Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Survey of Existing E-mail Spam

Data in Heterogeneous Databases

Highlights

Regression Algorithms in Python

Accuracy Analysis of Continuance

Discovering Thoughts, Inventing Future

VOLUME 18 ISSUE 2 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 18 Issue 2 (Ver. 1.0)

Open Association of Research Society

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

Global Journals Inc.

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

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals Pvt Ltd
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, 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: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Excluding Air Parcel Charges):

Yearly Subscription (Personal & Institutional)
250 USD (B/W) & 350 USD (Color)

Editorial Board

Global Journal of Computer Science and Technology

Dr. Abdurrahman Arslanyilmaz

Computer Science & Information Systems Department
Youngstown State University
Ph.D., Texas A&M University
University of Missouri, Columbia
Gazi University, Turkey
Web: cis.ysu.edu/~aarslanyilmaz/professional_web

Dr. Chutisant Kerdvibulvech

Dept. of Inf. & Commun. Technol.,
Rangsit University
Pathum Thani, Thailand
Chulalongkorn University Ph.D. Thailand
Keio University, Tokyo, Japan

Dr. Sukhvinder Singh Deora

Ph.D., (Network Security), MSc (Mathematics),

Masters in Computer Applications

Dr. Sotiris Kotsiantis

Ph.D. in Computer Science, University of Patras, Greece
Department of Mathematics, University of Patras, Greece

Dr. Ramadan Elaiess

Ph.D.,

Computer and Information Science

Dr. Manpreet Singh

Ph.D.,
(Computer Science)

Nicla Romano

Professor in Cellular and Developmental Biology;
Cytology and Histology; Morfogenesis and Comparative Anatomy

Dr. Muhammad Abid

M.Phil,
Ph.D Thesis submitted and waiting for defense

Dr. K. Venkata Subba Reddy

Ph.D in Computer Science and Engineering

Loc Nguyen

Postdoctoral degree in Computer Science

Faisal Mubuke

M.Sc (IT), Bachelor of Business Computing, Diploma in Finanicial services and Business Computing

Jiayi Liu

Physics, Machine Learning,
Big Data Systems

Dr. Yuanyang Zhang

Ph.D in Computer Science

Asim Gokhan Yetgin

Design, Modelling and Simulation of Electrical Machinery;
Finite Element Method, Energy Saving, Optimization

Anup Badhe

Bachelor of Engineering (Computer Science)

Dr. S. Nagaprasad

M.Sc, M. Tech, Ph.D

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

The Method of Normalizing OWL 2 DL Ontologies

By Malgorzata Sadowska & Zbigniew Huzar

Wroclaw University of Science and Technology

Abstract- The paper proposes a method of normalizing OWL 2 DL ontologies. The method introduces rules aimed at refactoring OWL 2 constructs. The proposed transformations only use a subset of OWL 2 constructs and enable to present an input OWL 2 ontology in a new but semantically equivalent form. The normalization is motivated by the fact that normalized OWL 2 DL ontologies have a unified structure of axioms so that they can be compared in an algorithmic way.

Keywords: *OWL 2, normalization.*

GJCST-C Classification: *H.3.5*

Strictly as per the compliance and regulations of:

The Method of Normalizing OWL 2 DL Ontologies

Małgorzata Sadowska [α] & Zbigniew Huzar [σ]

Abstract - The paper proposes a method of normalizing OWL 2 DL ontologies. The method introduces rules aimed at refactoring OWL 2 constructs. The proposed transformations only use a subset of OWL 2 constructs and enable to present an input OWL 2 ontology in a new but semantically equivalent form. The normalization is motivated by the fact that normalized OWL 2 DL ontologies have a unified structure of axioms so that they can be compared in an algorithmic way.

Keywords: OWL 2, normalization.

I. Introduction

OWL 2 Web Ontology Language (OWL 2) [1] is a language of knowledge representation used for defining ontologies. The ontologies which satisfy syntactic conditions listed in the specification in Section 3 of [2] are called OWL 2 DL ontologies and have their semantics expressed in *SROIQ* description logic [1]. *SROIQ* was designed to provide additions to OWL-DL to guarantee decidability in reasoning [3, 4].

The domain ontologies are expected to provide a knowledge base about specific application area. Therefore they should be consistent. An ontology consistency check [5] is one of the reasoning problems that can be answered with the use of inference engines.

In our work, we would like to take advantage of and reuse the existing OWL 2 DL domain ontologies. We assume that the selected OWL 2 DL domain ontology is syntactically correct, consistent and adequately describes the notions from the needed domain. One can successfully conduct reasoning over the ontology with the use of one of the reasoning engines available. However, it is not obvious or conclusive how to effectively process other useful operations on ontologies, for example how to compare two or more ontologies. The problem of comparing two ontologies with the agreed vocabulary was faced in [6] in a method of semantic validation of UML class diagrams. In [6], in the beginning, the UML class diagram is transformed into an ontology expressed in OWL 2. Next, the two ontologies-the domain ontology and the ontological representation of the UML diagram-need to be compared against each other.

The question arises: how to correctly and automatically find out whether one ontology is compliant or contradictory concerning another one ? For the purpose of answering the question, we introduce such a form of normalization that allows for unifying the structure of axioms in the ontologies so that it is possible to automatically compare them. The method of normalization of ontologies and the method of semantic validation of UML class diagrams has been implemented in the prototype tool [8]. This paper presents the details of conducting the transformation of OWL 2 ontology to its normalized form. The important fact is that the presented transformations only change the structure but do not affect the semantics of axioms or expressions in the OWL 2 ontology.

We propose the following groups of transformations of OWL 2 constructs:

Group I. Extraction of declarations of entities: A declaration in OWL 2 associates an *Entity* with its type. If a declaration axiom for the selected *Entity* is missing from the ontology, it can be retrieved based on the usage of the *Entity*. In OWL 2, the declaration axiom can be specified for all types of entities: *Class*, *Datatype*, *ObjectProperty*, *DataProperty*, *AnnotationProperty* and *NamedIndividual*.

Group II. Removal of duplicates in data ranges, expressions, and axioms: Following [2], sets written in one of the exchange syntaxes (e.g., XML or RDF/XML) may contain duplicates. Therefore, duplicates (if applicable) are eliminated from axioms (e.g. *EquivalentClasses*), data ranges (e.g. *DataUnionOf*) and expressions (e.g. *DataUnionOf*).

Group III. Restructuration of data ranges and expressions: The proposed restructurations are intended (1) to flatten the nested structures of the data ranges and expressions, (2) to eliminate the weakest cardinality restrictions included in the data ranges or expressions, and (3) to refactor the data ranges and expressions which are connected with union, intersection and complement constructors, based on the rules of the De Morgan's laws.

Group IV. Removal of syntactic sugar in axioms: OWL 2 offers the so-called syntactic sugar [4]. The syntactic sugar makes some axioms easier to write and read for humans (e.g., *DisjointUnion* axiom) but does not lend itself so easily to processing conducted by tools. Due to this fact, the removal of syntactic sugar allows e.g., easier comparison of axioms expressing the same semantics but written with a different syntax.

Group V. Restructuration of axioms: Most of OWL 2 axioms which contain several class expressions can be

Author α σ: Faculty of Computer Science and Management, Wroclaw University of Science and Technology.
e-mails: m.sadowska, zbigniew.huzar@pwr.edu.pl

restructured into several axioms containing only two class expressions each, e.g., *DisjointClasses* and *EquivalentClasses* axioms. This is only worth to be applied for axioms whose order of internal expressions is not important.

Group VI. Removal of duplicated axioms: A correctly specified OWL 2 ontology cannot contain two identical axioms (see Section 3). However, duplicated axioms may appear as a result of applying transformations from groups IV and V. Therefore, the last step of the normalization algorithm is to remove all duplicate axioms from the output ontology.

We define *ontology normalization* as a process of transforming the input ontology into the ontology in its refactored form. In Section 4, we presented replacing and replaced OWL 2 constructs used in the process of normalizing OWL 2 DL ontologies. The details of the ontology normalization algorithm are presented in Section 5. The process consists of four phases, which are executed in the following order in the algorithm:

1. Extraction of declarations (group I).
2. Refactorization of data ranges and expressions through applying transformations from group II.
3. Restructuration of expressions and data ranges through applying transformations from group III.
4. Refactorization of axioms through applying transformations from groups II, IV, V and VI.

We consider the output ontology (obtained as a result of conducting the algorithm) as *normalized.* Because all transformations (of the replaced OWL 2 constructs to the replacing OWL 2 constructs) preserve semantics, the semantics of the normalized ontology is the same as the semantics of the input ontology.

In this paper, OWL 2 constructs are written with the use of functional-style syntax [2]. Additionally, the following convention is used:

- C – indicates a class,
- CE (possibly with an index) indicates a class expression,
- OP – indicates an object property,
- OPE (possibly with an index)– indicates an object property expression,
- DP – indicates a data property,
- DPE (possibly with an index) indicates a data property expression,
- DR – indicates a data range,
- a – indicates an individual,
- lt – indicates a literal,
- $\alpha = \beta$ – means the textual identity of α and β OWL 2 constructs.

II. Related Work

According to our investigation, a similar concept of normalization of OWL 2 ontologies has not yet been proposed. In this paper, the normalization is aimed at unifying the structure of axioms in the ontologies allowing for automatic processing of the ontologies. A different purpose (as well as a different kind of) ontology normalization has been proposed in [9], where ontology normalization was suggested to be a pre-processing step that aligns structural metrics with intended semantic measures. Additionally, in [10] and [11], normalization has been proposed as an aspect of ontology design that provides support for ontology reuse, maintainability, and evolution. In [10] and [11] the criteria for normalization are focused on engineering issues that make ontologies modular and understandable for knowledge engineers.

III. OWL 2 Ontology as a Set of Axioms

The structural specification of OWL 2 [2] is defined with the use of Unified Modeling Language (UML) [7], and the notation is compatible with Meta-Object Facility (MOF) [12]. OWL 2 ontologies consist of entities (classes, datatypes, object properties, data properties, annotation properties and named individuals), expressions (class expressions, data and object property expressions) and axioms (e.g., subclass axioms).

The main component of OWL 2 ontologies is axioms (see fig.1) which specify what is true in a specific domain. The expressions are used to represent complex notions in the described domain. Textually, expressions can be seen as components of axioms, for example, two or more class expressions are needed to specify *DisjointClasses* axiom (see fig. 2). Finally, entities constitute the vocabulary of an ontology.

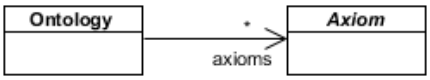


Fig.1: A relation between OWL 2 ontology and axioms (extract from Figure 1 of OWL 2 specification [2]).

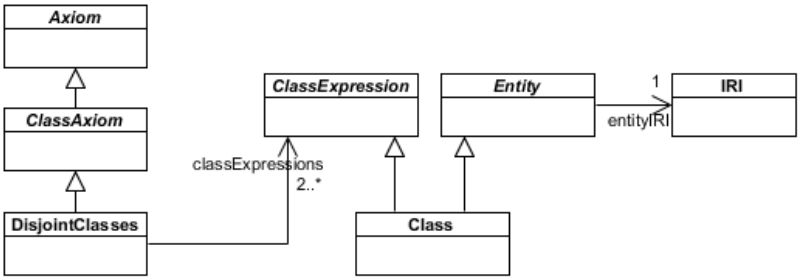


Fig.2: A relation between the selected class axiom, relevant expressions and entities (by OWL 2 specification [2]).

Because the association end named *axioms* (see fig. 1) is specified with the use of UML *Multiplicity Element* and a *Set* collection type (*is Ordered*=false and *is Unique*=true), a correct OWL 2 ontology cannot contain two axioms that are textually equivalent. In the normalization method, it is assured through applying the transformations from group VI.

Nevertheless, the ontology may have axioms which contain the same information. For example, it may include the following two axioms: *DisjointUnion* (:*Child* :*Boy* :*Girl*) and *DisjointClasses*(:*Boy*: *Girl*). The semantics of DisjointUnion [2] states that *Child* class is a disjoint union of *Boy* and *Girl* class expressions which are pairwise disjoint. Therefore, the additional information specified by *DisjointClasses* can be seen as redundant and will be refactored with the proposed transformation rules (here from groups II and IV).

The structural specification of OWL 2 [2] defines an abstract class *Axiom* (see fig. 3). The abstract class *Axiom* is specialized by the following classes: *ClassAxiom* (abstract), *ObjectPropertyAxiom* (abstract), *DataPropertyAxiom* (abstract), *Declaration*, *Datatype Definition* (abstract), *HasKey*, *Assertion* (abstract) and *AnnotationAxiom* (abstract).

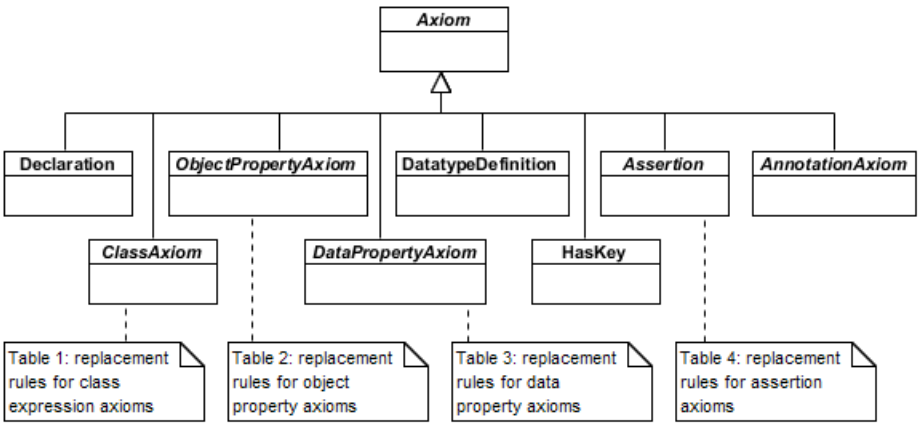


Fig. 3: The axioms of OWL 2 [2]and the tables which specify the proposed replacement rules.

AnnotationAxiom [2] axioms are mainly used to improve readability for humans. The axioms do not affect the semantics [2]. Therefore, they are not further restructured in this paper.

Declaration [2] axioms specify that entities are part of the vocabulary in ontology and are of a specific type, e.g., class, datatype, etc. OWL 2 DL ontology must explicitly declare all datatypes that occur in datatype definition, although in general, it is advisable to declare all entities for verification of the correctness of the usage of the entity based on its type. In the normalization method, if a declaration axiom is missing from the ontology, it is automatically retrieved based on the entity usage (transformation from the group I). This is applied toall types of entities but *AnnotationProperty*, because *AnnotationProperty* is only used to provide annotation and has no effect on the semantics.

Datatype Definition [2] axiom defines a new datatype as being semantically equivalent to a unary data range. The *Datatype Definition* axiom is defined in the form that does not need to be restructured. Nonetheless, the data ranges included in other axioms or expressions may require refactoring (transformation from group III). The replacement rules for data ranges are presented in Table 5.

HasKey[2] axiom states that each named instance of the specified class expression is uniquely identified by the specified object property and/or data property expressions. It is useful in querying about individuals which are uniquely identified. The *HasKey*

axiom itself is defined in the form that does not need to be restructured, but the class expression and object property expressions included in the axiom are restructured by the transformations presented in Tables 6 and 7 (transformation from group III).

To summarize, Section 4 presents replacement rules for *Class Axioms* in Table 1, for *Object Property Axioms* in Table 2, for *Data Property Axioms* in Table 3 and *Assertion* axioms in Table 4, Table 5 presents replacement rules for data ranges, Table 6 - for class expressions and Table 7 for object property expressions. Each row in Tables 1-7 contains the number of the transformation group (by Groups I-VI defined in Introduction). Additionally, all the transformations from Group III are marked with the sub-number (1)-(3) which defines a concrete refactorization within the group.

IV. OWL 2 Construct Replacements

Each replaced OWL 2 construct is semantically equivalent to the defined replacing construct(s). Most of the proposed transformations are our original proposals, but some of them come from the OWL 2 specification [2]. The specification defines the so-called syntactic sugar for selected axioms in more detail. This is for ease of writing of some popular patterns for humans. It is cited, where applicable, separately in each table.

a) Class expression axioms

OWL 2 class expression axioms define the relationships between class expressions. The abstract class *ClassAxiom* is specified by the following concrete classes: *SubClassOf*, *EquivalentClasses*, *DisjointClasses* and *DisjointUnion*. In Table 1, transformations of IDs: 3, 6 and 8 are defined in [2], the other transformations are proposed by us. The replacing axioms in ID 6 are semantically equivalent.

Table 1: Replaced and replacing class expression axioms.

ID	Group	Replaced axiom	Replacing axiom(s)
1	II	EquivalentClasses(CE_1 ... CE_i ... CE_j ... CE_N) and $1 \le i \le j \le N$ and $N \ge 3$ and $CE_i = CE_j$	EquivalentClasses(CE_1 ... CE_i ... CE_N) and $1 \le i \le N$ and $N \ge 2$
2	V	EquivalentClasses(CE_1 ... CE_N) and $1 \le i \le N$ and $N \ge 2$	EquivalentClasses (CE_i CE_j) and $i,j \in \{1,N\}$ and $i \ne j$ and $N \ge 2$
3	IV	EquivalentClasses(CE_1 CE_2)	SubClassOf(CE_1 CE_2) SubClassOf(CE_2 CE_1)
4	II	DisjointClasses(CE_1 ... CE_i... CE_j ... CE_N) and $1 \le i \le j \le N$ and $N \ge 3$ and $CE_i = CE_j$	DisjointClasses(CE_1 ... CE_i ... CE_N) and $1 \le i \le N$ and $N \ge 2$
5	V	DisjointClasses(CE_1 ... CE_N) and $N \ge 2$	DisjointClasses(CE_i CE_j) and $i,j \in \{1,N\}$ and $i \ne j$ and $N \ge 2$
6	IV	DisjointClasses(CE_1 CE_2)	SubClassOf (CE_1 ObjectComplementOf(CE_2)) SubClassOf (CE_2 ObjectComplementOf(CE_1))
7	II	DisjointUnion(C CE_1 ... CE_i ... CE_j ... CE_N) and $1 \le i \le j \le N$ and $N \ge 3$ and $CE_i = CE_j$	DisjointUnion(C CE_1 ... CE_i ... CE_N) and $1 \le i \le N$ and $N \ge 2$
8	IV	DisjointUnion(C CE_1 ... CE_N) and $N \ge 2$	EquivalentClasses(C ObjectUnionOf (CE_1 ... CE_N)) DisjointClasses(CE_1 ... CE_N) and $N \ge 2$

b) Object property axioms

OWL 2 object property axioms define the relationships between property expressions. The abstract class *ObjectPropertyAxiom* is specified by the following concrete classes: *SubObjectPropertyOf*, *EquivalentObjectProperties*, *Disjoint Object Properties*, *InverseObjectProperties*, *ObjectPropertyDomain*, *Object PropertyRange*, *ReflexiveObjectProperty*, *Irreflexive ObjectProperty*, *FunctionalObjectProperty*, *Inverse FunctionalObjectProperty*, *SymmetricObjectProperty*, *AsymmetricObjectProperty* and *TransitiveObjectProperty*. In Table 2, transformations of IDs: 3 and 6-14 are defined in [2], the other transformations are proposed by us. The replacing axioms in ID 6 are semantically equivalent.

Table 2: The replaced and replacing object property axioms.

ID	Group	Replaced axiom	Replacing axiom(s)
1	II	EquivalentObjectProperties(OPE_1 ... OPE_i ... OPE_j ... OPE_N) and 1 ≤ i ≤ j ≤ N and N ≥ 3 and $OPE_i = OPE_j$	EquivalentObjectProperties(OPE_1 ... OPE_i ... OPE_N) and 1 ≤ i ≤ N and N ≥ 2
2	V	EquivalentObjectProperties(OPE_1 ... OPE_N) and 1 ≤ i ≤ N and N ≥ 2	EquivalentObjectProperties(OPE_i OPE_j) and i,j ∈ {1,N} and i ≠ j and N ≥ 2
3	IV	EquivalentObjectProperties(OPE_1 OPE_2)	SubObjectPropertyOf(OPE_1 OPE_2) SubObjectPropertyOf(OPE_2 OPE_1)
4	II	DisjointObjectProperties(OPE_1 ... OPE_i... OPE_j ... OPE_N) and 1 ≤ i ≤ j ≤ N and N ≥ 3 and $OPE_i = OPE_j$	DisjointObjectProperties(OPE_1 ... OPE_i ... OPE_N) and 1 ≤ i ≤ N and N ≥ 2
5	V	DisjointObjectProperties(OPE_1 ... OPE_N) and 1 ≤ i ≤ N and N ≥ 2	DisjointObjectProperties(OPE_i OPE_j) and i,j ∈ {1,N} and i ≠ j and N ≥ 2
6	IV	InverseObjectProperties(OPE_1 OPE_2)	EquivalentObjectProperties(OPE_1 ObjectInverseOf(OPE_2)) EquivalentObjectProperties(OPE_2 ObjectInverseOf(OPE_1))
7	IV	ObjectPropertyDomain(OPE CE)	SubClassOf (Object Some Values From(OPE owl:Thing) CE)
8	IV	ObjectPropertyRange(OPE CE)	SubClassOf(owl:Thing ObjectAllValuesFrom(OPE CE))
9	IV	FunctionalObjectProperty(OPE)	SubClassOf(owl:Thing ObjectMaxCardinality(1 OPE))
10	IV	InverseFunctionalObjectProperty(OPE)	SubClassOf (owl:Thing ObjectMaxCardinality(1 ObjectInverseOf(OPE)))
11	IV	ReflexiveObjectProperty(OPE)	SubClassOf(owl:Thing ObjectHasSelf(OPE))
12	IV	IrreflexiveObjectProperty(OPE)	SubClassOf(ObjectHasSelf(OPE) owl:Nothing)
13	IV	SymmetricObjectProperty(OPE)	SubObjectPropertyOf(OPE ObjectInverseOf(OPE))
14	IV	TransitiveObjectProperty(OPE)	SubObjectPropertyOf (ObjectPropertyChain(OPE OPE) OPE))

c) *Data property axioms*

OWL 2 data property axioms define the relationships between property expressions. The abstract class *DataProperty Axiom* is specified by the following concrete classes: *SubDataProperty Of*, *EquivalentDataProperties*, *DisjointDataProperties*, *Data PropertyDomain*, *DataPropertyRange*, and *Functional DataProperty*.In Table 3, transformations of IDs: 3 and 6-8 are defined in [2], the remaining transformations are proposed by us.

Table 3: The replaced and replacing data properties axioms.

ID	Group	Replaced axiom	Replacing axiom(s)
1	II	EquivalentDataProperties (DPE_1 ... DPE_i... DPE_j ... DPE_N) and 1 ≤ i ≤ j ≤ N andN ≥ 3 and $DPE_i = DPE_j$	EquivalentDataProperties (DPE_1 ... DPE_i ... DPE_N) and 1 ≤ i ≤ N and N ≥ 2
2	V	EquivalentDataProperties(DPE_1 ... DPE_N) and 1 ≤ i ≤ N and N ≥ 2	EquivalentDataProperties(DPE_i DPE_j) and i,j ∈ {1,N} and i ≠ j and N ≥ 2
3	IV	EquivalentDataProperties(DPE_1 DPE_2)	SubDataPropertyOf(DPE_1 DPE_2) SubDataPropertyOf(DPE_2 DPE_1)
4	II	DisjointDataProperties(DPE_1 ... DPE_i... DPE_j ... DPE_N) and 1 ≤ i ≤ j ≤ N and N ≥ 3 and $DPE_i = DPE_j$	DisjointDataProperties (DPE_1 ... DPE_i ... DPE_N) and 1 ≤ i ≤ N and N ≥ 2
5	V	DisjointDataProperties(DPE_1 ... DPE_N) and 1 ≤ i ≤ N and N ≥ 2	DisjointDataProperties(DPE_i DPE_j) and i,j ∈ {1,N} and i ≠ j and N ≥ 2
6	IV	DataPropertyDomain(DPE CE)	SubClassOf(DataSomeValuesFrom (DPE rdfs:Literal) CE)
7	IV	DataPropertyRange(DPE DR)	SubClassOf(owl:Thing DataAllValuesFrom(DPE DR))
8	IV	FunctionalDataProperty(DPE)	SubClassOf(owl:Thing DataMaxCardinality(1 DPE))

d) Assertion axioms

OWL 2 *Assertion*[2] axioms are used to state facts about individuals. Following structural specification [2], the abstract class *Assertion* is specified by the following concrete classes: *SameIndividual*, *DifferentIndividuals*, *ClassAssertion*, *ObjectPropertyAssertion*, *NegativeObjectPropertyAssertion*, *DataPropertyAssertion*, *NegativeDataPropertyAssertion*. In Table 4, all transformations are proposed by us.

Table 4: The replaced and replacing assertion axioms.

ID	Group	Replaced axiom	Replacing axiom(s)
1	II	SameIndividual($a_1 \dots a_i \dots a_j \dots a_N$) and $1 \le i \le j \le N$ and $N \ge 3$ and $a_i = a_j$	SameIndividual($a_1 \dots a_i \dots a_N$) and $1 \le i \le N$ and $N \ge 2$
2	V	SameIndividual($a_1 \dots a_N$) and $1 \le i \le N$ and $N \ge 2$	SameIndividual($a_i\ a_j$) and $i,j \in \{1,N\}$ and $i \ne j$ and $N \ge 2$
3	II	DifferentIndividuals($a_1 \dots a_i \dots a_j \dots a_N$) and $1 \le i \le j \le N$ and $N \ge 3$ and $a_i = a_j$	DifferentIndividuals($a_1 \dots a_i \dots a_N$) and $1 \le i \le N$ and $N \ge 2$
4	V	DifferentIndividuals($a_1 \dots a_N$) and $1 \le i \le N$ and $N \ge 2$	DifferentIndividuals($a_i\ a_j$) and $i,j \in \{1,N\}$ and $i \ne j$ and $N \ge 2$

e) Data ranges

OWL 2 data ranges [2] are used in restrictions on data properties. The abstract class *DataRange* is specified by the following concrete classes: *DataComplementOf*, *DataUnionOf*, *DataOneOf*, *Datatype*, *DatatypeRestriction* and *DataIntersectionOf*. In Table 5, all transformations are our own proposals.

Table 5: The replaced and replacing data ranges.

ID	Group	Replaced data range	Replacing data range(s)
1	III (3)	DataComplementOf (DataComplementOf(DR))	DR
2	II	DataUnionOf($DR_1 \dots DR_i \dots DR_j \dots DR_N$) and $1 \le i \le j \le N$ and $N \ge 3$ and $DR_j = DR_i$	DataUnionOf($DR_1 \dots DR_i \dots DR_N$) and $1 \le i \le N$ and $N \ge 2$
3	III (1)	DataUnionOf (DataUnionOf($DR_1 \dots DR_{Ai} \dots DR_{AN}$) ... $DR_{Bj} \dots DR_{BM}$)) and $1 \le i \le N$ and $N \ge 2$ and $1 \le j \le M$ and $M \ge 2$	DataUnionOf ($DR_1 \dots DR_{Ai} \dots DR_{AN} \dots DR_{Bj} \dots DR_{BM}$)) and $1 \le i \le N$ and $N \ge 2$ and $1 \le j \le M$ and $M \ge 2$
4	II	DataIntersectionOf ($DR_1 \dots DR_i \dots DR_j \dots DR_N$) and $1 \le i \le j \le N$ and $N \ge 3$ and $DR_j = DR_i$	DataIntersectionOf($DR_1 \dots DR_i \dots DR_N$) and $1 \le i \le N$ and $N \ge 2$
5	III (1)	DataIntersectionOf (DataIntersectionOf($DR_1 \dots DR_{Ai} \dots DR_{AN}$) ... $DR_{Bj} \dots DR_{BM}$)) and $1 \le i \le N$ and $N \ge 2$ and $1 \le j \le M$ and $M \ge 2$	DataIntersectionOf ($DR_1 \dots DR_{Ai} \dots DR_{AN} \dots DR_{Bj} \dots DR_{BM}$)) and $1 \le i \le N$ and $N \ge 2$ and $1 \le j \le M$ and $M \ge 2$
6	III (3)	DataIntersectionOf (DataComplementOf(DR_1) ... DataComplementOf(DR_N)) and $1 \le i \le N$ and $N \ge 2$	DataComplementOf (DataUnionOf($DR_1 \dots DR_N$)) and $1 \le i \le N$ and $N \ge 2$
7	III (3)	DataUnionOf (DataComplementOf(DR_1) ... DataComplementOf(DR_N)) and $1 \le i \le N$ and $N \ge 2$	DataComplementOf (DataIntersectionOf($DR_1 \dots DR_N$)) and $1 \le i \le N$ and $N \ge 2$
8	II	DataOneOf($lt_1 \dots lt_i\ lt_j \dots lt_N$) and $1 \le i \le j \le N$ and $N \ge 1$ and $lt_j = lt_i$	DataOneOf($lt_1 \dots lt_i \dots lt_N$) and $1 \le i \le N$ and $N \ge 1$

f) Class expressions

OWL 2 class expressions are constructed of classes and properties. In structural specification [2] class expressions are represented by the *ClassExpression* abstract class. The abstract class *ClassExpression* is specified by the following concrete classes: *Class*, *ObjectIntersectionOf*, *ObjectUnionOf*, *ObjectComplementOf*, *ObjectOneOf*, *ObjectSomeValuesFrom*, *ObjectAllValuesFrom*, *ObjectHasValue*, *ObjectHasSelf*, *ObjectMinCardinality*, *ObjectMaxCardinality*, *ObjectExactCardinality*, *DataSomeValuesFrom*, *DataAllValuesFrom*, *DataHasValue*, *DataMinCardinality*, *DataMaxCardinality* and *DataExactCardinality*. In Table 6, the transformations of IDs: 9-14 and 19 are defined in [2], the other transformations are our proposal.

We exclude two general cases from further considerations - those of the existential and universal class expressions:

- *DataSomeValuesFrom(DPE_1 ... DPE_N DR)*, where N ≥ 2 and
- *DataAllValuesFrom(DPE_1 ... DPE_N DR)*, where N ≥ 2.

The reason is that in both class expressions the data range DR arity *MUST* be N (N≥ 2). However, the current version of OWL 2 specification [2] does not provide any constructor, which may be used to define data ranges of arity more than one (see Section 7 of [2]). If a future version of the specification provides such a constructor, one can consider removal of duplicates and further restructuration of the mentioned class expressions.

Table 6: The replaced and replacing class expressions.

ID	Group	Replaced class expression	Replacing class expression(s)
1	III (3)	ObjectComplementOf (ObjectComplementOf(CE))	CE
2	II	ObjectUnionOf(CE_1 ... CE_i ... CE_j ... CE_N) and 1 ≤ i ≤ j ≤ N and N ≥ 3 and CE_i = CE_j	ObjectUnionOf(CE_1 ... CE_i ... CE_N) and 1 ≤ i ≤ N and N ≥ 2
3	III (1)	ObjectUnionOf (ObjectUnionOf(CE_1 ... CE_{Ai} ... CE_{AN}) ... CE_{Bj} ... CE_{BM})) and 1 ≤ i ≤ N and N ≥ 2 and 1 ≤ j ≤ M and M ≥ 2	ObjectUnionOf (CE_1 ... CE_{Ai} ... CE_{AN} ... CE_{Bj} ... CE_{BM})) and 1 ≤ i ≤ N and N ≥ 2 and 1 ≤ j ≤ M and M ≥ 2
4	II	ObjectIntersectionOf (CE_1 ... CE_i... CE_j ... CE_N) and 1 ≤ i ≤ j ≤ N and N ≥ 3 and CE_i = CE_j	ObjectIntersectionOf(CE_1 ... CE_i ... CE_N) and 1 ≤ i ≤ N and N ≥ 2
5	III (1)	ObjectIntersectionOf (ObjectIntersectionOf) (CE_1 ... CE_{Ai} ... CE_{AN}) ... CE_{Bj} ... CE_{BM})) and 1 ≤ i ≤ N and N ≥ 2 and 1 ≤ j ≤ M and M ≥ 2	ObjectIntersectionOf (CE_1 ... CE_{Ai} ... CE_{AN} ... CE_{Bj} ... CE_{BM})) and 1 ≤ i ≤ N and N ≥ 2 and 1 ≤ j ≤ M and M ≥ 2
6	III (3)	ObjectIntersectionOf (ObjectComplementOf(CE_1) ... ObjectComplementOf(CE_N)) and 1 ≤ i ≤ N and N ≥ 2	ObjectComplementOf (ObjectUnionOf(CE_1 ... CE_N)) and 1 ≤ i ≤ N and N ≥ 2
7	III (3)	ObjectUnionOf (ObjectComplementOf(CE_1) ... ObjectComplementOf(CE_N)) and 1 ≤ i ≤ N and N ≥ 2	ObjectComplementOf (ObjectIntersectionOf(CE_1 ... CE_N)) and 1 ≤ i ≤ N and N ≥ 2
8	II	ObjectOneOf(a_1 ... a_i... a_j ... a_N) and 1 ≤ i ≤ j ≤ N and N ≥ 1 and a_i = a_j	ObjectOneOf(a_1 ... a_i ... a_N) and 1 ≤ i ≤ N and N ≥ 1
9	IV	ObjectSomeValuesFrom(OPE CE)	ObjectMinCardinality(1 OPE CE)
10	IV	ObjectAllValuesFrom(OPE CE)	ObjectMaxCardinality (0 OPE ObjectComplementOf(CE))
11	IV	ObjectHasValue(OPE a)	ObjectSomeValuesFrom (OPE ObjectOneOf(a))
12	IV	DataSomeValuesFrom(DPE DR)	DataMinCardinality(1 DPE DR)
13	IV	DataAllValuesFrom(DPE DR)	DataMaxCardinality (0 DPE DataComplementOf(DR))
14	IV	DataHasValue(DPE lt)	DataSomeValuesFrom (DPE DataOneOf(lt))
15	III (2)	ObjectUnionOf (ObjectMinCardinality(n_1 OPE CE) (ObjectMinCardinality(n_2 OPE CE) CE_i... CE_N) and 1 ≤ i ≤ N and N ≥ 3 and n_1≥ 0 and n_2≥ 0 and n_1 ≤ n_2	ObjectUnionOf (ObjectMinCardinality(n_1 OPE CE) CE_i... CE_N) and 1 ≤ i ≤ N and N ≥ 2 and n_1≥ 0
16	III (2)	ObjectIntersectionOf (ObjectMinCardinality(n_1 OPE CE) ObjectMinCardinality(n_2 OPE CE) CE_i... CE_N) and 1 ≤ i ≤ N and N ≥ 3 and n_1≥ 0 and n_2≥ 0 and n_1 ≤ n_2	ObjectIntersectionOf (ObjectMinCardinality(n_2 OPE CE) CE_i... CE_N) and 1 ≤ i ≤ N and N ≥ 2 and n_2≥ 0
17	III (2)	ObjectUnionOf (ObjectMaxCardinality(m_1 OPE CE)	ObjectUnionOf (ObjectMaxCardinality(m_2 OPE CE)

		ObjectMaxCardinality(m_2 OPE CE) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $m_1 \geq 0$ and $m_2 \geq 0$ and $m_1 \leq m_2$	CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $m_2 \geq 0$
18	III (2)	ObjectIntersectionOf (ObjectMaxCardinality(m_1 OPE CE) ObjectMaxCardinality(m_2 OPE CE) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $m_1 \geq 0$ and $m_2 \geq 0$ and $m_1 \leq m_2$	ObjectIntersectionOf (ObjectMaxCardinality(m_1 OPE CE) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $m_1 \geq 0$
19	IV	ObjectExactCardinality(n OPE CE) and $n \geq 0$	ObjectIntersectionOf (ObjectMinCardinality(n OPE CE) ObjectMaxCardinality(n OPE CE))
20	III (2)	ObjectUnionOf (DataMinCardinality(n_1 DPE DR) DataMinCardinality(n_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $n_1 \leq n_2$ and $n_1 \geq 0$ and $n_2 \geq 0$	ObjectUnionOf (DataMinCardinality(n_1 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $n_1 \geq 0$
21	III (2)	ObjectIntersectionOf (DataMinCardinality(n_1 DPE DR) DataMinCardinality(n_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $n_1 \geq 0$ and $n_2 \geq 0$ and $n_1 \leq n_2$	ObjectIntersectionOf (DataMinCardinality(n_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $n_2 \geq 0$
22	III (2)	ObjectUnionOf (DataMaxCardinality(m_1 DPE DR) DataMaxCardinality(m_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $m_1 \geq 0$ and $m_2 \geq 0$ and $m_1 \leq m_2$	ObjectUnionOf (DataMaxCardinality(m_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $m_2 \geq 0$
23	III (2)	ObjectIntersectionOf (DataMaxCardinality(m_1 DPE DR) DataMaxCardinality(m_2 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 3$ and $m_1 \geq 0$ and $m_2 \geq 0$ and $m_1 \leq m_2$	ObjectIntersectionOf (DataMaxCardinality(m_1 DPE DR) CE_i... CE_N) and $1 \leq i \leq N$ and $N \geq 2$ and $m_1 \geq 0$
24	IV	DataExactCardinality(n DPE DR) and $n \geq 0$	ObjectIntersectionOf (DataMinCardinality(n DPE DR) (DataMaxCardinality(n DPE DR))

g) Object property expressions

The following OWL 2 structural specification [2] object property expressions are represented by *ObjectPropertyExpression* abstract class. The abstract class *ObjectPropertyExpression* is specified by the following concrete classes: *ObjectProperty* and *InverseObjectProperty*. In Table 7, the transformation is our proposal.

Table 7: The replaced and replacing object property expressions.

ID	Group	Replaced object property expression	Replacing object property expression
1	III (3)	ObjectInverse of(ObjectInverse Of (OP))	OP

V. Ontology Normalization Algorithm

The following is an outline of the algorithm which transforms the syntactically correct and consistent OWL 2 DL ontology selected by the user – denoted by OWL_{ONT} – into the normalized ontology. The OWL_{ONT}' and OWL_{ONT}'' are intermediate ontologies required to process the input ontology into the output ontology. In the beginning, both OWL_{ONT}' and OWL_{ONT}'' are empty. On completion of the algorithm, the OWL_{ONT}'' represents the normalized ontology.

Algorithm: Outline of the ontology normalization algorithm

Input: Syntactically correct and consistent OWL 2 DL ontology

Output: Normalized OWL 2 DL ontology

BEGIN

1. Take the first axiom from OWL_{ONT}.
2. Take the first entity from the selected axiom.
3. If the entity is declared, add the declaration axiom to OWL_{ONT}'.If the entity is not declared, extract the declaration axiom for the entity based on its usage and add the new declaration axiom to OWL_{ONT}'.
4. Take the next entity from the selected axiom.
5. Repeat steps 3-4 until no more entities in the selected axiom are available.
6. Apply to the selected axiom allapplicable replacement rules defined in Tables 5-7, receiving a modified axiom.
7. Add the modified axiom to OWL_{ONT}'.
8. Take the next axiom from OWL_{ONT}.
9. Repeat steps 2-8 until no more axioms in OWL_{ONT} are available.
10. Take the first axiom from OWL_{ONT}'.
11. Apply to the axiom allapplicable replacement rules defined in Tables 1-4.
12. If transformations result in only one axiom, add the axiom to OWL_{ONT}". Otherwise, if as a result of transformations the axiom splits into two or more axioms, repeat step 11 for each split axiom independently.
13. Take the next axiom from OWL_{ONT}'.
14. Repeat steps 11-13 until no more axioms in OWL_{ONT}' are available.
15. Eliminate any of the duplicated axioms from OWL_{ONT}" ontology.
16. Return the OWL_{ONT}" as a normalized ontology.

END

Comments to the algorithm:

1. OWL 2 ontologies are built of axioms which may contain some expressions. Data ranges are contained in two axioms: *DatatypeDefinition* and *DataPropertyRange*, as well as in some expressions, e.g.,*DataAllValuesFrom*, *DataMinCardinality*, etc. Therefore, to perform fewer iterations of the normalization algorithm, first, we conduct all the transformations of the data ranges in axioms and expressions, as well as the expressions in axioms, and later on of the axioms themselves.
2. If the input ontology does not contain any duplicated axioms, the resulting ontology will contain at least the same number of axioms as the input ontology.
3. The order of the conducted transformations is not important because the resulting ontology will always be semantically equivalent. However, depending on the selected order, the resulting ontology may have a different textual form. The possible textual differences in the output ontology include: (1) the order of axioms and (2) the order of expressions in axioms (only if the order of expressions in the selected axiom is not important).
4. The resulting ontology may contain fewer kinds of axioms and expressions. In particular, the ontology will not contain the below-mentioned list of axioms and expressions because they are refactored and reduced in accordance with the presented transformations:
   - *Class axioms: EquivalentClasses*, *DisjointClasses*, *DisjointUnion*,
   - *Object property axioms: EquivalentObjectProperties*, *InverseObjectProperties*, *ObjectPropertyDomain*, *ObjectPropertyRange, InverseFunctionalObject Property*, *FunctionalObjectProperty*, *ReflexiveObject Property*, *IrreflexiveObjectProperty*, *SymmetricObject Property*,*TransitiveObjectProperty*,
   - *Data property axioms: EquivalentDataProperties*, *DataPropertyDomain*, *DataPropertyRange*, *Functional DataProperty*,
   - *Class expressions: ObjectSomeValuesFrom*, *Object AllValuesFrom*, *ObjectHasValue*, *ObjectExact Cardinality*, *DataSomeValuesFrom*, *DataAllValues From*, *DataHasValue*, *DataExactCardinality*.
5. The method of normalization and the defined transformations are unidirectional, which means that it is not possible to retrieve the original ontology from the normalized ontology.

VI. Example of Single Normalization

The example presents transformations conducted with the use of the normalization algorithm. The following is an input ontology, which contains one axiom:

EquivalentClasses(:FourLeafClover: FourLeafClover ObjectIntersectionOf(
ObjectMinCardinality(3: hasLeaf:Leaf) ObjectMaxCardinality(7 :hasLeaf :Leaf)
ObjectExactCardinality(4 :hasLeaf :Leaf)))

Steps 1-5 of the algorithm extract declarations of entities:

Declaration(Class (:FourLeafClover)) (1)
Declaration(Class (:Leaf)) (2)
Declaration(ObjectProperty (:hasLeaf)) (3)

Steps 6-9 of the algorithm result in the following transformations:

Rule 19 from Table 6 applied on the given axiom (4)
EquivalentClasses(:FourLeafClover :FourLeafClover ObjectIntersectionOf(
ObjectMinCardinality(3 :hasLeaf :Leaf)
ObjectMaxCardinality(7 :hasLeaf :Leaf)
ObjectIntersectionOf(ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))

Rule 5 from Table 6 applied on (4) (5)
EquivalentClasses(:FourLeafClover :FourLeafClover ObjectIntersectionOf(
ObjectMinCardinality(3 :hasLeaf :Leaf)
ObjectMaxCardinality(7 :hasLeaf :Leaf)
ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))

Rule 20 from Table 6 applied on (5) (6)
EquivalentClasses(:FourLeafClover :FourLeafClover
ObjectIntersectionOf(ObjectMaxCardinality(7 :hasLeaf :Leaf)
ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))

Rule 23 from Table 6 applied on (6) (7)
EquivalentClasses(:FourLeafClover :FourLeafClover
ObjectIntersectionOf(ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))

Steps 10-15 of the algorithm result in the following transformations:

Rule 1 from Table 1 applied on (7) (8)
EquivalentClasses(:FourLeafClover ObjectIntersectionOf(
ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))

Rule 2 from Table 1 applied on (8) (9)
SubClassOf(:FourLeafClover ObjectIntersectionOf(
ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))
SubClassOf(ObjectIntersectionOf(ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)) :FourLeafClover)

Steps 16-17 of the algorithm return the normalized ontology:

Declaration(Class (:FourLeafClover)) (1)
Declaration(Class (:Leaf)) (2)
Declaration(ObjectProperty (:hasLeaf)) (3)
SubClassOf(:FourLeafClover ObjectIntersectionOf((9A)
ObjectMinCardinality(4 :hasLeaf :Leaf)
ObjectMaxCardinality(4 :hasLeaf :Leaf)))
SubClassOf(ObjectIntersectionOf(ObjectMinCardinality(4 :hasLeaf :Leaf) (9B)
ObjectMaxCardinality(4 :hasLeaf :Leaf)) :FourLeafClover)

VII. Proofs of the Correctness of the OWL 2 Construct Replacements

This section aims at presenting proofs of correctness of the OWL 2 construct replacements presented in tables in Section 4. The replacing language constructs (right column in tables) are semantically equivalent to the replaced language constructs (left column in tables).

The proofs are based on direct model-theoretic semantics [13]for OWL 2, which is compatible with the description logic *SROIQ*. The following convention is used:

1. V_C is a set of classes containing at least the owl: Thing and owl: Nothing classes.
2. V_{OP} is a set of object properties containing at least the object properties owl: topObjectProperty and owl: bottomObjectProperty.
3. Δ_I is a nonempty set called the object domain.
4. $(\ )^C$ is the class interpretation function that assigns to each class $C \in V_C$ a subset $(C)^C \subseteq \Delta_I$ such that $(\text{owl: Thing})^C = \Delta_I$ and $(\text{owl: Nothing})^C = \emptyset$
5. $(\ )^{OP}$ is the object property interpretation function that assigns to each object property $OP \in V_{OP}$ a subset $(OP)^{OP} \subseteq \Delta_I \times \Delta_I$ such that $(\text{owl:topObjectProperty})^{OP} = \Delta_I \times \Delta_I$ and $(\text{owl: bottomObjectProperty})^{OP} = \emptyset$
6. $\alpha = \beta$ means semantic equivalence of α and β sets.
7. $\alpha \vDash B$ means that α formula is the semantic consequence of B set of formulas.

Proving equivalence comes down to the use of the interpretation definition and the rules of set theory. We selected two replacement rules for the proofs; all other ones could be proved analogically.

Proof 1 for construct replacements from Table 1 ID 6:
We have to prove that the interpretation of

DisjointClasses(CE_1 CE_2)

is equivalent to the interpretation of

SubClassOf(CE_1 ObjectComplementOf(CE_2))

The interpretation of

DisjointClasses(CE_1 CE_2)

is (1) [13]:

$$(CE_1)^C \cap (CE_2)^C = \emptyset \quad (1)$$

The interpretation of

ObjectComplementOf(CE_2)

is (2) [13]:

$$\Delta_I \setminus (CE_2)^C \quad (2)$$

The interpretation of

SubClassOf(CE_1 CE_3)

is (3) [13]:

$$(CE_1)^C \subseteq (CE_3)^C \quad (3)$$

Based on (2) and (3) the interpretation of

SubClassOf(CE_1 ObjectComplementOf(CE_2))

is (4):

$$(CE_1)^C \subseteq \Delta_I \setminus (CE_2)^C \quad (4)$$

We have to show that (4) is correct.

If we assume that (4) is false, it means that (5) is true:

$$(CE_1)^C \not\subseteq \Delta_I \setminus (CE_2)^C \quad (5)$$

It means that there exist:

$$x \in (CE_1)^C \wedge x \notin \Delta_I \setminus (CE_2)^C \Leftrightarrow$$
$$x \notin \Delta_I \setminus (CE_2)^C \Rightarrow x \in (CE_2)^C$$

Then:

$$x \in (CE_1)^C \wedge x \in (CE_2)^C \Leftrightarrow$$
$$x \in (CE_1)^C \cap (CE_2)^C$$

It means that:

$$(CE_1)^C \cap (CE_2)^C \neq \emptyset$$

We have received contradiction, which had to be proved.

Proof 2 for construct replacements from Table 6 ID 7:
We have to prove that the interpretation of

ObjectUnionOf(
ObjectComplementOf(CE_1)

...
ObjectComplementOf(CE_N))

where $1 \leq i \leq N$ and $N \geq 2$ is equivalent to the interpretation of

ObjectComplementOf(ObjectIntersectionOf(CE_1 ... CE_N))

where $1 \leq i \leq N$ and $N \geq 2$.

The interpretation of

ObjectUnionOf(CE_1 ... CE_N)

is (14) [13]:

$$(CE_1)^C \cup \ldots \cup (CE_n)^C \tag{14}$$

The interpretation of

ObjectIntersectionOf(CE_1 ... CE_n)

is (15) [13]:

$$(CE_1)^C \cap \ldots \cap (CE_n)^C \tag{15}$$

Based on De Morgan's law for sets, (2) and (14) the interpretation of

ObjectUnionOf(
ObjectComplementOf(CE_1)
...
ObjectComplementOf(CE_N))

is (16):

$$(\triangle_I \setminus (CE_1)^C) \cup \ldots \cup (\triangle_I \setminus (CE_N)^C) \tag{16}$$

(17) is a result of application of (16) to (17):

$$\triangle_I \setminus ((CE_1)^C \cap \ldots \cap (CE_N)^C) \tag{17}$$

Based on (2) and (15) interpretation of

ObjectComplementOf(ObjectIntersectionOf(CE_1 ... CE_N))

is (18):

$$\triangle_I \setminus ((CE_1)^C \cap \ldots \cap (CE_N)^C) \tag{18}$$

The equations (17) and (18) are equal, which had to be proved.

VIII. Conclusions

The paper introduces the concept of ontology normalization as a process of transforming the input OWL 2 ontology into the ontology in its refactored form. The process is defined through a group of OWL 2 construct replacements. Because all individual replacing constructs preserve the semantics of the replaced constructs, the resulting ontology does not change the semantics of the original ontology.

Thanks to the presented approach, users obtain the possibility to automate the processing of ontologies because the normalized ontologies have the structure of axioms unified. However, the normalized ontology has reduced readability from the point of view of human readers, which is caused especially by the transformations from the group IV, which remove the syntactic sugar from the ontology.

The presented normalization algorithm is implemented in a prototype tool [8] which additionally allows for comparing two ontologies with the agreed vocabulary. More specifically, the tool states whether or not two ontologies are compliant or contradictory by the method outlined in [6].

References Références Referencias

1. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. https://www.w3.org/TR/owl2-overview/. 2012.
2. OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax (Second Edition). W3C Recommendation 11 December 2012, http://www.w3.org/TR/owl2-syntax/. 2012.
3. I Horrocks, O. Kutz, and U. Sattler, "The Even More Irresistible SROIQ," Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2006). AAAI Press, pp. 57–67, 2006.
4. OWL 2 Web Ontology Language New Features and Rationale (Second Edition) W3C Recommendation 11 December 2012, https://www.w3.org/TR/owl2-new-features/. 2012.
5. OWL 2 Web Ontology Language Profiles (Second Edition). W3C Recommendation 11 December 2012. https://www.w3.org/TR/owl2-profiles/. 2012.
6. M. Sadowska and Z. Huzar, "Semantic Validation of UML Class Diagrams with the Use of Domain Ontologies Expressed in OWL 2," Software Engineering: Challenges and Solutions. Springer International Publishing, pp. 47–59, 2017.
7. OMG, Unified Modeling Language,Version 2.5, Doc. No: ptc/2013-09-05, http://www.omg.org/spec/UML/2.5. 2015.
8. M. Sadowska, "A Prototype Tool for Semantic Validation of UML Class Diagrams with the Use of Domain Ontologies Expressed in OWL 2," In Towards a Synergistic Combination of Research

and Practice in Software Engineering. Springer, Cham, pp. 49–62, 2018.
9. V. Denny and Y. Sure, "How to design better ontology metrics," The Semantic Web: Research and Applications, pp. 311–325, 2007.
10. A. L. Rector, "Normalisation of ontology implementations: Towards modularity, re-use, and maintainability," Proceedings Workshop on Ontologies for Multiagent Systems (OMAS) in conjunction with European Knowledge Acquisition Workshop, pp. 1–16, 2002.
11. A. L. Rector, "Modularisation of domain ontologies implemented in description logics and related formalisms including OWL," Proceedings of the 2nd international conference on Knowledge capture. ACM, pp. 121–128, 2003.
12. Meta Object Facility (MOF) Core Specification, version 2.0. Object Management Group, OMG, http://www.omg.org/spec/MOF/2.0/PDF/. 2006.
13. OWL 2 Web Ontology Language Direct Semantics (Second Edition) W3C Recommendation 11 December 2012, https://www.w3.org/TR/2012/REC-owl2-direct-semantics-20121211/. 2012.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Privacy Preserving Access of Outsourced Data in Heterogeneous Databases

By J.Bama & Dr. M.S.Thanabal
PSNA College of Engineering and Technology

Abstract- Privacy is main concern in the world, among present technological phase. Information security has become a dangerous issue since the information sharing has a common need. Recently, privacy issues have been increased enormously when internet is flourishing with forums, social media, blogs and e-commerce, etc. Hence research area is retaining privacy in data mining. The sensitive data of the data owners should not be known to the third parties and other data owners. To make it efficient, the horizontal partitioning is done on the heterogeneous databases is introduced to improve privacy and efficiency. we address the major issues of privacy preservation in information mining. In particular, we consider to provide protection between different data owners and to give privacy between them by partitioning the databases horizontally and the data's are available in the heterogeneous databases. Our proposed work is to center around the study of security saving on unknown databases and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than k-secrecy.

Keywords: *privacy preserving, homomorphic encryption, third parties.*

GJCST-C Classification: *H.2.5*

Strictly as per the compliance and regulations of:

Privacy Preserving Access of Outsourced Data in Heterogeneous Databases

J.Bama [α] & Dr. M.S.Thanabal [σ]

Abstract- Privacy is main concern in the world, among present technological phase. Information security has become a dangerous issue since the information sharing has a common need. Recently, privacy issues have been increased enormously when internet is flourishing with forums, social media, blogs and e-commerce, etc. Hence research area is retaining privacy in data mining. The sensitive data of the data owners should not be known to the third parties and other data owners. To make it efficient, the horizontal partitioning is done on the heterogeneous databases is introduced to improve privacy and efficiency. we address the major issues of privacy preservation in information mining. In particular, we consider to provide protection between different data owners and to give privacy between them by partitioning the databases horizontally and the data's are available in the heterogeneous databases. Our proposed work is to center around the study of security saving on unknown databases and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than k-secrecy. Symmetric homomorphic encryption scheme, which is significantly more efficient than the asymmetric schemes. Our proposed work helps the valid user can extract with key issue in partition data in automated approach and the data's are partitioned horizontally.

Keywords: privacy preserving, homomorphic encryption, third parties.

I. Introduction

Now a days, data's are the biggest assets. We can see that increasing number of organizations that collect data very often concerning about the individuals and used them for various purposes such as scientific research, medical data, marketing etc. Organization may also give access to the data they own or even release such data to third parties. Data once released are no longer under the control of the organization owning them. So, the organization owners cannot prevent the modification of the data. The main problem is addressed as preserving the privacy of the data being stored in the databases.

Data mining is the process of extracting the knowledge from the enormous set of databases. The data mining has various applications such as Market Analysis and Management, Corporate Analysis and Risk Management, Fraud Detection, Intrusion Detection (Intrusion Detection means any kind of action that threatens integrity, confidentiality or the availability of the network resources),Retail Industry, Biological Data Analysis, Financial Data Analysis, Telecommunication Industry. There are some major disadvantages of data mining are their privacy issues, security issues and others as the misuse of information.

So, data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Mining encompasses various algorithms such as Clustering, Classification and Association Rule Mining [10]. Data mining deals with the mining of kinds of patterns. The kinds of patterns can be done in two ways either descriptive or classification and prediction. Our project deals with the descriptive way (ie) Mining of Associations.

The preview of data mining which falls within the problem of finding association rules is also called as Knowledge discovery in databases. In the Retail stores the Associations are used to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules [3]. The main objective of data privacy is to protect the personally identifiable information. In case, if the general information is considered then it should be linked either directly or indirectly to an individual person. The personal data subjected to mining and then the attribute values related with the individuals should be kept private and must be protected from disclosure. Data miners can be able to learn from the global models instead from the characteristics of a particular individual. The objective is not only to protect the personally known data but also to identify the trends and patterns that are not supposed to be discovered. Some information requires special care and handling. Incase if the information is handled inappropriately then it results in penalties, identify theft, financial loss, invasion of privacy or unauthorized access by one or more individuals. The confidential data's sensitivity is high. For example, confidential data's are research details, library transactions, personal information, information covered by non-disclosure agreements, contracts, facilities, management information. The concept of partitioning is used in our project for ensuring the privacy of data's available in the databases. Partitioning is the process of physically or logically partitioning data into segments that are more easily accessed or maintained and also

Author α: PG SCHOLAR, Dept. of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu.
e-mail: Bamavani1993@gmail.com
Author σ: ASP/CSE, Dept. of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamil Nadu.
e-mail: Ms_thanabal@yahoo.com

the arrangement of allocating data to data sites which occurs within the same common data architecture. Partitioning can be done in two ways. They are vertical partitioning and horizontal partitioning .If the databases are partitioned in the column wise then it is called vertical partitioning of databases. If the databases are partitioned in the row wise then the partitioning is known to be horizontal partitioning of databases.

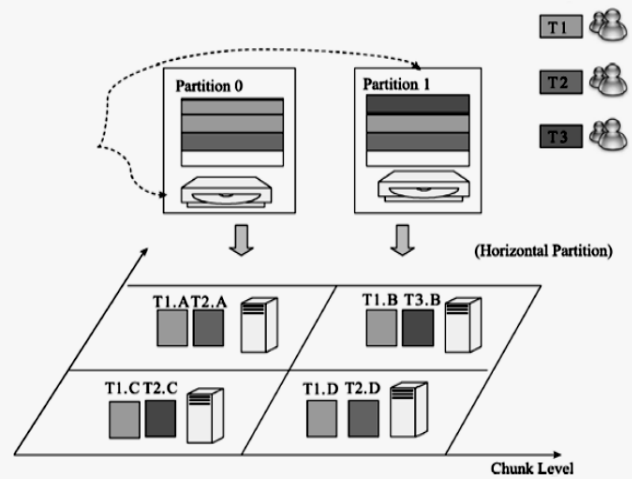


Figure 1

II. Background

Sensitive information is defined as data that is protected against unwanted disclosure. Access to sensitive information should be safeguarded. Sensitive information includes all data, in its original and duplicate form, for which there is either legal, ethical or contractual requirement that it be protected or access restricted. Also includes the unauthorized access of any data that is protected by the university policy. This information must be restricted to those with a legitimate business need for access. For example, public safety information, financial donor information, information security records, information file encryption keys. In order to find the association rules and their causal relationship between the set of items can be found using the association rule mining. So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together. It is machine learning - rule based approach for determining the relationship between the variables used in large databases. Association rules are largely used between the products in large scale transaction data being recorded by the point-of-sale (POS) systems in supermarkets which was done according to the concept of Rakesh Agarwal, Tomasz and Arun Swami. Association Rules are used in market analysis, intrusion detection, mining, bioinformatics, and continuous production . In spite of the sequence mining, they don't consider the order of items either across the transaction or within the transaction.

a) Association Rule Mining (Arm)

The popular research method used in data mining for discovering the interesting relations between variable in large databases. Association Rules (AR) are useful for analyzing and predicting the customer behavior. The IF-THEN statements are the association rules (AR) that help to uncover the relationship between unrelated data available in a relational database or other information Repository. Example If a customer buys bread then 80% of people are expected to buy butter. These association rules expresses about how the items or objects are related to each other and how they tend to group together.

The Association strength can be measured using the support, confidence values. Support is the ratio of the number of itemsets satisfying both antecedent and consequent to the total number of transactions. Confidence is derived from subset of transactions in which the two entities are related. Association Rule Mining can be done using three algorithms. They are Apriori algorithm, FP Growth algorithm, Éclat algorithm. In my work I have used FP Growth algorithm for discovering the frequent patterns in the database.

- Support(X) =No. of transactions contains X/total number of transactions.
- Confidence= (support(X union Y))/Support(X) .

b) FP Growth Algorithm

FP Growth algorithm means Frequent Pattern Growth Algorithm which is a scalable technique for mining the frequent pattern in the database. Frequent item set mining is possible without candidate generation so that the FP Growth algorithm is more efficient and a biggest improved over the Apriori Algorithm. This algorithm consumes less memory and a linear running time.

Procedure:

1. Build a compact data structure called the FP tree.
2. Extracts Frequent Itemsets directly from the FP tree.

III. Proposed Work

The proposed system introduced a privacy-preserving outsourced frequent itemset mining solution for horizontal partitioned from heterogeneous databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, we built a privacy-preserving outsourced association rule mining solution for horizontal partitioned databases for the unknown database and conceiving private refresh methods to database frameworks that backings thoughts of obscurity assorted than K-secrecy. Our solutions protect data owner's raw data from other data owners and the server. Our solutions also ensure the privacy of the mining results from the server is shown in Figure 3.1

Compared with most existing solutions, our solutions cannot leak less information about the data owners' raw data. Therefore, our solutions are suitable to be used by data owners wishing to outsource their databases to the cloud but require a high level of privacy without compromising on performance. Other than the settings of vertically partitioned databases and cloud/third-party-aided mining, privacy-preserving frequent itemset mining and association rule mining have been studied in the settings of horizontally partitioned databases data publishing and differential privacy.

The proposed system introduce a symmetric homomorphic encryption scheme using mediate Certificate less algorithm (using only modular additions and multiplications), which is significantly more efficient than asymmetric schemes. The scheme supports many homomorphic additions and limited number of homomorphic multiplications, and comprises the following three algorithms Key generation algorithm, Encryption algorithm, Decryption algorithm.

Advantage

The advantage of the proposed system is that the valid user can extract with key issue in partition data in automated approach.

a) Features of Proposed System

(i) The feature of the proposed system is managing data in horizontal partitioning.
(ii) The partition data is converted sensitive format by using Mediated Certificate less Algorithm.
(iii) If any valid user wants to review their original sources, they must submit valid attribute to extract heterogeneous databases.

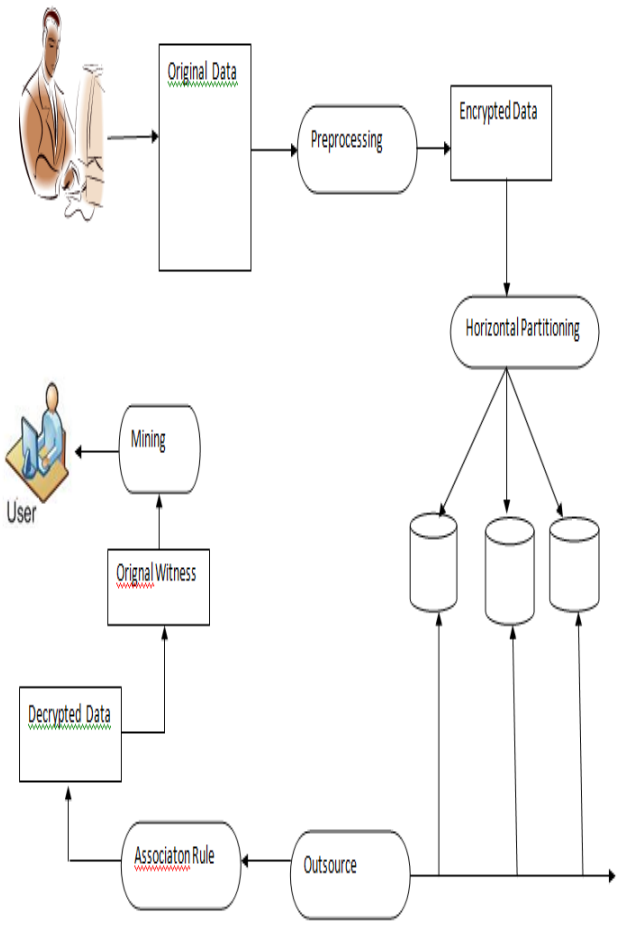


Figure 3.1: Architecture of proposed system

IV. Performance Evaluation

The horizontal partitioning algorithms used for comparison
As follow:

1. *Hash partition:* The data is evenly distributed to the predefined individual partitions, which ensures that the data of each partition has the same amount roughly;
2. *Schism partition:* Duplicate overlapped nodes; generate Partition according to the spanning graph.

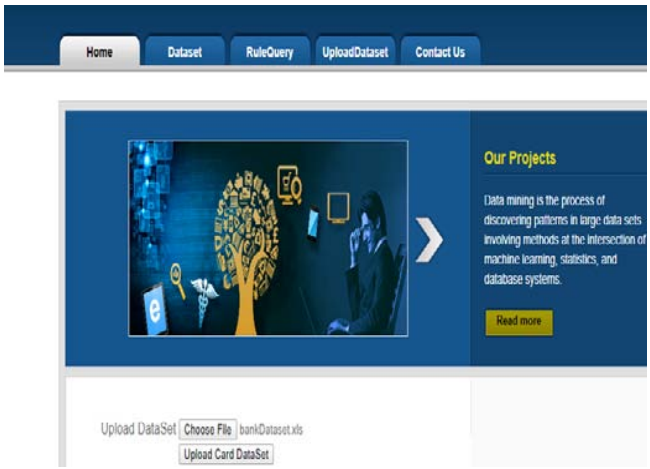
a) Partition Quality Evaluating

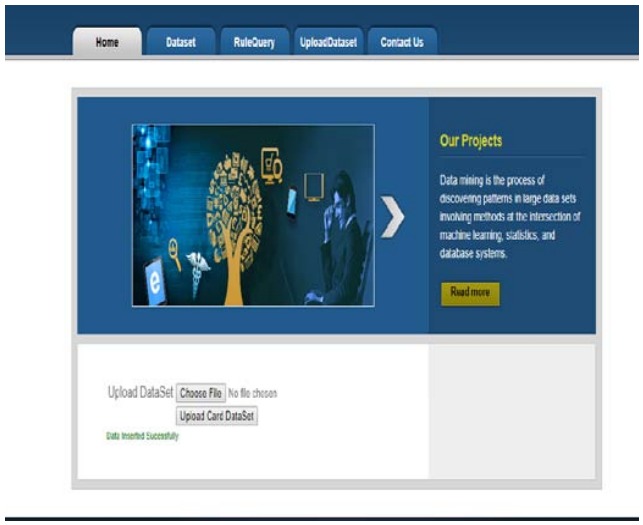
When a query accessed attributes on different partitions we call it a distributed query. The distributed queries cost more resources, so we regard the proportion of non distributed queries as the quality of the partitioning scheme.

The result shows that the proportion of non-distributed queries of larger than Hash and Schism, which indicates that with a certain number of partitions, the efficiency is better than Hash and Schism. In addition, Hash and Schism must know it before. The proposed Scheme can dynamically adapt to the coming

workload and predict the trend of the workload to give a better partition scheme, so proposed Scheme has better partition quality than Hash and Schism.

V. Result





VI. Conclusion

We proposed a privacy-preserving outsourced frequent itemset mining solution for horizontal partitioned databases. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. Based on this solution, we built a privacy-preserving outsourced association rule mining solution for horizontal partitioned databases. Our solutions also ensure the privacy of the mining results from the cloud. Compared with most existing solutions, our solutions leak less information about the data owners' raw data. Our evaluation has also demonstrated that our solutions are very efficient; therefore, our solutions are suitable to be used by data owners wishing to outsource their databases to the cloud but require a high level of privacy without compromising on performance.

References Références Referencias

1. Arun, V., Gowthami, S., & Padma, S. K. (2015, December)' Securely mining transactional databases for association rules using FDM' In Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on (pp. 340-345) IEEE.
2. Adhvaryu, R. V., & Domadiya, N. H. (2014)'Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database' International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 3(5).
3. Agrawal, R., & Srikant, R. (1994, September)' Fast algorithms for mining association rules' In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
4. Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., & Moser, S. A. (1998)'Association rules and data mining in hospital infection control and public health surveillance' Journal of the American medical informatics association, 5(4), 373-381.
5. Creighton, C., & Hanash, S. (2003)' Mining gene expression databases for association rules' Bioinformatics, 19(1), 79-86.
6. Dong, X., & Li, X. (2015, November)'A Novel Distributed Database Solution Based on MySQL' In Information Technology in Medicine and Education (ITME), 2015 7th International Conference on (pp. 329-333) IEEE.
7. Han, J., Pei, J., & Yin, Y. (2000, May)'Mining frequent patterns without candidate generation' In ACM sigmod record (Vol. 29, No. 2, pp. 1-12). ACM.
8. Kantarcioglu, M., & Clifton, C. (2004)'Privacy-preserving distributed mining of association rules on

horizontally partitioned data' IEEE transactions on knowledge and data engineering, 16(9), 1026-1037.
9. Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001, November)'Effective personalization based on association rule discovery from web usage data'In Proceedings of the 3rd international workshop on Web information and data management (pp. 9-15). ACM.
10. Rozenberg, B., & Gudes, E. (2006)'Association rules mining in vertically partitioned databases' Data & Knowledge Engineering, 59(2), 378-396.
11. Sheikhalishahi, M., and Martinelli, F. (2017, July)'Privacy preserving clustering over horizontal and vertical partitioned data' In Computers and Communications (ISCC), 2017 IEEE Symposium on (pp. 1237-1244) IEEE.
12. Yin, X., & Han, J. (2003, May) 'CPAR: Classification based on predictive association rules' In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 331-335). Society for Industrial and Applied Mathematics.
13. Zhan, J., Matwin, S., & Chang, L. (2005, August)'Privacy-preserving collaborative association rule mining' In DBSec (pp. 153-165).
14. Zaki, M.J. (2000)'Scalable algorithms for association mining'IEEE Transactions on Knowledge and Data Engineering, 12(3), 372-390.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques

By Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas & Jinat Ara

Southeast University

Abstract- E-mail is one of the most secure medium for online communication and transferring data or messages through the web. An overgrowing increase in popularity, the number of unsolicited data has also increased rapidly. To filtering data, different approaches exist which automatically detect and remove these untenable messages. There are several numbers of email spam filtering technique such as Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes and so on. This paper illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. However, here we present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.

Keywords: e-mail spam; unsolicited bulk email; spam filtering methods; machine learning; algorithm.

GJCST-C Classification: H.1.2

Strictly as per the compliance and regulations of:

A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques

Hanif Bhuiyan [α], Akm Ashiquzzaman [σ], Tamanna Islam Juthi [ρ], Suzit Biswas [ω] & Jinat Ara[¥]

Abstract- **E-mail is one of the most secure medium for online communication and transferring data or messages through the web. An overgrowing increase in popularity, the number of unsolicited data has also increased rapidly. To filtering data, different approaches exist which automatically detect and remove these untenable messages. There are several numbers of email spam filtering technique such as Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes and so on. This paper illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. However, here we present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.**

Keywords: e-mail spam; unsolicited bulk email; spam filtering methods; machine learning; algorithm.

I. Introduction

In recent years, internet has been created several platforms for making human life become more secure. Among these; e-mail is a substantial platform for user communication. Email is nothing; simply it's called an electronic messaging framework which transmits the message from one user to another [1]. Nowadays, e-mail has turned into a typical medium [2] because of its several branches like Yahoo mail [3], Gmail [4], Outlook [5] etc, which are completely free for all web user by following some administration [6, 7]. At present, Email called a secure worldwide communication medium for its several functions. But sometimes email becomes more hazardous for some "Spam Email".

Generally, Spam email called as junk email or unsolicited message which sent by spammer through Email. The process is, collected the address on the web and sends the message through domain's username. Actually, it has been produced for financial profits using the assortment of procedures [8] and instruments that incorporate spoofing, bonnets, open intermediaries, mail transfers, bulk mail instruments called mailers, and so forth. Spam filtering is a challenging undertaking for an assortment of reasons. For spam email, users are facing several problems like abuse of traffic, limit the storage space, computational power, become a barrier for finding the additional email, waste users time and also threat for user security [9, 10]. So, becoming email more secure and effective, appropriate Email filtering is essential.

Several types of researches have been performed on email filtering, some acquired good accuracy and some are still going on. According to researcher's overview, Email filtering is a process to sort email according to some criteria. As there are various methods exist for email filtering, among them, inbound and outbound filtering is well known. Inbound filtering is the process to read a message from internet address and outbound filtering is to read the message from the local user. Moreover, the most effective and useful email filtering is Spam filtering which performs through anti-spam technique. As spammers are proactive natures and using dynamic spam structures which have been changing continuously for preventing the anti-spam procedures and thus making spam filtering is a challenging task [9, 10].

Spam filtering is a process to detect unsolicited massage and prevent from entering into user's inbox. Now days, various systems have been existed to generate anti-spam technique for preventing unsolicited bulk email. Most of the anti-spam methods have some inconsistency between false negatives (missed spam) and false positives (rejecting good emails) which act as a barrier for most of the system to make successful anti-spam system. Therefore, an intelligent and effective spam-filtering system is the prime demand for web users.

Among various approach, Fiaidhi et al. [11] and Arora et al. [12] proposed method evaluate that, 70% today's business email's are spam [13]. Spam filtering has two major section; "Knowledge engineering" and "Machine learning". Knowledge engineering is an arrangement of guidelines to determine the spam

Corresponding Author α: Computer Science and Engineering Department, University of Asia Pacific, Dhaka, Bangladesh.
e-mail: hanifbhuiyan.c@gmail.com,
Author σ ρ: Computer Science and Engineering Department, University of Asia Pacific, Dhaka, Bangladesh.
e-mails: zamanashiq3@gmail.com, juthi.islam09@gmail.com
Author ω: Computer Science and Engineering Department, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.
e-mail: bsuzit@gmail.com
Corresponding Author ¥: Department of Computer Science and Engineering, South east University, Dhaka, Bangladesh.
e-mail: aracse2014@gmail.com

emails. In contrast, Machine learning is more efficient than knowledge engineering. It does not require any predefined rules. Naive Bayes, Support Vector Machines, Neural Networks, K-nearest neighbor, Rough sets, and artificial immune system are some prominent technique of Machine learning for spam filtering those are works by matching the regular expression, keywords from message text and so on.

II. Several Email Spam Filtering Methods

At present, number of spam email has increased for several criteria such as an advertisement, multi-level marketing, chain letter, political email, stock market advice and so forth. For restricting spam email, several methods or spam filtering system has been constructed by using various concept and algorithms. This section concluded by describing few of spam filtering methods to understand the process of spam filtering and its effectiveness.

a) Standard Spam Filtering Method

Email Spam filtering process works through a set of protocols to determine either the message is spam or not. At present, a large number of spam filtering process have existed. Among them, Standard spam filtering process follows some rules and acts as a classifier with sets of protocols. Figure.1 shows that, a standard spam filtering process performed the analysis by following some steps [14]. First one is content filters which determine the spam message by applying several Machines learning techniques [8, 10, 15-18]. Second, header filters act by extracting information from email header. Then, backlist filters determine the spam message and stop all emails which come from backlist file. Afterward, "Rules-based filters" recognize sender through subject line by using user defined criteria [19]. Next, "Permission filters" send the message by getting recipients pre-approvement. Finally, "Challenge-response filter" performed by applying an algorithm for getting the permission from the sender to send the mail.

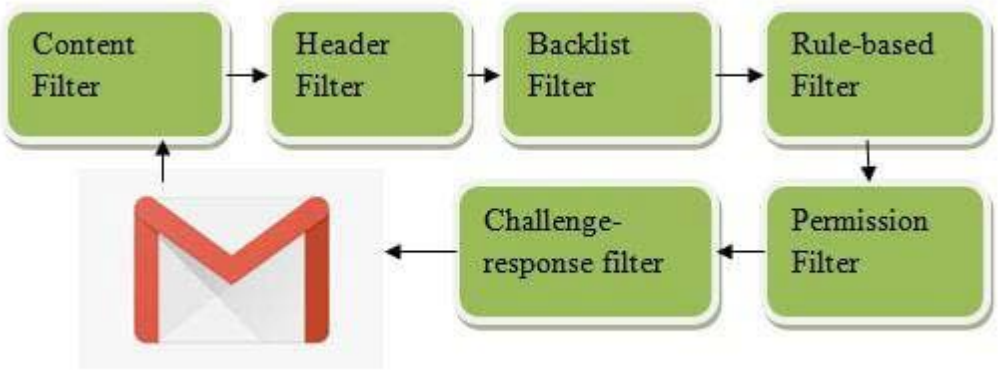


Figure 1: A standard process of Email spam filtering system

b) Client Side and Enterprise Level Spam Filtering Methods

A client can send or receive an email by just one clicking through an ISP. Client level spam filtering provides some frameworks for the individual client to secure mail transmission. A client can easily filter spam through these several existing frameworks by installing on PC. This framework can interact with MUA (Mail user agent) and filtering the client inbox by composing, accepting and managing the messages [2].

Enterprise level spam filtering is a process where provided frameworks are installing on mail server which interacts with the MTA for classifying the received messages or mail in order to categorize the spam message on the network. By this system, a user on that network can filter the spam by installing appropriate system [21, 22] more efficiently. By far most; current spam filtering frameworks use principle based scoring procedures. An arrangement of guidelines is connected to a message and calculate a score based principles that are valid for the message. The message will consider as spam message when it exceeds the threshold value. As spammers are using various strategies, so all functions are redesigned routinely by applying a list-based technique to automatically block the messages. Figure 2 represents the method of client side and enterprise level spam filtering [7].

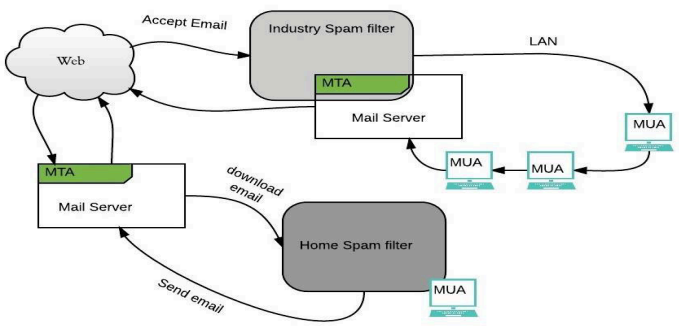


Figure 2: Client Side and Enterprise level Email spam filtering system

c) *Case Base Spam Filtering Method*

Among several spam filtering methods; case base or sample base filtering is one of the prominent method for Machine Learning Technique.

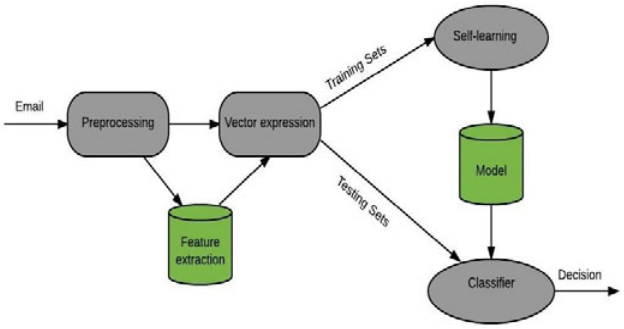


Figure 3: Case Base Spam Filtering System

Here, describes a sample of case base spam filtering architecture by applying Machine learning techniques [Fig. 3] in detail. The full process perform through several steps which followed by the figure 3.

At the first step, extracted all email (spam email and legitimate email) from individual users email through collection model. Then, the initial transformation starts with the pre-processing steps through client interface, highlight extraction and choice, email data classification, analyzing the process and by using vector expression classifies the data into two sets.

Finally, machine learning technique is applied on training sets and testing sets to determine email whether it is spam or legitimate. The final decision makes through two steps; through self observation and classifier's result to make decision whether the email is spam or legitimate.

III. Overview of Several Existing Email Spam Filtering Systems for Machine Learning Technique

Mohammed et al. [2] [2013] proposed an approach for Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques with the help of spam filtering which performs the work by creating a spam-ham dictionary from the given training data and applying data mining algorithm to filter the training and testing data. After applying various classifier on1431 dataset, the approach predicts that, Naïve Bays and SVM classifiers are the prominent classifier for spam filtering or classification.

Subramaniam et al. [23] [2012] implemented Naïve Bayesian Anti-spam Filtering Technique on Malay Language to investigate the utilization of Naïve Bayesian procedure to combat spam issue. An experiment conducted through Naïve Bayesian method for filtering Malay language spam and the result depicts that, propose approach has gained 69% accuracy. They realized that by reducing false positive and expanding training corpus the result would much better for classifying Malay language spam.

Sharma et al. [24] [2013] described Adaptive Approach for Spam Detection. This article consider SPAMBASE dataset and various machine learning technique such as Bays Net, Logic Boost, Random tree, JRip, J48, Multilayer Perception, Kstar, Random Forest, Random Committee are applied for classifying the spam. It measures the accuracy by grouping the spam/non-spam e-mails from labeled emails of a single account. The paper estimates that, total accuracy was 95.32% which depicts the quality of the proposed approach.

Banday et al. [25] [2008] discuss the procedures of statistical spam filters design by incorporating Naïve Bayes, KNN, SVM, and Bayes Additive Regression Tree. Here evaluates these procedures in terms of accuracy, recall, precision, etc. Though all machine learning classifiers are effective but according to this approach, CBART and NB classifiers has better capability to spam filtering. This approach estimates that during spam filtering calculations of false positive are more costly than false negative.

Awad et al. [1] [2011] proposed an ML- based approach on for Spam E-mail Classification. In this article present the most prominent machine learning strategies and its effectiveness regarding spam email classification. Here introduced Portrayals algorithms and the performance of Spam Assassin corpus. The result shows that, Naïve bays and rough sets methods are the promising algorithms for email classification. They perform their future research to improve the Nave Bays and Artificial immune system by hybrid system or by resolution the feature reliance issue.

Chhabra et al. [26] [2010] developed Spam Filtering using Support Vector Machine by considering Nonlinear SVM classifier with different kernel functions over Enron Dataset. Here considered six datasets and perform the analysis of datasets having diverse spam: ham ratio and makes satisfactory Recall and Precision Value.

Tretyakov et al. [27] [2004] discussed Machine Learning Techniques through Spam Filtering. In this article compared the precision between before eliminating false positive and after eliminating false positive. They represent the result that the result becomes more reliable considering both precision results (before eliminating and after eliminating false positive) either taking one.

Shahi et al. [28] [2013] developed Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine. The fundamental concern of this study was to look at the effectiveness of Naïve Bayesian and SVM Spam filters. The correlation of productivity between these Spam filters was done based

on the precision and recall. Approach showed that Nave Bays produce better accuracy than SVM.

Kaul et al. [29] [2004] implemented Filtering Spam E-mail with Support Vector Machines. Here in this paper they consider a virtual machine called SpamStop. SpamStop performs on the large dataset to produce more accurate result. It has a drawback such as SpamStop does not yet incorporate an assortment of standard pre-filtering mechanisms.

Suganya et al. [30] [2014] worked on short message and misspelling of data on online Social Networks (OSNs) user post. They used machine learning technique with content- based features for short message and Filtered Wall (FW) [31] to evaluate a system for filtering spam massage. They categorized the classification process into two levels; first-level classifier performs on Neutral and Non-neutral through hard binary categorization and second level classifier performs through RBFN model [32].

Rathi et al. [33] [2013] proposed an approach using Data mining technique for finding the best classifier for email classification. They analyzed various data mining technique for measuring the performance of several classifiers through "with feature selection algorithm" and "without feature selection algorithm". After selecting the Best feature selection algorithm, they considered the selected algorithm for their feature selection purpose. They experiment their data by using several algorithms such as Naïve Bayes, Bayes Net, Support vector machine, and Function tree, J48, Random Forest and Random Tree. The whole dataset consists of 58 attributes and 4601 instances. Considering Random Tree algorithm highest accuracy was 99.72% and the lowest accuracy was 78.94% for Naïve Bayes algorithm.

Mohammed et al. [11] [2013] presents an approach for filtering spam email using machine learning algorithms. At first, they filter Spam and Ham word from the training datasets by applying tokenization method based on these token create the testing and training table using various data mining algorithm. Then find the frequency of spam and ham tokens for measuring the probability which is suggested by Paul Graham [34]. For ham token, the probability value was 0 and for spam token probability value was 1. They used Nielson Email-1431 [35] dataset and emphasized that the Naïve Bayes and Support Vector Machine are the most effective classifier.

Singh et al. [36] [2018] discussed the solution and classification process of spam filtering and presented a combining classification technique to get better spam filtering result. With the help of Data mining, they collected all the information of previous failures, success and current problems of spam filtering. In this method, researchers used binary value where 1 for spam email and 0 for not spam emails. But its success rate was very poor. So they apply NB, KNN, SVM, Artificial Neural Network classification method and find their accuracy. Based on these two techniques (machine learning and knowledge engineering) effectiveness, they adopt a classification technique for spam filtering. Moreover, here first collect data from user training set, compared and find the spam email and then use a global training set to optimize the classification technique. Using this technique increases the precision rate at least 2%.

Abdulhamid et al. [37] [2018] introduced a performance analysis based approach by using some classification techniques such as Bayesian Logistic Regression, Hidden Naïve Bayes, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Naïve Bayes, Radial Basis Function (RBF) Network, Voted Perceptron, Lazy Bayesian Rule, Multilayer Perceptron, Random Tree and J48. The competence of these techniques classified through Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error, Receiver Operator Characteristics Area and Root Relative Squared Error using Spam base dataset and WEKA data mining tool. For conducting the performance and comparison, datasets are considered from UCI Machine Learning Repository. Considering Rotation Forest algorithm acquired the highest accuracy was 0.942 and the REP Tree algorithm showed the lowest accuracy was 0.891. They applied the F-measure method for finding precision and recall. The highest F-measure considered from Rotation forest algorithm and lowest F-measure considered from Naïve Bayes algorithm. For finding the probability use ROC curves on randomly selected positive and negative instance and for Rotation forest algorithm the ROC curves carried the highest score was 0.98. In contrast, Random Tree having the lowest score which was 0.905. For finding the statistics result, they use kappa Statistics and the result was much better for Rotation Forest algorithm which approximately 0.879. This paper showed that, Rotation Forest classifier gained the best result with 0.942 accuracies, then J48 with 0.923, Naïve Bayes with 0.885 and Multilayer Perception with 0.932.

Sah et al. [38] [2017] proposed a method for detecting of malicious spam through feature selection and improve the training time and accuracy of malicious spam detection system. They also showed the comparison of difference classifier as Naïve Bayes (NB) and Support Vector Machine (SVM) based on accuracy and computation time. The proposed approach completed by four steps such as preparing the text data, creating word dictionary, Feature extraction process and training the classifier. For preparing text data researchers split the dataset into the training set (702 mails) and a test set (260 mails) and divided into spam and ham mails. Performed feature selection process by generating feature vector matrix. According

to the approach, Naïve Bayes selected as good classifiers among others.

Verma et al. [39] [2017] proposed a method for spam detection using Support Vector Machine algorithm and feature extraction. This methodology works through several steps such as Email collections, preprocessing, feature extraction, SVM training, test classifier, top word predictors, test email and result. First they take a dataset from Apache Public corpus. In preprocessing section, they remove all special symbol, URL and HTML tags and also unnecessary alphabet. Then they mapped all word from the dictionary using Vocab file. SVM classifier applied on the training dataset. The Accuracy of the system was 98%.

Rusland et al. [40] [2017] perform the analysis using Naïve Bayes algorithm for email spam filtering on two datasets which are evaluated based on the accuracy, recall, precision and F-measure. Naïve Bayes algorithm is a probability-based classifier and the probability is counting the frequency and combination of values in a dataset. This research performed through three phases such as pre-processing, Feature Selection, and implementation through Naïve Bayes Classifier. First they remove all conjunction words, articles from the email body in pre-processing section. Made two datasets through WEKA tool; one is a Spam Data and another is the SpamBase dataset. The average accuracy was 8.59% by considering two datasets where Spam data get 91.13% and the SpamBase data get 82.54% accuracy. The average precision for SpamBase was 88% and for Spam data was 83%. They proposed that, Naïve Bayes classifier performs better on SpamBase data compared with Spam Data.

Yuksel et al. [41] [2017] use Support Vector Machine and Decision tree for spam filtering. The Decision tree used in data mining and the support vector machines as a supervised learning model which can analyze the data for spam classification. First data was divided into two sections; one is training and other is test data, then the algorithm was trained and evaluated through Microsoft Azure platform which provides tools for machine learning and compared results with decision tree and support vector machine algorithm. The result of SVM method was 97.6% and for Decision tree the result was 82.6%. The result estimate that, SVM classifier performed better than DT.

Choudhary et al. [42] [2017] presented a novel approach using machine learning classification algorithm for finding and classifying SMS spam by using Short Message Service (SMS). The first step in this approach is feature selection and for that, they work on presence of mathematical symbols: UGLs, Dots, special symbols, emotions, Lowercased words and Uppercased words, mobile number, keyword specific and the message length in the SMS. After that they created a system design and collected a dataset which contained 2608 emails out of 2408 collected SNS Spam Corpus. The SMS Spam Corpus v.0.1 consists two sets of messages as SMS Spam Corpus v.0.1 Small and SMS Spam Corpus v.0.1 Big. Using "WEKA tools" for five machine learning approaches; such as Naive Bayes, Logistic Regression, J48, Decision Tree and Random Forest. Evaluating result uses with True Positive Rate (TP) and True Negative Rate (TN). False Positive Rate (FP), False Negative Rate (FN), Precision, Recall, F-measure and Receiver Operating Characteristics (ROC) area achieved 96.5% true positive rate and 1.02% false positive rate with Random Forest machine learning algorithm and it performs better algorithm with high rate accuracy.

DeBarr et al. [43] [2009] use Random Forest algorithms for classification of spam email then refining the classification model using active learning. They take data from RFC 822(Internet) email message and divided each email into two sections and converted each message to term frequency and inverse document frequency (TF/IDF) features. Here select an initial set of email message using clustering technique to label as training examples and for clustering used Partitioning Around Medoids (PAM) algorithm. After considering the cluster prototype messages for training they experiment with some algorithm Random Forest, Naive Bayes, SVM and kNN. Here Random Forest algorithm performs the best classifier with 95.2% accuracy.

IV. Summary of Existing E-mail Spam Classification Approaches

Since last few decades, researchers are trying to make email as a secure medium. Spam filtering is one of the core features to secure email platform. Regarding this several types of research have been progressed reportedly but still there are some untapped potentials. Over time, still now e-mail spam classification is one of the major areas of research to bridge the gaps. Therefore, a large number of researches already have been performed on email spam classification using several techniques to make email more efficient to the users. That's why, this paper tried to arrange the summarized version of various existing Machine Learning approaches. In addition, in order to evaluates the most of the approaches like Random Forest, Naive Bayes [11, 23, 43], SVM [8, 10, 18], kNN [27, 36], and Random Forest [15, 16] used reliable and well known dataset for benchmarking performance such as SpamData [16], The Spam Assassin [44], The Spambase, Ecml-pkdd 2006 challenge dataset [45], PU corpora dataset [15], Enron dataset [46],Trec 2005 dataset [47]. Some of these dataset are in a prepared structure e.g. ECML and data accessible in Spambase UCI archive [20]. Among them, some of the classifiers also used novel methods applied in the feature selection for improving classification such as [1, 11].

Table I: Summary of different existing email spam classification approaches regarding Machine Learning Techniques

Sr. No.	Author	Algorithms	Corpus or Datasets	Accuracy/ Performance
1	Mohammed et al.	Naive Bayes, SVM, KNN,Decision Tree, Rules	Email-1431	85.96% Accuracy Achieved
2	Subramaniam et al.	Naive Bayesian	Collection of spam emails from Google's Gmail Account	96.00% Accuracy Achieved
3	Sharma et al.	Various Machine Learning Algorithms Adaptions	SPAMBASE	94.28% Accuracy Achieved
4	Banday et al.	Naive Bayes, K-Nearest Neighbor, SVM, classification Bayes Additive Regression Tree	Real life data set	96.69% Accuracy Achieved
5	Awad et al.	Naive Bayes, SVM, k-Nearest Neighbor, Artificial Neural Networks, Rough Sets	Spam Assassin	99.46% Accuracy Achieved
6	Chhabra et al.	Nonlinear SVM classifier.	Enron dataset	For Dataset 3, spam: real, the ratio is 1:3, for satisfactory Recall and Precision Values
7	Tretyakov	Bayesian classification, k-NN, ANNs, SVMs	PU1 corpus	94.4% Accuracy Achieved
8	Shahi et al.	Naïve Bayes, SVM	Nepali SMS	92.74% Accuracy Achieved
9	Kaul et al	SVM	Sample emails	90% ~ 95%Accuracy Achieved
10	Suganya et al.	Rule Baseed Method	Online Social Networks (OSNs) user post	Excellence Accuracy for Given Datasets
11	Rathi et al.	Naive Bayes, Bayes Net,SVM, and Random Forest	Custom Collection	99.72% Accuracy Rate
12	Mohammed et al.	Word Filterization by Tokenization, Appling	Nielson Email-1431	Reported Satisfactory Accuracy for Proposed Method
13	Singh et al.	Naive Bayes, k-Nearest Neighbor, SVM, Artificial Neural Network.	Custom Collection	Reported Improvement of precision rate at least 2%
14	Abdulhamid et al.	Various Machine Learning Algorithms	UCI Machine Learning Repository	94.2% Accuracy Achieved
15	Sah et al.	Naïve Bayes, SVM	& Custom Collection	Reported good Accuracy overall
16	Verma et al.	Customised SVM	Apache Public Corpus	98% Accuracy Rate Reported
17	Rusland et al.	Modified Naive Bayes withselective features	SpamBase, SpamData	SpamBase get 88%Precision Rate and SpamData get 83%
18	ksel et al.	Microsoft Azure platform defined decision tree and SVM	Custom Collection	SVM Accuracy 97.6% Decision Tree Accuracy 82.6%
19	Choudhary et al.	Feature Engineered Naive Bayes	The SMS Spam Corpus v.0.1	96.5% True Positive Rate Accuracy
20	DeBarr et al.	Random Forest algorithm	Custom Collection	95.2% Accuracy

V. Discussion

From the observation, it seems that, the majority of email spam filtering process performed through Machine learning technique using Naïve Bayes and SVM algorithm. Most of the approaches adopt different dataset such as "ECML" data and Spam base UCI archive [20]. Among several papers, Mohammad et al. introduce a classifier for feature selection which regarded as the most novel classifier for feature selection [1, 11]. Rathi et al proposed an approach considering "Naïve Bayes", "Bayes Net", "SVM" and "Random forest" algorithm and obtain the higher accuracy than others which approximately crossed 99.72% accuracy [32]. Another one is, Awad et al. which proposed an approach considering "Naïve Bayes", "SVM", "K-Nearest Neighbor", "Artificial neural Networks", "Rough sets" algorithm and obtain 99.46% accuracy which seems good on their effectiveness [1]. After the analysis it should predict that, "Naïve Bayes" and "SVM" algorithm is the most effective algorithm in machine learning technique and have the ability to better classification of email spam.

VI. Conclusion

This survey paper elaborates different Existing Spam Filtering system through Machine learning techniques by exploring several methods, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding several parameters. Moreover, all the existing methods are effective for email spam filtering. Some have effective outcome and some are trying to implement another process for increasing their accuracy rate. Though all are effective but still now spam filtering system have some lacking which are the major concern for researchers and they are trying to generate next generation spam filtering process which have the ability to consider large number of multimedia data and filter the spam email more prominently.

References

1. Awad, W. A., & ELseuofi, S. M. (2011). Machine Learning methods for E-mail Classification. International Journal of Computer Applications, 16(1).
2. Saad, O., Darwish, A., & Faraj, R. (2012). A survey of machine learning techniques for Spam filtering. International Journal of Computer Science and Network Security (IJCSNS), 12(2), 66.
3. Chen, Y., Jain, S., Adhikari, V. K., Zhang, Z. L., & Xu, K. (2011, April). A first look at inter-data center traffic characteristics via yahoo!datasets. In INFOCOM, 2011 Proceedings IEEE (pp. 1620-1628). IEEE.
4. Barlow, K., & Lane, J. (2007, October). Like technology from an advanced alien culture: Google apps for education at ASU. In Proceedings of the 35th annual ACM SIGUCCS fall conference (pp. 8-10). ACM.
5. Fisher, D., Brush, A. J., Gleave, E., & Smith, M. A. (2006, November). Revisiting Whittaker & Sidner's email overload ten years later. In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (pp. 309-312). ACM.
6. Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1), 63-92.
7. Cunningham, P., Nowlan, N., Delany, S. J., & Haahr, M. (2003, May). A case-based approach to spam filtering that can track concept drift. In The ICCBR (Vol. 3, pp. 03-2003).
8. Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.
9. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).
10. Wang, Q., Guan, Y., & Wang, X. (2006). SVM-Based Spam Filter with Active and Online Learning. In TREC.
11. Mohammed, S., Mohammed, O., Fiaidhi, J., Fong, S. J., & Kim, T. H. (2013). Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques.
12. Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A. (2014, February). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on (pp. 153-155). IEEE.
13. Scholar, M. (2010). Supervised learning approach for spam classification analysis using data mining tools. organization, 2(8), 2760-2766.
14. Christina, V., Karpagavalli, S., & Suganya, G. (2010). A study on email spam filtering techniques. International Journal of Computer Applications, 12(1), 0975-8887.
15. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes? In CEAS (Vol. 17, pp. 28-69).
16. Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive bayesian anti-spam filtering. arXiv preprint cs/0006013.
17. Hovold, J. (2005, July). Naive Bayes Spam Filtering Using Word-Position-Based Attributes. In CEAS (pp. 41-48).
18. Hidalgo, J. M. G. (2002, March). Evaluating cost-sensitive unsolicited bulk email categorization.

In Proceedings of the 2002 ACM symposium on Applied computing (pp. 615-620). ACM.
19. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.
20. Fawcett, T. (2003). In vivo spam filtering: a challenge problem for KDD. ACM SIGKDD Explorations Newsletter, 5(2), 140-148.
21. Wu, C. T., Cheng, K. T., Zhu, Q., & Wu, Y. L. (2005, September). Using visual features for anti-spam filtering. In Image Processing, 2005. ICIP 2005. IEEE International Conference on (Vol. 3, pp. III-509). IEEE.
22. Cormack, G. V., Gómez Hidalgo, J. M., & Sánz, E. P. (2007, November). Spam filtering for short messages. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 313-320). ACM.
23. Subramaniam, T., Jalab, H. A., & Taqa, A. Y. (2010). Overview of textual anti-spam filtering techniques. International Journal of Physical Sciences, 5(12), 1869-1882.
24. Sharma, S., & Arora, A. (2013). Adaptive approach for spam detection. International Journal of Computer Science Issues, 10(4), 23-26.
25. Banday, M. T., & Jan, T. R. (2009). Effectiveness and limitations of statistical spam filters. arXiv preprint arXiv: 0910.2540.
26. Chhabra, P., Wadhvani, R., & Shukla, S. (2010). Spam filtering using support vector machine. Special Issue IJCCT, 1(2), 3.
27. Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. In Data Mining Problem-oriented Seminar, MTAT (Vol. 3, No. 177, pp. 60-79).
28. Shahi, T. B., & Yadav, A. (2013). Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. International Journal of Intelligence Science, 4(01), 24.
29. Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.
30. Suganya, T., Sridevi, K., & ArulPrakash, M. Detection of Spam in Online Social Networks (OSN) Through Rule-based System.
31. Rahane, U., Lande, A., Bavikar, O., Chavan, S., & Shedge, K. N. International Journal of Engineering Sciences & Research Technology Advanced Filtering System to Protect OSN user Wall From Unwanted Messages.
32. Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. Neural computation, 1(2), 281-294.
33. Rathi, M., & Pareek, V. (2013). Spam mail detection through data mining-A comparative performance analysis. International Journal of Modern Education and Computer Science, 5(12), 31.
34. Graham, P. (2002). A plan for spam (http://www.paulgraham.com/spam.html).
35. Kang, N., Domeniconi, C., & Barbará, D. (2005, November). Categorization and keyword identification of unlabeled documents. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.
36. Singh, V. K., & Bhardwaj, S. (2018). Spam Mail Detection Using Classification Techniques and Global Training Set. In Intelligent Computing and Information and Communication (pp. 623-632). Springer, Singapore.
37. Shafi'i Muhammad Abdulhamid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection.
38. Sah, U. K., & Parmar, N. (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers.
39. Verma, T. (2017). E-Mail Spam Detection and Classification Using SVM and Feature Extraction.
40. Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. In IOP Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012091). IOP Publishing.
41. Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. (2017). Design of a Machine Learning Based Predictive Analytics System for Spam Problem. Acta Physica Polonica, A., 132(3).
42. Choudhary, N., & Jain, A. K. (2017). Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique. In Advanced Informatics for Computing Research (pp. 18-30). Springer, Singapore.
43. DeBarr, D., & Wechsler, H. (2009, July). Spam detection using clustering, random forests, and active learning. In Sixth Conference on Email and Anti-Spam. Mountain View, California (pp. 1-6).
44. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.
45. Mavroeidis, D., Chaidos, K., Pirillos, S., Christopoulos, D., & Vazirgiannis, M. (2006). Using tri-training and support vector machines for addressing the ECML/PKDD 2006 discovery challenge. In Proceedings of ECMLPKDD 2006 Discovery Challenge Workshop (pp. 39-47).
46. Klimt, B., & Yang, Y. (2004, July). Introducing the Enron Corpus. In CEAS.

47. Bratko, A., & Filipic, B. (2005, November). Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track. In TREC.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Accuracy Analysis of Continuance by using Classification and Regression Algorithms in Python

By Swayanshu Shanti Pragnya

Centurion University

Abstract- Reinforcement rate of technics and appositeness towards the convenience of the human being is a perennial mechanism. Mathematics has always been in the root towards the implementation of any algorithm or analysis regarding statistics or language. Extracting more about the data and analyzing them to solve a particular problem is the reason behind any analysis. Scrutiny itself has the different number of outcome which can be predictive or descriptive. Now prediction is how far accurate is tested by using various techniques. The enhancement in problem-solving capability leads to come up with a new aptitude concerning machine learning algorithms. But before prediction of data set collection, exploration, feature extraction, model building, accuracy testing are primarily required to invent. So for explaining all these processes, concept learning is essential. In this paper different algorithms like SVM, Linear and Logistic Regression, Decision tree, and Random forest algorithms will be used to demonstrate the accuracy in titanic data from Kaggle Website with all the required steps by using Python language.

Keywords: data analysis, machine learning, linear regression, logistic regression, random-forest, SVM, pandas and seaborn library, confusion matrix, ROC, precision-recall curve.

GJCST-C Classification: I.1.2

Strictly as per the compliance and regulations of:

Accuracy Analysis of Continuance by using Classification and Regression Algorithms in Python

Swayanshu Shanti Pragnya

Abstract- Reinforcement rate of technics and appositeness towards the convenience of the human being is a perennial mechanism. Mathematics has always been in the root towards the implementation of any algorithm or analysis regarding statistics or language. Extracting more about the data and analyzing them to solve a particular problem is the reason behind any analysis. Scrutiny itself has the different number of outcome which can be predictive or descriptive. Now prediction is how far accurate is tested by using various techniques. The enhancement in problem-solving capability leads to come up with a new aptitude concerning machine learning algorithms. But before prediction of data set collection, exploration, feature extraction, model building, accuracy testing are primarily required to invent. So for explaining all these processes, concept learning is essential. In this paper different algorithms like SVM, Linear and Logistic Regression, Decision tree, and Random forest algorithms will be used to demonstrate the accuracy in titanic data from Kaggle Website with all the required steps by using Python language.

Keywords: data analysis, machine learning, linear regression, logistic regression, random-forest, SVM, pandas and seaborn library, confusion matrix, ROC, precision-recall curve.

I. Introduction

Now a day's statistical analysis in any data is performed just to analyze the data little bit more by using mathematical terms. But only resolving a data is not sufficient when it comes to analysis that too by using statistics. So at this point, predictive audit comes which is nothing but a part of inferential statistics. Here we try to infer any outcome based on analyzing patterns from previous data to predict for the next dataset when it comes to prediction first buzzword came, i.e., machine learning. So machine learning combine's statistical analysis and computer science for the prediction purpose. Machine learning also introduced to self-learning process from particular data. This learning reduces the gap between computer and statistics. Alarge amount of data prediction can be possible by human interaction as a human brain can analyze the situation with various aspects. Here the partition of algorithms occur, i.e., Supervised (used for labeled data) and unsupervised (data with no tag for learning) algorithm. As the name itself says that machine will learn, but the question arises how that is by using data. In general, by performing mistakes, we learn anything so in Machine learning these mistakes are the data which will be given to the machine to learn. But only learning is not sufficient for a model as again we need to test whatever that machine learned is it accurate or not. Here accuracy testing is required which we are going to measure by creating confusion matrix.

Before building any model in machine learning first, we need to collect the data then few pre-processing is required. Feature extraction is essential to know which features are vital in our model building. After getting the features we can build our model by using different algorithms, depending on our problem statement. Once the model is built, now we need to check its accuracy. Here we will know all the process carried out in model building. Different algorithms used like SVM, K- means, Decision tree, Random Forest, Linear and Logistic regression, from statistics standard deviation, variance analysis, Mean usability, displacement calculation and so on. All the concepts will execute by Python language and code will implement by using Jupyter Notebook.

a) *The Need of Classification and Regression*

Both classification and regression are frequently used in Data mining techniques. Regression comes into eye view when we need to predict dependant (Rely upon other attributes) variable which has relation with other data.

Example- In our given Titanic data the number of survived passenger is somehow dependent upon which class the passenger is traveling as well as which cabin they were sitting. So for predicting which person survived is relative upon all these attributes so here we will use regression technique to predict.

As the name itself defines Classification is all about the categorization of data based on condition.

Support Vector Machine algorithm can give high accuracy when the data set is small and as well as less missing values in the given dataset.

Tools

Python: Open source as well as easy to understand, the syntax is easy for beginners and used for statistical data analysis.

Author: Centurion University, Department of Computer Science Engineering, Hyderabad, 7500033, India.
e-mail: Swayanshu1997@gmail.com

Pandas: Highly used library for data analysis. Easy to understand. Open source as well as easy to use in data manipulation.

Numpy: Used for scientific computing with python.

Matplotlib: It is a mathematical extension from Numpy (Library for mathematical calculation) as well as primarily used for plotting graphs.

II. Method

Linear and logistic regression [3] both used for prediction purpose. But what's the difference is much more important to know. These are the following attributes to perceive the difference between these two regression algorithms.

Outcome after regression: In linear regression, the result we got is continuous whereas logistic regression has limited number of possible values.

Dependent variable: Logistic regression used for the instance of true/false, yes/no, 0/1 which are categorical in nature but linear regression used in case of a continuous variable like a number, weight, height, etc.[4]

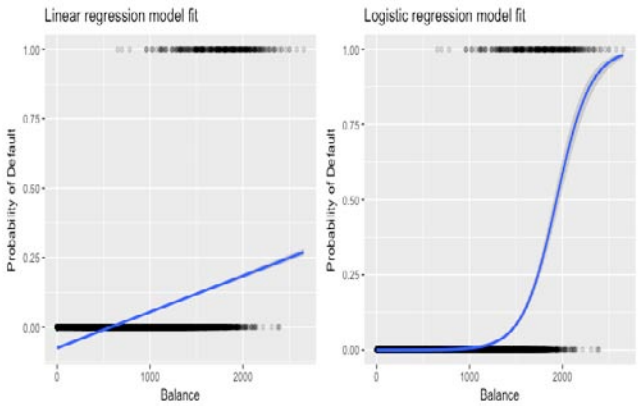


Fig. 1: Linear and logistic regression

Equation:

Linear regression gives a linear equation in the form of Y = aX + B, means degree 1 equation But, logistic regression gives curved association which is in the form of Y = e^X/1 + e^-X

Minimization of error:

Linear regression **(LR)** uses ordinary least squares method which minimizes the error and, Logistic Regression [5] use the least square method which reduces the error quadratic-ally.

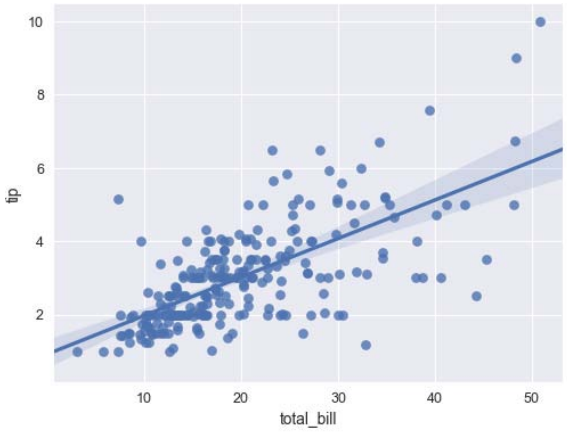


Fig. 2: Linear regression with nearest data [3]

III. Support Vectors

These are the vectors (magnitude and direction) which take support for classification purpose near to the hyper plane. [2]

Hyper-plane: Generally plane forms in 2 dimensions but more than 2D it is called the hyper-plane. Though support vectors drawn in more than two extent that's why it splits data through hyper-plane [2].

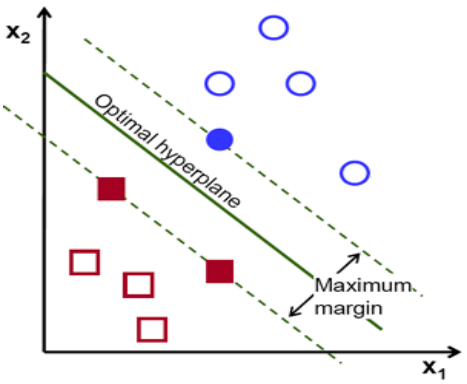


Fig. 3: SVM

In the above example we saw the set of blue and red dots separated, but in the next picture, the splitting is done via hyper-plane to segregate data set in two different clusters.

Way to find right hyper-plane: Nearest data point and hyper-plane distance are known as margin. So when the margin is less the chance of correct segregation is more. [5]

Decision Trees: It is a decision sequence which designed in such a tree-like structure. It includes Yes or No type of answers. In our given data set the Passenger either survive or will die.

Random Forest: Tree will be the combination of the Decision tree.

IV. Code and Explanation

Step 1- Irrespective of any regression or classification algorithm initially need to import libraries like Pandas, Numpy, Matplotlib, Seaborn and from Scikit-learnlinear, logistic regression and SVM module.

Step 2 – Loading data in CSV file format as the data has been taken from Kaggle Titanic competition. Where train and test data set were grasped for regression.[10]

Step 3- Select required columns in X (mostly independent variable) and in Y take dependant column as per here number of passengers survived is dependant that's why clasped in Y.

Step 4- Data cleaning and fill null values to prepare data.

Step 5- For knowing which column is influenced (value related to other column in data) more on the output column, we need to plot graphs by using regression type. [9]

Step 6 – Split the data set into train and test by using Scikit-learn(free software for Machine learning libraries for Python programming).

Step 7- Fill all the null values using Mean or Dummy Values.

Step 8- Finally call regression function whether it is linear, logistic or SVM, KNN, Decision tree.[7].

Step 9- Calculate accuracy of all the algorithms and print it.

Step 10- By importing confusion matrix calculate precision and Recall to Plot the graph.

a) Complete Python Code for algorithms

```
# linear algebra
Import numpy as np
# data processing
import pandas as pd
# data visualization
Import seaborn as sns
%matplotlib inline
from matplotlib import pyplot as plt
# Algorithms
from sklearn.ensemble import
RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC, LinearSVC
test_df = pd.read_csv("test.csv")
train_df = pd.read_csv("train.csv")
train_df.describe()
#what data is actually missing
total =
train_df.isnull().sum().sort_values(ascending=False)
percent_1 =
train_df.isnull().sum()/train_df.isnull().count()*100
percent_2 = (round(percent_1,
1)).sort_values(ascending=False)
missing_data = pd.concat([total, percent_2], axis=1,
keys=['Total', '%'])
missing_data.head(5)
ax.legend()
_ = ax.set_title('Male')
```

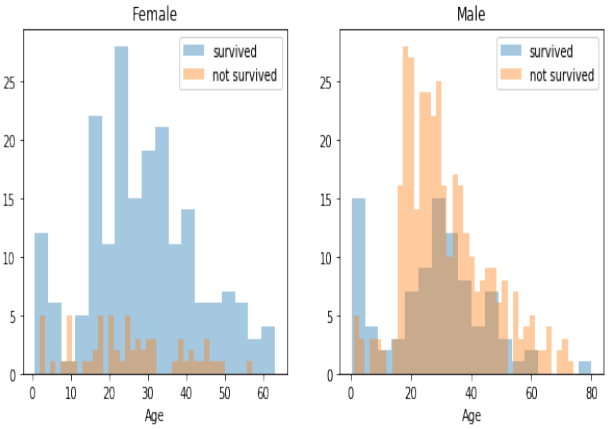


Fig. 4: Gender wise survival representation

```
#Embarked seems to be correlated with survival,
sns.barplot(x='Pclass', y='Survived', data=train_df)
for dataset in data:
    # extract titles
dataset['Title'] = dataset.Name.str.extract(' ([A-Za-
z]+)\.', expand=False)
dataset['Title'] = dataset['Title'].replace('Mlle', 'Miss')
    # convert titles into numbers
dataset['Title'] = dataset['Title'].map(titles)
    # filling NaN with 0, to get safe
dataset['Title'] = dataset['Title'].fillna(0)
# Let's take a last look at the training set, before we start
training the models.
train_df.head(5)
```

b) Building Machine Learning Models

```
X_train = train_df.drop("Survived", axis=1)
Y_train = train_df["Survived"]
X_test  =test_df.drop("PassengerId", axis=1).copy()
# Random Forest
random_forest =
RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, Y_train)
Y_prediction = random_forest.predict(X_test)
random_forest.score(X_train, Y_train)
```

```
acc_random_forest = round(random_forest.score(X_train, Y_train) * 100, 2)
print(round(acc_random_forest,2,), "%")
92.82 %
# Logistic Regression
logreg = LogisticRegression()
logreg.fit(X_train, Y_train)
Y_pred = logreg.predict(X_test)
acc_log = round(logreg.score(X_train, Y_train) * 100, 2)
print(round(acc_log,2,), "%")
82.04 %
# KNN
knn = KNeighborsClassifier(n_neighbors = 3)
knn.fit(X_train, Y_train)
Y_pred = knn.predict(X_test)
acc_knn = round(knn.score(X_train, Y_train) * 100, 2)
print(round(acc_knn,2,), "%")
85.75 %
# Decision Tree
decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, Y_train)
Y_pred = decision_tree.predict(X_test)
acc_decision_tree = round(decision_tree.score(X_train, Y_train) * 100, 2)
print(round(acc_decision_tree,2,), "%")
92.82 %
results = pd.DataFrame({
    'Model': ['Support Vector Machines', 'KNN', 'Logistic Regression',
              'Random Forest', 'Naive Bayes',
              'Decision Tree'],
    'Score': [acc_linear_svc, acc_knn, acc_log,
acc_random_forest, acc_gaussian, acc_perceptron,
acc_sgd, acc_decision_tree]})
result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
O/P-
Model Score:
Random Forest 92.82
Decision Tree  92.82
KNN85.75
Logistic Regression 82.04
Support Vector Machines 77.89
```

c) Confusion Matrix

```
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import confusion_matrix
predictions = cross_val_predict(random_forest, X_train, Y_train, cv=3)
confusion_matrix(Y_train, predictions)
O/P-
array([[490,  59],
[ 87, 255]])
print("Precision:", precision_score(Y_train, predictions))
print("Recall:",recall_score(Y_train, predictions))
Precision: 0.812101910828
Recall: 0.745614035088
```

d) Precision Recall Curve

```
From sklearn.metrics import precision_recall_curve
# getting the probabilities of our predictions
y_scores = random_forest.predict_proba(X_train)
```

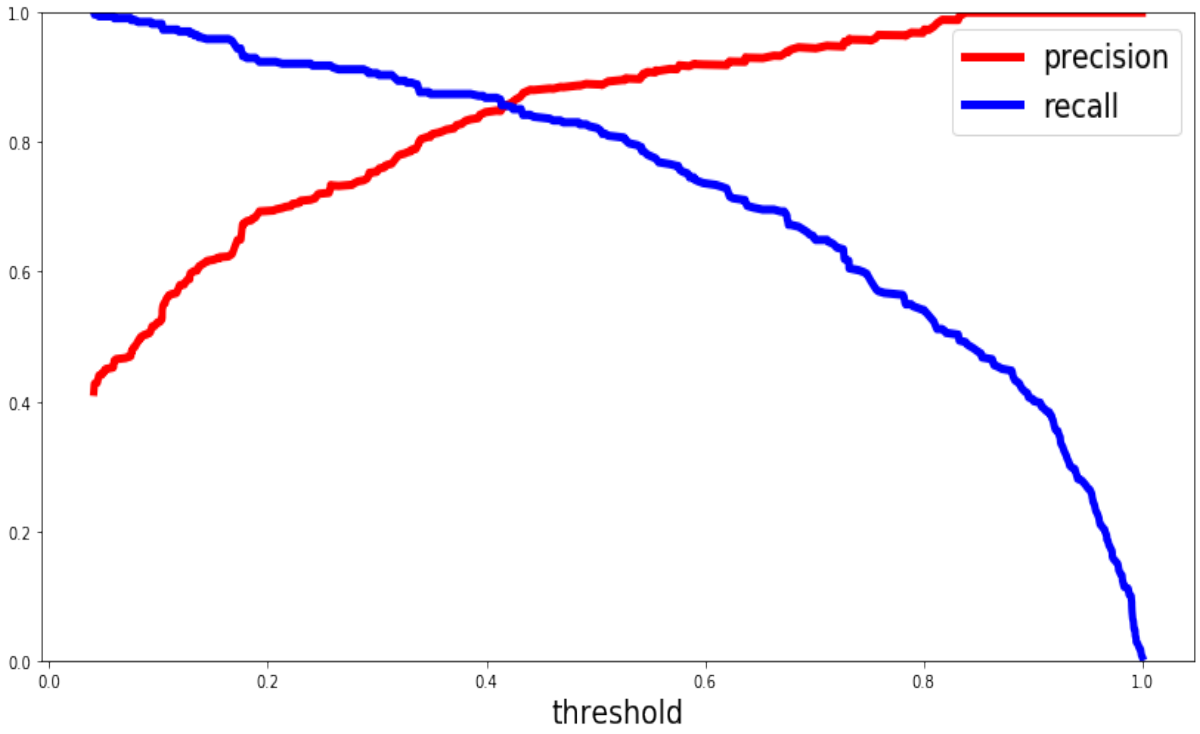


Fig. 5: Precision and Recall graph

```
defplot_precision_vs_recall(precision, recall):
plt.ylabel("recall", fontsize=19)
plot_precision_vs_recall(precision, recall
plt.show()
```

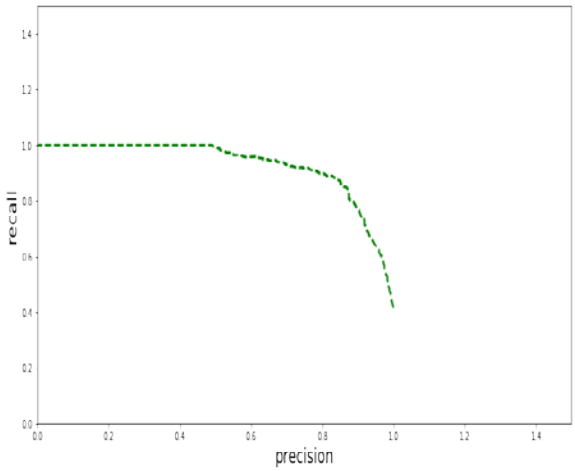


Fig. 6: ROC

V. Result Analysis

By using the above code, we have already calculated the accuracy of each algorithm. Now by using confusion matrix, we will reckon how many numbers are correctly.

n=192	**Predicted: 0**	**Predicted: 1**	
Actual: 0	TN = 118	FP = 12	130
Actual: 1	FN = 47	TP = 15	62
	165	27	

Fig. 7: Confusion Matrix Example

Predicted i.e. Precision will be = TP /(TP + FP) and Recall will be TP /(TP + FN).

Where, TP = Total positive prediction, FP = False positive and FN = False negative.

As per our result, we got Precision as 0.812101910828 and Recall as 0.745614035088. So our models have predicted 81% accurately. From the results, we got both random forest, and decision tree is giving high accuracy.

Algorithms name	Accuracy in %
Random Forest	92.82
Decision Tree	92.82
KNN	85.75
Logistic Regression	82.04
Support Vector Machines	77.89

Fig. 8: Algorithm and Percentage of accuracy

VI. Conclusion

Here we have studied the basic about machine learning, linear regression, logistic regression, SVM, KNN, Decision tree and Random forest tree algorithm. We have executed the code by using python language and got the output successfully by using Confusion matrix, Precision-recall curve. At the end, we have calculated Random forest, and decision tree model are giving a higher accuracy of 92.82 % of data by using modules from scikit learn. As the objective was for knowing all these five algorithms and code execution which is computed with accuracy. We have also performed confusion matrix, for result analysis and got the result by getting the Precision and Recall value.

References Références Referencias

1. The Tragedy of Titanic: A Logistic Regression Analysis. Dina Ahmed Mohamed Ghandour1 and May Alawi Mohamed Abdalla2.
2. A Comparative Analysis on Linear Regression and Support Vector Regression Kavitha S Assistant Professor Computer Science and Engineering Bannari Amman Institute of Technolgy Sathyamangalamkvth.sgm@gmail.com
3. An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain Park, Hyeoun-Ae College of Nursing and System Biomedical Informatics National Core Research Center, Seoul National University, Seoul, Korea.
4. Bagley, S. C., White, H., &Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology, 54(10), 979-985. Bewick, V., Cheek, L., & Ball, J. (2004).
5. Statistics review 13: Receiver operating characteristic curves. Critical Care (London, England), 8(6), 508512. http://dx.doi.org/10.1186/cc3000
6. Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. Higher Education: Handbook of Theory and Research, 8, 379–410. 2. Bagley, S. C., White, H., & Golomb, B. A. (2001).
7. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology, 54(10), 979-9853
8. Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms Tryambak Chatterjee* Department of Management Studies, NIT Trichy, Tiruchirappalli, Tamilnadu, India.
9. GE, "Flight Quest Challenge," Kaggle.com. [Online]. Available: https://www.kaggle.com/c/flight2-final. [Accessed: 2-Jun-2017].
10. "Titanic: Machine Learning from Disaster," Kaggle.com. [Online]. Available: https://www.kaggle.com/c/titanic-gettingStarted. [Accessed: 2-Jun-2017]. [3] Wiki, "Titanic." [Online]. Available: http://en.wikipedia.org/wiki/Titanic.[Accessed: 2-Jun-2017].
11. Kaggle, Data Science Community, [Online]. Available: http://www.kaggle.com/[Accessed:2-Jun-2017.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 2 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Requirement Elicitation Model (REM) in the Context of Global Software Development

By Muhammad Yaseen & Umar Farooq

University of Engineering & Technology

Abstract- *Contxext:* Requirement elicitation is difficult and critical phase of requirement engineering and the case is worst in global software development (GSD). The study is about requirement elicitation in the context of GSD.

Objective: Development of requirement elicitation model (REM) which can address the factors that have positive impact and the factors that have negative impact during elicitation in GSD. The propose model will give solutions and practices to the challenges during elicitation.

Method: Systematic literature review (SLR) and empirical research study will be used for achieving the goals and objectives.

Expected Outcomes: The expected results of this study will be REM that will help vendor organizations for better elicitation during GSD.

Keywords: *systematic literature review, requirement engineering, global software development, requirement elicitation model.*

GJCST-C Classification: *D.2.1*

Strictly as per the compliance and regulations of:

Requirement Elicitation Model (REM) in the Context of Global Software Development

Muhammad Yaseen [α] & Umar Farooq [σ]

Abstract- *Context:* Requirement elicitation is difficult and critical phase of requirement engineering and the case is worst in global software development (GSD). The study is about requirement elicitation in the context of GSD.

Objective: Development of requirement elicitation model (REM) which can address the factors that have positive impact and the factors that have negative impact during elicitation in GSD. The propose model will give solutions and practices to the challenges during elicitation.

Method: Systematic literature review (SLR) and empirical research study will be used for achieving the goals and objectives.

Expected Outcomes: The expected results of this study will be REM that will help vendor organizations for better elicitation during GSD.

Keywords: *systematic literature review, requirement engineering, global software development, requirement elicitation model.*

I. Introduction

Requirement engineering (RE) is the systematic and discipline way of collecting user requirements for a software system and to manage it[1, 2]. The purpose of RE process is satisfaction of user needs and what customer wants from a software product [2-5].Requirement elicitation is the first phase of RE during software development life cycle[6]. During elicitation phase we do direct communication with users or customers and gather requirements by applying various elicitation techniques. The quality of software system is more depended on the quality of how better the requirements are gathered[1, 7]. RE is very difficult when we do it locally but the case very difficult when the development is done globally where clients and vendors are separated by distance and face challenges like culture issues, time zone difference and difference in languages and terminologies. Due to extra challenges in GSD proper elicitation process is affected[8, 9].

II. Motivation and Related Work

Miguel Romero [10]discuss that culture and time differences are big challenges in GSD. For reducing the effect of challenges it is necessary for people to share knowledge about requirements and have a proper knowledge management system. The author describes relevant skills for elicitation in GSD are; English language skills, understanding of cultures of others, computer mediated communication skills, use of proper communication protocols, ability to resolve conflicts and teamwork skills are needed for effective elicitation.

Nosheen Sabahat[6] after doing survey and interviews explains the effectiveness of elicitation techniques. She concluded that in GSD the most effective elicitation techniques are prototyping, scenarios and interview. The prototype is considered to be the best technique in GSD because prototypes represents product earlier and customers are more satisfied. Scenarios are best in case where prototypes are difficult to make. Questionnaires and other traditional techniques are not suitable in GSD. The author propose iterative model where elicitation and analysis works iteratively at same time.

Bin Wen [11] discuss that traditional elicitation techniques are not enough to apply in GSD during collecting of requirements, collaborative techniques like social intelligent for networked software and semantic web technology should be selected during elicitation.

Gabriela N. Aranda [12]in his paper suggests the strategies to overcome the challenges like lack of face to face meetings and culture issues in GSD. Culture differences cannot be ignored but stakeholders can learn about these differences and training is the best solution for that. The author has explained in detail about the trainings and its strategies. Use of ontologies is suggested as best way to reduce language differences. Using ontologies can clarify the structure of knowledge and as well as reduce conceptual and terminological confusion. Technology selection must be discussed in team while doing elicitation because technology selection process is carried out by studying and confronting the personal preferences of people who need to work together.

Fabio Calefato [13] discuss that as face to face communication during elicitation in GSD is difficult but still it is possible to have systems and technologies that can be alternative to face to face communication. After empirical studies the author design computer mediated tool for synchronous text solution in case where face to face communication is difficult. The tool contains audio

Author α: University of Engineering & Technology, Peshawar.
e-mail: Yaseen_cse11@yahoo.com
Author σ: Universiti Tun Hussein Onn Malaysia.
e-mail: umarfarooq.ktk@gmail.com

and video conferencing facility. The tool was further evaluated through case study using students.

Neetu Kumari. S [3] proposes the model which will address the issues with elicitation in GSD but the levels of this model are not defined. The model is limited to find the challenges only not the solutions. The characteristics of model are not explained. Further the methodology of collecting data from literature is not systematic so we need advance model which can address the challenges, critical success factors and the solutions for problems and challenges.

III. Objectives

The objective of this research study is to develop REM in order to support vendor organizations in better elicitation of requirements in the context of GSD. This model will address challenges and success factors during requirement elicitation. The propose model will address the solutions and practices needed to better implement success factors during elicitation and to reduce the effect of the challenges. For achieving the objectives, we will do the SLR and empirical study to find factors that are important during elicitation of requirements in GSD. For the implementation of factors, practices and solutions will be extracted through SLR and then questionnaire survey will be conducted to validate success factors, challenges and practices. Survey will be conducted in software industry to mention some new practices not mentioned in literature before. After finding the practices and solutions the propose model will be developed. The aim is to reduce the gap between researchers and software development vendors. Other researchers also adopted the same methodology in other fields to suggest such models[14].

IV. Research Questions

The work reported in this paper is based on the five research questions which have posted in the following way:

RQ1: What are the factors as identified from literature that have positive impact during requirement elicitation phase of RE in the context of GSD?

RQ2: What are the factors in real practice that have positive impact during requirement elicitation phase of RE in the context of GSD?

RQ3: What are the factors and challenges as identified from literature that have negative impact during requirement elicitation phase of RE in the context of GSD?

RQ4: What are the factors in real practice that have negative impact during requirement elicitation phase of RE in the context of GSD?

RQ5: Are there differences between the factors identified through the literature and the real-world practice?

RQ6: Is the REM practically successful in terms of finding and alleviate implementation factors and challenges faced by vendor organizations during elicitation in GSD?

Through SLR, RQ1 and RQ3 will be address i.e. studying what factors (CSFs and Cs) have already been reported in the literature. In future in order to facilitate vendors in implementing factors important for success of requirement elicitation, we will program/code the REM in the form of software. Moreover for overcoming challenges faced by vendor organizations in GSD the REM will suggest solutions. This tool will produce different assessment reports and do different activities for the software vendor organizations.

V. Design of REM and Research Methodology

The methodology of the REM consists of the following three phases.

Phase#1: SLR will be conducted for data collection.

Phase#2: Empirical study will be conducted to validate the result of SLR and to find the practices for the mentioned factors.

Phase#3: For evaluation and validation of REM, Case study will be conducted. To explain the aforesaid three phases the following subsections are added.

a) Collection of data and its analysis

i. *CSFs (critical success factors):* Factors that have a positive impact during requirement elicitation in GSD.

ii. *CRs (critical Risks):* Factors that have a negative impact during requirement elicitation in GSD.

iii. *Practices:* For implementing CSFs, practices will be extracted and used.

SLR will be used to identify factors (CSFs and Cs). Through SLR we will extract, analyse and will explore data relevant to our research questions. SLR is different from ordinary literature reviews being formally planned and more systematic. According to Kitchenham [15] SLR is divided into 3 phases. First phase is planning the review, second is conducting the review and implementation is the last phase. Before conducting SLR a protocol will be designed which include all the steps needed for SLR. From research questions a search string will be constructed for different libraries accordingly. Search procedure and plan will be defined and then protocol will be executed. After execution, inclusion and exclusion criteria will be define to tell which paper to include in final list.

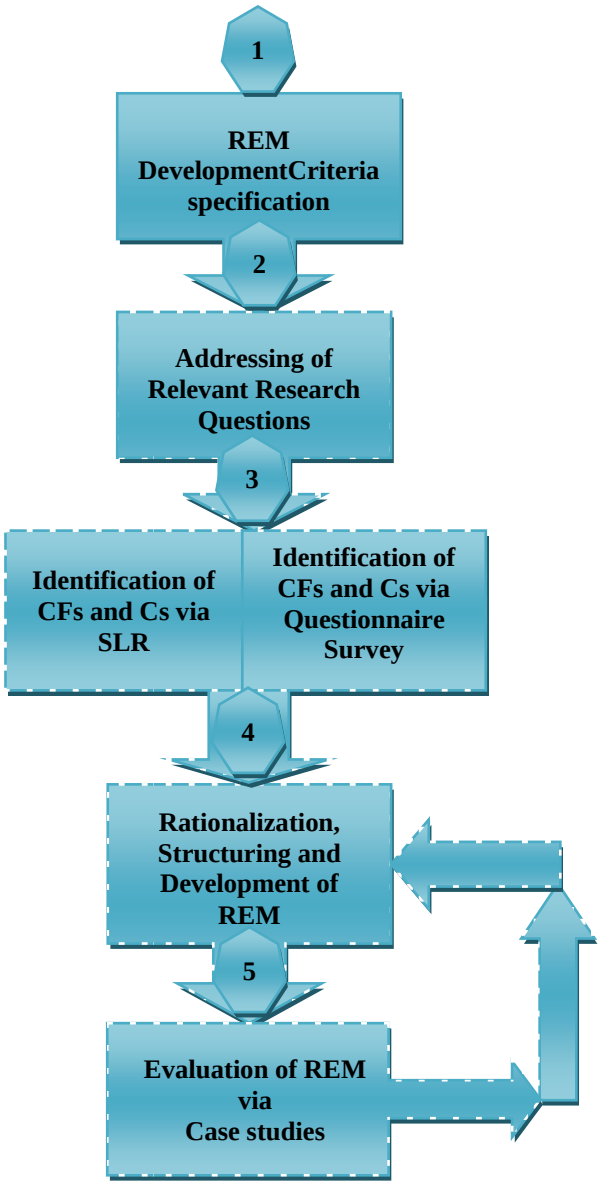


Figure 1: REM Development Cycle Activities

To successfully implement factors (CSFs and Cs), a questionnaire survey will be conducted with experts working in the software industry. The purpose of this survey is:

- Validation of the results of SLR.
- To find new factors (CSFs and Cs) which are not previously identified.
- To identify practices for the success implementation of (CSFs and Cs).

After REM is successfully design then for evaluation in real world environment case studies will be conducted in software industry.

b) Development of REM

For the design of REM we have used the five stages as shown in Figure 2. A similar approach has also been used by another researcher[16, 17].

The development of REM is the first stage; it is used to set criteria for REM success. The below mentioned two criteria will be used for the development and assessment of REM.

A. *User satisfaction:* This criteria focus on the satisfaction of end user from the result of REM. He/ She should be able to use the REM without any confusion or ambiguity to promote objectives according to their requirements and assumptions.

B. *Usability:* This criteria emphasis on the structure easiness of the REM. It states that the structure of the REM should be flexible and easy to understand because organizations do not accept complex models and standards which require resources, training and effort.

Data collection and analysis is the stage 2. Rationalization and structuring of results will be performed in stage 3. Development of REM based on the results of empirical studies in stage4. Evaluation and validation of the REM via case studies will be performed in final stage i.e. stage 5.

Planned structure of the REM is shown in Figure 3. Relationship between REM levels is also shown, factors/risks and practices used to address risks and success implementation of RE process.

c) REM structure

We will build the structure of the REM on the bases of following three extensions.

- REM levels
- Factors (CSFs, Cs) in each level
- Practices and solutions for the implementation of factors

Classification of CSFs and Cs in different categories will be the base for defining the levels of REM. Each level will be consisting of different factors (CSFs and Cs). For each factor in the particular level, practices will be given for its proper implementation. Like CMMI and other models, for organizations to achieve certain level they must address and should follow the practices for each CSFs and Cs under that particular level.

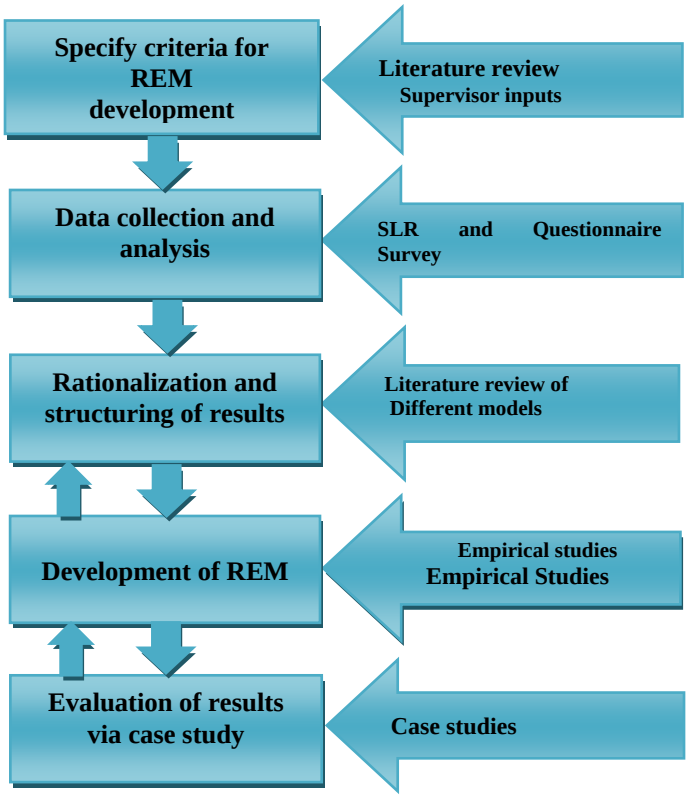


Figure 2: REM development stages

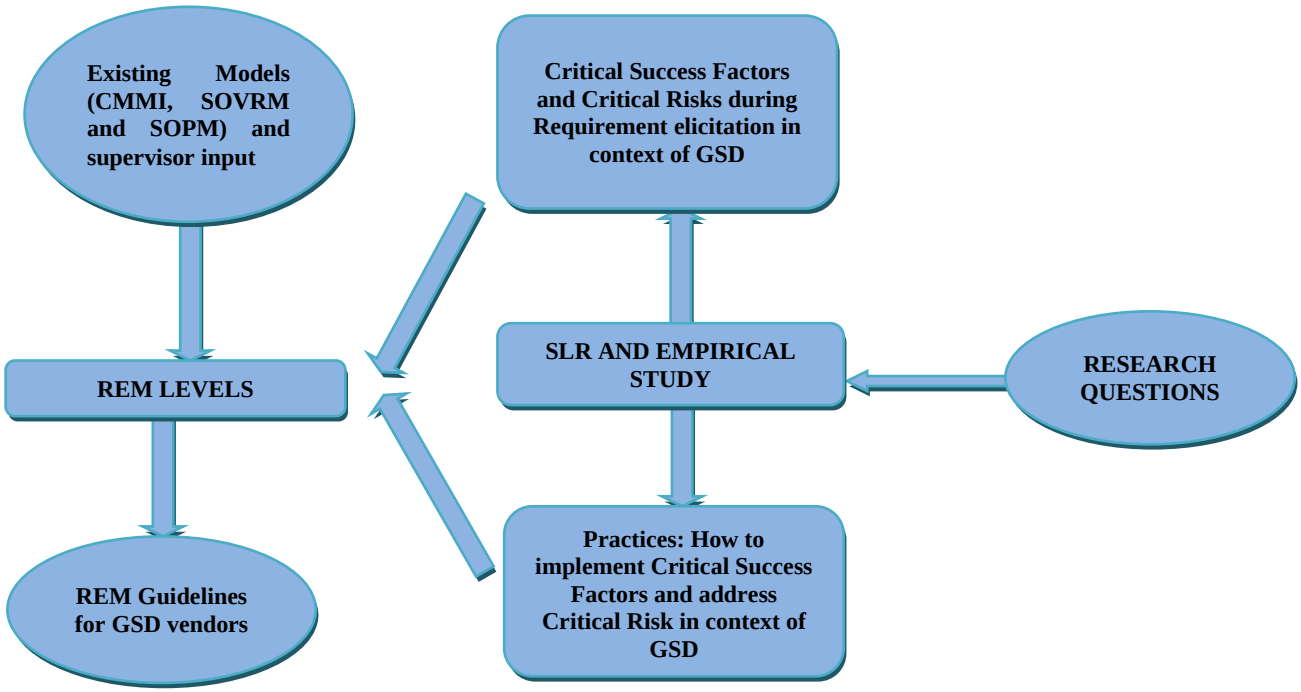


Figure 3: REM Structure

d) REM evaluation

REM will be validated through industrial case studies. For case studies maximum of five organizations will be enough. Case study is the best tool for evaluating any model in real environment. A focus group session will be arranged to get feedback from the participant about REM. The criteria will be ease of use and user satisfaction as discussed in section 5.2. In focus group evaluation two or more people interact and generate ideas without getting help from researcher. Focus group is more open as compared to individual interview.

VI. Research Carried up to Date

We have done the following research work so far:

- Identification of problem and objectives
- Research questions specification
- Selection of research methodology
- Defining structure of REM
- Evaluation method selection
- Conduct of SLR

VII. Conclusion and Future Work

In our paper we have presented the structure of REM with different levels and phases. We have discussed how this model will help vendors in better implementation of success factors during elicitation in GSD. Detail methodology for the development of REM was introduced. This model will be used as tool for software developers and will produce different assessment reports in different situations.

Acknowledgment

We are thankful to all members of the Software Engineering Research Group (SERG-UOM) at University of Malakand for their constructive reviews and also thankful to the anonymous reviewers of the conference.

References Références Referencias

1. Pandey, U. Suman, and A. Ramani, "An effective requirement engineering process model for software development and requirements management," presented at Advances in Recent Technologies in Communication and Computing (ARTCom), 2010 International Conference on, 2010.
2. W. J. Lloyd, M. B. Rosson, and J. D. Arthur, "Effectiveness of elicitation techniques in distributed requirements engineering," presented at Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on, 2002.
3. S. Neetu Kumari and A. S. Pillai, "A survey on global requirements elicitation issues and proposed research framework," presented at Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on, 2013.
4. M. Kauppinen, M. Vartiainen, J. Kontio, S. Kujala, and R. Sulonen, "Implementing requirements engineering processes throughout organizations: success factors and challenges," Information and Software Technology, vol. 46, pp. 937-953, 2004.
5. B. Arthi, "Distributed Requirements Negotiations Using Mixed Media," Intâ€™l Journal of Eng. and Technol, vol. 1, 2009.
6. N. Sabahat, F. Iqbal, F. Azam, and M. Y. Javed, "An iterative approach for global requirements elicitation: A case study analysis," presented at Electronics and Information Engineering (ICEIE), 2010 International Conference On, 2010.
7. A. Ahmad, A. Shahzad, V. K. Padmanabhuni, A. Mansoor, S. Joseph, and Z. Arshad, "Requirements prioritization with respect to Geographically Distributed Stakeholders," presented at Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on, 2011.
8. M. Yaseen, S. Baseer, and S. Sherin, "Critical challenges for requirement implementation in context of global software development: A systematic literature review," presented at 2015 International Conference on Open Source Systems & Technologies (ICOSST), 2015.
9. T. Illes-Seifert, A. Herrmann, M. Geisser, and T. Hildenbrand, "The Challenges of Distributed Software Engineering and Requirements Engineering: Results of an Online Survey," presented at WORKSHOP P, 2007.
10. M. Romero, A. VizcaÃ-no, and M. Piattini, "Teaching requirements elicitation within the context of global software development," presented at Computer Science (ENC), 2009 Mexican International Conference on, 2009.
11. B. Wen, Z. Luo, and P. Liang, "Distributed and Collaborative Requirements Elicitation based on Social Intelligence," presented at Web Information Systems and Applications Conference (WISA), 2012 Ninth, 2012.
12. G. N. Aranda, A. VizcaÃ-no, A. Cechich, and M. Piattini, "Strategies to minimize problems in global requirements elicitation," CLEI electronic journal, vol. 11, 2008.
13. F. Calefato, D. Damian, and F. Lanubile, "Computer-mediated communication to support distributed requirements elicitations and negotiations tasks," Empirical Software Engineering, vol. 17, pp. 640-674, 2012.
14. S. Ali and S. U. Khan, "SOFTWARE OUTSOURCING PARTNERSHIP MODEL," Science International, vol. 26, 2014.
15. B. Kitchenham and C. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering, Keele University and Durham University Joint Report," EBSE 2007-001, 2007.
16. S. Ali and S. U. Khan, "Software Outsourcing Partnership (SOP): A Systematic Literature Review Protocol with Preliminary Results," International Journal of Hybrid Information Technology, vol. 7, pp. 377-392, 2014.
17. M. Yaseen, S. Baseer, S. Ali, and S. U. Khan, "Requirement Implementation Model (RIM) in the Context of Global Software Development."

Global Journals Guidelines Handbook 2018

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards "FARSC" title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

- The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.

Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSS can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSS, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.

The FARSS will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSS member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.
Once you are designated as FARSS, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSS members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSS member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSS is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSS can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSS member can decide its price and we can help in making the right decision.

The FARSS member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN SCIENCE (MARSS)

The ' MARSS ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.

The “MARSS” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSS or William Walldroff, M.S., MARSS.

MARSS accrediting is an honor. It authenticates your research activities. After becoming MARSS, you can add 'MARSS' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSS designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSS, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders,Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Auxiliary Memberships

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.

The author fees of such paper may be waived off up to 40%.

The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

After nomination of your institution as "Institutional Fellow" and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

"

- In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
- In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
- In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone.

"

Preferred Author Guidelines

We accept the manuscript submissions in any standard (generic) format.

We typeset manuscripts using advanced typesetting tools like Adobe In Design, CorelDraw, TeXnicCenter, and TeXStudio. We usually recommend authors submit their research using any standard format they are comfortable with, and let Global Journals do the rest.

Alternatively, you can download our basic template from https://globaljournals.org/Template.zip

Authors should submit their complete paper/article, including text illustrations, graphics, conclusions, artwork, and tables. Authors who are not able to submit manuscript using the form above can email the manuscript department at submit@globaljournals.org or get in touch with chiefeditor@globaljournals.org if they wish to send the abstract before submission.

Before and during Submission

Authors must ensure the information provided during the submission of a paper is authentic. Please go through the following checklist before submitting:

1. Authors must go through the complete author guideline and understand and *agree to Global Journals' ethics and code of conduct,* along with author responsibilities.
2. Authors must accept the privacy policy, terms, and conditions of Global Journals.
3. Ensure corresponding author's email address and postal address are accurate and reachable.
4. Manuscript to be submitted must include keywords, an abstract, a paper title, co-author(s') names and details (email address, name, phone number, and institution), figures and illustrations in vector format including appropriate captions, tables, including titles and footnotes, a conclusion, results, acknowledgments and references.
5. Authors should submit paper in a ZIP archive if any supplementary files are required along with the paper.
6. Proper permissions must be acquired for the use of any copyrighted material.
7. Manuscript submitted *must not have been submitted or published elsewhere* and all authors must be aware of the submission.

Declaration of Conflicts of Interest

It is required for authors to declare all financial, institutional, and personal relationships with other individuals and organizations that could influence (bias) their research.

Policy on Plagiarism

Plagiarism is not acceptable in Global Journals submissions at all.

Plagiarized content will not be considered for publication. We reserve the right to inform authors' institutions about plagiarism detected either before or after publication. If plagiarism is identified, we will follow COPE guidelines:

Authors are solely responsible for all the plagiarism that is found. The author must not fabricate, falsify or plagiarize existing research data. The following, if copied, will be considered plagiarism:

- Words (language)
- Ideas
- Findings
- Writings
- Diagrams
- Graphs
- Illustrations
- Lectures

- Printed material
- Graphic representations
- Computer programs
- Electronic material
- Any other original work

Authorship Policies

Global Journals follows the definition of authorship set up by the Open Association of Research Society, USA. According to its guidelines, authorship criteria must be based on:

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

Changes in Authorship

The corresponding author should mention the name and complete details of all co-authors during submission and in manuscript. We support addition, rearrangement, manipulation, and deletions in authors list till the early view publication of the journal. We expect that corresponding author will notify all co-authors of submission. We follow COPE guidelines for changes in authorship.

Copyright

During submission of the manuscript, the author is confirming an exclusive license agreement with Global Journals which gives Global Journals the authority to reproduce, reuse, and republish authors' research. We also believe in flexible copyright terms where copyright may remain with authors/employers/institutions as well. Contact your editor after acceptance to choose your copyright policy. You may follow this form for copyright transfers.

Appealing Decisions

Unless specified in the notification, the Editorial Board's decision on publication of the paper is final and cannot be appealed before making the major change in the manuscript.

Acknowledgments

Contributors to the research other than authors credited should be mentioned in Acknowledgments. The source of funding for the research can be included. Suppliers of resources may be mentioned along with their addresses.

Declaration of funding sources

Global Journals is in partnership with various universities, laboratories, and other institutions worldwide in the research domain. Authors are requested to disclose their source of funding during every stage of their research, such as making analysis, performing laboratory operations, computing data, and using institutional resources, from writing an article to its submission. This will also help authors to get reimbursements by requesting an open access publication letter from Global Journals and submitting to the respective funding source.

Preparing your Manuscript

Authors can submit papers and articles in an acceptable file format: MS Word (doc, docx), LaTeX (.tex, .zip or .rar including all of your files), Adobe PDF (.pdf), rich text format (.rtf), simple text document (.txt), Open Document Text (.odt), and Apple Pages (.pages). Our professional layout editors will format the entire paper according to our official guidelines. This is one of the highlights of publishing with Global Journals—authors should not be concerned about the formatting of their paper. Global Journals accepts articles and manuscripts in every major language, be it Spanish, Chinese, Japanese, Portuguese, Russian, French, German, Dutch, Italian, Greek, or any other national language, but the title, subtitle, and abstract should be in English. This will facilitate indexing and the pre-peer review process.

The following is the official style and template developed for publication of a research paper. Authors are not required to follow this style during the submission of the paper. It is just for reference purposes.

Manuscript Style Instruction (Optional)

- Microsoft Word Document Setting Instructions.
- Font type of all text should be Swis721 Lt BT.
- Page size: 8.27" x 11'", left margin: 0.65, right margin: 0.65, bottom margin: 0.75.
- Paper title should be in one column of font size 24.
- Author name in font size of 11 in one column.
- Abstract: font size 9 with the word "Abstract" in bold italics.
- Main text: font size 10 with two justified columns.
- Two columns with equal column width of 3.38 and spacing of 0.2.
- First character must be three lines drop-capped.
- The paragraph before spacing of 1 pt and after of 0 pt.
- Line spacing of 1 pt.
- Large images must be in one column.
- The names of first main headings (Heading 1) must be in Roman font, capital letters, and font size of 10.
- The names of second main headings (Heading 2) must not include numbers and must be in italics with a font size of 10.

Structure and Format of Manuscript

The recommended size of an original research paper is under 15,000 words and review papers under 7,000 words. Research articles should be less than 10,000 words. Research papers are usually longer than review papers. Review papers are reports of significant research (typically less than 7,000 words, including tables, figures, and references)

A research paper must include:

a) A title which should be relevant to the theme of the paper.
b) A summary, known as an abstract (less than 150 words), containing the major results and conclusions.
c) Up to 10 keywords that precisely identify the paper's subject, purpose, and focus.
d) An introduction, giving fundamental background objectives.
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.
f) Results which should be presented concisely by well-designed tables and figures.
g) Suitable statistical data should also be given.
h) All data must have been gathered with attention to numerical detail in the planning stage.

Design has been recognized to be essential to experiments for a considerable time, and the editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned unrefereed.

i) Discussion should cover implications and consequences and not just recapitulate the results; conclusions should also be summarized.
j) There should be brief acknowledgments.
k) There ought to be references in the conventional format. Global Journals recommends APA format.

Authors should carefully consider the preparation of papers to ensure that they communicate effectively. Papers are much more likely to be accepted if they are carefully designed and laid out, contain few or no errors, are summarizing, and follow instructions. They will also be published with much fewer delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and suggestions to improve brevity.

Format Structure

It is necessary that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

All manuscripts submitted to Global Journals should include:

Title

The title page must carry an informative 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) where the work was carried out.

Author details

The full postal address of any related author(s) must be specified.

Abstract

The abstract is the foundation of the research paper. It should be clear and concise and must contain the objective of the paper and inferences drawn. It is advised to not include big mathematical equations or complicated jargon.

Many researchers searching for information online will use search engines such as Google, Yahoo or others. By optimizing your paper for search engines, you will amplify the chance of someone finding it. In turn, this will make it more likely to be viewed and cited in further works. Global Journals has compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Keywords

A major lynchpin of research work for the writing of research papers is the keyword search, which one will employ to find both library and internet resources. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining, and indexing.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy: planning of a list of possible keywords and phrases to try.

Choice of the main keywords is the first tool of writing a research paper. Research paper writing is an art. Keyword search should be as strategic as possible.

One should start brainstorming lists of potential keywords before even beginning 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 a research paper?” Then consider synonyms for the important words.

It may take the discovery of only one important paper to steer in the right keyword direction because, in most databases, the keywords under which a research paper is abstracted are listed with the paper.

Numerical Methods

Numerical methods used should be transparent and, where appropriate, supported by references.

Abbreviations

Authors must list all the abbreviations used in the paper at the end of the paper or in a separate table before using them.

Formulas and equations

Authors are advised to submit any mathematical equation using either MathJax, KaTeX, or LaTeX, or in a very high-quality image.

Tables, Figures, and Figure Legends

Tables: Tables should be 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. Authors must submit tables in an editable format and not as images. References to these tables (if any) must be mentioned accurately.

Figures

Figures are supposed to be submitted as separate files. Always include a citation in the text for each figure using Arabic numbers, e.g., Fig. 4. Artwork must be submitted online in vector electronic form or by emailing it.

Preparation of Eletronic Figures for Publication

Although low-quality images are sufficient for review purposes, print publication requires high-quality images to prevent the final product being blurred or fuzzy. Submit (possibly by e-mail) EPS (line art) or TIFF (halftone/ photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Avoid using pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings). 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: Authors are advised 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. Also, you can email your editor to remove the color fee after acceptance of the paper.

Tips for writing a good quality Computer Science Research Paper

Techniques for writing a good quality computer science research paper:

1. Choosing the topic: In most cases, the topic is selected by the interests of the author, but it can also be suggested by the guides. You can have several topics, and then judge which you are most comfortable with. This may be done by asking several questions of yourself, like "Will I be able to carry out a search in this area? Will I find all necessary resources to accomplish the search? Will I be able to find all information in this field area?" If the answer to this type of question is "yes," then you ought to choose that topic. In most cases, you may have to conduct surveys and visit several places. Also, you might have to do a lot of work to find all the rises and falls of the various data on that subject. Sometimes, detailed information plays a vital role, instead of short information. Evaluators are human: The first thing to remember is that evaluators are also human beings. They are not only meant for rejecting a paper. They are here to evaluate your paper. So present your best aspect.

2. Think like evaluators: If you are in confusion or getting demotivated because your paper may not be accepted by the evaluators, then think, and try to evaluate your paper like an evaluator. Try to understand what an evaluator wants in your research paper, and you will automatically have your answer. 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.

3. Ask your guides: If you are having any difficulty with your research, then do not hesitate to share your difficulty with your guide (if you have one). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work, then ask your supervisor to help you with an alternative. He or she might also provide you with a list of essential readings.

4. Use of computer is recommended: As you are doing research in the field of computer science then this point is quite obvious. Use right software: Always use good quality software packages. If you are not capable of judging good software, then you can lose the quality of your paper unknowingly. There are various programs available to help you which you can get through the internet.

5. Use the internet for help: An excellent start for your paper is using Google. It is a wondrous search engine, where you can have your doubts resolved. You may also read some answers for the frequent question of how to write your research paper or find a model research paper. You can download books from the internet. If you have all the required books, place importance on reading, selecting, and analyzing the specified information. Then sketch out your research paper. Use big pictures: You may use encyclopedias like Wikipedia to get pictures with the best resolution. At Global Journals, you should strictly follow here.

6. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right? It is a good habit which helps to not lose your continuity. You should always use bookmarks while searching on the internet also, which will make your search easier.

7. Revise what you wrote: When you write anything, always read it, summarize it, and then finalize it.

8. Make every effort: Make every effort to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in the introduction—what is the need for a particular research paper. Polish your work with good writing skills and always give an evaluator what he wants. Make backups: When you are going to do any important thing like making a research paper, you should always have backup copies of it either on your computer or on paper. This protects you from losing any portion of your important data.

9. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several unnecessary diagrams will degrade the quality of your paper by creating a hodgepodge. So always try to include diagrams which were made by you to improve the readability of your paper. Use of direct quotes: When you do research relevant to literature, history, or current affairs, then use of quotes becomes essential, but if the study is relevant to science, use of quotes is not preferable.

10.Use proper verb tense: Use proper verb tenses in your paper. Use past tense to present those events that have happened. Use present tense to indicate events that are going on. Use future tense to indicate events that will happen in the future. Use of wrong tenses will confuse the evaluator. Avoid sentences that are incomplete.

11. Pick a good study spot: Always try to pick a spot for your research which is quiet. Not every spot is good for studying.

12. Know what you know: Always try to know what you know by making objectives, otherwise you will be confused and unable to achieve your target.

13. Use good grammar: Always use good grammar and words that will have a positive impact on the evaluator; use of good vocabulary does not mean using tough words which the evaluator has to find in a dictionary. Do not fragment sentences. Eliminate one-word sentences. Do not ever use a big word when a smaller one would suffice.

Verbs have to be in agreement with their subjects. In a research paper, do not start sentences with conjunctions or finish them with prepositions. When writing formally, it is advisable to never split an infinitive because someone will (wrongly) complain. Avoid clichés like a disease. Always shun irritating alliteration. Use language which is simple and straightforward. Put together a neat summary.

14. 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 for your topic. You may also maintain your arguments with records.

15. Never start at the last minute: Always allow enough time for research work. Leaving everything to the last minute will degrade your paper and spoil your work.

16. Multitasking in research is not good: Doing several things at the same time is a bad habit in the case of research activity. Research is an area where everything has a particular time slot. Divide your research work into parts, and do a particular part in a particular time slot.

17. Never copy others' work: Never copy others' work and give it your name because if the evaluator has seen it anywhere, you will be in trouble. Take proper rest and food: No matter how many hours you spend on your research activity, if you are not taking care of your health, then all your efforts will have been in vain. For quality research, take proper rest and food.

18. Go to seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

19. Refresh your mind after intervals: Try to give your mind a rest by listening to soft music or sleeping in intervals. This will also improve your memory. Acquire colleagues: Always try to acquire colleagues. No matter how sharp you are, if you acquire colleagues, they can give you ideas which will be helpful to your research.

20. Think technically: Always think technically. If anything happens, search for its reasons, benefits, and demerits. Think and then print: When you go to print your paper, check that tables are not split, headings are not detached from their descriptions, and page sequence is maintained.

21. Adding unnecessary information: Do not add unnecessary information like "I have used MS Excel to draw graphs." Irrelevant and inappropriate material is superfluous. Foreign terminology and phrases are not apropos. One should never take a broad view. Analogy is like feathers on a snake. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Never oversimplify: When adding material to your research paper, never go for oversimplification; this will definitely irritate the evaluator. Be specific. Never use rhythmic redundancies. Contractions shouldn't be used in a research paper. Comparisons are as terrible as clichés. Give up ampersands, abbreviations, and so on. Remove commas that are not necessary. Parenthetical words should be between brackets or commas. Understatement is always the best way to put forward earth-shaking thoughts. Give a detailed literary review.

22. Report concluded results: Use concluded results. From raw data, filter the results, and then conclude your studies based on measurements and observations taken. An appropriate number of decimal places should be used. Parenthetical remarks are prohibited here. Proofread carefully at the final stage. At the end, give an outline to your arguments. Spot perspectives of further study of the subject. Justify your conclusion at the bottom sufficiently, which will probably include examples.

23. Upon conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print for 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 of 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 criteria peer reviewers will use for grading the final paper.

Final points:

One purpose of organizing a research paper is to let people interpret your efforts selectively. The journal requires the following sections, submitted in the order listed, with each section starting on a new page:

The introduction: This will be compiled from reference matter and reflect the design processes or outline of basis that directed you to make a study. As you carry out the process of study, the method and process section will be constructed like that. The results segment will show related statistics in nearly sequential order and direct reviewers to similar intellectual paths throughout the data that you gathered to carry out your study.

The discussion section:

This will provide understanding of the data and projections as to the implications of the results. The use of good quality references throughout the paper will give the effort trustworthiness by representing an alertness to 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 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 avoid:

- Insertion of a title at the foot of a page with subsequent text on the next page.
- Separating a table, chart, or figure—confine each to a single page.
- Submitting a manuscript with pages out of sequence.
- In every section of your document, use standard writing style, including articles ("a" and "the").
- Keep paying attention to the topic of the paper.
- Use paragraphs to split each significant point (excluding the abstract).
- Align the primary line of each section.
- Present your points in sound order.
- Use present tense to report well-accepted matters.
- Use past tense to describe specific results.
- Do not use familiar wording; don't address the reviewer directly. Don't use slang or superlatives.
- Avoid use of extra pictures—include only those figures essential to presenting results.

Title page:

Choose a revealing title. It should be short and include the name(s) and address(es) of all authors. It should not have acronyms or abbreviations or exceed two printed lines.

Abstract: This summary should be two hundred words or less. It should clearly and briefly explain the key findings reported in the manuscript and must have precise statistics. It should not have acronyms or abbreviations. It should be logical in itself. Do not cite 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 approaches 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. Use comprehensive sentences, and do not sacrifice readability for brevity; you can maintain it succinctly by phrasing sentences so that they provide more than a 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 limit the initial two items to no more than one line each.

Reason for writing the article—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 for this; results of any numerical analysis should be reported. Significant conclusions or questions that emerge from the research.

Approach:

- Single section and succinct.
- An outline of the job done is always written in past tense.
- Concentrate on shortening results—limit background information to a verdict or two.
- Exact spelling, clarity 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 of comprehending and calculating the purpose of your study without having to refer to other works. The basis for the study should be offered. Give the most important references, but avoid making a comprehensive appraisal of the topic. Describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will give no attention to your results. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here.

The following approach can create a valuable beginning:

- Explain the value (significance) of the study.
- Defend the model—why did you employ this particular system or method? What is its compensation? Remark upon its appropriateness from an abstract point of view as well as pointing out sensible reasons for using it.
- Present a justification. State your particular theory(-ies) or aim(s), and describe the logic that led you to choose them.
- Briefly explain the study's tentative purpose and how it meets 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 for every section. If you make the four points listed above, you will need at least four paragraphs. Present surrounding information only when it is necessary to support a situation. The reviewer does not desire to read everything you know about a topic. Shape the theory 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 soundly written procedures segment allows a capable scientist to replicate 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 to give the least amount of information that would permit another capable scientist to replicate 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 section, mention the specific item describing the way, but draw the basic principle while stating the situation. The purpose is to show 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:

Materials may be reported in part of a section or else they may be recognized along with your measures.

Methods:

- Report the method and not the 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—detail how procedures were completed, not how they were performed on a particular day.
- If well-known procedures were used, account for the procedure by name, possibly with a reference, and that's all.

Approach:

It is embarrassing to use vigorous voice when documenting methods without using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result, when writing up the methods, most authors use third person passive voice.

Use standard style in this and 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 as 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. Use statistics and tables, if suitable, to present consequences most efficiently.

You must clearly differentiate material which would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matters should not be submitted at all except if requested by the instructor.

Content:

- Sum up your conclusions in text and demonstrate them, if suitable, with figures and tables.
- In the manuscript, explain each of your consequences, and point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation of an exacting study.
- Explain results of control experiments and give 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 manuscript.

What to stay away from:

- Do not discuss or infer your outcome, report surrounding information, or try to explain anything.
- Do not include raw data or intermediate calculations in a research manuscript.
- Do not present similar data more than once.
- A manuscript should complement any figures or tables, not duplicate information.
- Never confuse figures with tables—there is a difference.

Approach:

As always, use past tense when you submit 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 section.

Figures and tables:

If you put figures and tables at the end of some details, make certain that they are visibly distinguished from any attached appendix materials, such as raw facts. Whatever the position, each table must be titled, numbered one after the other, and include a heading. All figures and tables must be divided from the text.

Discussion:

The discussion is expected to be the trickiest segment to write. A lot of papers submitted to the journal are discarded based on problems with the discussion. There is no rule for how long an 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 implications of the study. The purpose here is to offer an understanding of your results and support all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of results should be fully 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 the prospect, and let it drop at that. Make a decision as to whether each premise is supported or 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 an experiment might be personalized to accomplish a new idea.
- Give details of all of your remarks as much as possible, focusing on mechanisms.
- Make a decision as to whether the tentative design sufficiently addressed the theory and whether or not it was correctly restricted. Try to present substitute explanations if they are sensible alternatives.
- One piece of 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 other available information. Present work done by specific persons (including you) in past tense.

Describe generally acknowledged facts and main beliefs in present tense.

The Administration Rules

Administration Rules to Be Strictly Followed before Submitting Your Research Paper to Global Journals Inc.

Please read the following rules and regulations carefully before submitting your research paper to Global Journals Inc. to avoid rejection.

Segment draft and final research paper: You have to strictly follow the template of a research paper, failing which your paper may get rejected. You are expected to write each part of the paper wholly on your own. The peer reviewers need to identify your own perspective of the concepts in your own terms. Please do not extract straight from any other source, and do not rephrase someone else's analysis. Do not allow anyone else to proofread your manuscript.

Written material: You may discuss this with your guides and key sources. Do not copy anyone else's paper, even if this is only imitation, otherwise it will be rejected on the grounds of plagiarism, which is illegal. Various methods to avoid plagiarism are strictly applied by us to every paper, and, if found guilty, you may be blacklisted, which could affect your career adversely. To guard yourself and others from possible illegal use, please do not permit anyone to use or even read your paper and file.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350