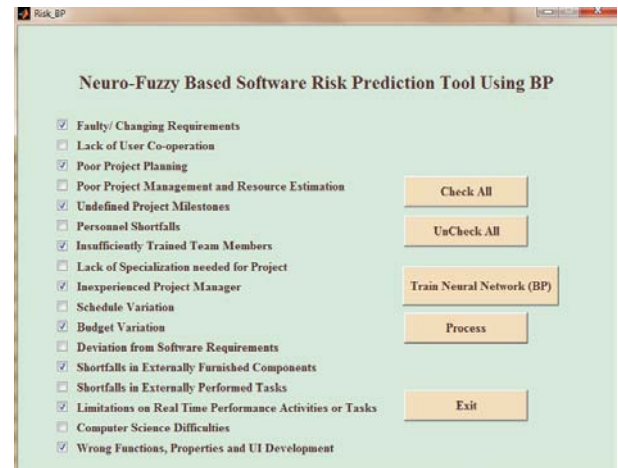


Figure 4 : Neuro- Fuzzy Based Software Risk Prediction Tool using BP Algorithm

Finally 3rd GUI (Graphical User Interface) is created by using LM algorithm as shown below in fig 5

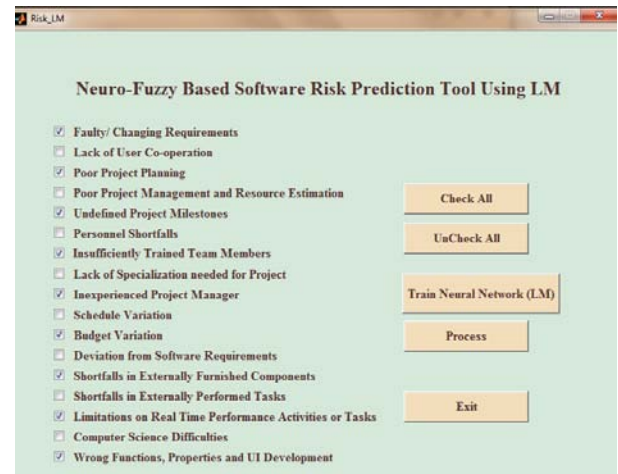


Figure 5 : Neuro- Fuzzy based Software Risk Prediction Tool using LM Algorithm

V. RESULTS AND COMPARISON

Neural Network is trained with three different algorithms: BR, BPA & LM and outputs are obtained. From the table 1. The comparison among three different algorithms can be seen. In the table 17 inputs parameters are used and corresponding Security vulnerability output is computed for BR, BP and LM algorithms. The comparison shows that BR provides the better results than BP and LM algorithms. The results provides by BR are accurate where as BP and LM are over fitting the values for the same dataset.

The table1: Summarizes the results achieved by these three different algorithms over the same dataset. Some short terms are used in the table for input parameters are as follows.

1. FR : Faulty/Changing Requirements.
2. LUC : Lack of user Co-operation.
3. PPP: Poor Project Planning.

4. PPMRE: Poor Project Management and Resource Estimation.
5. UPM: Undefined Project Milestones.
6. PS: Personnel Shortfalls.
7. ITTM: Insufficiently Trained Team Members.
8. LOS: Lack of Specialization.
9. IPM: Inexperienced Project Manager.
10. SV: Schedule variation.
11. BV: Budget variation
12. DFSR: Deviation From Software Requirements
13. SEFC: Shortfalls in Externally Furnished Components.
14. SEPT: Shortfalls in Externally Performed Tasks.
15. LRTPA: Limitations on Real Time Performance Activities or Tasks.
16. CSD: Computer Science Difficulties.
17. WFPUID: Wrong Functions, Properties and UI(User Interface) Development.
18. Regarding Security vulnerability Output the following short terms are used.
19. SVBR: Security Vulnerability Using BR (Bayesian Regulation).
20. SVBP: Security Vulnerability using BP (Back propagation).
21. SVLM: Security Vulnerability using LM (Liebenberg-Marquardt).

Table1 : Risk Estimation by using Three Different Training Algorithms

FR	LU C	PP P	PP MR E	U P M	PS	IT T M	LO S	IP M	SV	BV	DF SR	SE FC	SE PT	LRTPA	CSD	WF UID	SV BR	SV BP	SV LM
T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	94.49%	122.00%	107.50%
F	T	T	T	T	F	T	T	T	T	T	T	T	T	T	T	T	92.90%	126.42%	116.44%
T	F	T	T	F	T	T	T	T	T	T	T	T	T	T	T	F	89.29%	62.04%	79.91%
T	T	T	T	T	T	F	T	F	T	T	F	T	T	F	T	T	84.77%	145.45%	96.73%
T	T	F	T	T	T	T	T	T	F	T	T	F	T	F	F	T	79.81%	-20.94%	73.39%
F	T	T	F	T	F	T	F	T	T	F	T	T	F	T	T	T	72.06%	90.22%	73.43%
T	F	T	F	F	T	F	F	T	T	F	T	T	T	T	T	F	68.96%	36.64%	38.46%
F	T	T	T	F	T	F	T	F	T	F	T	T	T	F	T	F	65.51%	75.38%	47.59%
T	F	F	T	F	T	F	T	T	F	T	F	T	F	T	F	T	59.30%	34.81%	61.16%
T	T	F	F	T	F	F	T	T	F	T	F	F	T	F	T	T	53.03%	47.23%	57.09%
F	F	F	T	F	T	F	F	F	T	T	F	F	T	T	T	T	47.27%	100.01%	68.02%
T	F	T	F	T	F	T	F	F	T	F	T	T	F	T	F	F	43.63%	3.58%	23.40%
T	T	T	F	F	F	T	F	F	T	F	T	T	F	F	T	F	37.41%	54.94%	29.29%
F	T	F	F	F	F	F	T	T	F	F	F	T	T	T	F	T	33.26%	-14.34%	18.33%
F	F	F	T	F	T	T	F	T	F	T	T	F	T	F	F	F	28.80%	19.98%	27.09%
F	F	F	F	F	T	T	F	T	F	T	T	F	T	F	F	F	21.32%	20.63%	31.53%
T	F	F	F	T	T	T	F	F	F	T	F	F	F	F	F	F	17.35%	26.50%	23.10%
F	F	F	F	F	F	F	F	F	T	F	T	T	F	T	F	F	12.64%	35.06%	12.13%
F	F	F	F	F	F	F	F	F	F	F	F	T	F	T	F	F	5.96%	21.22%	2.85%
F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	2.23%	10.60%	4.48%

VI. CONCLUSION

Software Risk Prediction is one of the most important tasks for the development of secure and reliable system. It should be preferred that during the early stages of software development life cycle to find and fix the security flaws. Neuro-fuzzy approach based risk prediction tool is developed using MATLAB. After creation of Fuzzy Inference System, Neural Network is trained with three different algorithms using 'trianbr', 'traingdx' and 'trainlm'. From the results it is concluded that BR algorithm performs better than other algorithms. With the use of BR algorithm better accuracy level is achieved then other algorithms.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Lee, T., Jung, S. O., In, H. P., & Lee, H. J. (2007, August). Cyber Threat Trend Analysis Model Using HMM. In Information Assurance and Security, 2007. IAS 2007. Third International Symposium on (pp. 177-182). IEEE.
2. Greitzer, Frank L., Andrew P. Moore, Dawn M. Cappelli, Dee H. Andrews, Lynn A. Carroll and Thomas D. Hull. "Combating the insider cyber threat." Security & Privacy, IEEE 6, no. 1 (2008): 61-64.
3. Ingalsbe, Jeffrey A., Louis Kunimatsu, Tim Baeten and Nancy R. Mead. "Threat modeling: diving into the deep end." Software, IEEE 25, no. 1 (2008): 28-34.
4. Oongsakorn, P., Turney, K., Thornton, M., Nair, S., Szygenda, S. & Manikas, T. (2010, April). Cyber threat trees for large system threat cataloging and analysis. In Systems Conference, 2010 4th Annual IEEE (pp. 610-615). IEEE.
5. Gegick, Michael and Laurie Williams. "Matching attack patterns to security vulnerabilities in software-intensive system designs." In ACM SIGSOFT Software Engineering Notes, vol. 30, no. 4, pp. 1-7. ACM, 2005.
6. Gandotra, V., Singhal, A., & Bedi, P. (2009, October). Threat mitigation, monitoring and management plan-A new approach in risk management. In Advances in Recent Technologies in Communication and Computing, 2009.

- ARTCom'09. International Conference on (pp. 719-723). IEEE.
7. Tang, X. & Shen, B. (2009, July). Extending Model Driven Architecture with Software Security Assessment. In Secure Software Integration and Reliability Improvement, 2009. SSIRI 2009. Third IEEE International Conference on (pp. 436-441). IEEE.
8. Dhillon, Danny. "Developer-Driven Threat Modeling: Lessons Learned in the Trenches." *Security & Privacy*, IEEE 9, no. 4 (2011): 41-47
9. Okubo, T., Kaiya, H. & Yoshioka, N. (2011, August). Effective security impact analysis with patterns for software enhancement. In Availability, Reliability and Security (ARES), 2011 Sixth International Conference on (pp. 527-534). IEEE.
10. Stango, A., Prasad, N. R. & Kyriazanos, D. M. (2009, June). A threat analysis methodology for security evaluation and enhancement planning. In Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on (pp. 262-267). IEEE.
11. Frantz, T. L., & Carley, K. M. (2009, July). Information assurances and threat identification in networked organizations. In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on (pp. 1-5). IEEE.
12. Debar, H., Becker, M., & Siboni, D. (1992, May). A neural network component for an intrusion detection system. In Research in Security and Privacy, 1992. Proceedings. 1992 IEEE Computer Society Symposium on (pp. 240-250). IEEE.
13. Mukkamala, S., Janoski, G., & Sung, A. (2002). Intrusion detection using neural networks and support vector machines. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on (Vol. 2, pp. 1702-1707). IEEE.
14. Gong, R. H., Zulkernine, M. & Abolmaesumi, P. (2005, May). A software implementation of a genetic algorithm based approach to network intrusion detection. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on (pp. 246-253). IEEE.
15. Haslum, K., Abraham, A. & Knapskog, S. (2008, May). Hinfra: Hierarchical neuro-fuzzy learning for online risk assessment. In Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on (pp. 631-636). IEEE.
16. Ngai, E. W. T., & Wat, F. K. T. (2005). Fuzzy decision support system for risk analysis in e-commerce development. *Decision support systems*, 40(2), 235-255.
17. Gandotra, V., Singhal, A. & Bedi, P. (2010, April). A step towards secure software system using fuzzy logic. In Computer Engineering and Technology (ICCET), 2010 2nd International Conference on (Vol. 1, pp. V1-427). IEEE.
18. Li, S. (2000). The development of a hybrid intelligent system for developing marketing strategy. *Decision Support Systems*, 27(4), 395-409.
19. Shieh, C. S., & Lin, C. T. (2000). Direction of arrival estimation based on phase differences using neural fuzzy network. *Antennas and Propagation, IEEE Transactions on*, 48(7), 1115-1124.
20. Wong, J., Ho, D., & Capretz, L. F. (2008). Calibrating function point backfiring conversion ratios using neuro-fuzzy technique. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16(06), 847-862.
21. Haslum, K., Abraham, A. & Knapskog, S. (2008, May). Hinfra: Hierarchical neuro-fuzzy learning for online risk assessment. In Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on (pp. 631-636). IEEE.
22. Boehm, B. W. (1991). Software risk management: principles and practices. *Software*, IEEE, 8(1), 32-41.
23. Hu, Y., Zhang, X., Sun, X., Zhang, J., Du, J. & Zhao, J. (2010, November). A unified intelligent model for software project risk analysis and planning. In Information Management, Innovation Management and Industrial Engineering (ICIII), 2010 International Conference on (Vol. 4, pp. 110-113). IEEE.
24. Bragina, T. & Tabunshchyk, G. (2011, February). Fuzzy model for the software projects design risk analysis. In CAD Systems in Microelectronics (CADSM), 2011 11th International Conference. The Experience of Designing and Application of (pp. 335-341). IEEE.



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING

Volume 13 Issue 6 Version 1.0 Year 2013

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals Inc. (USA)

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

Data Preprocessing in Multi-Temporal Remote Sensing Data for Deforestation Analysis

By Dr. Manjula. K.R, Dr. Jyothi. Singaraju
& Prof. Anand Kumar Varma. Sybiyal

SASTRA University, Tamilnadu

Abstract - In recent years, the contemporary data mining community has developed a plethora of algorithms and methods used for different tasks in knowledge discovery within large databases. Furthermore, algorithms become more complex and hybrid as algorithms combining several approaches are suggested, the task of implementing such algorithms from scratch becomes increasingly time consuming. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data pre-processing steps are often necessary to prepare data for further usage. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results.

Keywords : data preprocessing, data mining, remote sensing images, deforestation analysis.

GJCST-C Classification : J.1



DATA PREPROCESSING IN MULTI-TEMPORAL REMOTE SENSING DATA FOR DEFORESTATION ANALYSIS

Strictly as per the compliance and regulations of:



RESEARCH | DIVERSITY | ETHICS

Data Preprocessing in Multi-Temporal Remote Sensing Data for Deforestation Analysis

Dr. Manjula. K.R ^α, Dr. Jyothi. Singaraju ^σ & Prof. Anand Kumar Varma. Sybiyal ^ρ

Abstract - In recent years, the contemporary data mining community has developed a plethora of algorithms and methods used for different tasks in knowledge discovery within large databases. Furthermore, algorithms become more complex and hybrid as algorithms combining several approaches are suggested, the task of implementing such algorithms from scratch becomes increasingly time consuming. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data pre-processing steps are often necessary to prepare data for further usage. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results.

This paper introduces and defines the study area and throws light on the data preprocessing on both collateral and image data. Under data preprocessing, the non spatial data are preprocessed with normalization, generalization and other techniques. For the satellite image, the preprocessing is done both at the image dissemination and during feature extraction process. These data are preprocessed to fill data gaps and correct data anomalies. This paper provides a brief description of local maximum likelihood method, pepper salt method, boundary clean method and edge matching methods which are used while classifying the image.

Keywords : data preprocessing, data mining, remote sensing images, deforestation analysis.

1. INTRODUCTION

The technical progress in computerized data acquisition and storage results in the growth of vast databases. With continues increase and accumulation, the huge amount of the computerized data have far exceeded human ability to completely interpret and use. Users need adequate search tools in order to quickly access and filter relevant information. The development of novel technique and tools in assist for humans aiding in the transformation of data into useful knowledge, has been the heart of the comparatively new and interdisciplinary research areas called "Knowledge Discovery in Databases (KDD)". With

rapid growth in development of research in data mining order to quickly access and filter relevant information. The development of novel technique and tools in assist for humans aiding in the transformation of data into useful knowledge, has been the heart of the comparatively new and interdisciplinary research areas called "Knowledge Discovery in Databases (KDD)". With rapid growth in development of research in data mining and data warehouse, many systems were emerged in those fields.

It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results. What variables should be selected? What measurement framework, such as Euclidean space or non-metric network space, should be used? What spatial relations or contextual information should be considered? Can the chosen data adequately represent the complexity and nature of the problem?

a) Study Area

The setting of this study spans an area of 5000 Square Kilometers and it includes the mandals of Chittoor such as Thirupathi, Kalahasthi, Yerpedu, Renigunta and major portion of Kadapa Mandals such as Nandalur, Chitvel, Rajampet, Pullampet, Obulavari Palli, Kodur, and Nellore District mandals such as Venkatagiri, Rapur, Kaluya and Takkili. The study area boundary in lat-long is E 79 39" to E 78 45" and N 13 35" to N 14 33". The study area district outline is specified in the Figure 1:



Figure 1.1 : Map Showing the District Outline Containing the Study Area

Beside the undamaged natural environment in some parts, a big part of the area has been changed by

Author ^α : SAP, School of Computing, Dept. of CSE, SASTRA University, Tirumalaisamudram, Tamil Nadu.

E-mail : manju_sakvarma@yahoo.co.in

Author ^σ : Professor & BOS Chairperson, Dept. of CS, SPMVV, Tirupati, Chittoor District, Andhra Pradesh.

E-mail : jyothi.spmvv@gmail.com

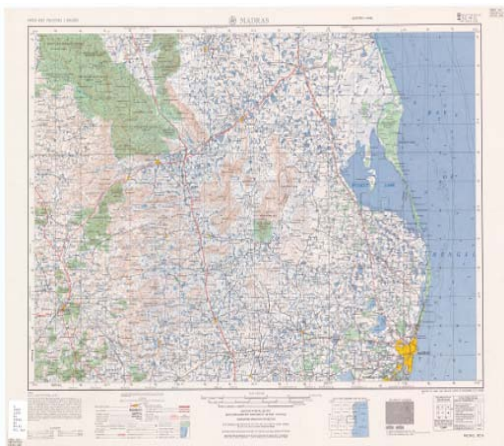
Author ^ρ : Professor, Dept. of Civil Engineering, SIETK, Puttur.

E-mail : manju_sakv@yahoo.co.in

agriculture and grazing activities. The following figures 1.2(a) and (b), 1.3.(a), (b) and (c) and 1.4 represents top sheets, satellite images path row of the study area and scene and satellite image along with scanned mandal boundary map of Cuddapah district.



(a)

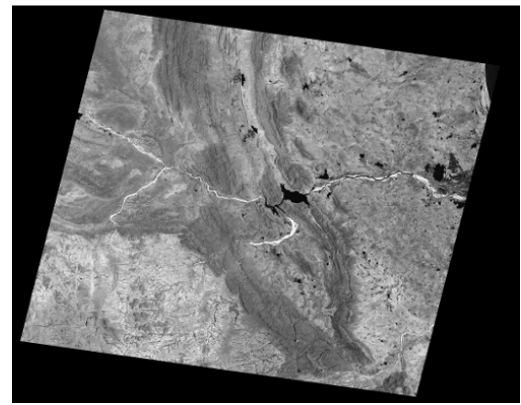


(b)

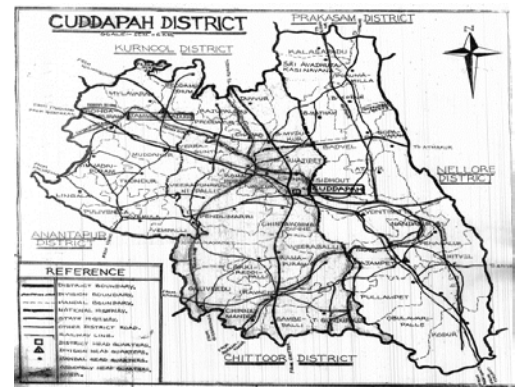
Figure 1.2 : Topography and Forest Area of Study area is shown in the Map



(a)



(b)



(c)

Figure 1.3 : Reference Map of the study: (a) Path Row boundary, (b) Scene of the study area, (c) Scanned paper map of Cuddapah District

i. Data Collection

Spatial information on Land use / Land cover is a necessary prerequisite in planning, utilizing and management of the natural resources. The study area is based on the secondary data, the satellite imagery which is downloaded from the Global Land cover Mapping web site and National Remote Sensing Agency and Survey of India. Three kinds of data are used for this study:

- Satellite data,
- Topographic and thematic maps and
- Descriptive data.

The necessary *satellite data* is selected and acquired after visits to the three test areas. The main criteria for selection of satellite data are:

- Date of acquisition of images according to the local calendar;
- Weather conditions (cloud cover) during the acquisition of images;
- Spectral and spatial resolution of images.
- After examine the above list, the following satellite images available for study area, one Landsat Thematic Mapper, Enhanced Thematic Mapper and IRS P6 LISS III images are ordered.

Table 1.1 : Various Inputs used in the Study

S.No	Type of Map	Resolution/Scale	Date/Year of Acquisition	Source
1	Topsheet	1:50,000	57 O/5 – 1973-79	SOI
		1:50,000	57 J/11 – 1973	SOI
		1:50,000	57 O/6 – 1973	SOI
		1:50,000	57 N/9 – 1973	SOI
2	Landsat – TM	28.5 mt	1991	GLCF
3	Landsat – ETM+LISS 3	Medium 250,000	05 th April, 2001	NRSC
4	IRS P6. LISS 3 101 – 63	Medium 250,000	6 th Feb, 2010	NRSC
5	Mandal Maps	A3 Size	-	Mandal HQ

ii. Review of Literature

Amos Storkey [2] proposed various data preprocessing methods applied on any data before applying data mining techniques to ensure the quality of decision making. Aleksandar Lazarevic et al [3] proposed the software system for spatial data analysis and modelling (SDAM) which provide flexible machine learning tools for supporting an interactive knowledge discovery process in large centralized or distributed spatial databases. Caroline M. Bruce and David W. Hilbert [4] suggested a Pre-processing methodology for application to Landsat 7 TM/ETM+. This report details the various pre-processing techniques either to derive multitemporal and multispatial image classifications or to use in biophysical/geochemical modelling. P.S. Roy et al [10] proposed a multilevel land use land cover classification system, wherein LULC information can be accessed Nationwide, State wide and at the intrastate, regional or municipal level. Stefan Erasmi et al [11] evaluated available satellite data sets and established a transparent work flow for the monitoring of past and future land cover dynamics at a regional scale based on medium resolution satellite data while mapping deforestation and land cover conversion at the rainforest margin in central Sulawesi, Indonesia.

II. DATA PREPARATION

One of the methods for change detection using satellite images is to compare the results of classified images. The advantage of the classified-map comparison method is that not only the location but also the nature and type of the changes are determined in the study area. In this method, first, the images of different times are classified according to the purpose of change detection. Afterward, by overlaying these classified images with a proper overlay condition, the location and amount of these changes that are interested is determined. As the goal is to determine the

deforestation, the only two classes that are considered are the forest and non-forest.

III. DATA PREPROCESSING

Under data preprocessing, the non spatial data are preprocessed with normalization, generalization and other techniques. For the satellite image, the preprocessing is done both at the image dissemination and during feature extraction process. These data are preprocessed to fill data gaps and correct data anomalies. This paper provides a brief description of various preprocessing methods that is applied on the collected images in order to achieve the data quality of the study while classifying the image.

a) Part I - Collateral Data Preprocessing

Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Spatial data sets often contain large amounts of data arranged in multiple layers. These data may contain errors and may not be collected at a common set of coordinates. Therefore, various data preprocessing steps are often necessary to prepare data for further usage. The following figure explores the preprocessing steps generally used for all type of data [3].

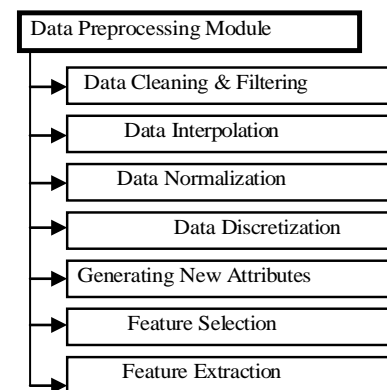


Figure 3.1 : Data Preprocessing Functions

i. Data Cleaning and Filtering

Due to the high possibility of measurement noise present in collected data sets, there is a need for data cleaning. Data cleaning consists of removing duplicate data points, and removing value outliers, as well spatial outliers. Data can also be filtered or smoothed by applying a median filter with a window size specified by the user.

ii. Data Interpolation

In many real life spatial domain applications, the resolution will vary among data layers and the data will not be collected at a common set of spatial locations. Therefore, it is necessary to apply an interpolation procedure to the data to change data resolution and to compute values for a common set of locations. Deterministic interpolation techniques such as inverse distance and triangulation can be used but they

do not take into account a model of the spatial process or variograms.

iii. Data Normalization

The system supports two normalization methods: the transformation of data to a normal distribution and the scaling of data to a specified range. In this work, normalization is applied for the image data while georeferencing the three time period based images.

iv. Data Discretization

This step is necessary in some modeling techniques like association rules, decision tree learning and all classification problems and includes different attribute and target splitting criteria. In this work discretization is applied for collateral data that includes population data with diversity in data. So this data is discretized into three ranges of groups as 'High', 'Low' and 'Medium'.

Table 3.1 : Discretization of population in the ranges

Population Size & Growth(Difference)	Density = Population / Area	Range Label
500 <	<100	Low
>=500<1000	>=100 and < 200	Medium
>=1000	> 200	High

v. Generating New Attributes

Users can generate new attributes by applying supported operators to a set of existing attributes. The density range, population range etc., are created as new attributes for the study.

vi. Feature Selection

In domains with a large number of attributes this step is often beneficial for reducing attribute space by removing irrelevant attributes. Several selection techniques (Forward Selection, Backward Elimination, Branch and Bound) and various criteria (inter-class and probabilistic selection criteria) are supported in order to identify a relevant attribute subset.

In this thesis, while preparing a single table input for association rule mining, some of the attributes in individual tables are removed as irrelevant. For deriving rules the attributes such as Gridcode, Area etc are removed as it does not give any meaningful information while deriving rules.

vii. Feature Extraction

In contrast to feature selection where a decision is target-based, variance-based dimensionality reduction through feature extraction is also supported. The transformed data can be plotted in d-dimensional space and resulting plots can be rotated, panned and zoomed to better view possible data groupings.

viii. Data Partitioning

Partitioning allows users to split the data set into more homogenous data segments, thus providing better modeling results.

b) Part – II - Image Data Preprocessing

Availability, Accessibility, and Affordability of Remote Sensing Data, a range of airborne and space-borne sensors has acquired remote sensing data, with the number of sensors and their diversity of capability increasing over time. Ideally, the following image characteristics are required for studying deforestation [1][4].

- Cloud free and clear atmosphere during the time of data acquisition;
- Availability of imagery for the optimum date or dates;
- Spatial resolution fine enough for accurate mapping and course enough so image size is manageable;
- Band selection (band width, placement, and number of bands) optimized to identify features of interest;
- Study area covered on a single image;
- Same sensor and sun position when images were acquired similar atmospheric conditions.
- Pragmatically, it is rather difficult to acquire the data with the above characteristics. Instead, the following problems are common in data acquisition process:
 - Unavailability of data for specific time period;
 - Persistent cloud coverage throughout the year and for many years;
 - Cost is too high specially for commercial satellite data;
 - Availability of data in usable format (digital or hard copy);
 - Cost of processing, in producing value added product,
 - Lack of expertise, equipment/software for analysis;
 - Significant improvements have been made in terms of spectral, spatial, temporal and radiometric resolutions. More specifically, improvements have been observed in
 - Visibility and clarity that includes more detailed image of a smaller piece of land;
 - Clear definition involving more precisely the specific colours or light responses reflecting off of the field; and
 - Frequent data acquisition on a regular interval of every other day or every 5-7 days.
 - The background environment reflected through the remote sensing image obtained in different instant is different because of the influence of various factors in the acquisition process. These factors can be divided into two categories: remote sensing system factors and environmental factors.
 - The remote sensing system factors are: the impact of temporal, spatial, spectral and radiation resolution.
 - The environmental factors are: The impact of atmospheric conditions, soil moisture and phonological characteristics.

The impact at different times and the influence of these factors on the images must be fully taken into account in the change detection. The influence may be eliminated as much as possible by the geometric registration and radiometric correction on the remote sensing images.

Preprocessing and Analysis of the Satellite Images

Prior to data analysis, initial processing on the raw data is usually carried out to correct for any distortion due to the characteristics of the imaging system and imaging conditions. Depending on the user's requirement, some standard correction procedures may be carried out by the ground station operators before the data is delivered to the end-user. Figure 3.20 derives the processing procedure applied to image data [1] [4] [10].

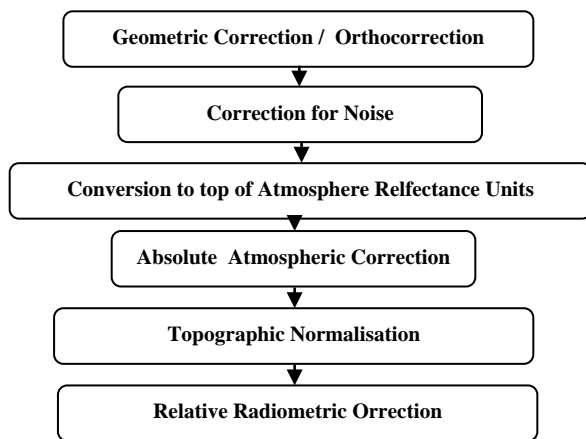


Figure 3.2 : Processing Procedures Applied to Imagery Data

Usually three types of errors occur when a satellite image is generated by the satellite sensor. The first is the sensor error. The second is the error created by the atmospheric parameters, which affect the amount of radiation received by the sensor. The third one is the geometric errors related to the curvature of the Earth surface, the Earth rotation, elevation differences, location and situation of the satellite etc. Therefore, these errors should be considered and managed before using the data:

1. *Sensor Errors:* The two images used were already corrected by their providers. Therefore, there was no need for any processing in this regard.
2. *Radiometric Correction:* The Earth atmosphere scatters the shorter wavelengths in a selective manner and this reduces the contrast of the image. The numerical value of each pixel in the image is not a realistic representation of the amount of radiation from the ground surface. These values are changed either by atmospheric absorption or by scattering throughout the atmosphere.

3. *Atmospheric Effects:* Scattering and absorption of EM radiation by the atmosphere have significant effects that impact sensor design as well as the processing and interpretation of images. When the concentration of scattering agents is high, scattering produces the visual effect we call haze. Haze increases the overall brightness of a scene and reduces the contrast between different ground materials. In general, atmospheric errors are discussed in three parts: the Haze, Sun angle and Skylight errors.

Atmospheric corrections are required in the following situations:

- When user want to compare the images related to different times.
- When using methods such as image subtraction and image division for change detection, the effect of atmosphere on the two images related to different times are quite different.
- When the ratio of two bands of an image is needed to be calculated, because the atmosphere has different effects on different wavelengths.
- When user want to study spectral characteristics of different phenomena.

If user wanted to use the division or subtraction of images for determining the changes in forest land use, then user would have to correct for the haze, sun angle and skylight errors [4]. In this approach the results of the land use classification maps extracted from the three images are compared. The classification of land use can be done better and more accurate with the raw (unprocessed) images. Therefore, there was no need for the above corrections for images used in this study.

- *Geometric Corrections:* The process and analysis of multi-temporal data can be done only when they are geo-referenced similarly, or in another words, when they are geo-referenced to each other [11]. The images of this study had to be geo-referenced to each other with an accuracy of one pixel. Otherwise, the error coming from different coordinates for similar objects in the two images can be wrongly accepted as a land use change. To prevent such a problem, in comparison of multi-temporal images, geo-reference one of the images using the available topographic maps and then geo-referencing the other images according to the first one, i.e. using image-to-image registration is done.

In photo/image registration (geo-referencing), the most important task is the proper selection of control points, especially when there is a long time period between the map and the image. In this work, the first order polynomial equations for geo-referencing of the images is used, which remove the errors related to the rotation and scaling of the image. The image may also be transformed to conform to a specific map projection system. Furthermore, if accurate geographical location

of an area on the image needs to be known, ground control points (GCP's) are used to register the image to a precise map (geo-referencing).

In this study, the ETM+ image of the year 2001 was first geo-referenced using the information in its header approximately. Then, it was geo-referenced accurately using the available 1:25000 digital maps and the digitized features of the 1:50000 maps of the area. Afterward, the TM image of 1991 was geo-referenced using the already registered TM image. For geo-referencing the 2001 image 18 control points were used initially. Every control point with an RMSE or residual error bigger than a pixel size was removed from the calculation and the process of registration was repeated with the rest of the control points. Finally, 10 points with the average error of 0.100 meters remained and were used for registration. For image-to-image registration of the 1991 image 20 control points were initially used. Finally, 6 points were removed and the image was geo-referenced using the remained 14 points with the RMSE of 0.92 meters.

All images and aerial photographs were rectified to UTM zone 39 N with at least 25 well distributed ground control points. At first geometric correction was carried out using topographic maps with the scale of 1:25000 to geo code aerial photos. Also for geometric correction of the 2001 IRS-1C land sat image, topographic maps with the scale of 1:25000 were used and then this rectified image was employed to register the 2011 LISS-III image. Geometric correction of Land sat TM image of 1990 was carried out by the use of IRS-P6 LISS-III image. Finally, a first-order polynomial model was applied and all data were resampled to a 30 m pixel size using the nearest neighbour method. After geometric correction of aerial photos, all photos for each year were mosaic ked to prepare one image for land cover mapping.

➤ *Image Enhancement:* In order to aid visual interpretation, visual appearance of the objects in the image can be improved by image enhancement techniques such as grey level stretching to improve the contrast and spatial filtering for enhancing the edges [4]. The goal of image enhancement is to improve the visual interpretability of an image by increasing the distinction between features. In this study, two false colour composites (FCC) are produced for selecting training samples. Also image fusion was done to increase spatial resolution of the LISS-III image. LISS-III image was fused with IRS-1C PAN image to generate an image with high spatial resolution [1]. Land sat TM enhanced false colour composites RGB (red, green, blue) 4,5,3; 5,3,2; 4,5,7 and 4,3,2 are used for the interpretation and delimitation of the land cover classes [11]. Interpretation and vectorization on the screen, available in Arc Info format was the preferred methodology because polygons created have

vector format and can be directly transformed to a land cover map.

- *Neighborhood Filling:* This method has been used to clean and fill the missing cell in the image while doing image classification.
- *Edge Matching:* This features is carried out to maintain the continuity of classes between adjoining mandals/districts/states. Generation of seamless geo data set at district/state level, creation of metadata, class wise area statistics are prepared.
- *Aerial Photos Interpretation:* Land cover pattern is interpreted visually on black and white aerial photographs and simultaneously digitized with the Arcmap software. Identifying features in aerial photos is performed based on tone, texture, pattern, size and shape.
- *Post-Classification Change Detection:* Post-classification comparison change detection algorithm is used to determine changes in urban areas in 3 decades from 1991 to 2011. Finally, due to anthropogenic activity, changes such as the reduced vigour of forest vegetation, urbanization, mining etc are noticed in the area.

IV. CONCLUSION

Data mining is data-driven but also, more importantly, human-centered, with the user controlling the selection and integration of data, cleaning and transformation of the data, choice of analysis methods, and the interpretation of results. The abundance of spatial data provides exciting opportunities for new research directions but also demands caution in using these data. The data are often from different sources and collected for different purposes under various conditions, such as measurement uncertainty, biased sampling, varying area unit, and confidentiality constraint. It is important to understand the quality and characteristics of the chosen data. Careful selection, preprocessing, and transformation of the data are needed to ensure meaningful analysis and results. Pre-processing improves performance, but massive data volumes associated with encoding spatial relationships for all combinations of geographic objects prohibits the storage of all spatial relationships.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Arms ton J.D, Danaher, T.J, Goulevitch, B.M and Byrne, M.I, "Geometric Correction of Land sat MSS,TM and ETM+Imagery for Mapping of Woody Vegetation Cover and Change Detection in Queensland", Climate Impacts and Natural Resource Systems, ISBN0-9581366-0, www.nrm.-qld.gov.au/slats, 2002.
2. Amos Storey, "Data Mining and Exploration: Preprocessing", School of Informatics, [http://www. -](http://www.)

- inf.ed.ac.uk/teaching/ courses/dme/, January 23, 2006.
3. Aleksandra Lazarevic, Tim Fiez and Zoran Obradovic, *"A Software System for Spatial Data Analysis and Modeling"*, INEEL University Research Consortium project No: C94-175936, www.ist.temple.edu/~zoran/papers/lazarevic00.pdf.
 4. Caroline M. Bruce and David W. Hilbert, *"Pre-processing Methodology for Application to Land sat TM/ETM+ Imagery of the Wet Tropics"*, Cooperative Research Centre for Tropical Rainforest Ecology and Management. Rainforest CRC, Cairns. (44 pp.), ISBN: 0864437609, www.rainforest-crc.jcu.edu.au, March 2006.
 5. Principles of Remote Sensing- Centre for Remote Imaging, Sensing and Processing, CRISP, www.crisp.nus.edu.sg/~research/tutorial/rsmain.html.
 6. Hutchinson C, *"Techniques for Combining Land sat and Ancillary Data for Digital Classification Improvement"*, Photogrammetric Engineering and Remote Sensing, Vol.48, No.1, 123-130, 1982.
 7. Luis Otavo Alvares, Gabriel Oliveira, Vania Bogorny, *"A Framework for Trajectory Data Preprocessing for Data Mining"*, http://www.inf.ufsc.br/~vania/artigos/seke2009_6.pdf.
 8. Lilles and TM and Keifer W, *"Remote Sensing and Image Interpretation"*, New York: John Wiley, 1994.
 9. Loveland T.R, Sohl T.L, Stedman S.V, Gallant A.L, Saylor K.L and Nap ton D.E, "A strategy for estimating the rates of recent United States land-cover changes. *Photogrammetric Engineering and Remote Sensing*", 68, 1091–1100, 2002.
 10. Roy P.S, Dwivedi R.S and Vijay an P, *"Remote Sensing Applications-Land Use Land Cover Analysis"*, National Remote Sensing Centre, 2011.
 11. Stefan Erasmi, Andre Twele, Muhammad Ardiansyah, Adam Malik and Martin Kappas, *"Mapping Deforestation and Land Cover Conversion at The Rainforest Margin in Central Sulawesi, Indonesia"*, EAR SeL proceedings 3, 2004.





This page is intentionally left blank



Deriving Association between Student's Comprehension and Facial Expressions using Class Association Rule Mining

By Dr. M. Mohamed Sathik & G. Sofia

Bharathiar University, India

Abstract - The scope of this study was to discover the association between facial expressions of students in an academic lecture and the level of comprehension shown by their expressions. This study focused on finding the relationship between the specific elements of learner's behavior for the different emotional states and the relevant expression that could be observed from individual students. The experimentation was done through surveying quantitative observations of the lecturers in the classroom in which the behavior of students are recorded and were statistically analyzed. The main aim of this paper is to derive association rules that represent relationships between input conditions and results of domain experiments. Hence the relationship between the physical behaviors that are linked to emotional state with the student's comprehension is being formulated in the form of rules. We present Predictive Apriori algorithm that is able to find all valid class association rules with high accuracy. The rules derived by Predictive Apriori are pruned by objective and subjective measures.

Keywords : *class association rules, predictive apriori algorithm, pruning, facial expression, objective measure, subjective measure.*

GJCST-C Classification : *H.2.8*



DERIVING ASSOCIATION BETWEEN STUDENTS COMPREHENSION AND FACIAL EXPRESSIONS USING CLASS ASSOCIATION RULE MINING

Strictly as per the compliance and regulations of:



RESEARCH | DIVERSITY | ETHICS

Deriving Association between Student's Comprehension and Facial Expressions using Class Association Rule Mining

Dr. M. Mohamed Sathik ^α & G. Sofia ^σ

Abstract - The scope of this study was to discover the association between facial expressions of students in an academic lecture and the level of comprehension shown by their expressions. This study focused on finding the relationship between the specific elements of learner's behavior for the different emotional states and the relevant expression that could be observed from individual students. The experimentation was done through surveying quantitative observations of the lecturers in the classroom in which the behavior of students are recorded and were statistically analyzed. The main aim of this paper is to derive association rules that represent relationships between input conditions and results of domain experiments. Hence the relationship between the physical behaviors that are linked to emotional state with the student's comprehension is being formulated in the form of rules. We present Predictive Apriori algorithm that is able to find all valid class association rules with high accuracy. The rules derived by Predictive Apriori are pruned by objective and subjective measures.

Keywords : class association rules, predictive apriori algorithm, pruning, facial expression, objective measure, subjective measure.

1. INTRODUCTION

Today's learning community focus on the vision of faculty and students working collaboratively towards deep, meaningful, high quality learning. The achievements of digital communication lead learning communities into a new dimension. There is an increase in virtual schools worldwide as education mediated by computer is considered very important for the future [12]. Nowadays, Learning Management Systems (LMS) are being installed more and more by universities, community colleges, schools, businesses, and even individual instructors in order to add web technology to their courses and to supplement traditional face-to-face courses [10]. LMS systems accumulate a vast amount of information which is valuable for analyzing the students' behavior and could create a gold mine of educational data [7].

Teacher student Interaction plays a vital role in the classroom environment. [5] In the classroom, lecturers and students both consciously and students

Students' behavior and could create a gold mine of educational data [7].

Teacher student Interaction plays a vital role in the classroom environment. [5] In the classroom, lecturers and students--both consciously and unconsciously--send and receive nonverbal cue several hundred times a day. Lecturers should be aware of nonverbal communication in the classroom for two basic reasons: to become better receivers of student's messages and to gain the ability to send positive signals that reinforces students' learning. Lecturers should be skilled at avoiding negative signals that stifle their learning.

Studies have evaluated that student emotional states are expressed with specific behaviors that can be automatically detected [17]. A preliminary study, carried out as the first part of this research, proved that the communicative impact of the face is so powerful in interaction. The most expressive way students display emotions is through facial expressions. Facial expressions are the primary source of information, next to words, in determining the student's emotional feelings to express their comprehension. It also strongly recommends that there is a direct connection between the facial expressiveness of the students and their level of comprehension. Momentary expressions that signal emotions include muscle movements such as raising the eyebrows, wrinkling the forehead, shrinking or enlarging the eyes or curling the lip [9].

This research specifically focused on studying the relationship between facial expressions of the students in an academic lecture and the level of comprehension shown by their expressions. The aim was to identify physical behaviors that are linked to emotional states, and to identify how these emotional states are linked to student's comprehension. The significance of the study was statistically interpreted. Hence it derives the Association rules which show the relationship between facial expressions of students in an academic lecture and the level of comprehension shown by their expressions.

The remainder of this paper is organized as follows. The Concepts implemented in this paper are explained in section II. Methods adopted in this paper are presented in section III. The experimental results are discussed in section IV. Finally, a conclusion

Author α: Principal, Sadakathullah Appa College, Tirunelveli, India.
E-mail : mmdsadiq@gmail.com

Author σ: Research and Development Centre, Bharathiar University, Coimbatore, India. E-mail : joesofi@gmail.com

and directions for future work are briefly covered in the last section.

II. CONCEPTS

a) Association Rule Mining

Data mining is the analysis of observational data sets to find the relationships among the data and to summaries it in novel ways that are both understandable and useful to the data owner [4]. The mining of association rules is a typical data mining task that works in an unsupervised manner. A major advantage of association rules is that they are theoretically capable of revealing all interesting relationships, called associations. It discovers relationships among attributes, producing if-then statements concerning attribute-values [1]. An association rule $X \Rightarrow Y$ expresses that in those transactions where X occurs; there is a high probability of having Y as well. X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by its support and confidence. The confidence of the rule is the percentage of transactions with X in the dataset that contain the consequent Y also. The support of the rule is the percentage of transactions in the dataset that contain both the antecedent and the consequent.

Definition of Association Rule: Let $I = \{i_1, i_2, \dots, i_m\}$ be set of items, D be task relevant data of transactions, T be each transaction, a set of items, such that $T \subset I$ where \subset denotes proper subset and TID be the Transaction Identifier. An Association Rule is defined as an implication of type $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \Phi$. The Rule hold in D with confidence C and support S , where C : Confidence ($A \Rightarrow B$) = $P(A \cup B)$, S : Support ($A \Rightarrow B$) = $P(B | A)$ where P is probability. [20] If B be a dataset with n items, then the support of an item set X is the number of instances which satisfy X given by the formula:

$$\text{sup}(X) = \frac{|\{t \in B | X \subseteq t\}|}{|B|} \quad (1)$$

The confidence of an association rule is a percentage value that shows how frequently the consequent part occurs among all the groups containing the rule antecedent part:

$$\text{conf}(X \rightarrow Y) = \frac{|\{t \in B | X \cup Y \subseteq t\}|}{|\{t \in B | X \subseteq t\}|} \quad (2)$$

Association rule mining has been applied to e-learning systems for traditional association analysis (finding correlations between items), such as discovering interesting relationships from student's usage information in order to provide feedback to course author [11], finding out the relationships between each pattern of learner's behaviour [18] etc. Association

rule mining also has been applied to the learning of sequential patterns mining, which is a restrictive form of association rule mining in the sense that not only the occurrences themselves, but also the order between the occurrences of the items is taken into account. The extraction of sequential patterns has been mainly used in e-learning for evaluating the learners' activities and can be used in adapting and customizing resource delivery [19]; discovering and comparison with expected behavioural patterns specified by the instructor that describes an ideal learning path [8]; classification [2].

Classification using association rules combines association rule mining and classification, and is therefore concerned with finding rules that accurately predict a single target (class) variable. The key strength of association rule mining is that all interesting rules are found. The number of associations present in even moderate sized databases can be, however, very large – usually too large to be applied directly for classification purposes. Therefore, any classification learner using association rules has to perform three major steps: Mining a set of potentially accurate rules, evaluating and pruning rules, and classifying future instances using the found rule set.

a) Class Association Rules

Normal association rule mining does not have any target. It finds all possible rules that exist in data, i.e., any item can appear as a consequent or a condition of a rule. However, in some applications, the user is interested in some targets. Let T be a transaction data set consisting of n transactions. Each transaction is also labeled with a class y . Let I be the set of all items in T , Y be the set of all class labels and $I \cap Y = \emptyset$. A class association rule (CAR) is an implication of the form $X \rightarrow y$, where $X \subseteq I$, and $y \in Y$. The definitions of support and confidence are the same as those for normal association rules.

A class Association rule is defined to be an implication with a pre-specified target (a value of target attribute) as its consequence and its support and confidence are above given thresholds from a dataset respectively. Given a target attribute, minimum support σ and minimum confidence ψ , a complete class association rule set is a set of all class association rules, denoted by $Rc(\sigma, \psi)$.

Conceptually, class association rules differ from standard association rules in their consequence. The objective is to generate the complete set of class association rules that satisfy the minimum support as well as the minimum confidence constraints and to build a classifier from the class association rule set. To this aim, one combines the prediction of all rules which satisfy the example: if there is only one rule, the consequent of this rule is taken to be the predicted class for the example; if there is no rule satisfying the example, then a default class is taken to be the

predicted class; and if there are multiple rules satisfying the example, then their predictions must be combined. Our goal is to find the minimum subset of the complete class association rule set that has the same prediction power as the complete association rule set [6].

b) Pruning

Association rule mining algorithms normally discover a huge quantity of rules and do not guarantee that all the rules found are relevant [3]. Support and confidence factors can be used for obtaining interesting rules which have values for these factors greater than a threshold value. Although these two parameters allow the pruning of many associations, another common constraint is to indicate the attributes that must or cannot be present in the antecedent or consequent of the discovered rules. Hence the solution is to evaluate, and post-prune the obtained rules in order to find the most interesting rules for a specific problem. A pruning technique is used for removing redundant or insignificant rules.

For practical applications the number of mined rules is usually too large to be exploited entirely. This is why the pruning phase is more essential in order to build accurate and compact classifiers. The smaller the number of rules a classifier needs to approximate the target concept satisfactorily and the human can interpret the result easily. Pruning strategies try to close the gap between the mining of a large number of class association rules and a small and powerful set of classification rules. Hence Pruning is an imperative step in mining association rules which helps in accurate classification.

III. METHODS

In this research a study was conducted for observing the facial expressions of the students in academic lecture-environments. The scope of this study was to establish whether there is a relationship between the student's facial expressions and the comprehension of the students. Also to examine whether facial expression of the students is a tool for the lecturer to interpret comprehension level of students in virtual classroom.

In order to perform the experiment for the study, survey was taken using stratified sampling technique with a questionnaire. Questionnaire was given to 100 Lecturers from 10 academic institutions, and their responses were collected. It focuses on the role of facial expressions in non-verbal communication. It ranks the order in which the lecturer interprets the level of comprehension in the classroom through various nonverbal communication modes. It also measures the frequency of the expressions exhibited by the action units of face for the purpose of communication. Finally, how the expressions are correlated with the emotions of the students was analyzed. Experimental data in the

domain is integrated into a dataset after statistical interpretation to serve as the basis for analysis.

The goal of association rule mining is to find all rules satisfying some basic requirement such as minimum support and the minimum confidence. A set of association rules for the purpose of classification is called predictive association rule set. Predictive association rules are based on attribute values where the consequences of rules are pre-specified categories. A class association rule set is a subset of Predictive association rules with the specified targets (classes) as their consequences [6]. Hence mining predictive association rules undergoes the following two steps. Find all class association rules.

Prune and organize the found class association rules and return a sequence of predictive association rules. Here in this paper all the class association rules are derived by Predictive Apriori Algorithm and the derived rules are pruned by objective and subjective measures.

a) Mining Class Association Rules

The mining of association rules is a typical data mining task that works in an unsupervised manner. A major advantage of association rules is that they are theoretically capable of revealing all interesting relationships in a set of data.

The improved version of the Apriori algorithm is the Predictive Apriori algorithm [13], which automatically resolves the problem of balance between two parameters, maximizing the probability of making an accurate prediction for the dataset. In order to achieve this, a parameter called the exact expected predictive accuracy is defined and calculated using the Bayesian method [15], which provides information about the accuracy of the rule found. In this way the user only has to specify the maximal number or rules to discover.

Apriori mines considerably more rules than predictive Apriori but most of them are pruned in the final set of classification rules. The advantage of predictive Apriori is that it generates fewer rules right from the start [14].

b) Predictive Apriori Algorithm

The Predictive Apriori algorithm [13] generates frequent item sets, but it uses a dynamically increasing minimum support threshold. It searches with an increasing support threshold for the best rules concerning a support-based corrected confidence value. A rule is added if: the expected predictive accuracy of this rule is among the "n" best and it is not subsumed by a rule with atleast the same expected predictive accuracy.

This scheme is an adapted version from Scheffer [13]

1. Predictive Apriori Algorithm:

2. Input the number of desired association rules and a dataset D with class attribute C.
3. Set optimal class association Rule set $R = \{\emptyset\}$.
4. Set the support threshold as 1.
5. Determine all frequent item sets whose support value is greater than the support threshold.
6. With such frequent item sets generate rules with high predictive accuracy.
7. Select the strong rules and include them in R.
8. Repeat the generation of rules till you get the desired number of association rules.
9. Output the optimal class association rule set R.
10. An important improvement in the Predictive Apriori (PA) is that there is no need to specify any of the parameters. Its objective is to find the best N association rules, being N a fixed number. An optimum set of class association rules are the output of this algorithm.

c) Predictive Accuracy

The Predictive Apriori algorithm differs from standard apriori in such a way that it employs a different measure of interesting of an association rule[14]. Predictive apriori evaluates the confidence of rules depending on their support. Its measure of interestingness is to maximize the expected accuracy an association rule will have on unseen data. This suits the requirements of the classification task we want to perform afterwards.

Scheffer [13] uses a Bayesian framework to calculate the predictive accuracy out of the support and confidence of a rule. In doing so the support is a rough guideline of how much we should mistrust the confidence. The higher the support, the more the confidence converges to the expected accuracy on future data.

This algorithm uses the Bayesian method to propose a solution that quantifies the expected predictive accuracy of an association rule with a given confidence and the support of the rule's body (left side of the rule). Scheffer [13] defines predictive accuracy as: Let $X \Rightarrow Y$ is an association rule. The predictive accuracy, $C(X \Rightarrow Y) = \Pr(r \text{ satisfies } Y \mid r \text{ satisfies } X)$ is the conditional probability of $Y \subseteq r$ given that $X \subseteq r$ when the distribution of r (records) is governed by $P(\text{Process})$. The confidence $\text{conf}(X \Rightarrow Y)$ of the association rule $X \Rightarrow Y$ is the relative frequency of the predictive accuracy in the data. Hence the confidence value is optimistically biased if one wants to use it for a predictive task.

The predictive accuracy describes whether the predicted values match the actual values of the target field due to statistical fluctuations and noise in the input data values. Hence it refers the ability of the model to correctly predict the class label of new or previously unseen data.

Using Bayesian formula the expected accuracy E of a rule r , $X \Rightarrow Y$ given its confidence conf and the support of the rule body $s(X)$ is calculated as

$$E(c(r) | \text{conf}(r), s(X)) = \frac{\int cB[c, s(X)](\text{conf}(r))P(c)dc}{\int B[c, s(X)](\text{conf}(r))P(c)dc} \quad (3)$$

This equation calculates the expected accuracy over unseen instances given the support of the rule body and the confidence of the rule, given that the instances are independent and identically distributed. This expectation value is called predictive accuracy.

d) Pruning

Traditionally for Pruning, [16] the use of objective interesting measures such as Predictive accuracy, support and confidence, Laplace, chi-square statistic, correlation coefficient, Entropy gain, Gini index, conviction, etc can be used for ranking the obtained rules in order. Subjective measures can also be used based on subjective factors controlled by the user. The subjective approaches involve user participation in order to express, in accordance with his or her previous knowledge, which rules are of interest so that the user can select the rules with highest values in the measures that he/she is more interested. The number of rules can be decreased by only applying these objective and subjective measures.

In this paper the class association rules derived by Predictive Apriori are pruned by applying the objective measure, Accuracy Rule Ranking followed by the subjective measure Expert Domain Knowledge.

e) Pruning using Objective Measure

For Pruning using objective measure, the obtained class association rules are to be ranked first. The ranking of class association rules can be done using the objective measure of interestingness. Predictive apriori considers predictive accuracy for ranking and so it sorts rules according to their predictive accuracy. Threshold accuracy is considered and the rules with predictive accuracies below the threshold accuracy will be pruned.

f) Pruning using Subjective Measure

However, the rules discovered by Predictive Apriori Algorithm and pruned by Accuracy Rule Ranking method may not all be useful with respect to the domain. Hence it is essential to prune the rules guided by Expert domain knowledge. Also, some interesting rules may not be found from experimental data. Thus it is advisable to extend the Association Analysis to other sources such as the related literature in the domain, to enhance the Knowledgebase.

Prior experience and domain knowledge [3] of the persons play an important role in ranking the rules. Experts use linguistic values to indicate their knowledge about the matter in response through relationships

among the attributes in the dataset. To improve the comprehension of the rules, incorporate Expert domain knowledge and semantics.

Steps to Prune using Basic Domain Knowledge

1. Consider rules derived using the Predictive Apriori Algorithm which is ranked by objective measure.
2. Use domain expert opinion to determine obvious and uninteresting rules.
3. If a derived rule matches an obvious rule or identified as uninteresting, then prune the derived rule.
4. Store obvious rules in a rule base for future use. These represent interesting information.
5. Repeat this process until all rules discovered are considered interesting in the domain.

Before running the class association rule mining algorithm, the relevant knowledgebase on the dataset in accordance with statistical interpretation associated with expert's response have to be prepared. In the context of Virtual educational environments, we can identify some common attributes that is observed from the students as seen in table1. Attributes are evaluated and ranked using Gain Ranking Filter in Weka. Ranking exhibits the extent of the attribute in expressing the comprehension level.

Table 1 : Attributes common for student's facial expressions

Rank	Attribute	Instance
1	Eye	Neutral/Enlarge/Shrink
2	Eyebrow	Neutral /Raised/Lowered
3	Forehead	Wrinkles/No Wrinkles
4	Mouth	Neutral /Curl/Stretch

Finally, we use the knowledgebase as a basis of rule repository in which subjective analysis is performed and associations are identified to discover the rules. In this context the use of standard metadata about the action units represent the facial expressions of students and allows the creation and maintenance of a common knowledge base with a common vocabulary as shown in Table1.

IV. EXPERIMENTAL RESULTS

Experimental data in the domain is integrated into a dataset to serve as the basis for analysis. On analyzing the experimental data, association between facial expressions of students in an academic lecture and the level of comprehension shown by their expressions could be observed and the rules that were sufficient to answer any question with respect to the problem domain could be derived.

Table 2 : Student's Facial Expression Dataset

Row id	Attributes				Class Label
	Eye	Mouth	Forehead	Eyebrow	
1	Neutral	Neutral	NoWrinkles	Neutral	UD
2	Neutral	Smile	NoWrinkles	Neutral	UD
3	Neutral	Curled	NoWrinkles	Neutral	IC
4	Shrink	Neutral	Wrinkles	Lowered	IC
5	Shrink	Curled	Wrinkles	Lowered	IC
6	Neutral	Neutral	Wrinkles	Raised	IC
7	Enlarge	Neutral	NoWrinkles	Raised	C
8	Enlarge	Smile	NoWrinkles	Raised	C

C-Comprehensible, IC-Incomprehensible, UD-Undecided.

Using the statistical measure of interestingness such as correlation and mean on the attributes, data is cleaned, grouped and categorized to form a dataset as shown in Table2 as good data preparation is the key to produce valid and reliable model.

Prior Experiments and statistical analysis on this research strongly suggested that facial expression is the most frequently used nonverbal communication mode used by the students in the classroom and student's expressions are significantly correlated to their emotions which can help to recognize their comprehension in the lecture. In particular, the more expressive the student is, more the lecturer recognizes the comprehension of the students. Facial Expressions that signal emotions include muscle movements such as raising eyebrows, wrinkling the forehead, rolling the eyes or curling the lip. So the action units of face such as eyes, mouth, eyebrow and forehead are the emotion indicators. Here we analyzed whether the emotional feelings of the students with respect to comprehension are indicated through expressions of facial action units. Experiments were made with survey and analysis was done through SPSS.

In order to find the association between the facial expressions of students in an academic lecture and the level of comprehension shown by their expressions, Predictive apriori algorithm is applied on the above dataset and the class association rules are being derived using Weka tool.

The discovered rules are sorted and ranked according to their Predictive accuracy. Irrespective of the number of rules to be predicted set as 100, Objective pruning got the optimal number of rules by applying the threshold accuracy 0.46584 as shown in Table4. Hence the number of class association rules generated was 14 as shown in Table3.

Table 3 : Generated Class Association Rules

Rule No.	Rule
1	Forehead=Wrinkles ==> Class=IC
2	Eye=Shrink ==> Class=IC
3	Eye=Enlarge ==> Class=C
4	Mouth=Curled ==> Class=IC
5	Eyebrow=Lowered ==> Class=IC
6	Forehead=NoWrinkles Eyebrow=Raised ==> Class=C
7	Eyebrow=Neutral ==> Class=UD
8	Eyebrow=Raised ==> Class=C
9	Eye=Neutral Forehead=NoWrinkles ==> Class=UD
10	Eye=Neutral ==> Class=UD
11	Eye=Neutral ==> Class=IC
12	Mouth=Neutral ==> Class=IC
13	Forehead=NoWrinkles ==> Class=UD
14	Forehead=NoWrinkles ==> Class=C

Table 4 : Calculated Predictive accuracy

Rule No.	Predictive Accuracy
1	0.98292
2	0.9729
3	0.9729
4	0.9729
5	0.9729
6	0.9729
7	0.61331
8	0.61331
9	0.61331
10	0.49952
11	0.49952
12	0.49952
13	0.46584
14	0.46584

For further simplification of rules subjective pruning was done on the pruned rules to get the best optimal set of rules. The obtained results or rules are interpreted and evaluated by the domain expert's knowledge for further actions. Rules with similar Predictive accuracies are grouped and some interesting rules that may not be found from experimental data are identified and included. The final objective is to put the results into use in form of if-then rules as shown below.

1. If Forehead=Wrinkles then
Class =Incomprehensible
2. If Eye=Shrink or Mouth=Curled or
Eyebrow= Lowered then
Class =Incomprehensible
3. If Eye=Enlarge then
Class=Comprehensible
4. If Forehead=No Wrinkles and Eyebrow=Raised
and Eye=Enlarge then
Class= Comprehensible
5. If Eyebrow=Neutral and Eye =Neutral and Mouth
= Neutral and Forehead=Nowrinkles then
Class=Undecided
6. If Eyebrow=Raised and Eye=Enlarge then
Class=Comprehensible

7. If Forehead=No Wrinkles and Eye=Neutral and
Eyebrow=Neutral then
Class= Undecided
8. If Forehead=Wrinkles and Eyebrow=Raised and
Eye=Neutral and Mouth Neutral then
Class= Incomprehensible

Lecturers use the above discovered rules for making decisions about the comprehension of the student in the virtual classrooms in order to improve the student's learning.

These interesting Class Association Rules are useful for predictive analysis and are used to populate a knowledgebase. They represent the knowledge that a domain expert discovers on learning from experimental data and literature surveys which can be used for decision support and classification.

V. CONCLUSION

Recent research tells teachers and students use facial expressions to form impressions of another. Facial Expression plays a vital role in identification of Emotions and Comprehension of the students in the virtual classrooms. This study derived the association between the specific elements of learner's behaviour for the different emotional states and the relevant expression that could be observed from individual students. This paper derived association rules that represent the relationship between the physical behaviors that are linked to emotional state with the student's comprehension and it was being formulated in the form of rules. The effectiveness of this method will be improved by correlating more features from different action units of the face which would improve the classification process.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Agrawal R., Imielinski, T., Swami, A.N. (1993): Mining Association Rules between Sets of Items in Large Databases. In: Proc. of SIGMOD 207-216.

2. Castro, F., Vellido, A., Nebot, A. and Mugica, F. (2007): Applying Data Mining Techniques to e-Learning Problems: a Survey and State of the Art. Evolution of Teaching and Learning Paradigms in Intelligent Environment. Springer, 183-221.
3. Enrique García, Cristóbal Romero, Sebastián Ventura, Toon Calders, (2007) Drawbacks and solutions of applying association rule mining in learning management systems, Proceedings of the International Workshop on Applying Data Mining in e-Learning.
4. Hand D., Mannila H. and P. Smyth. (2001), Principles of Data Mining. MIT Press, Cambridge, Massachusetts, USA.
5. Hutoria, Building a Harmonious Classroom Atmosphere, Articles Base, Cutesier.
6. Jiuyong Li, Hong Shen, Rodney Topor, Minimal Optimal Class Association Rule Set, School of Computing and Information Technology Grith University, Australia.
7. Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., and Heiner, C. (2005): An educational data mining tool to browse tutor-student interactions: Time will tell! In: Proc. of the Workshop on Educational Data Mining, 15-22.
8. Pahl, C., Donnellan, C.: (2003) Data mining technology for the evaluation of web-based teaching and learning systems. In: Proc.of Int. Conf. E-learning, 17.
9. Perry, Bruce, (2000), Can some people read minds?, Science World, Sep 4,
10. Rice, W.H.: (2006). Moodle E-learning Course Development. A complete guide to successful learning using Moodle. Packet publishing.
11. Romero, C., Ventura, S., Bra, P. D.: (2004) Knowledge discovery with genetic programming for providing feedback to courseware author. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 425-464.
12. Russell G., Holkner B., (2000), Virtual Schools, in Futures, Volume 32, Issues 9-10, November 2000, pp 887-897.
13. Scheffer T. (2001) Finding Association Rules That Trade Support Optimally against Confidence. Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), pages 424-435. Springer-Verlag.
14. Stefan Mutter,, Mark Hall, and Eibe Frank, (2004), Using Classification to Evaluate the Output of Confidence-Based Association Rule Mining, Australian Conference on Artificial Intelligence.
15. Stefan Mutter, (2004), Classification using Association Rules. (Thesis).
16. Tan P., Kumar V.: (2000) Interesting Measures for Association Patterns: A Perspective. TR00-036. Department of Computer Science. University of Minnesota. 1-36.
17. Toby Dragon, Ivon Arroyo, Beverly P. Woolf, Winslow Burleson, (2008) Viewing Student Affect and Learning through Classroom Observation and Physical Sensors, Proceeding of the 9th International conferences on Intelligent Tutoring Systems, Pages 29 - 39, Springer-Verlag Berlin, Heidelberg.
18. Yu, P., Own, C., Lin, L.: (2001) on learning behavior analysis of web based interactive environment. In: Proc. of the Int. Conf. on Implementing Curricular Change in Engineering Education 1-10.
19. Zaiane, O., Luo, J.: (2001) Web usage mining for a better web-based learning environment. In: Proc. of Int. Conf. on advanced technology for education 60-64.
20. Zemirline, L. Lecornu, B. Solaiman and A. Ech-cherif, (2008) An Efficient Association Rule Mining Algorithm for Classification, Artificial Intelligence and Soft Computing – ICAISC, 717-727.





This page is intentionally left blank



Data Leakage Detection by using Fake Objects

By Rama Rajeswari Mulukutla & P. Poturaju

Grandhi Varalakshmi VenkataRao Institute of Technology, India

Abstract - Modern business activities rely on extensive email exchange. Email leakage have become widespread throughout the world, and severe damage has been caused by these leakages it constitutes a problem for organization. We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). If the data distributed to the third parties is found in a public/private domain then finding the guilty party is a nontrivial task to a distributor. Traditionally, this leakage of data has handled by water marking technique which requires modification of data. If the watermarked copy is found at Some unauthorized site then distributor claim his ownership. To overcome the disadvantage of using watermark, data allocation strategies are used to improve the probability of identifying guilty third parties. The distributor must assess the likelihood that the leaked data come from one or more agents, as opposed to having been gathered from other means. In this project, we implement and analyze a guilt model that detects the agents using allocation strategies without modifying the original data .the guilt agent is one who leaks a portion of distributed data. We propose data “realistic but fake” data records to further improve our chances of detecting leakage and identifying the guilty party. And Algorithms implemented using fake objects will improve the distributor chance of detecting the guilt agent. It is observed that by minimizing the sum objective the chance of detecting guilt agents will increase. We also develop a framework for generating fake objects.

Keywords : allocation strategies, data leakage, data privacy, fake records, leakage model.

GJCST-C Classification : H.2.m



DATA LEAKAGE DETECTION BY USING FAKE OBJECTS

Strictly as per the compliance and regulations of:



RESEARCH | DIVERSITY | ETHICS

Data Leakage Detection by using Fake Objects

Rama Rajeswari Mulukutla^a & P. Poturaju^σ

Abstract - Modern business activities rely on extensive email exchange. Email leakage have become widespread throughout the world, and severe damage has been caused by these leakages it constitutes a problem for organization. We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). If the data distributed to the third parties is found in a public/private domain then finding the guilty party is a nontrivial task to a distributor. Traditionally, this leakage of data has handled by water marking technique which requires modification of data. If the watermarked copy is found at Some unauthorized site then distributor claim his ownership. To overcome the disadvantage of using watermark, data allocation strategies are used to improve the probability of identifying guilty third parties. The distributor must assess the likelihood that the leaked data come from one or more agents, as opposed to having been gathered from other means. In this project, we implement and analyze a guilt model that detects the agents using allocation strategies without modifying the original data. the guilt agent is one who leaks a portion of distributed data. We propose data "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party. And Algorithms implemented using fake objects will improve the distributor chance of detecting the guilt agent. It is observed that by minimizing the sum objective the chance of detecting guilt agents will increase. We also develop a framework for generating fake objects.

Keywords : allocation strategies, data leakage, data privacy, fake records, leakage model.

I. INTRODUCTION

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data.

According to Demanding market conditions encourage many companies to outsource certain business processes (e.g. marketing, human resources) and associated activities to a third party. This model

referred as business process outsourcing (BPO) and it allow the companies to focus on their core competency by subcontracting with other activities to specialists, resulting in reducing the operational costs, and increasing the productivity. Security and business assurance are essential for BPO.

In many cases the service provider needs access to the company intellectual property and other confidential information to carry out their services. For example a human resources BPO vendor may need access to employee databases with sensitive information (social security numbers), a patenting law firm to some research results, a marketing service vendor to contact information for customers or a payment service provider may need to access the credit card numbers or bank account numbers.

The main security problem in BPO is that the service provider may not be fully trusted or may not be securely administered. Business agreements for BPO try to regulate how the data will be handled by service providers, but it is almost impossible to truly enforce or verify such policies across different administrative domains. Due to digital nature, relational databases are easy to duplicate and in many cases a service provider may have financial incentives to redistribute commercially valuable data or may simply handle it properly. Hence, we need powerful techniques that can detect and deter such dishonest.

We study unobtrusive techniques for detecting leakage of a set of objects or records. Specifically, we study the following scenario: After giving a set of objects to agents, the distributor discovers some of those same objects in an unauthorized place. (For example, the data may be found on a website, or may be obtained through a legal discovery process.) At this point, the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means.

We develop a model for assessing the "guilt" of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding "fake" objects to the distributed set.

II. PROBLEM DEFINITION

Suppose a distributor owns a set $T = \{t_1, t_m\}$ of valuable data objects. The distributor wants to share some of the objects with a set of agents U_1, U_2, \dots, U_n . An agent U_i receives a subset of objects R_i which belongs to T , determined either by a sample request or

Author ^a : M.Tech(CSE), Department of Computer Science & Engineering, Grandhi Varalakshmi VenkataRao Institute of Technology Affiliated to JNTUK. E-mail : rama_mulukutla2003@yahoo.com

Author ^σ : Associate Professor, M.Tech, Department of Computer Science & Engineering, Grandhi Varalakshmi VenkataRao Institute of Technology Affiliated to JNTUK. E-mail : raju1poturaju1@gmail.com

an explicit request, Sample request $R_i = \text{SAMPLE}(T, m_i)$: Any subset of m_i records from T can be given to U_i . Explicit request $R_i = \text{EXPLICIT}(T, \text{cond}_i)$: Agent U_i receives all T objects that satisfy cond_i . The objects in T could be of any type and size, e.g., they could be tuples of relation, or relations in a database. After giving objects to agents, the distributor discovers that a set S of T has leaked. This means that some third party, called the target, has been caught in possession of s . For example, this target may be displaying S on its website, or perhaps as part of a legal discovery process, the target turned over s to the distributor. Since the agents U_1, \dots, U_n have some of the data, it is reasonable to suspect them leaking the data. However, the agents can argue that they are innocent, and that the S data were obtained by the target through other means.

a) Agent Guilt Model

Suppose an agent U_i is guilty if it contributes one or more objects to the target. The event that agent U_i is guilty for a given leaked set S is denoted by $G_i|S$. The next step is to estimate $\Pr\{G_i|S\}$, i.e., the probability that agent G_i is guilty Given evidence S . To compute $\Pr\{G_i|S\}$, estimate the probability that values in S can be "guessed" by the target.

For instance, say that some of the objects in t are e-mails of individuals. Conduct an experiment and ask a person with approximately the expertise and resources of the target to find the e-mail of, say, 100 individuals, the person may only discover 20, leading to an estimate of 0.2. We call this estimate p_t , the probability that object t can be guessed by the target.

The two assumptions regarding the relationship among the various leakage events. Assumption 1. For all $t, t' \in S$ such that $t' \neq t$ the provenance of t is independent of the provenance of t' . The term "provenance" in this assumption statement refers to the source of a value t that appears in the leaked set. The source can be any of the agents who have t in their sets or the target itself (guessing). Assumption 2. An object $t \in S$ can only be obtained by the target in one of the two ways: A single agent U_i leaked t from its own R_i set. The target guessed (or obtained through other means) t without the help of any of the n agents.

To find the probability that an agent U_i is guilty, given a set S , Consider. The target guessed t_1 with probability p , and that agent leaks t_1 to S with probability $1 - p$. First compute the probability that he leaks a single object t to S . To compute this, define the set of agents $V_t = \{U_i \mid t \in R_i\}$ that have t in their data sets. Then using assumption 2 and known probability p , we have $\Pr\{\text{some agent leaked } t \text{ to } S\} = 1 - P_1 \dots \dots \dots (1.1)$. Assuming that all agents that belong to V_t can leak t to S with equal probability and using Assumption 2 we obtain, $\Pr\{U_i \text{ leaked } t \text{ to } S\} = \{1 - p / |V_t|, 0, \text{ if } U_i \notin V_t, \text{ o} \dots \dots (1.2)$ Otherwise. Given that agent U_i is guilty if he leaks at least one value to S , with

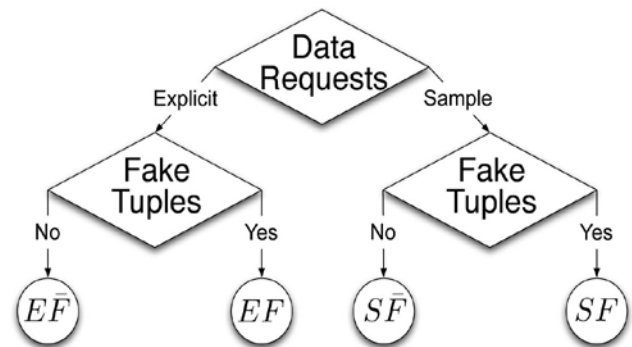
Assumption 1 and equation 1.2 compute the probability $\Pr\{G_i|S\}$, agent U_i is guilty, $\Pr\{G_i|S\} = 1 - \pi_{t \in S \cap R_i} (1 - 1 - p / |V_t|)$.

b) Data Allocation Problem

The distributor "intelligently" gives data to agents in order to improve the chances of detecting a guilty agent. There are four instances of this problem, depending on the type of data requests made by agents and whether "fake objects" are allowed.

Agent makes two types of requests, called sample and explicit. Based on the request the Fake objects are added to the data list. Fake objects are objects generated by the distributor that are not in set T . The objects are designed to look like real objects, and are distributed to agents together with T objects, in order to increase the chances of detecting agents that leak data.

Figure 1 : Leakage Instance Problems



The Figure represents our four problem instances with the names EF, EF, SF, and SF, where E stands for explicit requests, S for sample requests, F for the use of fake objects, and F for the case where fake objects are not allowed.

The distributor may be able to add fake objects to the distributed data in order to improve in his effectiveness in detecting guilty agents. Since, fake objects may impact the correctness of what agents do, So they may not be allowable. Use of fake objects may be inspired by the use of "trace" records in mailing lists. The distributor creates and add fake objects to the that he distributes to the agents. In many cases, the distributor may be limited how many fake objects he can create.

In EF problems, objectives values are initialized by agent's data requests. Say for example, that $t = \{t_1, t_2\}$ and $R_2 = \{t_1\}$. The distributor cannot remove or alter the R_1 or R_2 to the data to decrease the overlap $R_1 \cap R_2$. However, say the distributor can create one fake object ($B=1$) and both agents can receive one fake objects ($b_1=b_2=1$). If the distributor is able to create fake objects, he could improve further objective.

III. OPTIMIZATION PROBLEM

The distributor's data allocation to agents has one constraint and one objective. The distributor's constraint is to satisfy agents' requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His objective is to be able to detect an agent who leaks any portion of his data. We consider the constraint as strict. The distributor may not deny serving an agent request and may not provide agents with different perturbed versions of the same objects. The fake object distribution as the only possible constraint relaxation. The objective is to maximize the chances of detecting a guilty agent that leaks all its data objects. The $\Pr\{G_j|S = R_i\}$ or simply $\Pr\{G_j|R_i\}$ is the probability that agent U_j is guilty if the distributor discovers a leaked table S that contains all R_i objects. We define the difference functions $\Delta(i,j)$ is defined as $\Delta(i,j) = \Pr\{G_i|R_i\} - \Pr\{G_j|R_j\}$.

a) Problem Definition

Let the distributor have data request from n agents. The wants to give tables R_1, \dots, R_n to agents U_1, \dots, U_n respectively, so that Distributor satisfies agent's requests; and Maximizes the guilt probability differences $\Delta(i,j)$ for all $i,j=1, \dots, n$ and $i \neq j$. Assuming that the R_i sets satisfy the agent's requests, we can express the problem as a multi-criterion.

b) Optimization Problem

Maximize $(\dots, \Delta(i,j), \dots)_{i \neq j, \dots, (1.5)}$ (over R_1, \dots, R_n). The approximation [3] of objective of the above equation does not depend on the agent probabilities and therefore minimize there alive overlap among the agents as Minimize $(\dots, |R_i \cap R_j| / |R_i|, \dots)_{i \neq j, \dots, (1.6)}$ over (R_1, \dots, R_n) . This approximation valid if minimizing the relative overlap, $|R_i \cap R_j| / |R_i|$ maximizes (i,j) .

IV. OBJECTIVE APPROXIMATION

In case of sample request, all request are fixed size. Therefore, maximize the chance of detecting a guilt agent that leaks all his data by minimizing, $|R_i \cap R_j| / |R_i|$ is equivalent to minimizing $|R_i \cap R_j|$. The minimum value of $|R_i \cap R_j|$ maximizes $\prod |R_i \cap R_j|$ and $\Delta(i,j)$ since $\pi |R_i|$ is fixed. If agents have explicit data requests, that overlaps $|R_i \cap R_j|$ are defined by their own requests and $|R_i \cap R_j|$ are fixed. Therefore minimizing $|R_i|$ is equivalent to maximizing $|R_i|$ (with the addition of fake objects). The maximum value of $|R_i|$ minimizes $\prod |R_i|$ and maximizes $\Delta(i,j)$, since $\prod (R_i \cap R_j)$ is fixed. Our paper focus on identifying the leaker. So we propose to trace the ip address of the leaker. The file is send to the agents in the form of email attachments which need a secret key to download it. This secret key is used to generate random function and send to the agent either on the mobile number used at registration or to the other

global email service such as gmail. Whenever secret key mismatch takes place the fake file gets downloaded. To further enhance our objective approximation ip address tracking is done of the system where fake object is downloaded. Various commands are available for getting ip address information. ping, tracertr etc may one be used to get it. The ip address traced with time so as to overcome problem of dynamic ip addressing. But as we are doing in organization there is no problem of dynamic ip. Or else looking for the ip address universally it is unique that period of time therefore it can be traced to the unique system of the leaker.

V. ALLOCATION STRATEGIES

In this section, the allocation strategies that solve exactly or approximately the scalar versions of equation 1.7 for the different instances presented in Fig.1. In this Section. A deals with problems with explicit data requests, and in Section B with problems with sample data requests.

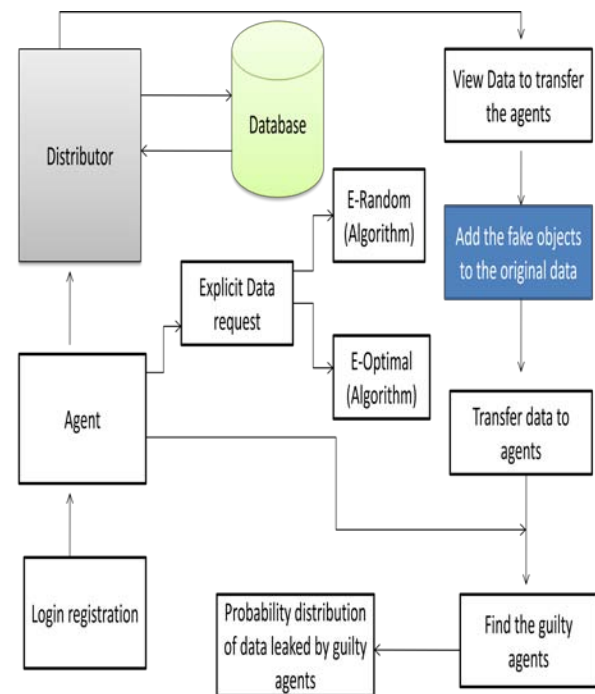


Figure 2: Architecture of Distributor

a) Explicit Data Requests

In case of explicit data request with fake not allowed problem to add fake objects to the distributed data. So, the data allocation is fully defined by the agents' data requests. In case of explicit data request with fake allowed, the distributor cannot remove or alter the request R from the agent. However distributor can add fake object. In algorithm for data allocation for explicit request, the input to this is a set of request R_1, \dots, R_n from n agents and different conditions for requests. The e-optimal algorithm finds the agent that is

eligible to receiving fake objects. Then create one fake object for iteration and allocate it to the agent selected. The e-optimal algorithm minimizes every term of the objective summation by adding maximum number b_i of fake objects to every set R_i yielding optimal solution.

Step 1: Calculate total fake records as sum of fake records allowed.

Step 2: While total fake objects > 0 .

Step 3: Select the agent that yield the greatest improvement in sum objective i.e; $i = \text{argmax}(1/|R_i| - 1/|R_i| + 1) \sum_j R_i \cap R_j$.

Step 4: Create fake record.

Step 5: Add this fake record to the agent and also to fake record set.

Year 2013

38

b) Sample Data Requests

With sample data requests, each agent U_i may receive any T subset out of $(|T|_m)$ different ones. Hence, there are $\prod_{i=1}^n (|T|_m)$ different object allocations. In every allocation, the distributor can permute T objects and keep the same chances of guilty agent detection. The reason is that the guilt probability depends only on which agents have received the leaked objects and not on the identity of the leaked objects.

Therefore, from the distributor's perspective, there are $\prod_{i=1}^n (|T|_m) / |T|$ different allocations. An object allocation that satisfies requests and ignores the distributor's objective is to give each agent a unique subset of size m . The s-max algorithm allocates to an agent the data record that yields the minimum increase of the maximum relative overlap among any pair of agents. The s-max algorithm as follows:

Step 1: Initialize $\text{Min_overlap} \leftarrow 1$, the minimum out of the maximum relative overlaps that the allocation of different objects to U_j .

Step 2: For $k \in \{k' | t_k \in R_i\}$ do.

Step 3: For all $j = 1, \dots, n$: $j = 1$ and $t_k \in R_j$ do.

Calculate absolute overlap as

$\text{abs_ov} \leftarrow |R_i \cap R_j| + 1$

Calculate relative overlap as:

$\text{rel_ov} \leftarrow \text{abs_ov} / \min(m_i, m_j)$

Step4: Find maximum relative as

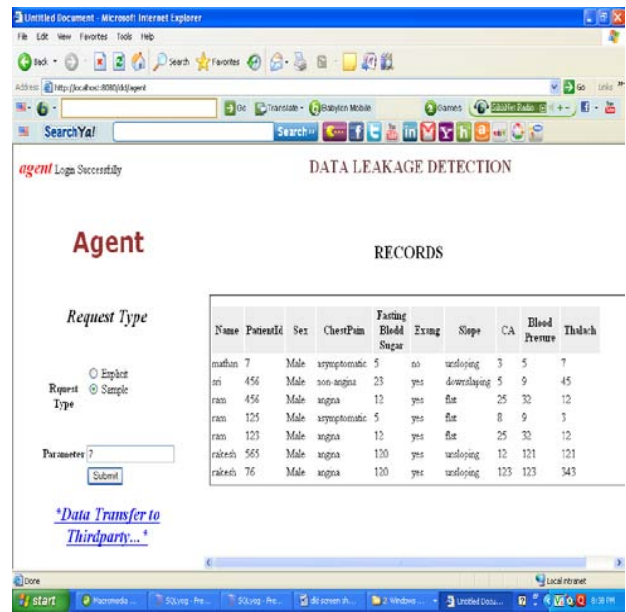
$\text{max_rel_ov} \leftarrow \text{Max}(\text{max_rel_ov}, \text{rel_ov})$

If $\text{max_rel_ov} \leq \text{min_overlap}$ then

$\text{min_overlap} \leftarrow \text{max_rel_ov}$

$\text{ret_k} \leftarrow k$

Return ret_k



VI. RELATED WORK

The guilt detection approach we present is related to the data provenance problem tracing the lineage of S objects implies essentially the detection of the guilty agents. Suggested solutions are domain specific, such as lineage tracing for data warehouses, and assume some prior knowledge on the way a data view is created out of data sources. Our problem formulation with objects and sets is more general and simplifies lineage tracing, since we do not consider any data transformation from R_i sets to S .

As far as the data allocation strategies are concerned, our work is mostly relevant to watermarking that is used as a means of establishing original ownership of distributed objects. Watermarks were initially used in images video and audio data whose digital representation includes considerable redundancy. Our approach and watermarking are similar in the sense of providing agents with some kind of receiver identifying information. However, by its very nature, a watermark modifies the item being watermarked. If the object to be watermarked cannot be modified, then a watermark cannot be inserted.

In such cases, methods that attach watermarks to the distributed data are not applicable. Finally, there are also lots of other works on mechanisms that allow only authorized users to access sensitive data through access sensitive data. Such approaches prevent in some sense data leakage by sharing information only with trusted parties. However, these policies are restrictive and may make it impossible to satisfy agents' requests.

VII. CONCLUSION AND FUTURE WORK

In spite of these difficulties, we have shown that it is possible to assess the likelihood that an agent is

responsible for a leak, based on the overlap of his data with the leaked data and the data of other agents, and based on the probability that objects can be “guessed” by other means. Our model is relatively simple, but we believe that it captures the essential trade-offs. The algorithms we have presented implement a variety of data distribution strategies that can improve the distributor’s chances of identifying a leaker.

We have shown that distributing objects judiciously can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive. Our future work includes the investigation of agent guilt models that capture the leakages.

ACKNOWLEDGEMENT

I would like to express my sincere thanks to my guide and my authors for their consistence support and valuable suggestions.

REFERENCES RÉFÉRENCES REFERENCIAS

1. R. Agrawal and J. Kiernan, “Watermarking Relational Databases”, *proc.28th Intl conf. very Large Data Bases (VLDB’02)*, VLDB Endowment, pp, 155-166, 2002.
2. P. Bonatti, S. D. C. di Vimercati and P. Samarati, “An Algebra for composing Access Control Policies”, *ACM Trans. Information and System Security*, vol.5, no.1, pp.1-35, 2002.
3. P. Buneman and W.C.T. an, “Provenance in Databases”, *proc.ACM SIGMOD*, pp. 1171-1173, 2007.
4. F. Hartung and B. Girod, “Watermarking of Uncompressed and Compressed Video”, *Signal processing*, vol.66, no.3, pp.283-30, 1998.
5. B. Mungamuru and H. Garcia-Molina, “Privacy, Preservation and Performance: The 3 P’s of Distributed Data Management”, technical report, Stanford Univ., 2008.
6. P. Papadimitriou and H. Garcia-Molina, “Data Leakage Detection”, technical report Stanford Univ., 2008.
7. L. Sweeney “Achieving K-Anonymity Privacy Protection Using Generalization and Suppression”, <http://en.scientificcommons.org/4319613>, 2002.
8. Y. Li, V. Swarup and S. Jajodia, “Fingerprinting Relational Databases: Schemes and Specialties”, *IEEE trans. Dependable and Secure Computing*, vol.2, pp.33-45, jan-mar. 2005.



This page is intentionally left blank



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING

Volume 13 Issue 6 Version 1.0 Year 2013

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals Inc. (USA)

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

Ipod System's Usability: An Application of the Fuzzy Logic

By Luciana Jácome Basto Cordeiro, Luiz Fernando Ribeiro Parente Filho,
Rodrigo Costa dos Santos, Walter Gassenferth & Maria Augusta Soares Machado

K.L. University, India

Abstract - In order for the initial result of this research to be obtained scientifically, a decision was made on setting the limits of the universe of users to be studied. For this end, the research was centered on the users of a Rio de Janeiro, Brazil, college and the EXCEL and MATLAB software were used. The Fuzzy Logic was employed to assess iPod usability.

The methodology presented here by is unprecedented and was developed by a 25-student study group of the above-mentioned college in their research.

Keywords : *fuzzy logic; usability; nbr.*

GJCST-C Classification : *H.2.m*



Strictly as per the compliance and regulations of:



Ipod System's Usability: An Application of the Fuzzy Logic

Luciana Jácome Basto Cordeiro ^α, Luiz Fernando Ribeiro Parente Filho ^σ, Rodrigo Costa dos Santos ^ρ,
Walter Gassenferth ^ω & Maria Augusta Soares Machado [¥]

Abstract - In order for the initial result of this research to be obtained scientifically, a decision was made on setting the limits of the universe of users to be studied. For this end, the research was centered on the users of a Rio de Janeiro, Brazil, college and the EXCEL and MATLAB software were used. The Fuzzy Logic was employed to assess IPod usability.

The methodology presented here by is unprecedented and was developed by a 25-student study group of the above-mentioned college in their research.

Keywords : fuzzy logic; usability; nbr.

I. INTRODUCTION

Due to technological advances, the world is getting more and more dynamic and competitive. Making use of automated and more complex systems becomes indispensable in this scenario. Thus, more qualified professionals and more user-friendly systems are required.

This paper assessed the usability of the MP3 and MP4 IPod software. The metrics for the analyses are as follows: ease of use, efficiency, and effectiveness to accomplish their tasks and the satisfaction the system provides its user. The methodology can be used for assessing any other systems and equipment with the same results. With such results, an excellent basis is achieved that allows for possible changes and improvements in the desired system.

Initially, a review of the literature on the usability of the fuzzy logic is presented; after that, the methodology used in the research. Closing the article is a presentation of the results and conclusions.

II. SYSTEMS USABILITY

The usability of a system is a relevant factor to motivate the user (client) into maximization of fidelity. Nevertheless, if the system does not make the user an ally, the user will certainly search for another system (PRESSMAN, 1992). The system must provide its user with ease of interaction, in an effective, efficient and satisfying manner.

Many techniques exist to assess the usability of systems such as those based on questionnaires assigned to the users; on formal models; knowledge-based ones; checklists; interaction essays; or monitoring systems (CYBIS, 2003). The Fuzzy logic has come to innovate these techniques.

III. THE FUZZY LOGIC

The first notions of the Fuzzy logic were developed by Jan Lukasiewicz (1878 – 1956) in 1920. Instead of using strict rules as well as line of logical thinking based on premises and conclusions, Lukasiewicz ascribes levels of pertinence $\{0, \frac{1}{2}, 1\}$ to classify vague and inaccurate concepts. Eventually, he expanded this set for all the values contained in the interval $[0, 1]$. However, the first publication on the Fuzzy logic dates back to 1965 by Lotfi Asker Zadeh, a professor of the University of California, Berkeley (CEZAR, MACHADO and OLIVEIRA JR., 2006).

The Fuzzy logic is based upon the theory of the Fuzzy Sets. This term is a generalization for the Traditional Sets theory to solve the paradoxes derived from the “true or false” classification of the Classical Logic. The Fuzzy Sets and the Fuzzy Logic provide the basis for generating powerful techniques towards problem-solving with broad applicability, especially in the fields of Control Engineering and decision-making.

The strength of the Fuzzy Logic stems from its ability to infer conclusions and generate answers based upon vague, ambiguous and qualitatively incomplete and inaccurate information. In this regard, the Fuzzy systems have the ability to ‘think’ like humans do. Its behavior is represented in very simple and natural way, leading to the construction of comprehensible and easy to maintain systems.

The Fuzzy Logic is based on the Fuzzy Sets theory. This is some generalization to the theory of Traditional Sets to solve the paradoxes arising from the “true or false” classification according to the Classical Logic. Traditionally, a logical proposition has two extremes, either “completely true” or “completely false”. However, according to the Fuzzy Logic, a premise ranges in the level of truth from 0 to 1, thus causing it to be either partially true or partially false. By incorporating the “level of truth” concept, the Fuzzy Sets theory provides some expansion to the Traditional sets theory,

Author ^{α ρ} : Faculdades IBMEC-RJ, Brasil.

Author ^ω : MSc, Faculdades IBMEC-RJ, Brasil.

E-mail : wgassen@quantiac.com

Author [¥] : DSc, Faculdades IBMEC-RJ, Brasil.

E-mail : mmachado@ibmecrj.br

whereby the groups are labeled qualitatively (by using such linguistic terms as: high, warm, active, small, near etc.) and the elements of these sets are so characterized by varying the level of pertinence (a value that indicates the level at which an element belongs to a set). For instance, temperatures between 30° (thirty degrees) and 40° (forty degrees) belong to the "high temperatures" set, although the 40° temperature has a greater level of pertinence in this set (OLIVEIRA JR. et al, 2007).

The level of association is not probability, but a measure of compatibility of the object with the concept represented by the Fuzzy Set. For example, number 0.7 is the compatibility of temperature 35° with the definition of the high temperatures Fuzzy Set. This figure (0.7) is not a probability of 35° being a high temperature, since this temperature is already defined as 35° (CEZAR, MACHADO and OLIVEIRA JR., 2006).

The conventional systems theory is based upon algebraic equations, either differential or difference ("crisp" mathematical models). For some types of systems, mathematical models can be obtained such as the electromechanical systems, since the laws of physics backing the process are well understood and defined. However, on a daily basis, we come across countless practical problems, and this makes it difficult to achieve a stable level of information required for the physics modeling to be accomplished. A great part of such systems can only be obtained through the knowledge of experts who directly take part in the process under analysis. Many times, such knowledge may be vague or inaccurate to be expressed by mathematical models.

IV. METHODOLOGY

This paper is based upon an applied research, since it aims at using a real case to support its analysis. The applied research is driven by the need to solve concrete problems with a practical purpose (VERGARA, 2000).

The system selected was the iPod. The name iPod refers to a series of digital audio players designed and sold by Apple Inc. The players of the iPod family provide a simple, click wheel-based, interface for the use. The largest of the iPod models stores media in an attached hard drive, whereas the smaller models, like the iPod shuffle and the iPod Nano, use a flash memory. Like the majority of digital portable players, the iPod may act as a data storage device when connected to a computer.

Usability is one of the items considered in the standards established by the ISO (International Standardization Organization) related to quality. The ABNT (Brazilian Standards Association) is the official entity accountable for the discussion and publication of technical standards in Brazil. It is the ISO representative in Brazil.

However, the ISO standards do not include a set of criteria or metrics for assessing systems usability. For that reason, the metrics used in this study are those presented by Santos (2007), who established a set of metrics for assessing system usability on the basis of a review of the literature on Brazilian scientific bases between 1995 and 2006. These system usability metrics are based on the ISO 9126 and on the assessment criteria according to some authors such as Shackel, Nielsen, Bastien & Scapin, Jordan, Shneiderman and Quesenberry.

According to Santos (2007), the metrics considered for usability assessment as well as those used in this research are as follows:

- Ease of learning
- Ease of remembering
- Error control
- Efficiency
- Effectiveness
- Satisfaction

'Ease of learning', or intelligibility, according to the ISO 9126 (2003), is the capability of the software to enable the user to learn how to handle it.

Such metrics is being assessed according to the following constructs:

- a) User's ease to accomplish a task for the first time;
- b) The user's first impression upon using the system;
- c) Number of attempts in order to learn how to accomplish a task;
- d) Time required for learning how to accomplish a task successfully;
- e) Interaction with the system's interface;
- f) Accomplishment of tasks in the system regarding message clarity, error recovery etc.;
- g) Ease of learning a task;
- h) Number of different possibilities the system provides in order to accomplish the same task, for example: standard path versus shortcut keys, shorter paths, macros, specific buttons etc.;
- i) Gain and productivity regarding the quickest way the user can accomplish a task in comparison with the standard way the system provides by default;
- j) The flexibility the system provides to perform the tasks in different ways, for example: personalization of shortcuts, values, menus, macros etc.;

'Ease of remembering', according to Nielsen, assesses system functionalities so they are easy to remember even after the user does not use them for some time, without the need of new coaching.

Such metrics is being assessed according to the following constructs:

- a) The capability of the system to guide through its execution with hints, help, notices etc.;
- b) Briefness to successfully accomplish a task in the system for the first time;

- c) Remembering how to perform a task after a certain period of time without using the system.

'Error Control', or operability, according to the ISO 9126, is the capability the software has to enable the user to operate and control it.

Such metrics is being assessed according to the following constructs:

- Ease of remembering the utilization of the system;
- The agility to remember how to use the system after a period of time without using it;
- The amount of errors caused by the system;
- Time elapsed before the system resumes normal operation when an error occurs;
- The feeling regarding the amount of errors caused by the system;
- Re-work due to the amount of errors caused by the system, resulting in some loss of information;
- The time spent to resume the execution of the task from the point where it was interrupted when an error occurs.

'Efficiency', also called operability, according to the ISO 9126, is the capability the software has to enable the user to operate and control it.

Such metrics is being assessed according to the following constructs:

- Satisfaction regarding recovery from the error by the system, undo, redo, back, save before closing etc.;
- System performance;
- The velocity for accomplishing the tasks;
- System productivity.

'Effectiveness', according to Quesenbery, assesses how the tasks were exactly accomplished and how often they produce errors.

Such metrics is being assessed according to the following constructs:

- Keeping the system under control;
- The amount of steps required to accomplish a task;
- The time taken to accomplish any task in the system.

'Satisfaction', or attractiveness, according to the ISO 9126, is the capability the software has to attract the user, to be pleasant.

Such metrics is being assessed according to the following constructs:

- The amount of steps required to accomplish a task in the system;
- The clarity of the error messages presented by the system;
- The user's feeling for using the system as a whole.

After collecting data and consolidating users' opinions, a Fuzzy methodology was applied in order to achieve a fuzzy triangular figure resulting from the frequency of users' opinions for the sets making up the constructs of the metrics being assessed.

The fuzzy triangular numbers are special fuzzy numbers with two very important features, namely: MODA and RANGE. The Moda represents the value of the fuzzy number the pertinence of which is equal to 1 (one). Range is half of the basis of the fuzzy number and represents the confidence interval for the number. Range is inverse to the confidence of the pertinence function: the lower the range, the higher the confidence on the data; the higher the range, the lower the confidence on the data (BRAGA, BARRETO & MACHADO, 1995).

The Likert scale was used to answer each question. This is a scale whereby the respondents are requested not only to agree with or disagree from the assertions, but also to inform their level of concordance or discordance (MATTAR, 1997). A five-point Likert scale was used for measuring usability: (Very Low ^ Very High).

V. RESULTS

The statistics for the opinions expressed in the questionnaire responded by the twenty-three users is presented below along with the description of the basic statistics followed by the resulting triangular fuzzy numbers for each metrics and the interpretation for them.

a) Statistical Description of the Sample

For the 'ease of learning' metric, Figure 1:

Considering that this system is used for teaching statistics for beginners, it can be observed that it is relatively easy for users to learn.

For the 'ease of remembering' metric, Figure 2:

By observing the table singly, users find no difficulty in remembering.

For the 'error control' metric, Figure 3:

Apparently, users are completely satisfied.

For the 'efficiency' metric, Figure 4:

Apparently, users are relatively satisfied.

For the 'effectiveness' metric, Figure 5:

Apparently, users are relatively satisfied.

For the 'satisfaction' metric, Figure 6:

Apparently, users are relatively satisfied.

b) Fuzzy Numbers

The calculations required for the fuzzy analysis of the results achieved with the metrics were made by using the MatLab mathematical software, which generated the results in graphical mode for each studied metric.

The graph presented for each metric represents two sets. The first set, depicted by a line with square markers, represents the fuzzy number (FN) for the mean of all frequencies found for the metric matters. The second set, depicted by a line with asterisk markers, represents the fuzzy number (FN) in the triangular shape

that mostly resembles the first set, which is the final result for the assessed metric.

i. *Ease of Learning*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 1.

It can be observed that the average opinion is 4 (average satisfaction) with range 3, thus indicating an average dispersion in interviewees' opinions.

Result still undefined; sampling to be increased.

ii. *Ease of Remembering*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 2.

It can be observed that the average opinion is 6 (good satisfaction) with range 4, thus indicating a high dispersion in interviewees' opinions.

It can be ascertained that the software is adequate regarding ease of remembering.

Even with this result, the sample has to be increased in order to reduce the range of the result.

iii. *Error Control*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 3.

It can be observed that the average opinion is 6 (high satisfaction) with range 3, thus indicating an average dispersion in interviewees' opinions.

It can be ascertained that the software is moderately adequate regarding error control.

Even with this result, the sample has to be increased in order to reduce the range of the result.

iv. *Efficiency*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 4.

It can be observed that the average opinion is 4 (average satisfaction) with range 2, thus indicating a low dispersion in interviewees' opinions.

It can be ascertained that the software is moderately adequate regarding efficiency.

Even with this result, the sample has to be increased.

v. *Effectiveness*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 5.

It can be observed that the average opinion is 6 (high satisfaction) with range 3, thus indicating an average dispersion in interviewees' opinions.

It can be ascertained that the software is adequate regarding effectiveness.

Even with this result, the sample has to be increased in order to reduce dispersion.

vi. *Satisfaction*

The triangular fuzzy number achieved for measuring the ease of learning is presented in Graph 6.

It can be observed that the average opinion is 4 (average satisfaction) with range 3, thus indicating an average dispersion in interviewees' opinions.

It can be ascertained that the software is adequate regarding satisfaction.

Even with this result, the sample has to be increased in order to reduce the range of the result.

VI. CONCLUSIONS

The results achieved through this study aim at presenting a new criterion for assessing software usability. Depending on the system being analyzed, this result can be interpreted as customer retention, profit increase, production increase, employee satisfaction, amongst other benefits. In this case, the methodology was applied to MP3 and MP4 software on a sample of students; that is, the largest target public to the system. The software must be easy to learn and remember, it cannot cause errors and complete its tasks efficiently and effectively. All these factors imply the user's end satisfaction towards the product.

After analyzing the partial results of this research, the usability of the iPod was obtained as per below:

- Users showed an average ease of learning level on how to use the software;
- Users showed good ease to remember level on how to use the software when they do not use the system for some time;
- Users showed average satisfaction as to error control;
- Users found the software has average efficiency;
- Users found the software effectiveness was good;
- Users were satisfied with using the software.
- Assessment of these partial results shall be carried out through a collection of data from the internet for other users throughout the country.

REFERENCES RÉFÉRENCES REFERENCIAS

1. RAGA, MARIO J. F.; BARRETO, JORGE M.; MACHADO, MARIA AUGUSTA S. Conceitos da Matemática Nebulosa na Análise de Risco. Rio de Janeiro: Artes & Rabiskus, 1995.
2. CEZAR, BRENO L., MACHADO, MARIA AUGUSTA S., OLIVEIRA JR, HIME A. Sistema de Apoio à Decisão na Concessão de Crédito Pessoal usando Lógica Fuzzy. Anais do Simpósio de Excelência em Gestão e Tecnologia (SEGET), Rio de Janeiro, 2006.
3. ISO 9126-1. Engenharia de software – Qualidade de produto. Parte 1: Modelo de qualidade. NBR ISO/IEC 9126-1. Rio de Janeiro: ABNT, 2003.
4. MATTAR, F.. Pesquisa de Marketing. São Paulo: Editora Atlas, 1997.
5. NIELSEN, J. Usability Engineering. Boston, MA: Academic Press, 1993.
6. OLIVEIRA JR., HIME A. Lógica Difusa: Aspectos Práticos e Aplicações. Rio de Janeiro: Interciência, 1999.

7. OLIVEIRA JR., HIME A., CALDEIRA, ANDRÉ M., MACHADO, MARIA A. S., SOUZA, REINALDO, TANSCHKEIT, RICARDO. Inteligência Computacional Aplicada à Administração, Economia e Engenharia em Matlab. Rio de Janeiro, Thompson, 2007.
8. PRESSMAN, R. S. Software Engineering – A Practitioner's Approach, 3 ed., McGraw-Hill, 1992.
9. WIKIPÉDIA. Disponível em: <http://pt.wikipedia.org/wiki/IPod>. Acesso em: 08 ago. 2007.
10. SANTOS, Rodrigo C.; MACHADO, Maria Augusta S. Development of a Methodology for Systems Usability Evaluation Using Fuzzy Logic Based on ISO. In: SIMPOI POMS 2007-X Simpósio de Administração da Produção, Logística e Operações Internacionais, 2007. Rio de Janeiro: FGV-EAESP, 2007.
11. VERGARA, SYLVIA C. Projetos e Relatórios de Pesquisa em Administração. 3 ed. São Paulo: Atlas, 2000.

TABLES AND FIGURES

	Total Insatisfaction			Total Satisfaction		Total
	1	2	3	4	5	
1 User's ease to accomplish a task for the first time	0%	0%	30%	30%	39%	100%
2 The user's first impression upon using the system	0%	0%	9%	52%	39%	100%
3 Number of attempts in order to learn how to accomplish a task	0%	13%	17%	48%	22%	100%
4 Time required for learning how to accomplish a task successfully	0%	9%	17%	39%	35%	100%
7 Ease of learning a task	0%	0%	13%	61%	26%	100%
8 Options for different ways to accomplish the same task	17%	35%	30%	13%	4%	100%
9 Gain and productivity regarding the quick way in relation to the standard way provided	4%	22%	35%	39%	0%	100%
10 Flexibility to perform the tasks in different ways	0%	30%	35%	30%	4%	100%
11 The capability of the system to guide through its execution with hints, help, notices etc.	22%	26%	22%	26%	4%	100%
12 Briefness to successfully accomplish a task in the system for the first time	0%	0%	30%	43%	26%	100%

Figure 1 : Sample results for the 'ease of learning' metric

	Total Insatisfaction			Total Satisfaction		Total
	1	2	3	4	5	
13 Remembering how to perform a task after a certain period of time without using the system.	0%	0%	9%	35%	57%	100%
14 Ease of remembering the utilization of the system	0%	0%	4%	52%	43%	100%
15 Agility to remember how to use the system after a period of time without using it	0%	0%	4%	39%	57%	100%

Figure 1 : Sample results for the 'ease of remembering' metric

	Total Insatisfaction			Total Satisfaction		Total
	1	2	3	4	5	
16 The amount of errors caused by the system	0%	17%	22%	26%	35%	100%
17 Time elapsed before the system resumes normal operation when an error occurs	0%	4%	35%	35%	26%	100%
18 The feeling regarding the amount of errors caused by the system	0%	13%	17%	35%	35%	100%
19 The amount of errors caused by the system, resulting in some loss of information	0%	0%	13%	43%	43%	100%
20 The time spent to resume the execution of the task caused by the system	0%	17%	17%	48%	17%	100%
21 Satisfaction regarding recovery from the error by the system, undo, redo, back, save before closing etc.	0%	13%	26%	35%	26%	100%
29 The clarity of the error messages presented by the systems	4%	17%	26%	39%	13%	100%

Figure 2 : Sample results for the 'error control' metric

	Total Insatisfaction			Total Satisfaction		
	1	2	3	4	5	Total
22 System performance	0%	0%	13%	65%	22%	100%
23 The velocity for accomplishing the tasks	0%	0%	13%	61%	26%	100%
24 System productivity	0%	0%	13%	70%	17%	100%
25 Keeping the system under control	0%	9%	13%	52%	26%	100%

Figure 3 : Sample results for the 'efficiency' metric

	Total Insatisfaction			Total Satisfaction		
	1	2	3	4	5	Total
26 The amount of steps required to accomplish a task	0%	0%	13%	61%	26%	100%
27 The time taken to accomplish any task in the system	0%	0%	13%	70%	17%	100%
28 The amount of steps required to accomplish a task in the system	0%	9%	13%	52%	26%	100%

Figure 4 : Sample results for the 'effectiveness' metric

	Total Insatisfaction			Total Satisfaction		
	1	2	3	4	5	Total
5 Interaction with the system's interface	0%	0%	22%	48%	30%	100%
6 Accomplishment of tasks in the system regarding message clarity, error recovery etc.	0%	17%	39%	30%	13%	100%
30 The user's feeling for using the system as a whole	0%	0%	13%	52%	35%	100%



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING

Volume 13 Issue 6 Version 1.0 Year 2013

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals Inc. (USA)

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

Object Serialization Formats and Techniques a Review

By Surbhi & Rama Chawla

Kurukshetra University DIET, India

Abstract - Serialization is a process of converting an object into a stream of data so that it can be easily transmittable over the network or can be continued in a persistent storage location. This storage location can be a physical file, database or Network Stream. This paper concludes some the work that is going on in the field of Object Serialization.

This paper presents Object Serialization Techniques that can be useful for various purposes, including object serialization Minimization which can be used to decrease the size of Serialized data.

Keywords : *object serialization, compression techniques, object oriented design, performance analytics, soap.*

GJCST-C Classification : *D.1.5*



Strictly as per the compliance and regulations of:



Object Serialization Formats and Techniques a Review

Surbhi^α & Rama Chawla^σ

Abstract - Serialization is a process of converting an object into a stream of data so that it can be easily transmittable over the network or can be continued in a persistent storage location. This storage location can be a physical file, database or Network Stream. This paper concludes some the work that is going on in the field of Object Serialization.

This paper presents Object Serialization Techniques that can be useful for various purposes, including object serialization Minimization which can be used to decrease the size of Serialized data.

Keywords : object serialization, compression techniques, object oriented design, performance analytics, soap.

I. INTRODUCTION

Serialization is the process of converting complex objects into stream of bytes for storage. Deserialization is its reverse process that is unpacking stream of bytes to their original form. It is also known as Pickling, the process of creating a serialized representation of object.

The following steps are necessary to do to create a serializable class:

1. Create a custom class with assigned properties.
2. Define the serialization functions.
3. Create a Controller class and instantiate our Custom class.
4. Serialize the object to a named file.
5. De-serialize the values by reading it from the file.

Object serialization has been investigated for many years in the context of many different distributed systems.

II. MOST POPULAR SERIALIZATION FORMATS

There are various data serialization formats available for developers according to choose form, There are also various ways to convert complex objects to sequences of bits. It does not include markup languages used exclusively as document file formats.

- Binary Format Serialization
- XML Format Serialization
- XML-RPC Serialization^[1]

- JSON Serialization^[2]
- YAML[C] Serialization

The following are the basic advantages of serialization: First is to facilitate the transportation of an object through a network and secondly create a clone of an object that can be restored later on.

III. RELATED WORK

In the paper "Object Serialization and De-serialization Using XML"^[3] Inter operability of potentially heterogeneous databases has been an ongoing research issue for a number of years in the database community.

With the trend towards globalization of data location and data access and the consequent requirement for the coexistence of new data stores with legacy systems, the cooperation and data interchange between data repositories has become increasingly important. The emergence of the extensible Markup Language (XML) as a database independent representation for data offers a suitable mechanism for transporting data between repositories.

This paper describes a research activity within a group at CERN (called CMS) towards identifying and implementing database serialization and deserialization methods that can be used to replicate or migrate objects across the network between CERN and worldwide centers using XML to serialize the contents of multiple objects resident in object oriented databases.

The paper "Generic Pickling and Minimization"^[4] presents generic pickling and minimization mechanisms that are provided as services similar to garbage collection. Pickling is used to externalize and internalize data. Minimization means to maximize the sharing in arbitrary data structures. The paper introduces the notion of an abstract store as a formal basis for the algorithms, and analyzes design decisions for the implementation aspects of pickling and minimization. The mechanisms presented here are fully implemented in the Alice programming system.

We presented a generic pickling and minimization mechanism. We showed how Alice, as a conservative extension of Standard ML, uses pickling in a type safe way for its component system. To build a formal base for the algorithms, we introduced abstract stores as a universal memory model. Un-pickling and pickling are based on this model, allowing us to analyze

*Author α σ : Department of Computer Science & Engineering, Kurukshetra University DIET, Karnal, Haryana, India.
E-mails : er.surbhi88@gmail.com, ramachawla27@gmail.com*

and evaluate our design decisions such as bottom up versus top down un-pickling and right to left versus left to right traversal. Minimization can be used to decrease the size of pickled data. However, the general mechanism presented here seems suitable for other applications such as efficient representation of runtime types. Finally, we extended the system with support for concurrency as present in Alice. The authors analyzed how pickler and minimizer must behave in such a concurrent setting.

In the paper “**Why Object Serialization is Inappropriate for Providing Persistence in Java**”^[5] the author paper describes why Object Serialization is not appropriate for providing persistence in Java. With numerous code examples, Object Serialization is shown to be easy to work with initially which seduces the developer into relying on it for persistence within more complex applications.

The advanced use of object serialization requires significant work from the programmer, something that is not apparent at first. The use of object serialization together with static and transient fields and within multithreaded programs is discussed together with the “big inhale problem”: the need to read in the entire object graph before processing over it can commence.

This paper has shown, with numerous supporting examples, that using Java’s object serialization mechanism to provide object persistence is inappropriate. The system appears simple on the surface but there are many implications from relying on it as a persistence technology. The programmer must state the types that are candidates for persistence at compile time, whereas making this decision at runtime, on a per object basis, is more appropriate.

The serialization mechanism suffers from the **big inhale** problem where the whole graph must be read before it can be used; loading objects on demand is more efficient, reducing delay in starting an application. The serialization mechanism creates copies of objects that it writes and reads. This can break some code that makes assumptions about the hash code of an object.

μ s per object	32 int		4 int, 2 null		tree(15)	
	w	r	w	r	w	r
JDK serialization	346	1410	169	596	1192	1889
UKA-serialization	35	39	19	28	201	354
improvement %	90	97	89	95	83	81
explicit marshaling	228	920	81	308	396	647
slim type encoding	19	187	16	159	72	213
internal buffering	0	159	6	19	0	330
buffer accessibility	48	65	30	49	502	291
two types of reset	16	40	17	33	21	54

Figure 1 : Object serialization for various type of objects with encodings^[5]

The complexity of using object serialization within 1a distributed environment, when evolving

classes and when using specialized class loaders is also discussed. The paper compares the performance of serializing and de-serializing a byte array and binary tree of the same data size to and from an NFS mounted disk and two kinds of local disk. Alternative solutions to object persistence in Java are presented at the end of the paper.

Using Experiments carried out by author to draws four conclusions:

1. The absolute amount of time to read and write a store is large;
2. Reading a store is much slower than writing a store; and if an application is likely to exhibit more reading than writing,
3. An NFS mounted disk should be used;
4. The use of JIT technology significantly increases the speed of using Java object serialization.

In the “**Object Serialization in the .NET Framework**”^[6], the author describes using Serialization in .Net framework. He describes the two most important reasons are to persist the state of an object to a storage medium so an exact copy can be recreated at a later stage, and to send the object by value from one application domain to another.

It is also used by remoting to pass objects by value from one application domain to another. This paper provides an overview of the serialization used in the Microsoft .NET Framework.

The author gives Serialization Guidelines, one should consider serialization when designing new classes since a class cannot be made serializable after it has been compiled. Some questions to ask are: Do one have to send this class across application domains? Will this class ever be used with remoting? What will users do with this class? Maybe they derive a new class that needs to be serialized. When in doubt, mark the class as serializable. It is probably better to mark all classes as serializable unless:

- They will never cross an application domain. If serialization is not required and the class needs to cross an application domain, derive the class from Marshal by Ref. Object.
- The class stores special pointers that are only applicable to the current instance of the class. If a class contains unmanaged memory or file handles, for example, ensure these fields are marked as Non-Serialized or don't serialize the class at all.

“**Comparison between JSON and YAML for data serialization**”^[7], this report determines and discusses the primary differences between two different serialization formats, namely YAML and JSON. A general introduction to the concepts of serialization and parsing is provided first, which also explains how they can be used to transfer and store data. This is followed by an analysis of the YAML and JSON formats, where

functionality, primary use cases, and syntax is described. In addition to this the perceived performance of implementations for both formats will also be investigated by conducting a number of tests.

Using the combined background information and results from the tests, conclusions regarding the main differences between the two are then determined and discussed.

As has been concluded, it is clearly very easy to read thanks to the required usage of whitespace and the ability to skip surrounding quotes for strings. YAML also has the advantage of allowing comments in the document. Users can easily read and manipulate the output, which is one of the reasons as to why it's often used for configuration files.

This enables the straightforward definition of strongly-typed objects that match serialized structures, for example existing XML formats. Inheritable translation scopes group sets of object serialization binding definitions, and enable inheritance. The present system supports (compressed) XML for serialization, while future work will develop alternate translation schemes, such as type-length-value and JSON.

Execution times in seconds for Seriali-zation

Method	Simple	Complex
JSON.generate	0.1550s	0.5830s
JSON.pretty_generate	0.1470s	0.6060s
YAML.dump	2.4531s	3.4732s

Execution times in seconds for De-Serialization

Method	Simple	Complex
JSON.parse	0.0440s	0.0790s
YAML.load	0.2750s	0.3360s

Table 2 : JSON VS. YAML Serialization performance ^[7]

The execution times measured for the serialization/deserialization process shows their results, similar to the serialization process, which can be seen in table 2. Both implementations are much faster at generating data structures from a serialized string than doing the opposite. YAML is also slower.

IV. CONCLUSION

The primary design goals for Serialization, to provide a simple and effective data exchange, but also being easy to generate and load. It is widely used and is used natively available in the most common modern programming. Object Serialization as presented here is especially well suited for functional programming languages, where the closure semantics and the ability to serialize code is essential. Also a minimization technique helps reduce Serialization sizes considerably.

V. FUTURE SCOPE

To implement means by which Serialization and Deserialization of Objects can be done using modern formats XML and JSON after adding Compression or Encryption or possibly both to the Object Streams.

In future I also want to see how the Performance of Object Serialization is affected in a Normal CLR Binary VS. Native JIT compiled Binary. The perceived performance of XML or JSON can be determined from a custom benchmarking. A complex data set will also be used to test performance of the implementations for documents with deeper hierarchies.

Another direction for future work may be to modify the semantics of the serialization algorithm to improve performance.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Mark Allman. 2003. An evaluation of XML-RPC. SIGMETRICS Perform. Eval. Rev. 30, 4 (March 2003), 2-11.
2. Audie Sumaray and S. Kami Makki. 2012. A comparison of data serialization formats for optimal efficiency on a mobile platform. In Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication. ACM, New York, NY, USA, Article 48.
3. N. Bhatti & W. Hassan, et al, "Object Serialization and Deserialization Using XML", Tata McGraw-Hill, ADVANCES IN DATA MANAGEMENT, Vol 1, 2000.
4. G. Imre, et al, "A Novel Cost Model of XML Serialization", Science Direct, Electronic Notes in Theoretical Computer Science, vol. 261, Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary, 2010.
5. Guido Tack, Leif Kornstaedt, et al, "Generic Pickling and Minimization", Science Direct, Electronic Notes in Theoretical Computer Science, Programming Systems Lab Saarland University, Saarbrücken, Germany, 2006.
6. Huw Evans, "Why Object Serialization is Inappropriate for Providing Persistence in Java", Department of Computing Science, The University of Glasgow, Glasgow, G12 8RZ, UK.
7. MALIN ERIKSSON, VICTOR HALLBERG, "Comparison between JSON and YAML for data serialization", the School of Computer Science and Engineering Royal Institute of Technology, 2011.
8. Piet Obermeyer and Jonathan Hawkins, "Object Serialization in the .NET Framework", Microsoft Corporation, 2001.
9. Lukasz Opyrchal and Atul Prakash, "Efficient Object Serialization in Java", Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA.



This page is intentionally left blank



A Review of Clone Detection Techniques using Model Semantics

By Yachna Arora & Er. Sarita Choudhary

Abstract - A model clone is a set of similar or identical fragments in a model of the system. Understanding and Identifying model clones are important aspects in software evolution. During the Evolution of the Software product Cloning is often a strategic means for the same. Clone detection techniques play an important role in software evolution research where attributes of the same code entity are observed over multiple versions. To successfully create any method or technique for model clones detection we will have to study all the models defined in UML including internal and External Structure of UML This paper reviews some of the techniques available for the Model Clone Prevention and Detection.

Keywords : *UML, semantic clones, model clone, clone detection, parsing, XML.*

GJCST-C Classification : *D.2.9*



Strictly as per the compliance and regulations of:



A Review of Clone Detection Techniques using Model Semantics

Yachna Arora^α & Er. Sarita Choudhary^σ

Abstract - A model clone is a set of similar or identical fragments in a model of the system. Understanding and Identifying model clones are important aspects in software evolution. During the Evolution of the Software product Cloning is often a strategic means for the same.

Clone detection techniques play an important role in software evolution research where attributes of the same code entity are observed over multiple versions. To successfully create any method or technique for model clones detection we will have to study all the models defined in UML including internal and External Structure of UML. This paper reviews some of the techniques available for the Model Clone Prevention and Detection.

Keywords : UML, semantic clones, model clone, clone detection, parsing, XML.

I. INTRODUCTION

Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of object-oriented software engineering. The Unified Modeling Language includes a set of graphic notation techniques to create visual models of object-oriented software-intensive systems.

The Unified Modeling Language was developed by Grady Booch, Ivar Jacobson and James Rumbaugh at Rational Software in the 1990s.^[1] Unified Modeling Language is used to specify, visualize, modify, construct and document the artifacts of an object-oriented software-intensive system underdevelopment.^[2]

UML combines techniques from data modeling, business modeling, object modeling and component modeling. It can be used with all processes, throughout the software development life cycle and across different implementation technologies.^[3]

a) Definition of a Clone

Software clones are regions of source code which are highly similar; these regions of similarity are called clones, clone classes, or clone pairs. . Cloning is the unnecessary duplication of data whether it is at design level or at coding level.

b) Clone Types

i. Code Clones

(i.e., duplicate fragments of source code) have been identified as a major source of software quality issues. As a consequence, a large body of research has been developed on how to prevent, or spot and

eliminate code clones. The problem with code clones is of course that they are linked only by their similarity, i.e., implicitly rather than explicitly which makes it difficult to detect them.

ii. Model Clones

Model Clones are duplicate fragments of Architecture of Model of the project. It is difficult to formulate the actual definition of a model clone because of the abstract nature of a model. We can however define a model clone as a set of similar or identical fragments in a model.

c) Clone Detection

The copying of code has been studied within software engineering mostly in the area of clone analysis. Software clones are regions of source code which are highly similar; these regions of similarity are called clones, clone classes, or clone pairs.

While there are several reasons why two regions of code may be similar, the majority of the clone analysis literature attributes cloning activity to the intentional copying and duplication of code by programmers; clones may also be attributable to automatically generated code, or the constraints imposed by the use of a particular framework or library.

Cloning is the unnecessary duplication of data whether it is at design level or at coding level.

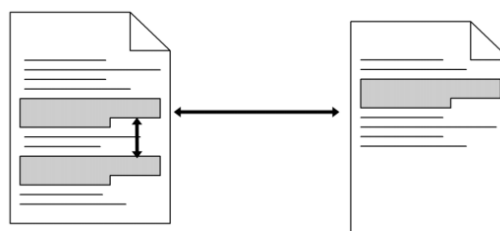


Figure 1 : An Example of Code Clones

It results to excessive maintenance costs as well. So cut paste programming form of software reuse deceivingly raise the number of lines of code without expected reduction in maintenance costs associated with other forms of reuse.

The reasons why programmers duplicate codes include the following reasons:

1. Making a copy of a code fragment is simpler and faster than writing the code from scratch. In addition, the fragment may already be tested so the introduction of a bug seems less likely.

Author^{ασ} : Dept. CSE, DIET.

2. Evaluating the performance of a programmer by the amount of code he or she produces gives a natural incentive for copying code.

Efficiency considerations may make the cost of a procedure call or method invocation seems too high a price. In industrial software development contexts, time pressure together with first and second points lead to plenty of opportunities for code duplication

II. RELATED WORK

Störrle Harald^[4], describes that, Code clones, have been identified as major source of software quality issues. Evidence suggests that this phenomenon occurs similarly in models, suggesting that model clones are as detrimental to model quality as they are to code quality.

However, programming language code and visual models have significant differences that make it difficult to directly transfer notions and algorithms developed in the code clone arena to model clones. In this article, we develop and propose a definition of the notion of "model clone" based on the thorough analysis of practical scenarios.

The Author proposes a formal definition of model clones, specify a clone detection algorithm for UML domain models, and implement it prototypically. The problem with code clones is of course that they are linked only by their similarity, i.e., implicitly rather than explicitly which makes it difficult to detect them.

The paper also discusses that the clones are a substantial problem for code based development, and model clones are increasingly becoming a problem for model-based development. However, currently, there is not much published work on model clones, and next to no work on UML model clones. Therefore, this article started out analyzing actual model clones in UML domain models, and proposed a terminological framework, a pragmatic definition, and a clone classification schema adapted from work on source code clones.

Liliane Jeanne Barbour^[5], describe that, Two identical or similar code fragments form a clone pair. Previous studies have identified cloning as a risky practice. A clone pair experiences many changes during the creation and maintenance of software systems. A change can either maintain or remove the similarity between clones in a clone pair.

If a change maintains the similarity between clones, the clone pair is left in a consistent state. However, if a change makes the clones no longer similar, the clone pair is left in an inconsistent state. The set of states and changes experienced by clone pairs over time form an evolution history known as a clone genealogy.

Specifically, two cases are most risky:

1. When a clone experiences inconsistent changes and then a re-synchronizing change without any modification to the other clone in a clone pair; and

2. When two clones undergo an inconsistent modification followed by a re-synchronizing change that modifies both the clones in a clone pair.

Cloning has been identified by previous researchers as a risky practice in software development and maintenance.

However, software projects have limited resources for reviewing and testing code. Identifying the clones most at risk of faults can help allocate the limited Resources.

Florian Deissenboeck, Benjamin Hummel Elmar Juergens, Michael Pfahler, Bernhard Schaetz^[6], describe that, Cloned code is considered harmful for two reasons:

1. Multiple, possibly unnecessary, duplicates of code increase maintenance costs.
2. Inconsistent changes to cloned code can create faults and, hence, lead to incorrect program behavior.

Likewise, duplicated parts of models are problematic in model-based development. Recently, we and other authors proposed multiple approaches to automatically identify duplicates in graphical models.

While it has been demonstrated that these approaches work in principal, a number of challenges remain for application in industrial practice. Moreover, we present tool support that eases the evaluation of detection results and thereby helps to make clone detection a standard technique in model based quality assurance.

In many application domains for embedded software systems, model-based development-the specification of the functionality of the software using (graphical) models and the automatic generation of production code from these models-is a state-of-the-art technique.

In this paper, techniques have been presented to improve scalability by an adapted subsystem detection, to improve relevance of detected by clones by providing use case specific rankings, and finally tool-support to ease inspection of the instances of the detected clones.

Finally, the detection approach can also be extended to other, less data flow oriented forms of models, e.g., state-oriented models like State Charts or State Flow, or process-oriented models like BPEL or ARIS, however, requiring adapted definitions of similarity as well as means of normalizing models and ranking clones.

Arun Lakhota, Junwei Li, Andrew Walenstein and Yun Yang^[7], describe that, Source code clones are copies or near-copies of other portions of code, often created by copying and pasting portion of source code. This working session is concerned with building a communal research infrastructure for clone detection. The intention of this working session is to try to build a

consensus on how to continue to build a benchmark suite and results archive for clone- and source comparison related research and development. Several automated and semi-automated clone detection techniques have been devised. Clone detection is done as an information retrieval (IR) task. The standard measure for IR techniques can be applied in the form of the detector's precision and recall. In the case of clone detection, "recall" refers the percentage of the clones that are found, and "precision" refers to the percentage of correct results as compared to "false positives" (code falsely reported to be clones). Different clone detectors report clones in different formats.

III. CONCLUSION

Cloning works at the cost of increasing lines of code without adding to overall productivity. Same software bugs and defects are replicated that reoccurs throughout the software at its evolving as well its maintenance phase.

Software clones are important aspects in software evolution. If a system is to be evolved, its clones should be known in order to make consistent changes. Cloning is often a strategic means for evolution.

Clone detection techniques play an important role in software evolution research where attributes of the same code entity are observed over multiple versions.

IV. FUTURE SCOPE

In UML Literature There exists a plenty information regarding Model clones. How Code clones can be identified and removed. However, a development of a system capable of detecting a complete Model is still in research domain.

As we know UML Models have two parallel structures; an external, visual representation as diagrams; and an internal, tree-like structure. The Tree like structure is usually presented in XML like modeling language, because of the free nature of XML format.

In future I would like to achieve a basis for comparisons of UML Domain Models. The process will be done using XML parsing of UML domain Models, with Referenced and Candidate Models.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Marc Hamilton, Software Development: A Guide to Building Reliable Systems, p.48, 1999.
2. FOLDOC (2001). Unified Modeling Language last updated 2002 01-03. Accessed 6 February, 2009.
3. Satish Mishra (1997). "Visual Modeling & Unified Modeling Language (UML): Introduction to UML". Rational Software Corporation. Accessed 9 November 2008.
4. Störrle Harald, "Towards clone detection in UML domain models", Springer-verlag, Softw Syst Model. 18 September 2011.
5. Liliane Jeanne Barbour, "EMPIRICAL STUDIES OF CODE CLONE GENEALOGIES", Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario, Canada. January, 2012.
6. Florian Deissenboeck, Benjamin Hummel Elmar Juergens, Michael Pfaehler, Bernhard Schaetz, "Model Clone Detection in Practice", Technische Universität München Garching b. München, Germany, fortiss GmbH München, Germany.
7. Lakhotia Arun, Li Junwei, Walenstein Andrew and Yang Yun, "Towards a Clone Detection Benchmark Suite and Results Archive", Software Research Laboratory Center for Advanced Computer Science University of Louisiana at Lafayette, IEEE, 2003.

GLOBAL JOURNALS INC. (US) GUIDELINES HANDBOOK 2013

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 GJCST 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 brevity.

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 $1.4 \times 10^{-3} \text{ m}^3$, or 4 mm somewhat than $4 \times 10^{-3} \text{ m}$. 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.

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.



Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish 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 through 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 an 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

A

Antecedent · 27, 28
Asterisk · 41
Autopsy · 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

B

Bayesian · 12, 14, 17, 28, 29
Benjamin · 50, 51

C

Consortium · 25
Cuddapha · 20

K

Kalahasthi · 19

L

Lazarevic · 21, 25
Litigious · 2

M

Markov · 12
Marquardt · 12, 14, 16, 17
Massachusetts · 32

N

Nebulosa · 43

P

Plethora · 19
Provenance · 34, 36
Purging · 10

R

Renigunta · 19

S

Sulawesi · 21, 25
Symposium · 11

T

Traingdx · 17

U

Ubiquitous · 48
Usando · 43

Y

Yerpedu · 19

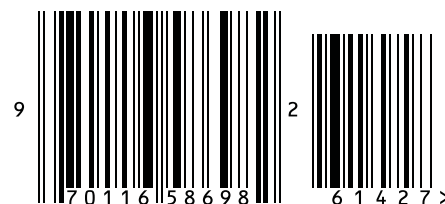


save our planet



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