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

# Global Journal

## of Computer Science and Technology: C

# Software & Data Engineering

Mobile Learning Framework

Representative Fictional Identities

Highlights

Critical Analysis of Solutions

Analyzing and Designing of Algorithms

Discovering Thoughts, Inventing Future

VOLUME 23 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 23 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. Stefano Berretti*

Ph.D. in Computer Engineering and Telecommunications, University of Firenze Professor Department of Information Engineering, University of Firenze, Italy

### *Dr. Abdurrahman Arslanyilmaz*

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

### *Dr. Lamri Sayad*

Ph.d in Computer science, University of BEJAIA, Algeria

### *Dr. Tauqeer Ahmad Usmani*

Ph.D in Computer Science, Oman

### *Dr. Hazra Imran*

Ph.D in Computer Science (Information Retrieval), Athabasca University, Canada

### *Dr. Magdy Shayboub Ali*

Ph.D in Computer Sciences, MSc in Computer Sciences and Engineering, BSc in Electronic Engineering, Suez Canal University, Egypt

### *Dr. Nurul Akmar Binti Emran*

Ph.D in Computer Science, MSc in Computer Science, Universiti Teknikal Malaysia Melaka, Malaysia

### *Dr. Asim Sinan Yuksel*

Ph.D in Computer Engineering, M.Sc., B.Eng., Suleyman Demirel University, Turkey

### *Dr. Anis Bey*

Dept. of Computer Science, Badji Mokhtar-Annaba University, Annaba, Algeria

### *Alessandra Lumini*

Associate Researcher Department of Computer Science and Engineering University of Bologna Italy

### *Dr. Rajesh Kumar Rolen*

Ph.D in Computer Science, MCA & BCA - IGNOU, MCTS & MCP - MIcrosoft, SCJP - Sun Microsystems, Singhania University, India

### *Dr. Rajneesh Kumar Gujral*

Ph.D in Computer Science and Engineering, M.TECH in Information Technology, B. E. in Computer Science and Engineering, CCNA Certified Network Instructor, Diploma Course in Computer Servicing and Maintenance (DCS), Maharishi Markandeshwar University Mullana, India

### *Dr. Aziz M. Barbar*

Ph.D. IEEE Senior Member Chairperson, Department of Computer Science AUST - American University of Science & Technology Alfred Naccash Avenue Ashrafieh, Lebanon

### *Dr. Federico Tramarin*

Ph.D., Computer Engineering and Networks Group, Institute of Electronics, Italy Department of Information Engineering of the University of Padova, Italy

### *Dr. Chutisant Kerdvibulvech*

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

### *Dr. Roheet Bhatnagar*

Ph.D in Computer Science, B.Tech in Computer Science, M.Tech in Remote Sensing, Sikkim Manipal University, India

# Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# A Novel Frequent Pattern Mining Algorithm for Evaluating Applicability of a Mobile Learning Framework

By D. D. M. Dolawattha & H. K. Salinda Premadasa
*University of Kelaniya*

*Abstract-* The applicability of a mobile learning system reflects how it works in an actual situation under diverse conditions. In previous studies, researches for evaluating applicability in learning systems using data mining approaches are challenging to find. The main objective of this study is to evaluate the applicability of the proposed mobile learning framework. This framework consists of seven independent variables and their influencing factors. Initially, 1000 students and teachers were allowed to use the mobile learning system developed based on the proposed mobile learning framework. The authors implemented the system using Moodle mobile learning environment and used its transaction log file for evaluation. Transactional records that were generated due to various user activities with the facilities integrated into the system were extracted. These activities were classified under eight different features, i.e., chat, forum, quiz, assignment, book, video, game, and app usage in thousand transactional rows.

*Keywords: system applicability, mobile learning, frequent pattern mining, apriori algorithm, fp growth algorithm.*

*GJCST-C Classification: ACM: H.2.8, H.3.3*


*Strictly as per the compliance and regulations of:*

# A Novel Frequent Pattern Mining Algorithm for Evaluating Applicability of a Mobile Learning Framework

D. D. M. Dolawattha [α] & H. K. Salinda Premadasa [σ]

*Abstract-* The applicability of a mobile learning system reflects how it works in an actual situation under diverse conditions. In previous studies, researches for evaluating applicability in learning systems using data mining approaches are challenging to find. The main objective of this study is to evaluate the applicability of the proposed mobile learning framework. This framework consists of seven independent variables and their influencing factors. Initially, 1000 students and teachers were allowed to use the mobile learning system developed based on the proposed mobile learning framework. The authors implemented the system using Moodle mobile learning environment and used its transaction log file for evaluation. Transactional records that were generated due to various user activities with the facilities integrated into the system were extracted. These activities were classified under eight different features, i.e., chat, forum, quiz, assignment, book, video, game, and app usage in thousand transactional rows. A novel pattern mining algorithm, namely Binary Total for Pattern Mining (BTPM), was developed using the above transactional dataset's binary incidence matrix format to test the system applicability. Similarly, Apriori frequent itemsets mining and Frequent Pattern (FP) Growth mining algorithms were applied to the same dataset to predict system applicability. The results reveal that the proposed pattern mining algorithm provides 82% assurance of system applicability with significant efficiency. In comparison, the Apriori and Frequent Pattern (FP)-Growth algorithms offer about 60% assurance of system applicability.

*Keywords: system applicability, mobile learning, frequent pattern mining, apriori algorithm, fp growth algorithm.*

## I. Introduction

Mobile learning (ML) emerged as an essential and successful learning method due to its flexibility of time and place for learning in the present globe. On the other hand, learners and teachers have an interactive and cost-effective learning environment for carrying out educational activities with cutting-edge technology advancements. However, at the same time, learners and teachers experience various limitations in this learning method, such as the quality of the learning applications. Therefore, it is vital to evaluate ML applications in various dimensions, such as their ML applicability [1]. The Cambridge Oxford Dictionary defines applicability as "the fact of affecting or relating to a person or thing." Applicability definitions for domains related to computer systems are minimal. Hence we found a specific definition for the public health sector domain, and it defines applicability as "the extent to which an intervention process could be implemented in another" [2]. On the other hand, transferability gives a meaning similar to applicability. It defines transferability as "the extent to which the measured effectiveness of an applicable intervention could be achieved in another setting". Rosemann and Vessey (2008) proposed the significance and capability to check the applicability of a research framework with three factors, i.e., importance, accessibility, and suitability [3]. Accordingly, we can recognize an applicable mobile learning system (MLS) as "the extent to which a MLS could be implemented effectively in a target environment as well as in another environment with a similar setting". This study's main objective is to evaluate the applicability of the proposed mobile learning framework (MLF). This framework consists of seven independent variables, and each variable has different influencing factors. Evaluating this MLF for applicability would help realize the proposed framework's applicability when it works in a real-world environment. Normally, pattern mining algorithms are used to describe patterns in usage behaviors [4]. Hence, we use pattern mining to study users' usage of different learning tools in the system, such as chat, forum, quizzes, assignments, etc. Proper understanding of these usages would help calculate how many users apply these tools in their academic activities in the proposed system. Considering each tool's usage would support predicting the system's overall applicability (Fig. 3: motivating example). Because if a learning system has better applicability, its users should fully practice the tools integrated into it. In this study, the authors intend to analyze the transactional log of the MLS, developed using the proposed MLF for evaluating applicability. According to the previous studies, the FP-Growth[5] efficient pattern mining and Apriori [6] popular association rule mining algorithms have been utilized to evaluate many automated systems. FP-Growth algorithm was used to evaluate systems to identify important segments in learning materials [7], learning

*Author α: University of Kelaniya, Dalugama, Kelaniya, Sri Lanka.*
*e-mail: salinda@ccs.sab.ac.lk*
*Author σ: Sabaragamuwa University of Sri Lanka, Belihuloya , Sri Lanka*

behaviors[8], best course modules[9], learners requirements[10], offensive activities, and illegal transactions[11]. In these studies, the FP-Growth algorithm helps to improve the learning process, course selection[9], course recommendation[10], security enhancement, and legal judgments [11] in learning systems and portals. On the other hand, the Apriori algorithm was used to evaluate systems to identify, information for decision making[12], tourists attractive places[13], course administration history[9], study preferences, archived cyber-attacks, and network hackings[14], requirements in software [15], health clinical information in clinical[16], and deformation states of landslides[17]. In these studies, the Apriori algorithm help in different activities such as better administrative decisions[12], improving tourist attraction[13], enhancing course facilities, offering productive subjects[9], preventing networks attacks[14], recommend software requirements[15], enhanced health decision and treatments[16], and predict landslides accurately[17]. But it is a very shortage of using FP Growth or Apriori algorithms in the evaluation applicability of learning systems. Also, difficult to find such applicability evaluation using any statistical approach too in previous studies. Moreover, these two well-established algorithms have no direct provisions for evaluating the applicability of the system. Nevertheless, using the rules and frequent patterns generated by Apriori and FP Growth algorithms can predict applicability. In this study, the authors proposed a novel frequent pattern mining algorithm for applicability evaluation directly in the proposed learning system to address this research gap. Also, FP Growth and Apriori algorithm are used to predict the applicability of the same system.

## II. Related Work

Pattern mining is a popular data mining tactic to emerge secret knowledge in data stores, and it is applied for solving issues arising in various scenarios. A Pattern mining approach was proposed to identify the most critical or complex learning segments in video tutorials. The proposed method integrates a learning model that learns the above-considered components of the analysed video transaction log. [7]. A pattern mining framework was proposed to identify learners' different learning behaviors and improve their learning process with the institutional learning system. This study's output reveals that the learners obtained significant study performance in different learning modes via various tools integrated into the system [8]. Another pattern mining framework was proposed to analyse huge databases by addressing existing pattern mining algorithms such as a large number of searching iterations, excessive space for processing, and too much time requirements. Results reveal that this approach can solve different kinds of pattern mining problems[18]. An online course recommender system was proposed using the FP-Growth algorithm to guide learners to select learning courses according to their preferences. The investigation displays that the system has better efficiency in instructing learners for selecting appropriate learning materials in their learning process [9]. FP-Growth-based data mining technique was used to promote educational services in educational institutes. Various variables related to learners and educational institutes were employed on FP-Tree to determine regular data items. Research reveals that the best selection of attributes in the data for the algorithm gives better results[19]. FP-Growth algorithm was used to elaborate users' access patterns in learning portals. The study revealed, such as learners' favorite courses, less and high navigation areas of the learning site, and recommendations for advancing both learner's gain and user-friendliness of the site[10]. Wu and Zhang (2019) researched to extract support information to prove offensive actions. An improved FP-Growth algorithm analysed data associated with illegal transactions and supported legal judgment with better efficiency and accuracy[11].

Another method for hidden information recovery from large databases is associate rule mining, and its applications are spread in various researches in the past. The E-learning system was evaluated by combining the association rule mining method and the fuzzy analytic hierarchy process[20]. Apriori algorithm-based association rule mining was used to analyse the Wi-Fi data in attractive tourist places. The study results give the association rules of travel patterns of tourists' movements and enable further enhancements of tourist's magnetism of travel destinations[13]. Learner assisting study guides approval mechanism was proposed using association rule mining with archived course administration data. This approach queries useful relationships in learners' learning subject preferences[9]. An improved Apriori algorithm is offered to find fresh network attacks using the data that was produced by previous episodes. This method has optimal accuracy with superior efficiency for discovering cyber strikes[14]. A customized Apriori algorithm was used to improve the audit system's security building association rules using fewer scanning cycles in logs with shorter processing time[21]. Apriori algorithm-based recommender system was proposed to enhance the accuracy of requirements in software development [15]. Apriori algorithm was coustomized to excavate intelligence information from virtual reality applications for effective decision making[22]. Apriori algorithm was executed successfully in analyzing the deformation states of landslides [17]. A heart disease prediction model was proposed by applying the Apriori algorithm in clinical datasets[23]. Here, we used the Apriori

associate rule mining and FP-Growth frequent pattern mining algorithm-based approaches as comparable methods to the novel proposed pattern mining algorithm for evaluation applicability of the proposed MLS.

## III. Preliminaries

In this section, the authors discuss the proposed MLF and its implemented MLS, which will be evaluated for applicability. Also, the critical theories undergo related to the proposed solution for evaluating the system's applicability, such as associate rule mining, Apriori, and FP-Growth algorithms.

### a) *Mobile Learning Framework for Higher Education*

We can refer to many frameworks that were implemented in the Moodle environment successfully in previous studies. Halvoník and Kapusta (2020) implemented an e-Learning material composing framework through Moodle LMS with the teachers' highest inclination[24]. A framework for evaluating student learning was performed through the Moodle platform to create useful tests for learners[25]. Karagiannis and Satratzemi (2017) proposed a Moodle implemented framework for e-learning content adaptation according to learner wishes and better usability with decent learning outcomes[26]. Hence, according to the literature, several conceptual frameworks have been implemented via the Moodle mobile learning environment (MMLE). So in our study, we implemented the proposed conceptual MLFrame through the MMLE.

In this study, we require to measure the applicability of the MLS implemented based on the proposed MLF for higher education (MLFrame). This framework consists of seven independent variables: learner, teacher, mobile ML devices, ML tools, ML contents, higher education institutes, and communication technology (Fig. 1). Each independent variable has several influencing factors: motivation, usefulness, interactivity, ease of use, etc. These factors were realized through various resources and facilities embedded in the MLS such as chat, forum, games, quizzes, assignments, etc.[27],[28]. This MLFrame was implemented in the MMLE by integrating new facilities and modifying the existing Moodle mobile application. Therefore, existing Moodle plugins were enhanced to develop these new ML facilities for the Moodle mobile environment[28]. Because most of the facilities available in the Moodle learning environment are not implemented in the Moodle mobile environment. Hence, when implementing the MLFrame in the Moodle mobile environment, we customized the Moodle ML application by upgrading relevant plugins to serve the facilities introduced in the MLFrame.

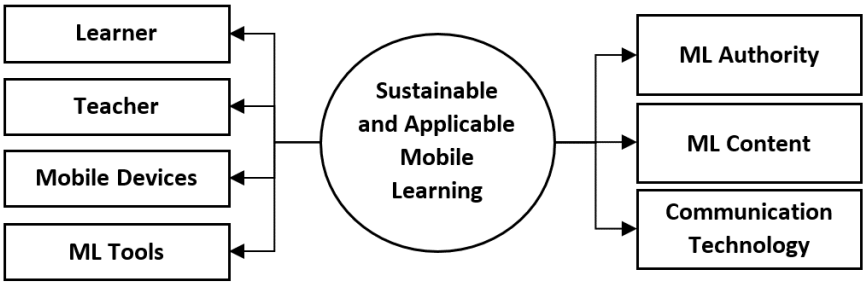



*Fig. 1:* Proposed MLF for higher education (MLFrame)

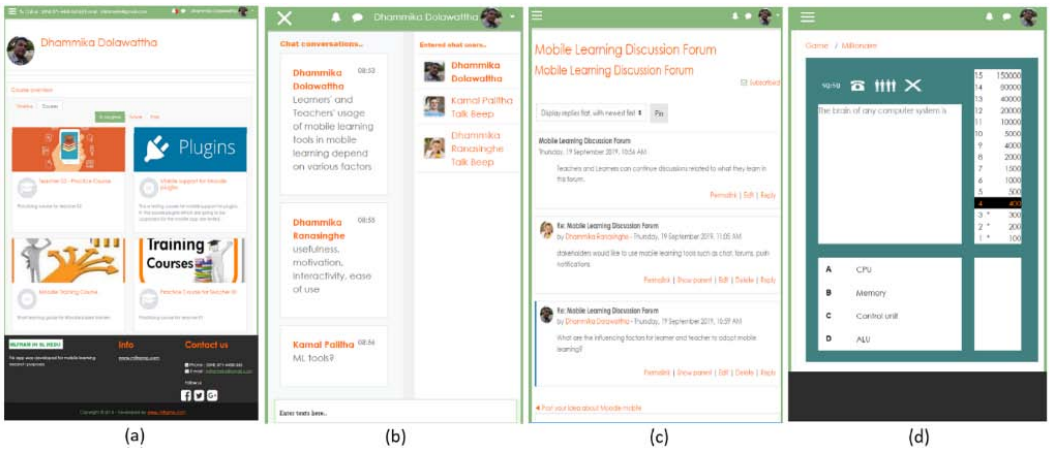



*Fig. 2:* Interfaces of Some of the Features Considered in the Study Such as (a) Course Overview, (b) Chat, (c) Forum, and (d) Game in the Modified Moodle Mobile App

*b) Motivating Example*

Fig. 3 depicts the concept associated with predicting the system applicability. Consider a MLS with facilities chatting, forum, quiz, assignment, book, video, game, and app_usage. If this app was allowed to use 100 particular users, then assume transaction log analysis as follows. 8,7,6,5 features out of all 8 features used by 60,10,15,5 users, respectively. Therefore, 70 users out of 100 users used at least 7 features out of all 8 features (or 88% features). Hence system applicability was 70% for 88% feature usage. Further, 85 users out of 100 users used at least 6 features out of 8 features (or 75% features). Hence system applicability is 85% for 75% features usage. Similarly, it can be realized that the system applicability is 90% for 63% features usage.

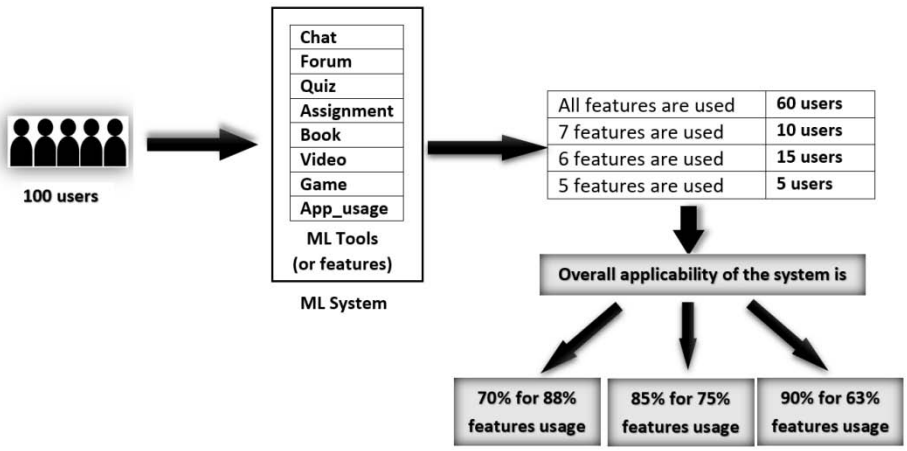



*Fig. 3:* Example for Predicting Applicability using Pattern Mining Algorithm

*c) Associate Rule Mining*

Association rule mining is used to find essential associations among stored data in large databases. 'Antecedent' and 'Consequent' are significant two fractions in association rules. These are data items finding in databases, and 'Antecedent' combines with the 'Consequent'. Moreover, finding frequent itemsets is a crucial requirement to mine association rules from databases. Additionally, two important factors, such as support and confidence, must be defined when finding association rules using these frequent itemsets[29].

*d) Apriori Algorithm*

Apriori is a Latin term which denotes "from what comes before". Bottom-up and breadth-first search strategies are taken into account. Agarwal and Srikant (1994) developed the Apriori algorithm for generating associate rules by frequent pattern mining. The main terminologies used in the Apriori algorithm are Min_supp, Min_conf, Frequent itemsets, Apriori Property, Join Operation, Join Step, and Prune Step [30].

*e) Frequent Pattern (FP) Growth Algorithm*

FP-Growth is a widespread pattern mining algorithm used in data mining. In this algorithm, frequent patterns are stored in a tree-like data structure called FP-tree. The algorithm's main steps are calculating each database item's support count by scanning, deleting irregular patterns, and order remains. Then FP-tree is constructed and frequent patterns are generated using FP-tree [5]. This improves the frequent pattern mining technique because it scans the database only twice. There is no candidate set generation, though it is not suitable for mining patterns in databases that are updated frequently[31].

## IV. Proposed Method

Details of the proposed novel frequent pattern mining algorithm, namely the Binary Total for Pattern Mining (BTPM) algorithm, which evaluates the proposed MLF's applicability, are described in this section.

*a) Process of the Proposed Algorithm*

The processing steps of the proposed algorithm implementation are described below.

*Procedure 1:* Purpose is generating a features list (extracting unique features, used in this evaluation, stored in the transaction log of the system)

*Step 01 & 09:* For loop iterates from 1st row to last row of the transactional database.

*Step 02 & 08:* For loop iterates from 1st item to the last item of a transactional database row.

*Step 03 & 07:* For loop iterates from 1st item to the last item of the array call Array Features. Array Features is an array that contains unique features of the transactional database.

*Step 04:* Check whether each item does not exist in the array Array Features ($f_m$).

*Step 05:* If the new item is not in the array, the item is saved in Array Features as the last item ($f_{s+1}$). Likewise,

Array Features consists of an array of distinct features in the transactional database.

*Procedure 2:* Purpose is creating BIMF Data Set(The usage transaction table needs to be converted into the *B*inary *I*ncidence *M*atrix *F*ormat (BIMF))

*Step 10:* Creating an array calls Array Data Set BIMF[] for saving the transactional database in binary incidence format array size is:
t x s; t means the number of transactions in the transactional database;s means the number of features in the array Features[]. Assign 0 for all the array elements.

*Step 11 & 19:* For loop iterates from 1st row to last row of the transactional database

*Step 12 & 18:* For loop iterates from 1st item to the last item of a transactional row

*Step 13 & 17:* For loop iterates from 1st item to the last item of the array called Array Features[].
(Note: Array Features[] contains distinct features of the transactional database)

*Step 14:* Check each item in the transaction database ($a_{ij}$) with feature items in Array Features ($f_m$)

*Step 15:* If a similar item is found, replace the value in the Array Dataset BIMF with 1. The position of the item (i.e. $d_{im}$) substituting in the Array Dataset BIMF is found by taking the associate item's row number of the transactional database ($i^{th}$row ) and column number of the Array Features ($m^{th}$ column).

Likewise, replace 0 with 1 in Array Dataset BIMF to denote similar features in the same column in Array Dataset BIMF for all items in the transactional database.

*Procedure 3:* Purposes are,

1) Saving BIMF dataset (ArrayDatasetBIMF[]) to an Array (ArrayTDVT[]) by creating tdvt and cnb. (tdvt: **T**otal **D**ecimal **V**alue of binary digit positions in entire **T**ransactions; cnb: **C**ount of the **N**on-zero **B**inary digits)
2) Selecting and Saving Patterns to ArrayPatterns

*Step 20:* Create arrays named ArrayTDVT[] and ArrayPatterns[]

*Step 21:* For loop iterates from 1st row to last row of the ArrayDatasetBIMF[]

*Step 22:* For loop iterate from 1st column to last column+2 ArrayDatasetBIMF[]

*Step 23:* Insert items (features) in Array DatasetBIMF[] to ArrayTDVT[] by row wise.

*Step 24:* Advance the inner loop

*Step 25:* Once particular row of Array DatasetBIMF[] is inserted to ArrayTDVT[], tdvt and cnb values are insertedto the same row in ArryTDVT[] as last tow column.

The second last element (*n+1*) of each row of the array is the **T**otal **D**ecimal **V**alue of binary digit positions in entire **T**ransactions ($tdvt_i$).

$tdvt_i = \sum_{j=0}^{n-1} d_{ij}\ 2^j$; where $d_{ij}$is the digit value of $j^{th}$ position in $i^{th}$ transaction i.e., $d_{ij}=\{0,1\}$

The last element (n+2) of each row of the array is the **C**ount of the *N*on-zero **B**inary digits in each transaction row ($cnb_i$).

$cnb_i = \sum_{j=0}^{n-1} d_{ij}$ ; where $d_{ij}$is the digit value of $j^{th}$ position in $i^{th}$ transaction i.e., $d_{ij}=\{0,1\}$

*Step 26:* At the end of each row, check whether the percentage of features used in a transaction is greater than or equal to minFUT. The minFUT is the given threshold value for the percentage of minimum features usage in a transaction. For instance, if minFUT = 70% for a 10 features transaction means, at least 7 features are used out of 10 features by a user.

(If j = n and $cnb_i(100/n)$ >= minFUT then)

If the above condition is satisfied, then save the transaction pattern to another array (i.e., ArrayPatterns).

In each row of ArrayPatterns[$tdvt_k$, pattern $Count_k$], $tdvt_k$ denotes total decimal values of binary digit positions in entire transactions. Feature usage transaction patterns can be derived by taking the binary conversion of the $tdvt_i$. And $patternCount_k$ gives the count of the number of occurrences of the pattern.

*Step 27:* For loop iterates from 1st row to last row of the ArrayPatterns[]

*Step 28:* Read existing tdvt values in the ArrayPatterns[]

*Step 30:* If the existing tdvt value equals the tdvt value, satisfy the conditions in Step 26, increase the patternCountvalues in the ArrayPatterns by one
ArrayPatterns[$tdvt_k$, $patternCount_k$= $patternCount_k$+ 1]

*Step 32:* If the existing tdvt value does not equal the tdvt value, satisfy the conditions in Step 26, add new row tothe ArrayPatterns

ArrayPatterns[$tdvt_{k+1}$ = $tdvt_i$, $patternCount_k$= 1]

*Procedure 4:* Purposes are generating patterns and their percentages

*Step 37:* For loop iterates from 1st row to last row of the ArrayPatterns[]

*Step 38:* Generate distinct transaction patterns by converting $tdvt_k$to its binary pattern whose percentage of featureusage in a transaction is greater than or equal to minFUT.

$k^{th}$ Pattern = CtoBinary($tdvt_k$)

$k^{th}$ pattern can be realize by considering both CtoBinary($tdvt_k$) and the ArrayFeatures[]. CtoBinary($tdvt_k$) and the ArrayFeatures[] array have same no of elements. We can identify the pattern by

replacing non-zero digits in CtoBinary($tdvt_k$) with the element in the same position of the ArrayFeatures[] and ignoring zeros.

*Step 39:* Percentage of each distinct usage transaction pattern can be calculated by using the equation,
Percentage of $k^{th}$ pattern = ($patternCount_k$/t) * 100

*Step 41:* Finally, the overall percentage of transactions whose FUT is greater than or equal to minFUT can becalculated by taking the percentage of summation of the element ArrayPatterns[$patternCount_k$]. i.e.

$$OPT = \left(\sum_{k=1}^{L} ArrayPatterns[patternCount_k]\right) \times \frac{1}{t} \times 100$$

Where L = No. of patterns = Number of rows in the ArrayPatterns, t = no. of transactions in the dataset

*Table 1:* Binary Total for Pattern Mining Algorithm (BTPM Algorithm)

**Algorithm – Binary Total for Pattern Mining Algorithm (BTPM Algorithm)**

Input *D:* = {$a_{ij}$} where i=1 to t (t denotes no of transactions or rows); j=1 to n (n denotes number of features in each transaction); $a_{ij}$ denotes $j^{th}$ feature of $i^{th}$ transaction in transactional database.
Input *minFUT*: Percentage of minimum features usage in a transaction
Output Patterns: Transaction patterns whose each feature usage is greater than or equal to minFUT
Output IPTs: Individual percentage of each transaction (IPT) pattern whose feature usage is greater than or equal to minFUT
Output OPT: Overall percentage of transactions whose individual transactions feature usage is greater than or equal to minFUT
Procedure 1: Generating Feature List
(1) For i=1 to t (t = no of transactions in D)
(2) For j=1 to n (n = no of features in $i^{th}$ transaction in D)
(3) For m=1 to s (s = no of items in ArrayFeatures)
(4) If $a_{ij}$<>ArrayFeature[$f_m$] then
(5) ArrayFeature[$f_{s+1}$=$a_{ij}$]
(6) End if
(7) Next m
(8) Next j
(9) Next i

Procedure 2: Creating binary incidence matrix format (BIMF) DataSet
(10) Create t x s ArrayDataSetBIMF[] with 0 value for each element
(11) For i=1 to t (t = no of transactions in D)
(12) For j=1 to n (n = no of features per transaction in D)
(13) For m=1 to s (s = no of items in ArrayFeatures)
(14) If $a_{ij}$ = ArrayFeatures[$f_m$] then
(15) update : ArrayDatasetBIMF[$d_{im}$=1]
(16) End if
(17) Next m
(18) Next j
(19) Next i

Procedure 3: Saving BIMF dataset (ArrayDatasetBIMF[]) to an Array (ArrayTDVT[]) by creating tdvt and cnb. Selecting and Saving Patterns to ArrayPatterns
(20) Create t x s+2 array names ArrayTDVT [] and create 1 x 2 array names ArrayPatterns []
(21) For i = 1 to t (t = no of rows in ArrayDatasetBIMF[])
(22) For j = 1 to s+2 (s = no of columns in ArrayDatasetBIMF[])
(23) insert item $d_{ij}$to ArrayTDVT[] on the position $i^{th}$ row and $j^{th}$ column
(24) Next j
(25) Insert $tdvt_i=\sum_{j=1}^{s} d_{ij}\, 2^{(j-1)}$, $cnb_i= \sum_{j=1}^{s} d_{ij}$ to ArrayTDVT as columns s+1 and s+2 in the $i^{th}$ row
(26) If j = s and $cnb_i$*(100/s) >= minFUT then
(27) For k = 1 to no of rows in ArrayPatterns
(28) Read ArrayPatterns[$tdvt_k$, $patternCount_k$]
(29) If ArrayPatterns[$tdvt_k$]= $tdvt_i$ then
(30) ArrayPatterns[$tdvt_k$, $patternCount_k$= $patternCount_k$+ 1]
(31) Else
(32) ArrayPatterns[$tdvt_{k+1}$ = $tdvt_i$, $patternCount_k$= 1]
(33) End if

(34) Next k
(35) End if
(36) Next i

Procedure 4: Generating patterns and their percentages
(37) For k=1 to number of raws in ArrayPatterns[tdvt$_k$,patternCount$_k$]
(38) $k^{th}$ Pattern = CtoBinary(tdvt$_k$)
(39) IPT of $k^{th}$ pattern = (patternCount$_k$/t) * 100
(40) Next k
(41) $\mathbf{OPT} = \left(\sum_{k=1}^{L} \boldsymbol{ArrayPatterns}[\boldsymbol{patternCount_k}]\right) \times \frac{\mathbf{1}}{\boldsymbol{t}} \times \mathbf{100}$;
Where L=number of patterns = row count of the ArrayPatterns.

*b) Algorithm Implementation with a Sample*

Fig. 4 depicts the generation of the required dataset by implementing the proposed algorithm to evaluate the applicability of MLFrame. In this example, five features were considered for the transaction database. Moodle tools usage and users tables of the log database of the system were used to create the transaction dataset.

*Step 01:* 'Transaction' table is created by copying user ids from the main user table and updating each user id from the tools usage log table in the database of the MLS.

*Step 02:* 'ArrayFeatures' table is created by inserting distinct features in the 'Transaction' table. As we use only 5 features, this ArrayFeature table consists of a row with five values.

*Step 03:* 'ArrayDatasetBIMF' consists 5 rows (transactions of 5 users) and 6 columns (the first column is for TID-transaction ids and the rest of the rows for features). The array consists of user ids and 5 tools (features) same order with the 'ArrayFeatures' in a particular row.

*Step 04:* The table 'ArrayTDVT' is created by adding two columns at the end to the 'ArrayDatasetBIMF'. New columns are 'TDVT,' and 'CNB' represents the decimal value of the binary values in the same row and the count of non-zero binary digits in the same column, respectively.

*Step 05:* Different feature usage patterns with frequency are stored in the array called 'ArrayPatterns'. Unique feature usage patterns are selected by satisfying the condition, each pattern's minimum number of features are greater than or equal to the minimum feature usage percentage (minFUT) supplied.

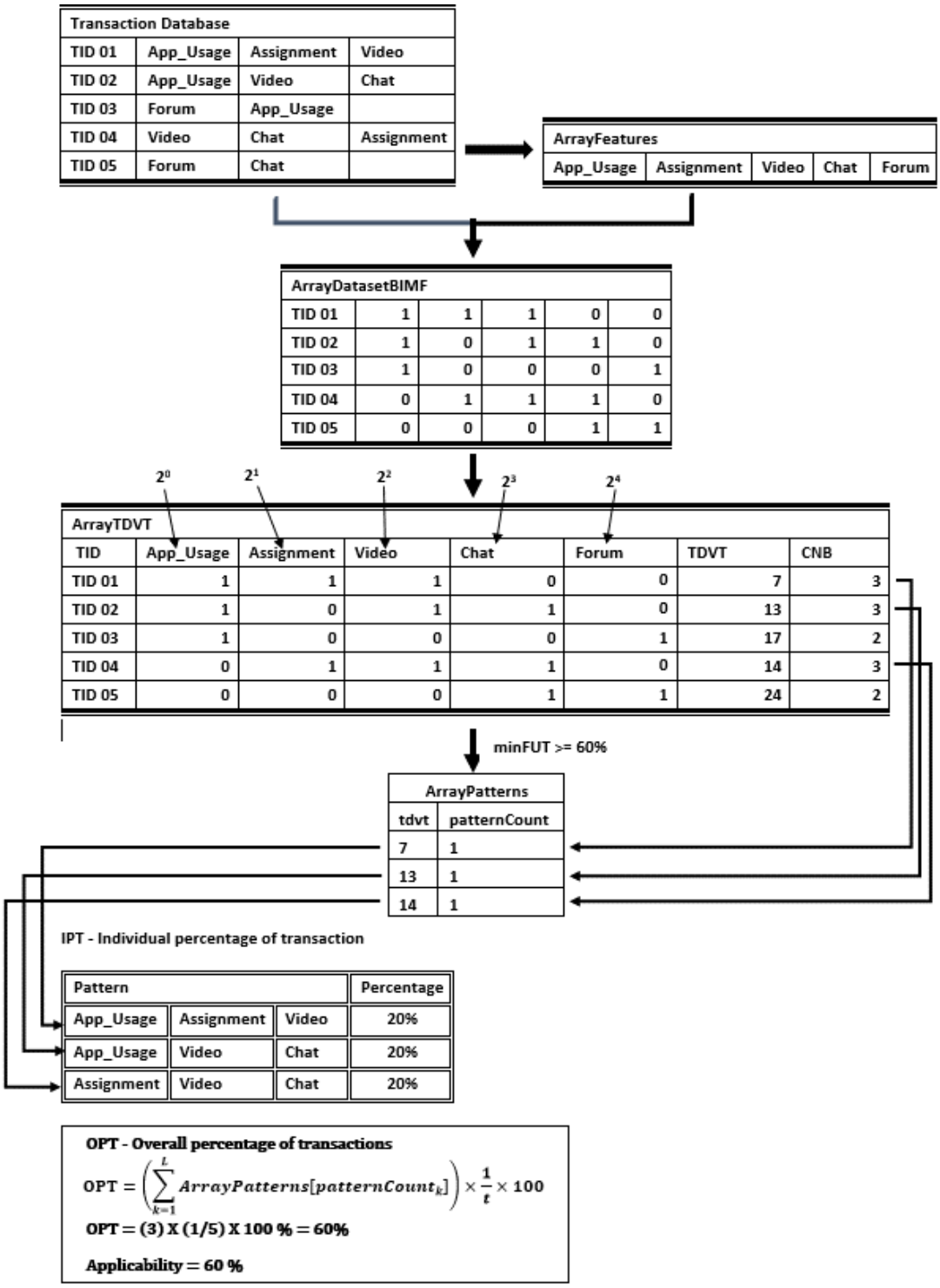



*Fig. 4:* A Sample Implementation of the Proposed Algorithm

*c) Comparison of Proposed BTPM Algorithm vs. Existing Pattern Mining Algorithms*

Popular algorithms for pattern mining such as Apriori and FP-Growth generate itemsets or candidate itemsets to find frequent itemsets. These itemsets are a group of transactional items that reside in the transaction database. Then, they use a minimum threshold value to minimize or prune itemsets to reduce data considered in mining. On the contrary, the BTPM algorithm uses an entirely mathematical technique, i.e., binary incidence matrix format of the transactional database with mathematical calculations. This reason minimizes memory usage and time for searching or traversals. Furthermore, in this method, the minimum

threshold value is used to reduce the transaction dataset simultaneously with the generation of required patterns. These causes reduce the computing load of the process.

## V. Methodology

The proposed MLS was given to precisely 1000 users, who are learners and teachers in the University of Kelaniya's four faculties. Among them, 220 students from each faculty of Science and Commerce & Management, and 200 students from each Faculties of Social Sciences and Humanities. Also, 160 teachers participated, and they consisted of 40 teachers in each faculty mentioned above. They were asked to use the system for around 50 days. This study was conducted according to the research framework illustrated in Fig. 5. The standard log file was extracted, and approximately half a million records were identified as different transactions related to the above user group on the given spell. Transactions were categorized into eight transactions with facilities integrated into the proposed ML application, i.e., chat, forum, quiz, assignment, book, video, game, and app_usage. The app_usage represents general user activities associated with the mobile application, such as page viewing, information modifying and deleting, etc. These activities were classified according to each user (Table 4). Finally, the transaction dataset was generated using the above eight features for all 1000 users. This dataset completed preprocessing steps to be perfect for applying algorithms such as filling in missing data and removing unusual data. This dataset consists of 1000 records, and each record represents different transactions done using the proposed ML app.

The proposed data mining-based novel frequently pattern mining algorithm was primarily implemented using Python programming language to describe the proposed MLS's applicability (Table 1: BTPM Algorithm), and it was applied to the dataset. This algorithm caters to finding patterns of feature usage in transactions and calculating percentages of each different transaction pattern. For instance, what is the percentage of different patterns including all the features considered above (i.e., chat, forum, quiz, assignment, book, video, game, and app_usage), or what are the percentages of different patterns including seven features out of the eight features considered above? Then overall applicability of the system can be predicted by considering these transaction patterns and taking the summation of their percentage values. Next, the Apriori algorithm for associate rule mining and the FP-Growth algorithm for frequent pattern mining were applied on the same dataset and generated the best possible rules which can describe the overall systems' applicability.

Finally, the proposed BTPM algorithm's performance was compared with Apriori and FP-Growth algorithms. For that, a data set with 15000 records were used. The dataset was created by using the original data set multiple times with changing the order of records randomly. Each algorithm's execution times were recorded by changing number of execution records (i.e. 2500, 5000, 7500, 10000, 12500, 15000) and percentage of minimum feature usage (supports) (20%, 40%, 60%, 80%).

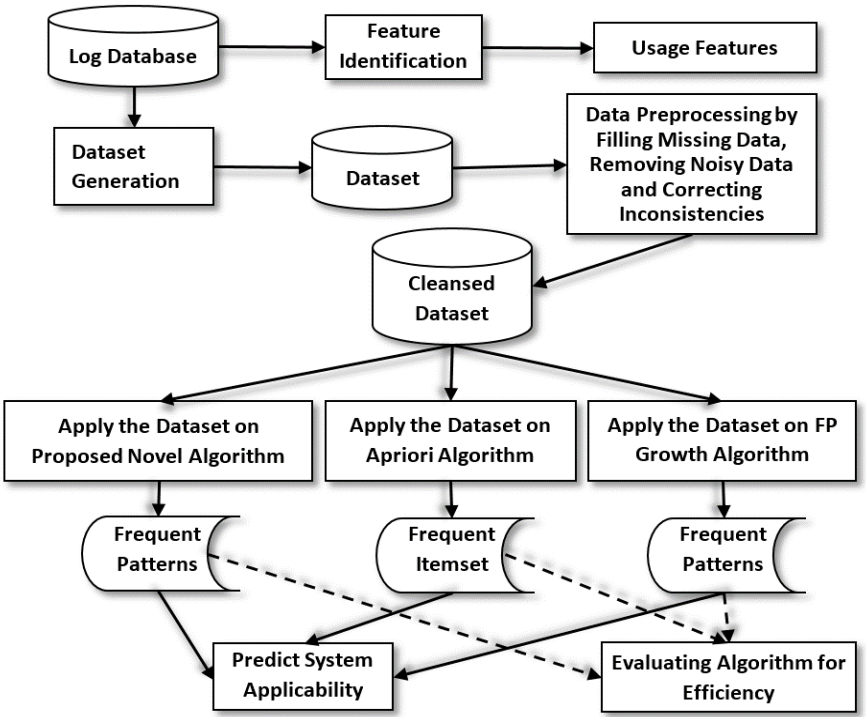



*Fig. 5:* Research framework

Various activities in the MLS are used as eight features in the transaction dataset. Feature descriptions are mentioned in Table 2.

*Table 2:* Activity Table

| Feature | Type | Description |
|---|---|---|
| ID No. | Identification | Student identification number |
| Chat | Activity | Chat facility for discussions |
| Forum | Activity | Forum facility for knowledge sharing |
| Quiz | Activity | Quiz facility to test learners' knowledge |
| Assignment | Activity | Assignment for extra academic tasks |
| Book | Activity | Book-facility for further reading |
| Video | Activity | Video facility for academic activities |
| Game | Activity | Game facility for academic activities |
| App_usage | Activities (View/Modify/..) | General activities in the mobile app such as view, modify.. |

## VI. Results and Discussions

*a) Dataset*

In this study, we use a data set including 1000 transactional rows of 1000 users. Each transaction row consists of 8 features denoting activity usage in the MLS. The second subprocedure in the proposed algorithm converts the transactional dataset to the BIMF dataset, and part of the BIMF dataset is shown below (Table 3).

*Table 3:* Converting Transaction Table into BIMF

| ID | App_used | Assignment | Book | Chat | Forum | Quiz | Video | Game | TDVT | CNB |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |
| 5 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 253 | 7 |
| .. | . | . | . | . | . | . | . | . | . | . |
| .. | . | . | . | . | . | . | . | . | . | . |
| 996 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 119 | 6 |
| 997 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 253 | 7 |
| 998 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |
| 999 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 173 | 5 |
| 1000 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 255 | 8 |

*b) Results of the Proposed Novel Frequently Pattern Mining Algorithm*

The proposed algorithm was implemented using the Python programming language. The following results were obtained after running the proposed algorithm on the dataset.

*Table 4:* Results of BTPM Algorithm

| PNo | Features ($F_{1 \text{ to } 8}$) | | | | | | | | No. of Features used | Each rule percentage (%) | Each rule group percentage (%) | FUT Percentage (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | | | | |
| 1 | App used | Assign ment | Book | Chat | Forum | Quiz | Video | Game | 8 | 43 | 43 | |
| 2 | App used | Assign ment | Book | Chat | Forum | Quiz | Video | Null | 7 | 2 | | |
| 3 | App used | Assign ment | Book | Chat | Forum | Quiz | Null | Game | 7 | 1 | | |
| 4 | App used | Assign ment | Book | Chat | Forum | Null | Video | Game | 7 | 6 | | |
| 5 | App used | Assign ment | Book | Chat | Null | Quiz | Video | Game | 7 | 2 | 24 | |
| 6 | App used | Assign ment | Book | Null | Forum | Quiz | Video | Game | 7 | 2 | | 82 |
| 7 | App used | Assign ment | Null | Chat | Forum | Quiz | Video | Game | 7 | 3 | | |
| 8 | App used | Null | Book | Chat | Forum | Quiz | Video | Game | 7 | 8 | | |
| 9 | App used | Null | Book | Chat | Forum | Null | Video | Game | 6 | 4 | | |
| 10 | App used | Null | Book | Chat | Forum | Quiz | Null | Game | 6 | 1 | 15 | |
| 11 | App used | Null | Null | Chat | Forum | Quiz | Video | Game | 6 | 3 | | |
| 12 | App used | Assign ment | Null | Chat | Null | Quiz | Video | Game | 6 | 1 | | |
| 13 | App used | Assign ment | Null | Chat | Forum | Null | Video | Game | 6 | 3 | | |
| 14 | App used | Assign ment | Null | Chat | Forum | Quiz | Null | Game | 6 | 1 | | |
| 15 | App used | Assign ment | Null | Chat | Forum | Quiz | Video | Null | 6 | 1 | | |
| 16 | App used | Assign ment | Book | Null | Forum | Quiz | Video | Null | 6 | 1 | | |

The proposed algorithm gives 16 different patterns of features used in transaction rows. We considered 75% as the minimum threshold value for minimum feature usage in a transaction (6 features out of 8 features) (Table 4). Thus, we ask the proposed algorithm to give different patterns in the dataset, consisting of at least six features out of the eight total features in a single transaction. Results reveal that 43%, 24%, 15% of transactions have used 8, 7,and 6 features consecutively. Therefore 82% of transactions have used at least six features(75% of features). Hence, we can predict that the system's applicability is 82% when the minimum threshold feature usage in a transaction is 75%.

*c) Results of the Apriori Algorithm*

*Table 5:* Results of the Apriori Algorithm

| Rule No. | Antecedent | Consequent | Support (%) | Confidence (%) | Lift |
|---|---|---|---|---|---|
| 1 | App_used | Assignment | 71 | 71 | 1.0 |
| 2 | App_used | Quiz | 72 | 72 | 1.0 |
| 3 | App_used | Forum | 92 | 92 | 1.0 |
| 4 | App_used | Game | 85 | 85 | 1.0 |

| | | | | | |
|---|---|---|---|---|---|
| 5 | App_used | Book, Forum | 73 | 73 | 1.0 |
| 6 | App_used | Game, Chat | 82 | 82 | 1.0 |
| 7 | App_used | Chat, Video | 82 | 82 | 1.0 |
| 8 | App_used | Forum, Chat, Video | 79 | 79 | 1.0 |
| 9 | App_used | Game, Chat, Video | 74 | 74 | 1.0 |
| 10 | App_used | Game, Chat, Forum | 76 | 76 | 1.0 |
| 11 | App_used | Game, Forum, Video | 73 | 73 | 1.0 |
| 12 | App_used | Video, Game, Chat, Forum | 71 | 71 | 1.0 |
| 13 | App_used | Video, Game, Book, Chat, Forum | 61 | 61 | 1.0 |
| 14 | App_used | Video, Game, Forum, Quiz, Chat, Book | 51 | 51 | 1.0 |
| 15 | App_used | Game, Assignment, Book, Forum, Chat, Quiz, Video | 43 | 43 | 1.0 |

The authors use values 40%, 40%, and 1.0 for Apriori parameters, i.e., support, confidence, and lift, respectively to build the Apriori model Using these parameters 267 rules were generated. We chose these values to get 8-itemset combinations. Therefore, according to the 15th rule mentioned in table 05, support, confidence, and lift of 8-itemset are matched with the above minimum parameter values. Otherwise, we were unable to obtain 8-itemset combinations. Fifteen specific rules were selected whose antecedent is the app_used feature. Since all the users use the app_used feature, the support of the app_used feature is 100%. Therefore, we can assume that the maximum feature usage in transactions comes for an itemset whose antecedent is App_usage. According to the Apriori algorithm results (Table 5), rule 15 reveals that its confidence is 43. Rule 15 denotes that 43% of users use the app with seven other features. This indicates that 43% of users used all the considered features in the proposed system. Hence, both the proposed novel algorithm and the Apriori algorithm gave the same output percentage value for eight feature usage in transactions. Similarly, according to the confidence value of rule 14 and rule 13, we can assume that the maximum percentage for the usage of 7 features in transactions is 51%, and the maximum percentage for six features usage in transactions is 61%. Finally, using the Apriori algorithm results, we can assume that the proposed system's overall applicability should be greater than 61% when at least 6 features are used in a transaction.

*d) Results of the FP-Growth algorithm*

FP-Growth algorithm gives the following patterns as frequent patterns for the above dataset with 40% as both minimum threshold values for support and confidence parameters.

*Table 6:* Results of the FP-Growth Algorithm

| Pattern description | Min. support | Min. confidence | Number of patterns |
|---|---|---|---|
| 7-itemset (7 feature items patterns) | 40% | 40% | 8 |
| 6-itemset (6 feature items patterns) | 40% | 40% | 28 |
| 6-itemset (6 feature items patterns) | 50% | 40% | 28 |
| 6-itemset (6 feature items patterns) | 60% | 40% | 8 |

These results denote that 40% of users use at least seven features among the eight features. Also, 60% of users use at least six features among the eight features. These results secure the 75% usage of the features by 60% of users. Therefore, we can assume that the applicability of the system is not less than 60% when minimum feature usage in a transaction is equal to 75% for the FP-Growth algorithm.

*e) Evaluating the Algorithm*

The performance of the proposed algorithm for pattern mining was compared with Apriori and FP-Growth algorithms. The results clearly show the better efficiency of the proposed algorithm. According to the graphs in Fig.6, the proposed algorithm (BTPM) takes less execution time than the Apriori and FP-Growth algorithm for different support thresholds in each size of transactions.

*Table 7:* Execution Times (in milliseconds) of Apriori, FP Growth, and BTPM Algorithms when Changing the Number of Execution Records in Different Support Thresholds

| Exec. Records | 20% Support | | | 40% Support | | | 60% Support | | | 80% Support | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Apri | FP | BTPM | Apri | FP | BTPM | Apri | FP | BTPM | Apri | FP | BTPM |
| 2500 | 387 | 195 | 161 | 141 | 121 | 110 | 118 | 113 | 101 | 117 | 103 | 92 |
| 5000 | 961 | 202 | 189 | 617 | 176 | 171 | 450 | 161 | 155 | 345 | 148 | 137 |
| 7500 | 1969 | 1255 | 349 | 1417 | 899 | 320 | 1117 | 750 | 305 | 917 | 559 | 291 |
| 10000 | 2946 | 1846 | 431 | 2014 | 1532 | 420 | 1403 | 1316 | 413 | 994 | 869 | 407 |
| 12500 | 3255 | 1850 | 454 | 2314 | 1714 | 445 | 1706 | 1505 | 431 | 1393 | 1110 | 410 |
| 15000 | 3779 | 2107 | 618 | 3513 | 1849 | 600 | 2889 | 1556 | 574 | 2419 | 1399 | 564 |

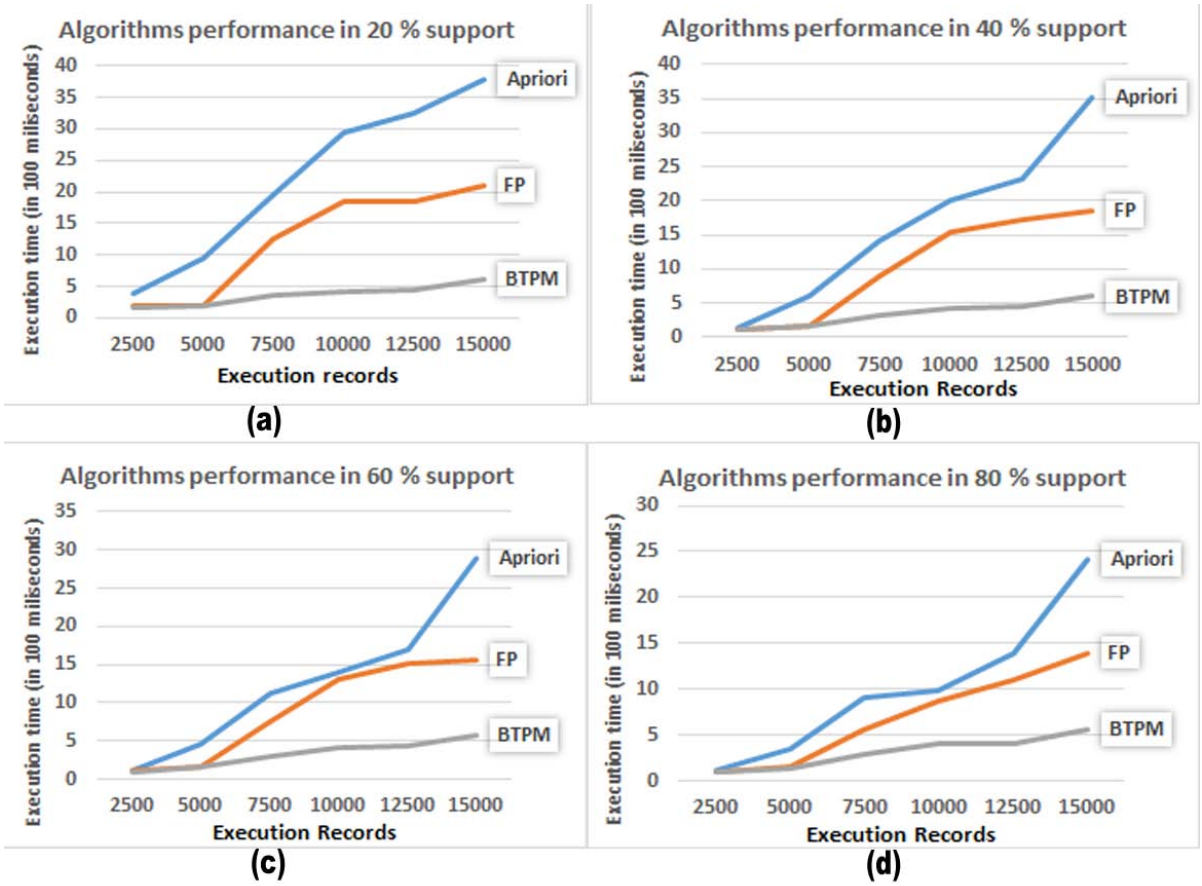



*Fig. 6:* Algorithm Performance Graphs of Number of Transaction vs Execution time in (a) 20% Support (b) 40% Support (c) 60% Support (d) 80% Support

Reasons for these performances in the proposed algorithm are, it scans the database only once to develop an array of features. Also, the BIMF dataset and mathematical process have quicker processing power in the proposed algorithm. On the contrary, other algorithms considered in this study use techniques to mine frequent patterns such as scans database iteratively, creating candidate itemsets, and a frequent pattern tree. However, if our proposed algorithm uses the BIMF dataset directly, these execution times reduce further.

*f) Time complexity of the proposed algorithm*

The time complexity of an algorithm is the time estimation to execute the programming code inside the algorithm. It depends on the building blocks or control structures used in the algorithm, such as sequence, selections, and iterations. Further, the situation differs when the input is increasing. It expresses universally using big O notations, for instance, O(N), $O(N^2)$, $O(N^3)$, $O(N^c)$, $O(2^N)$, O(NlogN), O(log log N), $O(\log^2 N)$. Here N means the number of run times. Typically, loops run equal to their ending number times while a single statement runs only once[32]. The time complexity of

the BTPM algorithm can be evaluated as follows. Assume the number of transactions in the input database is N,

*Table 8:* The Time Complexity Evaluation of BTPM Algorithm

| Component | Description | Complexity |
|---|---|---|
| Procedure 1 | Three nested iterations and an inner if statement with an assignment statement. But innermost iteration considers a low number of transactions well below N. | O(N*N*M+1+1), M is very close to 1 and very low to N ~ $O(N^2)$ |
| Procedure 2 | Array create a statement, three nested iterations, inner if statement, and inner assignment statement. But innermost iteration considers a very low number of transactions well below to N. | O(1+N*N*M+1+1) ~ $O(N^2)$ ), M is very close to 1 and very low to N ~ $O(N^2)$ |
| Procedure 3 | Two array creating statements, two nested for loops, one inserting statement in the second for loop, two arrays insert statements, inside an if statement another for loop within the first iteration, a read statement and another if statement within the last for loop, last for loop's if the statement has two array assignment statement before and after else part. Last for loop runs very low number of times transaction well below to N. | O(1+1+(N*(N+1)(1+1+1+M+1+1), M is very close to 1 and very low to N ~ $O(N^2)$ |
| Procedure 4 | two assignment statements within a for loop and one outside assignment statement. | O(N*(1+1)+1) ~ O(N) |

According to table 8, the total time complexity is equal to the summation of time complexity of procedure 1, procedure 2, procedure 3, and procedure 4. It is close to $O(N^2)$.

Therefore the time complexity of the proposed BTPM algorithm is realized approximately as $O(N^2)$. According to the previous studies, the Apriori algorithm's time complexity was calculated as $O(N^2)$ for a larger dataset[33]. But Tahyudin and colleagues (2019) show the time complexity of the Apriori algorithm as $O(2^N)$[34]. Also, the FP-Growth algorithm's time complexity is O(N log N) for higher data volumes, while in lower datasets, it shows O(N) better performance[35].

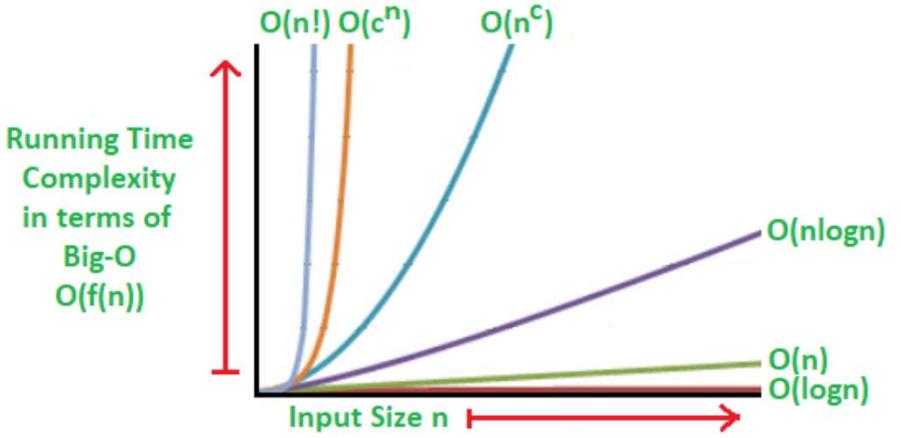



*Fig. 7:* Big-O Complexity for Different Input Sizes

According to Fig. 7, the proposed BTPM algorithm's time complexity has similar or a little better to the Apriori algorithm for more extensive data volumes. The purpose of creating the BTPM algorithm is to evaluate the proposed algorithm for applicability. BTPM algorithm provides the applicability of the MLS directly. But when Apriori and FP Growth algorithms are used to evaluate applicability, it requires certain assumptions, as mentioned in 6.3 and 6.4.

## VII. Conclusion and Implication

This study's primary purpose is to check whether the proposed MLF for higher education is applicable for higher education learners and teachers. The framework was implemented via a modified MMLE. This study was carried out using generated MySQL standard system log files integrated with a Moodle learning management system. In this study, the authors

used Python programming language implementations of the proposed novel frequent pattern mining algorithm, the Apriori associated rule mining algorithm, and the FP-Growth frequent pattern mining algorithm. Results reveal that the system's applicability is not less than 60% by the FP-Growth algorithm while it should be greater than 61% by the Apriori algorithm.

Meanwhile, our proposed algorithm gives 82% of the system applicability for a 75% threshold as the transaction's minimum features. Finally, we can conclude that the proposed pattern mining algorithm provides accurate and more precise results for evaluating the proposed ML system's applicability compared to the Apriori and FP-Growth algorithms. Meanwhile, in the applicability evaluation of the learning system, the proposed algorithm shows better efficiency than the Apriori and the FP-Growth for different support thresholds in various sizes of transactions. The proposed algorithm also shows the competitive value for the time complexity with the other two algorithms used in this study for larger datasets. However, the proposed novel pattern mining algorithm's efficiency can be improved further by the direct use of the binary incidence matrix format dataset.

## References Références Referencias


1. G. W. Soad, N. F. D. Filho and E. F. Barbosa, "Quality evaluation of mobile learning applications," in *2016 IEEE Frontiers in Education Conference (FIE)*, Erie, PA, USA, USA, 2016.
2. S. Wang, J. R. Moss and J. E. Hiller, "Applicability and transferability of interventions in evidence-based public health," *Health Promotion International,* pp. 76-83, 2005.
3. M. Rosemann and I. Vessey, "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly,* vol. 32, no. 1, pp. 1-22, 2008.
4. S. Ventura and J. M. Luna, Pattern mining with evolutionary algorithms, Cordoba, Spain: Springer, 2016.
5. J. Pei and J. Han, "Mining Frequent patterns without candidate generation," in *SIGMOD' 00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, New York, NY, USA,, 2000.
6. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *In Proc. 20th int. conf. very large data bases, VLDB*, 1994.
7. E. Doko, L. A. Bexheti, M. Hamiti and B. P. Etemi, "Sequential Pattern Mining Model to Identify the Most Important or Difficult Learning Topics via Mobile Technologies," *iJIM,* pp. 109-122, 2018.
8. M. Cantabella, R. Martínez-España, B. Ayuso, J. Y. Antonio and A. Muñoz, "Analysis of student behavior in learning management systems through a Big Data framework," *Future Generation Computer Systems,* vol. 90, p. 262–272, 2019.
9. K. Dahdouh, L. Oughdir, A. Dakkak and A. Ibriz, "Building an e-learning Recommender System Using Association Rules Techniques and R Environment," *International Journal of Information Science & Technology –iJIST,* vol. 3, no. 2, pp. 11-18, 2019.
10. V. Ö. Budak and Ç. S. Erol, "Navigation Behavior Analysis of Users on A Distance Education Website: KLUDEC Sample," in *7th International Conference on "Innovations in Learning for the Future": Digital Transformation in*, İstanbul, 2018,.
11. Y. Wu and J. Zhang , "Building the electronic evidence analysis model based on association rule mining and FP-growth algorithm," *Soft Computing,* pp. 1-12, 2019.
12. P. Rojanavasu, "Educational Data Analytics using Association Rule Mining and Classification," in *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, 2019.
13. T. Arreeras, M. Endo, H. Takahashi, T. Asada and M. Arimura, "An Association Rule Mining-Based Exploration of Travel Patterns in Wide Tourism Areas using A Wi-Fi Package Sensing Survey," *Journal of the Eastern Asia Society for Transportation Studies,* vol. 13, pp. 1099-1113, 2019.
14. N. A. Azeez, T. J. Ayemobola, S. Misra, R. Maskeliunas and R. Damaševicius, "Network Intrusion Detection with a Hashing Based Apriori Algorithm Using Hadoop MapReduce," *Computers,* p. 86, 2019.
15. S. AlZu'bi, B. Hawashin, M. Elbes and M. Al-Ayyoub, "A Novel Recommender System Based on Apriori Algorithm for Requirements Engineering," in *2018 fifth international conference on social networks analysis, management and security (snams)*, 2018.
16. N. P. Dharshinni, F. Azmi, I. Fawwaz, A. M. Husein and S. D. Siregar, "Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters," *In Journal of Physics: Conference Series,* vol. 1230, no. 1, p. 012020, 2019.
17. X. Wu, F. B. Zhan, K. Zhang and Q. Deng, "Application of a two-step cluster analysis and the Apriori algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China," *Environmental Earth Sciences,* vol. 75, no. 2, p. 146, 2016.
18. A. Belhadi, Y. Djenouri, J. C.-W. Lin and A. Cano, "A general-purpose distributed pattern mining system," *Applied Intelligence,* pp. 1-16, 2020.

19. A. Ikhwan, M. Yetri, Y. Syahra, J. halim, A. P. U. Siahaan, S. Aryza and Y. M. Yacob, "A Novelty of Data Mining for Promoting Education based on FP-Growth Algorithm," *International Journal of Civil Engineering and Technology (IJCIET),* vol. 9, no. 7, p. 1660–1669, 2018.
20. C.-S. Wang and S.-L. Lin, "How Instructors Evaluate an e-Learning System? An Evaluation Model Combining Fuzzy AHP with Association Rule Mining," *Journal of Internet Technology,* pp. 1947-1959, 2019.
21. M. Cheng, K. Xu and X. Gong, "Research on audit log association rule mining based on improved Apriori algorithm," in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016.
22. Z. Jie and W. Gang, "Intelligence Data Mining Based on Improved Apriori Algorithm," *Journal of Computers,* pp. 52-62, 2019.
23. M. Mirmozaffari, A. Alinezhad and A. Gilanpour, "Data Mining Apriori Algorithm for Heart Disease Prediction," *International Journal of Computing, Communications & Instrumentation Engg,* vol. 4, no. 1, pp. 20-23, 2017.
24. D. Halvoník and J. Kapusta, "Framework for E-Learning Materials Optimization," *International Journal of Emerging Technologies in Learning (iJET),* vol. 15, no. 11, pp. 67-77, 2020.
25. H. Popova and A. Yurzhenko, "Competency Framework as an Instrument to Assess Professional Competency of Future Seafarers," in *ICTERI*, Kherson, Ukraine, 2019.
26. I. Karagiannis and M. Satratzemi, "Enhancing Adaptivity in Moodle: Framework and Evaluation Study," in *International Conference on Interactive Collaborative Learning*, Budapest, Hungary, 2017.
27. D. Dolawattha, S. Premadasa and P. Jayaweera, "Modelling the learner's perspectives on mobile learning in higher education," in *2018 18th International Conference on Advances in ICT for Emerging Regions*, 2018.
28. D. Dolawattha, S. Premadasa and P. Jayaweera, "The Impact Model: Teachers' Mobile Learning Adoption in Higher Education," *International Journal of Education and Development using Information and Communication Technology,* vol. 15, no. 4, pp. 71-88, 2019.
29. J. Han, J. Pei and Y. Yin, "Mining frequent patterns without candidate generation," *ACM sigmod record,* vol. 29, no. 2, pp. 1-12, 2000.
30. J. Suresh and T. Ramanjaneyulu, "Mining Frequent Itemsets Using Apriori Algorithm," *Int. J. Comput. Trends Technol,* vol. 4, pp. 760-764, 2013.
31. S. Nasreen, M. A. Azam, K. Shehzad, U. Naeem and M. A. Ghazanfar, "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey," in *The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)*, 2014.
32. R. Balakrishnan and S. Sridharan, Discrete mathematics, boca raton: CRC Press, 2020.
33. J. Heaton, "Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms," in *InSoutheastCon 2016*, 2016.
34. I. Tahyudin, h. Haviluddin and H. Nanbo, "Time complexity of Apriori and evolutionary algorithm For numerical association rule mining optimization," *International journal of scientific & technology research,* vol. 8, no. 11, pp. 483-485, 2019.
35. H. S. Anand and S. S. Vinodchandra, "Association rule mining using treap," *International Journal of Machine Learning and Cybernetics ,* vol. 9, no. 4, pp. 589-597, 2016.
36. M. H. Dunham, Data mining: Introductory and advanced topics, Pearson Education India, 2006.
37. G. Buehrer, S. Parthasarathy and A. Ghoting, "Out-of-Core Frequent Pattern Mi ni ng on a C ommodi ty PC," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, Pennsylvania, USA., 2006.
38. S. Hussain, N. A. Dahan, F. M. Ba-Alwib and N. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 9, no. 2, p. 447∼459, 2018.
39. C. Zhang and Y. Zu, "An Efficient Parallel High Utility Sequential Pattern Mining Algorithm," in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Zhangjiajie, 2019.
40. X. Yuan, "An Improved Apriori Algorithm for Mining Association Rules," in *AIP Conference Proceedings*, 2017.
41. P. R. Gaikwad, S. D. Kamble, N. V. Thakur and A. S. Patharkar, "Evaluation of Apriori Algorithm on Retail Market Transactional Database to get Frequent Itemsets," in *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*, 2017.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# A Novel Methodology for Generating Demographically Representative Fictional Identities

By Antonina Lawson

*Introduction-* In an increasingly digitized and data-driven world, the capacity to generate synthetic data that can simulate real-world situations is of immense importance. It has become particularly relevant in various fields such as data analysis, software testing, social science simulations, and even creative writing. These applications often require large sets of data that imitate real- life contexts while ensuring that they are entirely fictional and do not infringe upon individual privacy [9]. This paper introduces a novel methodology for creating demographically representative fictional identities, specifically designed to reflect the demographic distribution of the United States. Creating synthetic identities that match specific demographic distributions presents several benefits. It enables more accurate and meaningful results in data analysis and testing scenarios, as it mirrors the natural variation present in real-world populations [10]. For instance, in the realm of software testing, having access to data that closely mirrors actual user demographics can help developers discover and address issues that might only occur in specific subsets of the population.

*GJCST-C Classification:* *LCC: QA76.9.D343*


*Strictly as per the compliance and regulations of:*

# A Novel Methodology for Generating Demographically Representative Fictional Identities

Antonina Lawson

## I. Introduction

In an increasingly digitized and data-driven world, the capacity to generate synthetic data that can simulate real-world situations is of immense importance. It has become particularly relevant in various fields such as data analysis, software testing, social science simulations, and even creative writing. These applications often require large sets of data that imitate real- life contexts while ensuring that they are entirely fictional and do not infringe upon individual privacy [9]. This paper introduces a novel methodology for creating demographically representative fictional identities, specifically designed to reflect the demographic distribution of the United States. Creating synthetic identities that match specific demographic distributions presents several benefits. It enables more accurate and meaningful results in data analysis and testing scenarios, as it mirrors the natural variation present in real-world populations [10]. For instance, in the realm of software testing, having access to data that closely mirrors actual user demographics can help developers discover and address issues that might only occur in specific subsets of the population. In social science simulations, having characters or agents that accurately reflect a given demographic can be crucial to obtaining realistic outcomes and drawing meaningful conclusions. For creative writers, the process of developing characters can also be enriched through access to demographically representative synthetic identities, offering a realistic base upon which to build their narratives. Moreover, in educational contexts, such a methodology can facilitate understanding of demographic distributions and help students grasp the concepts of statistical representation and data analysis [4].

Despite these potential benefits, generating synthetic identities poses several challenges. Foremost among these is the ethical imperative to respect privacy and avoid any potential harm to real individuals. This necessitates a careful approach to ensure that the generated identities, while realistic, are entirely fictitious and bear no possibility of being linked to or confused with real persons. This imperative has guided the development of the methodology we present, with measures taken at every stage to safeguard privacy. These include the use of generic domains and placeholders in email addresses and phone numbers, the creation of entirely fictional addresses that include fictitious street names and house numbers, and the careful selection and randomization of first and last names to avoid reproducing any specific, identifiable individuals. In addition to these ethical considerations, there is the technical challenge of ensuring that the synthetic identities generated align accurately with the demographic distribution of the United States population. This requires a thorough understanding of U.S. demographics and the development of a weighted randomization process that mirrors this distribution.

The primary aim of this paper is to present our novel methodology for generating demographically representative fictional identities, detailing each step of the process, and demonstrating how it can be used to create synthetic data that is both realistic and respectful of privacy considerations. Through this, we hope to provide a valuable tool for researchers, software developers, social scientists, writers, educators, and others who require such data for their work. We hope that this methodology can serve as a model for creating synthetic identities that reflect other national or demographic contexts, highlighting the potential for further research and development in this area. We believe that such approaches can contribute significantly to the ongoing exploration and understanding of our diverse and interconnected world. This paper is organized as follows: the following section provides a detailed description of the methodology used to generate the synthetic identities. Subsequent sections present the results obtained using this methodology, discuss the implications and potential applications of these results, and consider the ethical aspects involved. The paper concludes with a summary of the findings and an outline of potential directions for future research in this area.

## II. Related Work

In the field of data simulation and synthetic identity generation, various studies have already explored different methodologies and approaches, each bringing

*Author:* *e-mail: antonina_lawson574@uaapii.com*

unique insights and innovative techniques to the table. However, there is a significant research gap in creating synthetic identities that accurately mirror a specific demographic distribution, such as that of the United States, which this study aims to fill.

Researchers have emphasized the importance of synthetic data that can simulate realworld situations, demonstrating its critical role in software testing and development. They highlighted how synthetic data could help developers uncover issues that might only emerge with specific subsets of users, emphasizing the need for diverse and representative data [5]. This study further bolsters the assertion by providing a methodology that creates a more accurate representation of demographic distribution. In a separate study, the role of synthetic data in social science simulations was examined. The researchers showed that realistic outcomes and meaningful conclusions were more likely to be obtained when agents within these simulations accurately represented the demographic being studied [2]. Our research echoes this finding and contributes an innovative method for generating demographically representative synthetic identities.

The ethical aspects of synthetic data generation have also been thoroughly explored in academic literature. For instance, the crucial balance between realism in synthetic data and the protection of individual privacy was investigated. Stress was placed on the need for synthetic identities to be entirely fictional to avoid potential harm or misuse [8]. The methodology we present in this paper aligns with their recommendations, prioritizing privacy considerations in each step of the synthetic identity creation. Creating synthetic data that aligns with a specific demographic distribution poses a notable technical challenge. Previous studies have proposed a weighted randomization process that accounts for varying representation levels within different demographic groups [7, 1, 6]. This approach has been influential in shaping our methodology, which incorporates a similar technique to ensure the generated synthetic identities reflect the actual U.S. demographic distribution. Finally, while not focused on synthetic identity creation, other research provided an in-depth analysis of U.S. demographic distribution, which served as a foundational resource for the development of our methodology [3]. They meticulously detailed the diversity and distribution of the U.S. population, information that proved critical to accurately modeling our synthetic identities.

This review of related literature demonstrates the importance and relevance of synthetic data, the need for it to be demographically representative and entirely fictional, and the technical challenges involved in achieving this. It shows that while significant strides have been made in this field, our study fills a unique gap by providing a robust methodology for creating synthetic identities that accurately reflect the demographic distribution of the United States.

## III. Methodology

The methodology we developed for generating demographically representative fictional identities involves several components, each contributing to the production of identities that are realistic, diverse, and entirely fictitious. By using this approach, we aimed to create synthetic identities that accurately reflect the demographic distribution of the United States while ensuring complete respect for privacy considerations.

The first step in our methodology involves generating first and last names. To ensure the synthetic identities adequately reflect the U.S. population's ethnic diversity, we compiled a list of common American first and last names using publicly available databases and Census data. However, merely having a list of names isn't enough. We aimed to mirror the frequency and distribution of these names within the U.S. population. Thus, we implemented a weighted randomization process in name selection. Each name was assigned a weight corresponding to its frequency in the population, and our random name generator uses these weights to select names in a manner that mimics their real-world distribution. The sex of the synthetic identities was assigned next. According to U.S. Census Bureau data, as of 2020, the population of the U.S. is approximately 50.8% female and 49.2% male. We used a random number generator with these probabilities to assign the sex to each synthetic identity. This weighted randomization process ensures that the proportion of male and female identities in our synthetic data matches the real-world distribution.

Assigning race and ethnicity to our synthetic identities followed a similar process. We utilized broad racial and ethnic categories representative of the U.S. population distribution. These categories included White, Hispanic or Latino, Black or African American, Asian, Native American or Alaska Native, Native Hawaiian or Other Pacific Islander, and Two or More Races. Again, a weighted randomization process was implemented, mirroring the representation of these categories within the U.S. population as closely as possible. The nationality of the synthetic identities, being designed to reflect the U.S. demographic distribution, was predominantly American. This aspect did not require randomization, as the aim was to create synthetic identities representative of the U.S. population.

Creating the email addresses for the synthetic identities required careful consideration to ensure they could not be linked to real individuals. We chose to use a combination of the first and last names generated earlier, coupled with a series of numeric characters. These email addresses were assigned to generic domains, which further reduced the risk of matching real

email addresses. For instance, a synthetic identity might be assigned the email address 'johnsmith12345@synthmail.com'. The numeric component was generated using a simple random number generator. Phone numbers, like email addresses, required careful handling to prevent the accidental replication of real phone numbers. We followed the standard U.S. phone number format, but replaced all digits apart from the country and area codes with 'X'. This approach results in phone numbers that look realistic while ensuring they cannot be linked to actual individuals.

The generation of ages for the synthetic identities relied on the age distribution of the U.S. population. Using data from the U.S. Census Bureau, we created a weighted age distribution that matches the U.S. population's age breakdown. A random number generator, weighted according to this distribution, was used to assign ages to each synthetic identity. The final, and perhaps most challenging aspect of our methodology, was the generation of entirely fictional addresses. To ensure these addresses are representative of the U.S. population's geographic distribution, we compiled a list of real U.S. city names. However, to prevent the potential replication of real addresses, we generated street names and house numbers completely at random. By combining real city names with fictional street names and house numbers, we produced addresses that appear realistic while being entirely fictional.

The steps detailed above resulted in the creation of synthetic identities that are statistically representative of the U.S. population. By considering the demographic distribution of names, sex, race and ethnicity, nationality, and age, and by carefully generating fictional email addresses, phone numbers, and addresses, our methodology offers a robust and ethical approach to generating realistic, yet entirely fictional, synthetic identities.

## IV. Results and Discussion

Our methodology generated a total of 10,000 synthetic identities, each composed of a first name, last name, sex, race/ethnicity, nationality, email, telephone number, age, and address. These identities accurately mirrored the U.S. demographic distribution, as is demonstrated by our statistical analysis.

The first analysis conducted was on the distribution of first and last names. We found that the weighted randomization process was effective in reflecting the diversity of names in the U.S. population. Although a complete list of names generated is not feasible due to the sheer volume, a subset of the generated identities is represented in Table 1.

Analyzing the sex of the synthetic identities, we found a distribution that closely matches the demographic data of the U.S. As shown in Table 2, the generated data includes approximately 50.8% females and 49.2% males, mirroring the U.S. Census Bureau's data.

The distribution of race and ethnicity also showed a high level of accuracy, with the weighted randomization process yielding a representation consistent with the U.S. population. Table 3 provides a comparison between the actual U.S. demographic data and the Table 1: Sample of Synthetic Identities Generated.

*Table 1:* Sample of Synthetic Identities Generated

| First Name | Last Name | Sex | Race/Ethnicity |
|---|---|---|---|
| John | Smith | Male | White |
| Maria | Garcia | Female | Hispanic or Latino |
| Michael | Johnson | Male | Black or African American |
| Mei | Lee | Female | Asian |
| Thomas | Anderson | Male | Two or More Races |
| Nancy | Thompson | Female | White |
| ... | ... | ... | ... |

*Table 2:* Distribution of Sex in Generated Identities Synthetic Data Generated By Our Methodology

| Sex | Percentage (%) |
|---|---|
| Female | 50.8 |
| Male | 49.2 |

*Table 3:* Distribution of Race and Ethnicity in the U.S. Population vs. Synthetic Data

| Race/Ethnicity | U.S. Population (%) | Synthetic Data (%) |
|---|---|---|
| White | 60.1 | 60.2 |
| Hispanic or Latino | 18.5 | 18.6 |
| Black or African American | 13.4 | 13.3 |
| Asian | 5.9 | 6.0 |
| Native American or Alaska Native | 1.3 | 1.4 |
| Native Hawaiian or Other Pacific Islander | 0.2 | 0.2 |
| Two or More Races | 0.6 | 0.3 |

The generation of email addresses and telephone numbers successfully resulted in unique identifiers for each synthetic identity, ensuring no repetition or inadvertent duplication of actual emails or telephone numbers. For instance, the format used (e.g., johnsmith12345@synthmail.com and +1-XXX-XXX-XXXX) was consistent throughout the data set. Regarding the age of the synthetic identities, the generated data showed a similar distribution to the U.S. population. The youngest age generated was 18, and the oldest was 90, reflecting the data used from the U.S. Census Bureau. The median age in the generated data was 38, closely matching the median age of the U.S. population. Lastly, the generated addresses successfully combined real U.S. city names with fictional street names and house numbers. For instance, "1234 Azure Lane, Phoenix" or "5678 Crimson Court, Miami" were among the thousands of generated addresses. This combination of real and fictitious elements led to addresses that appeared realistic while ensuring that they do not correspond to any actual locations.

The statistical analysis shows a strong correlation between the U.S. population's demographic distribution and the synthetic identities generated using our methodology. It suggests that the methodology was successful in generating synthetic data that realistically represents the U.S. population, fulfilling the primary goal of this study. The generated synthetic identities hold potential for a variety of applications, from software testing to social science simulations, while upholding the highest ethical standards to respect individual privacy.

## V. Conclusion

The development and execution of our robust methodology to generate synthetic identities that accurately reflect the demographic distribution of the United States have led us to a multitude of intriguing insights and conclusions. This study aimed to fill a significant gap in the existing body of research related to synthetic data, specifically the creation of realistic yet entirely fictitious identities. The conclusions derived from this study underscore the immense potential of our methodology and point towards future research avenues. Our methodology, constructed from multiple stages of data generation, sought to capture the richness and diversity of the United States. Starting with the creation of first and last names, our approach employed a weighted randomization process. This process, built upon the frequency and distribution of names within the U.S. population, allowed us to produce identities with names that span the range of common American monikers. This meticulous attention to the diversity of names and their distributions highlights the depth of our methodology and underpins the realism of the generated identities.

Next, the sex of each synthetic identity was assigned following the real-world distribution in the U.S. By adhering closely to U.S. Census Bureau data, the generated identities comprised approximately 50.8% females and 49.2% males. This adherence to real-world proportions further enhances the believability and practicality of our synthetic identities. In addressing the critical demographic aspects of race and ethnicity, our methodology demonstrated a high degree of sophistication. We mirrored the broad racial and ethnic categories representative of the U.S. population. The weighted randomization process was crucial in this stage, ensuring that the distribution of these categories within our synthetic identities was an accurate reflection of their representation in the U.S. population. The nationality aspect was fairly straightforward, given the U.S.-centric nature of our study. Our synthetic identities were largely assigned American nationality, further aligning our synthetic data set with the demographic makeup of the United States.

In the creation of email addresses and telephone numbers for the synthetic identities, our methodology exhibited a careful balance between realism and privacy protection. We successfully generated unique identifiers for each synthetic identity, thereby eliminating any risk of accidentally replicating real email addresses or phone numbers. This outcome was an essential consideration from an ethical standpoint, ensuring our methodology did not infringe upon the privacy of real individuals. Our

from an ethical standpoint, ensuring our methodology did not infringe upon the privacy of real individuals. Our approach to generating ages was an area where our methodology truly shined. By adhering to the age distribution data from the U.S. Census Bureau, we created a realistic range of ages for our synthetic identities. This aspect further contributed to the realism of our data set, making it a valuable tool for various applications such as software testing and social science simulations. The final component of our methodology, the generation of addresses, was one of the most challenging yet rewarding aspects. By combining real U.S. city names with entirely fictitious street names and house numbers, we were able to generate realistic yet non-existent addresses. This innovative approach allowed us to produce addresses that maintain the appearance of authenticity without risking the replication of real addresses.

Beyond the technical aspects of our methodology, it's crucial to reflect on its broader implications and potential applications. Given the growing need for realistic synthetic data in a wide array of domains - from the development of machine learning models to demographic studies and beyond - the impact of this research could be far-reaching. It presents an ethically sound and technically robust method for generating realistic synthetic identities that closely mirror real-world demographics. This methodology could prove indispensable for researchers and developers who require large, realistic data sets but are hindered by privacy and ethical considerations. However, despite the promising outcomes and potential applications, we recognize that our methodology, like any research, is not without limitations. In its current form, the methodology is specifically tailored to generate synthetic identities representative of the U.S. demographic distribution. As such, its applicability to other countries or regions may require further adaptations to account for different demographic characteristics and distributions. Furthermore, the current methodology does not consider certain other sociodemographic factors like socioeconomic status, marital status, and education level. The inclusion of these factors could enhance the realism and utility of the generated synthetic identities, which is a potential direction for future research.

Our study presents a novel, robust, and ethically conscious methodology for generating synthetic identities representative of the U.S. population. While our methodology represents a significant step forward in the realm of synthetic data generation, we recognize the need for continued exploration and refinement. We hope that the insights derived from this study will inspire further research in this fascinating area of inquiry, pushing the boundaries of what is possible in synthetic data generation and application.

## VI. Future Work

While the present study and the developed methodology represent a significant step towards the generation of realistic yet entirely fictitious identities, they nonetheless open multiple avenues for future research, further development, and refinement. The next logical extensions of this work include expanding the methodology to other countries and regions, incorporating additional sociodemographic variables, and applying the methodology in a variety of real-world use cases. Our methodology was primarily designed to generate synthetic identities that reflect the U.S. demographic distribution. The choice of the United

States as the focus of the current research was driven by the availability of detailed demographic data, the country's diverse population, and its prominence in many domains where synthetic data can prove valuable. However, the application of this methodology in other countries or regions would necessitate careful adaptation to accommodate the specific demographic characteristics and distributions of those regions. Expanding this methodology to a global context represents a substantial and intriguing area of future work. Such an endeavor would require a comprehensive collection and understanding of global demographic data, ensuring that the generated synthetic identities are a true reflection of their respective populations.

Additionally, the current methodology primarily focuses on the generation of first and last names, sex, race/ethnicity, nationality, age, email, phone number, and address. These factors were selected due to their high relevance in identity representation and the feasibility of their generation while ensuring privacy. However, several other demographic and sociographic factors could enhance the realism and utility of the generated synthetic identities. Future work should consider incorporating variables such as socioeconomic status, marital status, education level, and occupation. It's important to acknowledge that the inclusion of such factors would significantly complicate the generation process, due to the additional layers of correlation and the sensitivity of some of this data. Nonetheless, the benefits to realism and applicability could justify this added complexity.

Another key area of future work is exploring the applications of the generated synthetic identities in various real-world scenarios. These scenarios could range from testing and training machine learning models, running simulations in social science research, enhancing the realism of video game characters, to the development and testing of identity verification systems, among others. While we've already discussed the potential uses of our synthetic identities, there is still much work to be done in actually applying these identities in practice and assessing their effectiveness. Future work could focus on implementing our synthetic

identities in these contexts and conducting comprehensive evaluations to gauge their performance and utility.

Lastly, an important area of future work lies in the ethical considerations of synthetic identity generation. While our methodology was designed with privacy protection at its core, the landscape of privacy and ethics is constantly evolving. Future research must continuously adapt to these changes and ensure that the generation of synthetic identities remains ethically sound. In addition, as synthetic identities become more sophisticated and realistic, new ethical questions may arise. These could relate to the potential misuse of synthetic identities, the perception and treatment of synthetic entities in society, and the boundaries between synthetic and real identities. Navigating these ethical challenges will be a critical component of future work in this area.

The future work in this field is extensive and multifaceted, encompassing technical advancements, geographical and demographic expansions, practical applications, and ethical considerations. Through continued research and development, the generation of synthetic identities holds immense potential for advancing numerous fields, contributing to methodological innovations, and pushing the boundaries of what is possible in the realm of synthetic data. The insights gained from this study represent a solid foundation for these future endeavors, and we look forward to the many exciting developments that lie ahead.

## References Références Referencias


1. Bob Anderson. *Machine Learning for Fraud Detection*. Tech Publishers, 2021.
2. Alberto Bartoli and Eric Medvet. Exploring the potential of gpt-2 for generating fake reviews of research papers. In AJ Tallon Ballesteros, editor, *FUZZY SYSTEMS AND DATA MINING VI*, volume 331 of *Frontiers in Artificial Intelligence and Applications*, pages 390–396, 2020. 6th International Conference on Fuzzy Systems and Data Mining (FSDM), ELECTR NETWORK, NOV 13-16, 2020.
3. Li Chen. Addressing data imbalance in fraud detection. In *Proceedings of the 5th International Conference on Data Science*, pages 200–210, 2022.
4. Arefeh Esmaili and Saeed Farzi. Effective synthetic data generation for fake user detection. In *2021 26TH INTERNATIONAL COMPUTER CONFERENCE, COMPUTER SOCIETY OF IRAN (CSICC)*. Comp Soc Iran, 2021. 26th International Computer Conference of the Computer-Society-of-Iran, ELECTR NETWORK, MAR 03-04, 2021.
5. Hyun Kim. Improving fraud detection with synthetic identities. In *Proceedings of the 10th International Conference on Machine Learning*, pages 500–510, 2022.
6. T Kuflik, B Shapira, Y Elovici, and A Maschiach. Privacy preservation improvement by learning optimal profile generation rate. In P Brusilovsky, A Corbett, and F DeRosis, editors, *USER MODELING 2003, PROCEEDINGS*, volume 2702 of *LECTURE NOTES IN ARTIFICIAL INTELLIGENCE*, pages 168–177. Univ Pittsburgh; User Modeling Inc, 2003. 9th International Conference on User Modeling, JOHNSTOWN, PENNSYLVANIA, JUN 22-26, 2003.
7. Hyung-Jin Mun and Kun-Hee Han. Blackhole attack: user identity and password seize attack using honeypot. *JOURNAL OF COMPUTER VIROLOGY AND HACKING TECHNIQUES*, 12(3, SI):185–190, AUG 2016.
8. Nisha Patel. A study on data privacy in the age of big data. *Data Privacy and Security Review*, 25(4):450–475, 2023.
9. Inwoo Ro, Boojoong Kang, Choonghyun Seo, and Eul Gyu Im. Detection method for randomly generated user ids: Lift the curse of dimensionality. *IEEE ACCESS*, 10:86020–86028, 2022.
10. John Smith. The synthetic identity generation: An overview. *Journal of Data Privacy*, 15(2):150–180, 2022.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Critical Analysis of Solutions to Hadoop Small File Problem

By Prof. Shwetha K. S. & Dr. Chandramouli H.

*Abstract-* Hadoop big data platform is designed to process large volume of data. Small file problem is a performance bottleneck in Hadoop processing. Small files lower than the block size of Hadoop creates huge storage overhead at Namenode's and also wastes computational resources due to spawning of many map tasks. Various solutions like merging small files, mapping multiple map threads to same java virtual machine instance etc have been proposed to solve the small file problems in Hadoop. This survey does a critical analysis of existing works addressing small file problems in Hadoop and its variant platforms like Spark. The aim is to understand their effectiveness in reducing the storage/computational overhead and identify the open issues for further research.

*GJCST-C Classification:* *LCC: QA76.585*

*Strictly as per the compliance and regulations of:*

# Critical Analysis of Solutions to Hadoop Small File Problem

Prof. Shwetha K. S. [α] & Dr. Chandramouli H. [σ]

*Abstract-* **Hadoop big data platform is designed to process large volume of data. Small file problem is a performance bottleneck in Hadoop processing. Small files lower than the block size of Hadoop creates huge storage overhead at Namenode's and also wastes computational resources due to spawning of many map tasks. Various solutions like merging small files, mapping multiple map threads to same java virtual machine instance etc have been proposed to solve the small file problems in Hadoop. This survey does a critical analysis of existing works addressing small file problems in Hadoop and its variant platforms like Spark. The aim is to understand their effectiveness in reducing the storage/computational overhead and identify the open issues for further research.**

## I. Introduction

Hadoop is an open source big data processing platform designed to process large volume of data. The data is kept in form of files in Hadoop distributed file system (HDFS). A map job is spawned on a java virtual machine (JVM) instance for each file in HDFS. The file data is copied to a memory block and the block is passed to map task. In addition, a object instance is created for each file in the Namenode of Hadoop to facilitate processing. When the file size is more than or equal to block size, maximum performance gain in achieved in terms of number of maps spawned and the meta data storage overhead at Namenode. In case of IoT applications, the data files are small (less than 2KB) and when these files are stored in HDFS for data processing, it affects the Hadoop performance [1-2]. On one hand, it drastically increases the storage overhead at Namenode for object bookkeeping [3]. On another hand it exhausts the computational resources by spawning multiple map tasks which only lasts for smaller duration to process small files. The time spent in bootstrapping the map task becomes higher than data processing time in case of small files. Various solutions have been proposed addressing the Hadoop small file problem. The existing solutions can be categorized as: (i) file merging solutions, (ii) file caching solutions, (iii) optimizing Hadoop cluster structure and (iv) Map task optimizations. In file merging solutions, pre-treatment of small files is done to form a big file and this big file is stored in HDFS. In file caching solutions, files are sent to a file queue, and when queue size crosses threshold files are sent to processing in a systematic manner. In Hadoop cluster structure optimization solutions, hierarchical memory structure is created combining cache and HDFS memory to reduce the overhead due to single name node. In map task optimization solution, number of JVM instances spawned for map tasks are reduced and shared.

This work does a critical analysis on various solutions in the above four categories of file merging, file caching, Hadoop cluster structure optimization and map task optimization. The effectiveness of each of the solutions in terms of storage and computation are analyzed and their open issues are identified. Based on the open issues, a prospective solution framework is designed and detailed.

## II. Survey

Ahad et al [4] proposed a dynamic merging strategy based on the file type for Hadoop. Dynamic variable size portioning is applied to blocks and the file contents are fitted to blocks using next fit allocation policy. By this way large file is created and saved to HDFS. In addition, authors also secured the block using Twofish cryptographic technique. The solution reduced name node memory, number of data blocks and processing time. Merging was done only based on file types without considering the context and their semantic relation. Siddiqui et al [5] proposed a cache based block management technique for Hadoop as a replacement for default Hadoop Archives (HAR). A logical chain of small files is built and transferred to data blocks. In addition, efficient read/write on blocks was facilitated using block manager. Though the solution achieved more than 92% space utilization of data blocks, small files are merged only based on size, without considering the semantic relations and content characteristics. Zhai et al [6] built a index based archive file to solve the small file problem in Hadoop. The small files are merged to large file and metadata record is created to retrieve each file content. Meta data records are arranged into buckets. An order preserving hash is created over metadata records. The hash and the metadata records are in turn written to a index file. The index files helps to retrieve the file contents for processing. This method is able to save atleast 11% disk space but the solution access efficiency becomes

*Author α: Ph.D Research Scholar Department of Computer Science & Engg., East Point College of Engineering and Technology, Bengaluru, Karnataka, India. e-mail: shwethaise.nhce@gmail.com*
*Author σ: Dr. Chandramouli H Professo Department of Computer Science & Engg., East Point College of Engineering and Technology, Bengaluru, Karnataka, India.*
*e-mail: hcmcool123@gmail.com*

lower with large number of small files. Also the indexing does not support streaming inputs. Cai et al [7] proposed a file merging algorithm based on two factors of distribution of the files and the correlation of the file. Correlation between files is built based on their history of access and the highly correlated files are kept in the same block. Through experiments, author found that placing highly correlated files in same block improved the speed up. The correlation is not based on content characteristics so over a period of time, performance can reduce. Choi et al [8] integrated combinedfileinputformat and JVM reuse to solve the small file problem. Small files are combined till block size and passed to map task. JVM instances are reused for the map task , so they overhead of JVM bootstrap is minimized. Though the integration reduces the computational overhead, the approach combined files in order without considering their semantics. Also the memory buildup due to JVM reuse can crash the tasks due to inefficient memory management. Peng et al [9] combined merging and caching techniques to solve the small file problem. User based collaborative filtering is applied to learn the correlation between the files. Files with higher correlation are merged into single large file. Remote procedure call (RPC) requests to fetch the block information about the files are reduced by caching the access requests and looking into cache for the blocks before placing RPC requests. By this way, authors were able to reduce the file access time by 50% and increase storage utilization by 25% compared to default Hadoop. The scheme does not works well for streaming data, as the correlation model proposed in this work is not adaptive to streaming data. Niazi et al [10] proposed a new technique called inode stuffing to solve the small file problem. For small files, the metadata and data block are combined and decoupling is maintained only for large files. The approach is not scalable as it increases the metadata storage overhead at Namenodes. Jing et al [11] proposed a dynamic queue method to solve the small file problem. The files are first classified using the period classification algorithm. The algorithm calculates similarity score based on sentence similarity between two documents. The similar files are then merged to large file using multiple queues for specific file sizes. Authors also used file pre-fetching strategy to improve the efficiency of file access. Analyzing similarity between pairs is a cumbersome task for large number of files. Sharma et al [12] proposed a dual merge technique called Hash Based-Extended Hadoop Archive to solve the small file problem in Hadoop. The small files are merged using two level compaction. This reduces the storage overhead at Namenode and increase the data block space utilization at Datanodes. File access is made efficient using two level hash function. The proposed solution is atleast 13% faster compared to default Hadoop. The files were merged without considering the content characteristics and their semantics. Wang et al [13] combined merging and caching to solve the small file problem in Hadoop. Authors proposed a equilibrium merger queue algorithm to merge small files to Hadoop block size and then merged file is saved to HDFS. Indexing is built to access small files. To reduce the communication overhead between the client and Namenode for small file access, pre-fetched cache is used. With the cache, the number of RPC calls to name node is reduced. The memory consumption at Namenode drastically reduced in the proposed solution compared to default Hadoop Archives. Contents were merged without considering their content characteristics and semantic correlation. Ali et al [14] proposed a enhanced best fit merging algorithm to merge small files based on type and size. The merging is done till Hadoop block size is reached and merged file is saved to HDFS. Author found that merging improved Hadoop storage utilization by 64% but the file access time was higher in this work. Prasanna et al [15] compressed many small files into a zip file to the size of Hadoop data block and saved to disk. This increased the disk utilization of data nodes and name nodes. But the computational overhead in compressing stage and decompressing during processing is higher. Huang et al [16] addressed the small file problem for the case of images in Hadoop. A two level model was proposed specific to medical images. The images were grouped at first level based on series and next level based on examination. The grouped images are saved to data blocks in HDFS. Indexing and pre-fetching is done to done is reduce the access time for small image files. The pre-fetching algorithm did not have higher cache hit. Renner et al [17] extended the Hadoop archive to appendable file format to solve the small file problem. Small files are appended to existing archive data files whose block size is not completely used. Authors used first fit algorithm to select the data blocks. In addition indexing is done to facilitate faster access. Red black tree structure is used for indexing for efficient lookup. Though this scheme improved the data block utilization, appending is done without considering content characteristics and semantic similarity. Liu et al [18] proposed a file merging strategy based on content similarity. Files are converted to vector space features and correlation between the features is measured using cosine similarity. When cosine similarity is greater than threshold, files are merged. In addition authors used pre-fetching and caching to speed up the file access. Constructing a global feature space for streaming data is difficult and thus this approach is not suitable for streaming data.Lyu et al [19] proposed an optimized merging strategy to solve small file problem. The small files are merged based on size in such that way block size is fully utilized. In addition authors used pre-fetching and caching to increase the access speed. Only block size utilization was considered as the only criteria for

merging without considering content characteristics and semantic relations. Similar to it Mu et al [20] proposed an optimization strategy to maximally fill the existing Hadoop archive by appending small files. In addition author also used secondary index to speed up the execution of file access. But here too merging was done without considering content characteristics and semantic relation. Wang et al [21] used probabilistic latent semantic analysis to determine the user access pattern and based on it small files are merged to a large file and placed in HDFS. In addition author also improved the pre-fetching hit ratio based user access transition pattern. Both the strategies improvised the speed of access and data block utilization. But this scheme is not suitable for multi user environment as for each user, a merging order must be kept and this increases the storage overhead. He et al [22] merging the small files based on balance of data blocks. The aim was to increase the data block utilization. Merging did not consider content characteristics and their semantic relation. Fu et al [23] proposed an flat storage architecture to handle the small files. In this scheme, both files and meta data are collocated with meta size fixed for any number of small files. This is facilitates by meta data having only pointer to related information in its index. But the scheme is not suited for Hadoop as collocation causes higher access overhead for large files. Tao et al [24] merged small files to large file and built a linear hash to small files to speed up access. File size was the only criteria considered for merging. Bok et al [25] integrated file merging and caching to solve the small file problem. Author used two level of cache for small files, so that access requests to –

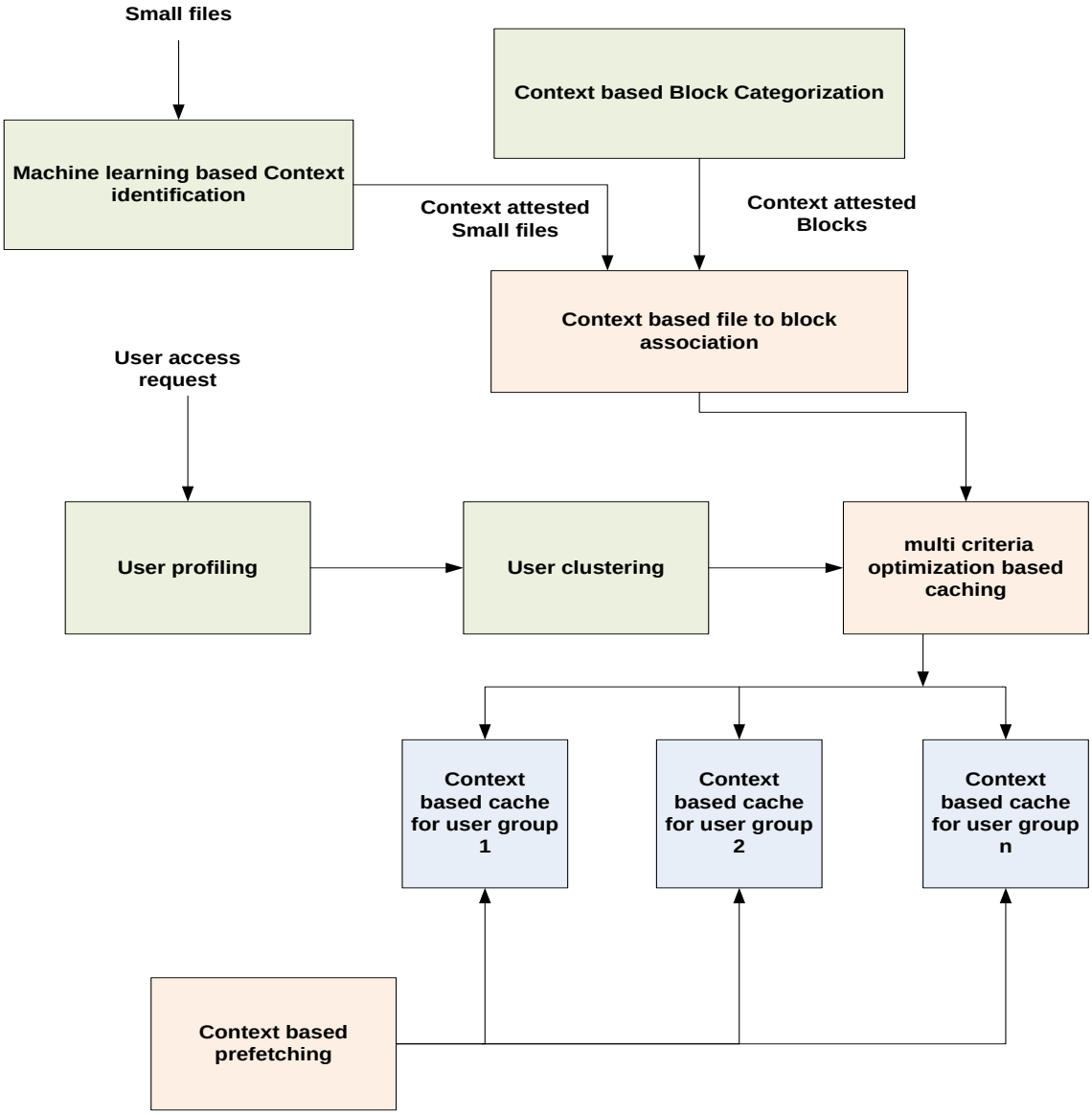



*Figure 1:* Research direction framework

*Table 1:* Survey Summary

| Work | Solution for Small file Problem | Gap |
|---|---|---|
| Ahad et al [4] | dynamic merging strategy based on the file type | Merging was done only based on file types without considering the context and their semantic relation |
| Siddiqui et al [5] | cache based block management technique | small files are merged only based on size, without considering the semantic relations and content characteristics |
| Zhai et al [6] | a index based archive file with order preserving hash for speedup | Does not support streaming |
| Cai et al [7] | file merging algorithm based on two factors of distribution of the files and the correlation of the file | The correlation is not based on content characteristics |
| Choi et al [8] | integrated combinedfileinputformat and JVM reuse to solve the small file problem | memory buildup due to JVM reuse can crash the tasks due to inefficient memory management |
| Peng et al [9] | combined merging and caching techniques to solve the small file problem | The scheme does not works well for streaming data, as the correlation model proposed in this work is not adaptive to streaming data |
| Niazi et al [10] | Coupling both meta data and small file together. | The approach is not scalable as it increases the metadata storage overhead at Namenodes |
| Jing et al [11] | Files classified using the period classification algorithm and merged based on similarity | Analyzing similarity between pairs is a cumbersome task for large number of files |
| Sharma et al [12] | Hash Based-Extended Hadoop Archive to solve the small file problem | The files were merged without considering the content characteristics and their semantics. |
| Wang et al [13] | combined merging and caching to solve the small file problem | Contents were merged without considering their content characteristics and semantic correlation |
| Ali et al [14] | enhanced best fit merging algorithm to merge small files based on type and size. | file access time was higher in this work |
| Huang et al [16] | A two level model was proposed specific to medical images | The pre-fetching algorithm did not have higher cache hit |
| Renner et al [17] | Small files are appended to existing archive data files | Appending is done without considering content characteristics and semantic similarity |
| Liu et al [18] | File content based merging | Constructing a global feature space for streaming data is difficult and thus this approach is not suitable for streaming data |
| Lyu et al [19] | optimized merging strategy to solve small file problem. | Only block size utilization was considered as the only criteria for merging without considering content characteristics and semantic relations |
| Wang et al [21] | probabilistic latent semantic analysis to determine the user access pattern and based on it small files are merged to a large file | scheme is not suitable for multi user environment as for each user, a merging order must be kept and this increases the storage overhead |
| He et al [22] | merging the small files based on balance of data blocks | Merging did not consider content characteristics and their semantic relation |
| Fu et al [23] | flat storage architecture collocating metadata and file in same object | the scheme is not suited for Hadoop as collocation causes higher access overhead for large files |
| Tao et al [24] | merged small files to large file and built a linear hash to small files to speed up access | File size was the only criteria considered for merging |
| Bok et al [25] | integrated file merging and caching to solve the small file problem | The merging was based only on size without considering the content characteristics and semantic similarity |

Namenode is totally minimized. Least recently used (LRU) mechanism is used to upgrade the cache. The merging was based only on size without considering the content characteristics and semantic similarity.

The summary of survey so far discussed is presented in Table 1.

## III. Open Issues

From the survey, following three open issues are identified.

1. Context specific merging
2. Personalized access
3. Streaming support

*Context specific merging:* In most of the existing approaches, merging was based only on size. Merging did not consider user access or application contexts, content characteristics and their semantic relation. In applications like recommendations based on user comments, it is necessary to co-locate user comments related to specific product characteristics in same blocks for application speedup.

*Personalized Access:* In most of the existing caching strategies, caching was based on least recently used at a global context without considering the user access context. But it is important to consider user access context as each user access behavior is different. Caching on global context can provide better performance for some users and can give worst performance for other users. To solve this access time discrepancy among the users, personalized caching strategy must be employed.

*Steaming Support:* Most of the merging schemes does not handle the steaming data effectively. Streaming data content similarity cannot be computed effectively using vector space modeling and their merging can become ineffective. Merging based on streaming arrival patterns has not been considered in earlier works.

## IV. Research Directions

Based on the open issues identified, a prospective framework for further research is presented in Figure 1.

The framework addresses three problem areas of context specific merging, personalized access and streaming support.

*Context Specific Merging:* It can be facilitated and made adaptive using machine learning. Based on the application contexts and inherent data characteristics the files to be merged can be found. Blocks can be categorized based on context and small files can be categorized based on context. Context based merging is the realized to merge files and blocks based on context similarity. Instead of flat context, hierarchical context can be learnt automatically from file summarization. File summarization strategies specific to file types can be proposed to identify the context to be associated with files and blocks.

*Personalized Access:* User can be clustered based on their content access patterns over a temporal duration and multiple caches can be maintained for each user group. Also the cache item management can be based on multi criteria optimization instead of LRU mechanisms. The items to pre-fetch can be identified based on context associated with files. By this way access speed up can be increased and optimized specific to each user group.

*Streaming Support:* To support streaming data, the context must be learnt dynamically in a light weight manner and association of small file to blocks must be done based on context. To learn context in a light weight manner, the streaming data characteristics and their arrival patterns must be used.

## V. Conclusion

This survey made a critical analysis of existing solutions for small file problem in Hadoop. The solutions were analyzed in four categories of file merging solutions, file caching solutions, optimizing Hadoop cluster structure and Map task optimizations. Based on the survey, three open issues of context specific merging, personalized access and streaming support are identified. Prospective solutions to these three open issues were identified and a solution roadmap for further exploration in this area was documented.

## References Références Referencias


1. Small size problem in Hadoop: http://blog.Cloudera.com/blog/2009/02/the-small-files-problem/
2. Solving Small size problem in Hadoop https://pastiaro.wordpress.com/2013/06/05/solving-the-small-files-problem-in-apache-hadoop-appending-and-merging-in-hdfs/
3. Bo Dong, Qinghua Zheng, Feng Tian, Kuo-Ming Chao, Rui Ma, Rachid Anane. (2012), An optimized approach for storing and accessing small files on cloud storage, Journal of Network and Computer Applications, 35 (2012) 1847-1862, Elsevier.
4. Ahad, Mohd & Biswas, Ranjit. (2018). Dynamic Merging based Small File Storage (DM-SFS) Architecture for Efficiently Storing Small Size Files in Hadoop. Procedia Computer Science. 132. 1626-1635. 10.1016/j.procs.2018.05.128.
5. Siddiqui, Isma & Qureshi, Nawab Muhammad Faseeh & Chowdhry, Bhawani & Uqaili, Mohammad. (2020). Pseudo-Cache-Based IoT Small Files Management Framework in HDFS Cluster. Wireless Personal Communications. 113. 10.1007/s11277-020-07312-3.
6. Zhai, Yanlong & Tchaye-Kondi, Jude & Lin, Kwei-Jay & Zhu, Liehuang & Tao, Wenjun & Du, Xiaojiang

& Guizani, Mohsen. (2021). Hadoop Perfect File: A fast and memory-efficient metadata access archive file to face small files problem in HDFS. Journal of Parallel and Distributed Computing. 156. 10.1016/j.jpdc.2021.05.011.
7. Cai, Xun & Chen, Cai & Liang, Yi. (2018). An optimization strategy of massive small files storage based on HDFS. 10.2991/jiaet-18.2018.40.
8. Choi, C., Choi, C., Choi, J. et al. Improved performance optimization for massive small files in cloud computing environment. Ann Oper Res 265, 305–317 (2018).
9. Peng, Jian-feng & Wei, Wen-guo & Zhao, Hui-min & Dai, Qing-yun & Xie, Gui-yuan & Cai, Jun & He, Ke-jing. (2018). Hadoop Massive Small File Merging Technology Based on Visiting Hot-Spot and Associated File Optimization: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings. 10.1007/978-3-030-00563-4_50.
10. S. Niazi, M. Ronström, S. Haridi, and J. Dowling, 'Size Matters : Improving the Performance of Small Files in Hadoop', presented at the Middleware'18. ACM, Rennes, France, 2018, p. 14.
11. Jing, Weipeng & Tong, Danyu & Chen, GuangSheng & Zhao, Chuanyu & Zhu, LiangKuan. (2018). An optimized method of HDFS for massive small files storage. Computer Science and Information Systems. 15. 21-21. 10.2298/CSIS171015021J.
12. V. S. Sharma, A. Afthanorhan, N. C. Barwar, S. Singh and H. Malik, "A Dynamic Repository Approach for Small File Management With Fast Access Time on Hadoop Cluster: Hash Based Extended Hadoop Archive," in *IEEE Access*, vol. 10, pp. 36856-36867, 2022
13. K. Wang, Y. Yang, X. Qiu and Z. Gao, "MOSM: An approach for efficient storing massive small files on Hadoop," *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, Beijing, China, 2017, pp. 397-401
14. A. Ali, N. M. Mirza and M. K. Ishak, "Enhanced best fit algorithm for merging small files," Computer Systems Science and Engineering, vol. 46, no.1, pp. 913–928, 2023.
15. L. Prasanna. Kumar, "Optimization Scheme for Storing and Accessing Huge Number of Small Files on HADOOP Distributed File System". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 2, Feb. 2016, pp. 315-9
16. Xin Huang, Wenlong Yi, Jiwei Wang, Zhijian Xu, "Hadoop-Based Medical Image Storage and Access Method for Examination Series", Mathematical Problems in Engineering, vol. 2021, Article ID 5525009, 10 pages, 2021.
17. Thomas Renner, Johannes Müller, Lauritz Thamsen, and Odej Kao. 2017. Addressing Hadoop's Small File Problem With an Appendable Archive File Format. In Proceedings of the Computing Frontiers Conference (CF'17). Association for Computing Machinery, New York, NY, USA, 367–372.
18. Liu, Jun. (2019). Storage-Optimization Method for Massive Small Files of Agricultural Resources Based on Hadoop. Journal of Advanced Computational Intelligence and Intelligent Informatics. 23. 634-640. 10.20965/jaciii.2019.p0634.
19. Y. Lyu, X. Fan, and K. Liu, "An optimized strategy for small files storing and accessing in HDFS," in Proc. IEEE Int. Conf. CSE, IEEE Int. Conf. EUC, Jul. 2017, pp. 611_614.
20. Q. Mu,Y. Jia, and B. Luo, ``The optimization scheme research of small files storage based on HDFS," in Proc. 8th Int. Symp. Comput. Intell. Design, Dec. 2015, pp. 431_434.
21. T. Wang, S. Yao, Z. Xu, L. Xiong, X. Gu, and X. Yang, ``An effective strategy for improving small _le problem in distributed file system," in Proc. 2nd Int. Conf. Inf. Sci. Control Eng., Apr. 2015, pp. 122_126
22. H. He, Z. Du, W. Zhang, and A. Chen, ``Optimization strategy of Hadoop small_le storage for big data in healthcare," J. Supercomput., vol. 72, no. 10, pp. 3696_3707, Aug. 2016
23. S. Fu, L. He, C. Huang, X. Liao, and K. Li, ``Performance optimization for managing massive numbers of small files in distributed file systems," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 12, pp. 3433_3448, Dec. 2015
24. W. Tao, Y. Zhai, and J. Tchaye-Kondi, "LHF: A new archive based approach to accelerate massive small _les access performance in HDFS", in Proc. 5th IEEE Int. Conf. Big Data Service Appl., Apr. 2019, pp. 40_48.
25. K. Bok, H. Oh, J. Lim, Y. Pae, H. Choi, B. Lee, and J. Yoo, ``An efficient distributed caching for accessing small files in HDFS," Cluster Comput., vol. 20, no. 4, pp. 3579_3592, Dec. 2017.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Literature Study on Analyzing and Designing of Algorithms

By Sneha Kumari & Aishwarya
*Ajeenkya D.Y. Patil University*

*Abstract-* The fundamental goal of problem solution under numerous limitations, such as those imposed by issue size, performance, and cost in terms of both space and time. Designing a quick, effective, and efficient solution to a problem domain is the objective. Certain problems are simple to resolve while others are challenging. To develop a quick and effective answer, much intelligence is needed. A new technology is required for system design, and the foundation of the new technology is the improvement of an already existing algorithm. The goal of algorithm research is to create effective algorithms that improve scalability, dependability, and availability in addition to cutting costs and turnaround times.

*Keywords:* *analysis, solution, time, algorithm, optimal, complexity, computing, application, space, design.*

*GJCST-C Classification:* *ACM: F.2.2, G.2.2*


*Strictly as per the compliance and regulations of:*

# Literature Study on Analyzing and Designing of Algorithms

Sneha Kumari [α] & Aishwarya [σ]

*Abstract-* The fundamental goal of problem solution under numerous limitations, such as those imposed by issue size, performance, and cost in terms of both space and time. Designing a quick, effective, and efficient solution to a problem domain is the objective. Certain problems are simple to resolve while others are challenging. To develop a quick and effective answer, much intelligence is needed. A new technology is required for system design, and the foundation of the new technology is the improvement of an already existing algorithm. The goal of algorithm research is to create effective algorithms that improve scalability, dependability, and availability in addition to cutting costs and turnaround times.

*Keywords: analysis, solution, time, algorithm, optimal, complexity, computing, application, space, design.*

## I. Introduction

Design and analysis of algorithms is referred to as DAA. It aids in the analysis of the answer prior to coding. Algorithms and documentation can be used to determine the space and time complexity. A clear image of the code you will write to address the problem is provided by algorithms and designs. It enables you to obtain the optimal time and spatial complexity for a shorter solution. The standards formeasuring algorithms before we can create effective ones. Algorithms are rated according to the amount of computing resources they need. The majority of these resources are running time and memory. Other factors may also be taken into consideration depending on the application, such as the volume of disc visits in a database programme or the amount of communication bandwidth in the networking application. The design of the algorithms must take into account a variety of challenges that arise in practice. Algorithms are instructions that you create in order to solve a complicated problem. You create these instructions by carrying out various computations, processing data, and scenario.

The methods are follows to solve a problem using descriptions of how to employ time and space resources are known as algorithms. Prior to implementing the actual code, you may use algorithms to learn more about the time and spatial complexity. Algorithms resemble technology in many ways. Although we all have the newest CPUs, we still need to run implementations of effective algorithms on that machine in order to get the full benefits of our investment in the most recent processor. When you develop the algorithms for the specific problem, you can determine the optimum solution. It is the most effective technique to illustrate any issue with the finest and most practical answers.

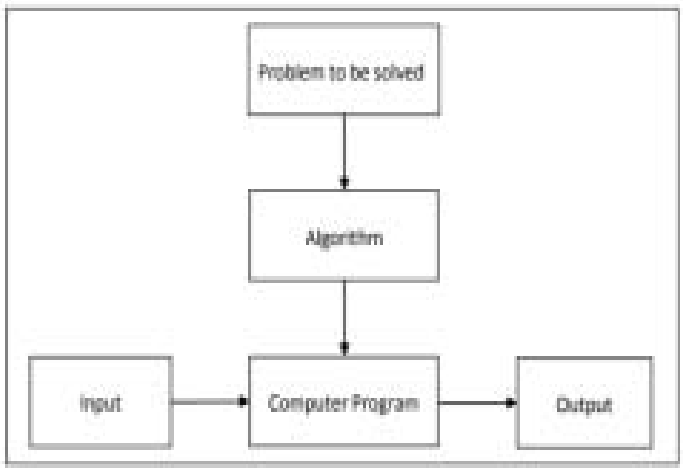



*Fig. 1:* The Notion of the Algorithm

*Author α: MCA (Data Science) Ajeenkya D.Y. Patil University Pune, Maharashtra, India. e-mail: sneha.kumari@adypu.edu.in*
*Author σ: Ajeenkya D.Y. Patil University Pune, Maharashtra, India. e-mail: Aishwarya@inurture.co.in*

## II. Need for Algorithmic Problem

In computer science, algorithms are vitally significant.

- Acquire a thorough grasp of the problem.
- To identify the best answer to the problem.
- Allows for scalability. As it helps with comprehending, break the problem down into smaller steps.
- Being aware of the design guiding concepts and algorithms.
- Make use of the greatest technology to obtain the best and most effective solution.
- Getting thorough knowledge about the problem is impossible without implementation.

## III. Procedural for Algorithmic Problem Solving

An initial input and a list of instructions are used by algorithms. The user's input, which may be expressed as words or numbers, is the first piece of information required to build judgements. The provided information is subjected to a series of computations, which may involve mathematical operations and moral assessments. The final step of an algorithm is called the output, and it is typically stated as more data. The picture below depicts the stages that go into creating and analyzing an algorithm.

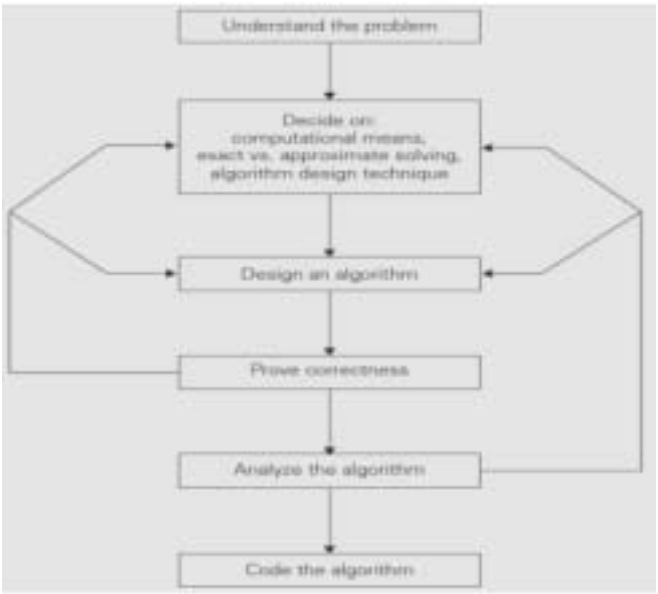



*Fig. 2:* Algorithm Design and Analysis Process [1]

*a) Problem Recognition*

Read the problem's description attentively to fully comprehend the problem statement; this is the 1st step in constructing an algorithm.

*b) Making Decisions*

*Decisions are made based on the following:*

a. Determining the Computational Device's Capabilities: In a RAM i; e random access machine, instructions are carried out one at a time (this is the underlying premise). As a result, algorithms created to run on these devices are known as sequential algorithms.
b. Selecting between exact and approximate problem-solving techniques. The choice between tackling the problem precisely or roughly is the next crucial option. An exact algorithm is one that solves a problem precisely and yields the desired outcome. When a problem is too complicated to have a precise solution, we must use a technique known as an approximation algorithm.
c. Techniques for designing algorithms: It is design methodology is a comprehensive approach for problem-solving that may be used to a variety of situations from different computing areas.

Programme = Algorithms + DS (Data Structures)

Although data structure & algorithm are separate concepts, programme is developed by combining them. Therefore, selecting the appropriate DS i; e data structures is necessary before constructing the algorithm. Algorithm implementation is only achievable with the aid of data structures and algorithms. Algorithmic strategy, methodology, and paradigm are a generic method for solving a variety of issues algorithmically. Examples include using "brute force," "divide and conquer," "dynamic programming," "greedy technique".

*c) Methods for Specifying an Algorithm*

An algorithm can be specified in three different ways. As follows: A flowchart, natural language &

pseudocode. The two methods that are most frequently used nowadays for describing algorithms are pseudocode and flowcharts.

*Natural Language*

Using plain language to describe an algorithm is really straightforward and simple. However, using normal language to describe an algorithm is not always straightforward, thus we only obtain a brief definition.

*Pseudocode*

It combines elements of normal language with those of programming languages. Natural English is frequently less exact than pseudocode.

*A flowchart*

Flowcharts were formerly the standard for expressing algorithms in the early days of computers, but this way of representation has since proven to be inconvenient. An algorithm is graphically represented by a flowchart. It is a way of representing an algorithm using a network of linked geometric forms that each include descriptions of an algorithm step.

*d) Proving the Accuracy of an Algorithm*

An algorithm's correctness must be established once it has been stated. An algorithm must provide a needed result in a finite period of time for each valid input.

*e) Analysis of an Algorithm*

The most crucial factor for an algorithm is efficiency. There are actually 2 types. They are efficiency in time, which measures how quickly the algorithm executes, and efficiency in space, which measures how much more memory it consumes. Therefore, the following criteria should be considered while analyzing an algorithm time efficiency, space efficiency, simplicity, and generality.

*f) Code for Algorithm*

An appropriate programming language is used to code or implement an algorithm. It is possible to make the conversion from an algorithm to a programme improperly or extremely inefficiently. An algorithm must be appropriately implemented. Writing efficient, optimized code is crucial if you want to lighten the load on the compiler.

## IV. Algorithm Characteristics

Each algorithm should possess the following six essential characteristics:

A) Input-One or more inputs may be present in an algorithm. The inputs are taken from a predetermined group of participants. Any form of file can be entered, including text, pictures, and images.

B) Output- It can provide one or more results i; e. output. It is essentially number that has a predefined relationship with the input.

C) Finiteness-An algorithm should end after a finite number of steps, then only it considers as computational method.

D) Certainty- An algorithm must have every step well described. For each scenario, the action that has to be taken must be vaguely described. Because the step is difficult to grasp, one would assume that it lacks definiteness. As a result, mathematical expressions are expressed in these situations in a way that is similar to how instructions are written in a computer language.

E) Efficiency- An algorithm is typically assumed to be efficient. means that the processes should be sufficiently simple that a man might be able to solve them.

F) Language Independence - An algorithm type should be languages-independent, meaning that its instructions or commands must function consistently regardless of the language in which they are implemented.

## V. Guidelines to Be followed for Development of Algorithm

The following guidelines must be adhered to while developing an algorithm:

- An algorithm will be surrounded by the symbols START (or BEGIN) and STOP (or END).
- The words OBTAIN, GET, READ, and INPUT are frequently employed to accept data from users.
- To display results or messages, statements like WRITE, PRINT, and DISPLAY are frequently used.
- Mathematical expressions are often denoted by the terms CALCULATE or COMPUTE and depending on the situation, appropriate operators may be used.

## VI. Anaysis of an Algorithm & its Method

A method for evaluating an algorithm's performance is algorithm analysis. The time and spatial complexity are the main variables on which the algorithms rely. Two algorithms are examined using asymptotic analysis to see how well they perform when the input size is changed (increased or reduced).

i. Worst Case Analysis- It is the algorithm's worst case, or the circumstance that causes the majority of operations to be carried out must be understood. In the worst case, we are able to determine an algorithm's upper bound running time. When the sought-after element (x) is not present in the array, linear search experiences its worst-case scenario. The search () function checks each member of arr[] independently if x is absent. The worst-case temporal complexity of the linear search would thus be O(n). The worst-case situation is represented by the Big O notation. Only the upper bound of time is computed when the procedure is applied.

ii. Best Case Analysis- The scenario in which an algorithm is run with the fewest number of operations is known as its "best case." It establishes the algorithms lower bound for execution time in the best-case scene. It is necessary to understand the situation that just executes a few activities. When x appears at the first position, the best case for the linear searching problem happens. In the best situation, the number of processes is fixed and independent of n. Thus, for time complexity, (1) is the best-case situation. The best scenario is represented by Omega notation. When the method is used, just the lowest bound of time is calculated.

iii. Average Case Analysis-It is an algorithm, which is the scenario in which the algorithm becomes aroused after a few operations. For example, when we execute a linear search technique in any data structure and find an element at the midway place, that scenario is referred to as the average case. When studying typical instances, Theta Notation is used. It establishes the complexity of time with the aid of the upper and lower bounds.

## VII. Advantages of an Algorithm

1. Since it shows a step-by-step approach to solving a particular problem, it is easy to understand.
2. An algorithm executes a predefined procedure.
3. Because each step has its own logical sequence, an algorithm is easy to debug.
4. An algorithm is used to divide the problem into smaller components or stages, which makes it simpler for programmers to turn the problem into usable software.
5. It is independent what programming language is used.

## VIII. Disadvantages of an Algorithm

1. Algorithms take a lot of time.
2. It's challenging to demonstrate looping and branching in algorithms.
3. Big challenges are hard to describe and even more challenging to write algorithms for.

## IX. Types of Algorithms

*a) Brute Force Algorithm*

The most fundamental algorithm that may be developed to address a problem is of this type. We must first identify at least one answer before attempting to enhance it in order to create the optimal one. The most simple and fundamental algorithm is one of them. Any problem can be solved using the brute force approach; however, it often doesn't add much time or space complexity.

*b) Recursive Algorithm*

It is the easiest algorithms to create since it doesn't need to individually consider each sub problem. When the problem scale is substantially decreased, the recursive algorithm procedure converts the problem into a smaller scale but similar form problem, which is then solved. Among these, the basic issue-solving process is characterized by self-reference at the level of recursive description, where the scenario and approach that may solve the problem directly are defined. Similar to mathematical induction, the fundamental concept of recursive process description involves quoting oneself in order to minimize the complexity of the problem. Recursion is a very efficient approach, but since it calls a recursive stack each time the recursion function is invoked, memory management must always be kept in mind. When the complexity is reduced to a given extent, the problem is then directly solved.

i. *Divide and Conquer Algorithm*

This is one of the techniques that programmers utilize the most. With this approach, the problems are divided into smaller ones, each of which is solved independently, before the combined solutions are used to determine the solution to the initial challenges. As it is relatively stable and ideal for the majority of the challenges posed, this algorithm is widely employed in a variety of problems. When deciding how to address a problem, the divide-and-conquer tactic is widely used. Strassen's Matrix Multiplication, Merge Sorting, Binary Search, Quick Sorting, etc. are a few typical issues that are resolved utilizing Divide and Conquer algorithms.


*Fig. 1:* Divide and Conquer Algorithm [15]

ii. *Dynamic Programming Algorithms*

This type of algorithm is most efficient ways of problem solutions, this algorithm is the most popular. This technique is very efficient in terms of time complexity since it just requires recalling previous results and applying it to future results that correspond. Since this type of procedure maintains the previously computed answer in order to avoid having to compute it repeatedly, it is also known as the recalled technique.

*There are two versions of this algorithm:*

*Bottom-Up Approach:* This method begins by resolving the smallest feasible subproblems first, building on the answers obtained from those subproblems to solve the larger problem.

*Top-Down Approach:* This method begins by resolving all of the problems until it reaches the necessary subproblem, which is then addressed utilizing subproblems that have already been resolved.

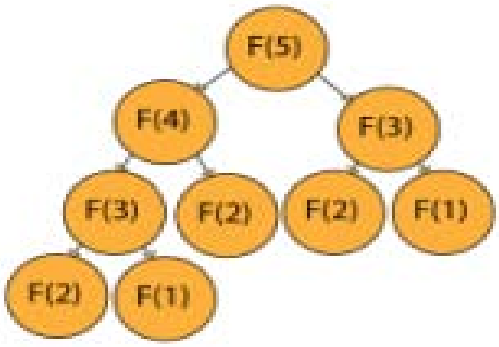



*Fig. 2:* Fibonacci Series in Dynamic Algorithm [12]

iii. *Greedy Algorithms*

This algorithm does not consider the future while making decisions; instead, it considers the situation at hand. It doesn't matter if the best outcome at the moment leads to the best result altogether. A greedy algorithm gradually assembles an approach, always choosing as the next step the element that offers the most apparent and immediate advantage. Greedy so works well with problems when choosing locally optimal also leads to a global solution. Although the algorithm that is greedy is not consistently successful, when it exists, it is fantastic! This method is typically the simplest since it is easy to develop. It's probable that this method won't work for all problems. But, if the problem has any of the following characteristics, we can decide if this approach can be applied to any of the problem cases.

*A Greedy Property Choice*

A greedy method can be used to tackle a problem if it is possible to make the best or most advantageous option at each stage without going back and changing the decision made at the prior stage. The "greedy choice property" is the name given to this trait.

*The Runner-Up Substructure*

If the most effective solution to the challenge is also the most effective solution to each of its subproblems, the problem can be solved using a greedy approach.

This trait is known as "optimal substructure". It achieves this because it is continuously working to get the best result possible locally. Examples of common cases or problems that the Greedy Algorithm solves includes the Kruskal's Algorithm, Prim's Algorithm, Dijkstra Shortest Path Algorithm, Huffman Coding, and others.

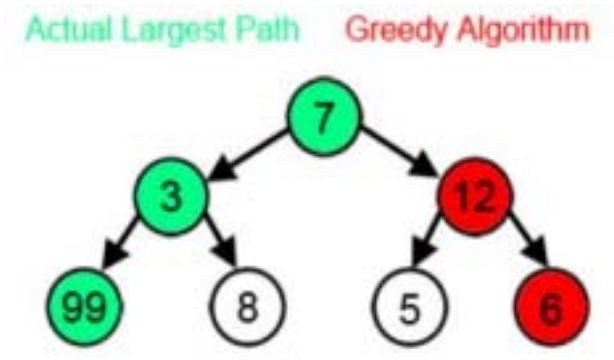



*Fig. 3:* Greedy Algorithm [15]

*Backtracking Algorithms*

It is based on a depth-first recursive search. It is an improvement over using raw force. Here, we choose one choice from the many that are available and try to solve the problem. The Brute force approach, which evaluates each potential answer, is used to choose the desired/best solutions. It is an algorithmic method for recursively addressing problems. The Backtracking Algorithm may be used to solve problems like the Hamiltonian Cycle, Rat in Maze Problem, the N Queen Problem, the M-Coloring Problem, etc.

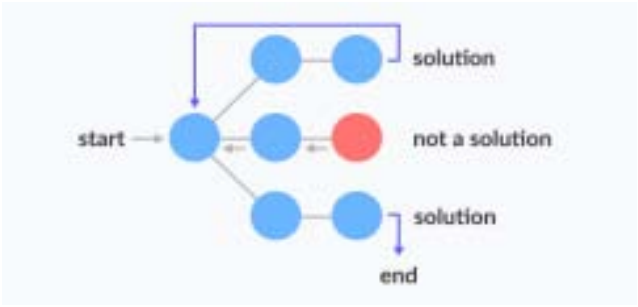



*Fig. 4:* Backtracking Algorithm [13]

*c) Randomized Algorithm*

This sort of algorithm bases its conclusions on random numbers, i.e., it incorporates random numbers into its reasoning. Selecting the desired result is helpful. the process of choosing a number at random that offers an instant benefit. One of the problems that the randomized Algorithm could fix is quicksort. The pivot in Quicksort is selected at random.

*d) Searching Algorithm*

A searching algorithm is a method for finding a certain key among a group of sorted or unordered data. A number of problems may be solved using the searching algorithm, such as the following: Binary search, sometimes referred to as linear search, is one form of search technique.

## X. Conclusion

Algorithms may be used by both individuals and machines to carry out routine activities. The primary distinction is that computers employ algorithms far more quickly and effectively than we can. A series of actions used to solve a problem is called an algorithm. In the field of information technology and computer science, building algorithms to tackle various sorts of problems requires careful planning and analysis. A problem that needs to be solved initiates the process of designing an algorithm, which is then followed by the classification of the problem's type into the categories listed above, the implementation of the algorithm, and finally an evaluation of the finished algorithm's efficiency (both in terms of time and space). The computer theory of complexity, which offers a theoretical estimate of the resources needed for an algorithm to effectively address a certain computer issue, includes algorithm analysis as a key component. Analysis is used to calculate how much space and time are needed to run a programme. Applying various algorithmic design techniques, such as divide-and-conquer, greedy, and others, to real-world issues. The capacity to comprehend and calculate the algorithm's performance. Algorithms are frequently simple to design, simple to implement, and quick to execute. Insidiously difficult mathematical proofs may be needed to demonstrate their correctness.

## References Références Referencias

1. https://www.brainkart.com/article/Algorithmic-problem solving_35898/.
2. Thomas H Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein. Introduction to algo-rithms, third edition. pp 360- 395.
3. Brassard, G. and Bratley, P. Fundamental of Algorithm- mics, Prentice-Hall, 1996.
4. Kashale Chimmanga, Josephat Kalezhi and Phillimon Mumba, “Application of best first search algorithm to demand control”, 2016 IEEE PES Power Africa Conference,pp. 51-55,IEEE 2016.
5. Sankar Peddapati and K.K. Phanisri Kruthiventi, ”A New Random Search Algorithm: Multiple Solution Vector Approach”, 2016 6th International Advanced Computing Conference,pp.187-190, IEEE 2016..
6. D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, New York: Addison-Wesley, 1989.
7. J. Cioffi, "The block-processing TF adaptive algorithm," IEEE Trans. Acoust., Speech, Signal Processing, vol. A SSP-34, no. 1, pp. 77-90, 1986. .
8. Ziewitz M (2015) Governing algorithms: Myth, mess, and methods. Science, Technology & Human Values 41(4): 3– 16.
9. M. Young, The Technical Writers Handbook, Mill Valley, CA: University Science, 1989.
10. Jiang Na, Yang Haiyan, Gu Qingchuan, Huang Jiya. Machine learning and its algorithm and development analysis [J]. Information and Computer Science (Theoretical Edition), 2019 (01): 83-84 + 87.

11. Montazeri and P. Duhamel, "A set of algorithms linking NLM S and RLS algorithms," in Proc. EU SP ICO-94, 1994, vol. 2, pp. 744-747.
12. https://stackabuse.com/dynamic-programming-in-java/.
13. https://www.programiz.com/dsa/backtra-cking-algo ithm.
14. https://favtutor.com/blogs/divide-and-conquer-algo rithm.
15. https://vikram-bajaj.gitbook.io/cs-gy-6033-i-design-and analysis-of-algorithms-1/chapter1.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Application of Meta-Programming Techniques for Accelerating Software Development and Improving Quality

By Amirali Kerimovs

*Annotation-* A contemporary software tool has been devised to evaluate software quality through metric analysis techniques. This tool calculates pertinent metrics utilizing quality indicators and establishes a composite quality indicator value for software products. The intricacies of software quality assessment processes have been elucidated, including the examination of software quality's standardization as well as the presentation level of its model. This enables the potential for enhancement through the formulation of suitable criteria for quality assessment, refining models for metric analysis, and quantitatively measuring quality across all phases of project implementation. Notably, the use of metric analysis to gauge software quality reveals a lack of standardized metrics, resulting in varying assessment methods and metrics from different measurement system providers. Interpreting metric values also proves challenging for most software users due to a lack of clarity and informativeness. Furthermore, it has been discovered that while decisions based on cost, development duration, and designer company reputation influence software implementation choices, they do not always guarantee optimal software quality.

*Keywords:* *software engineering, project management, software project, quality assessment criteria, software quality indicators, comprehensive quality indicator.*

*GJCST-C Classification:* *ACM Code: D.2.11*


*Strictly as per the compliance and regulations of:*

# Application of Meta-Programming Techniques for Accelerating Software Development and Improving Quality

Amirali Kerimovs

*Annotation-* A contemporary software tool has been devised to evaluate software quality through metric analysis techniques. This tool calculates pertinent metrics utilizing quality indicators and establishes a composite quality indicator value for software products. The intricacies of software quality assessment processes have been elucidated, including the examination of software quality's standardization as well as the presentation level of its model. This enables the potential for enhancement through the formulation of suitable criteria for quality assessment, refining models for metric analysis, and quantitatively measuring quality across all phases of project implementation. Notably, the use of metric analysis to gauge software quality reveals a lack of standardized metrics, resulting in varying assessment methods and metrics from different measurement system providers. Interpreting metric values also proves challenging for most software users due to a lack of clarity and informativeness. Furthermore, it has been discovered that while decisions based on cost, development duration, and designer company reputation influence software implementation choices, they do not always guarantee optimal software quality.

*Keywords: software engineering, project management, software project, quality assessment criteria, software quality indicators, comprehensive quality indicator.*

## I. Introduction

Considering the multi-faceted nature of software quality, a combination of these metrics is used for evaluation. Weighting factors, established by experts, are applied to individual metrics based on the dominant quality criteria. These combined indicators provide a comprehensive assessment of software quality. Extensive complexity metrics are particularly relevant during the design phase, while subsequent stages refine the value metrics.

In accordance with ISO [1] standards, quality pertains to the extent of alignment between relevant attributes and stipulated requirements. As defined by [10], quality signifies the entirety of features and traits within a product, process, or service, ensuring the capability to fulfil anticipated or declared needs. In accordance with [3], software quality refers to the extent of its possession of the requisite combination of attributes. Essentially, software quality reflects the degree to which software aligns with specified requirements.

The challenge is to ensure the desired software quality while recognizing that an unknown number of errors and defects persist within complex software systems, necessitating their containment or reduction to an acceptable level. Consequently, a pivotal objective within the modern software life cycle is the assurance of software product quality [4].

## II. Literature Review

Software quality is contingent upon the quality of methods and tools employed throughout its complete life cycle. Practical assessment of program quality is crucial not only upon completion but also during the design and development phases. The predicted or estimated quality of a software product comprises attributes evaluated or addressed at each life cycle stage, grounded in process quality and technological support [6].

The Software Development Life Cycle (SDLC) embodies a model depicting software creation and usage across various stages, commencing from the point of need identification and culminating in its retirement from user utilization. Numerous SDLC models exist, with three classified as foundational by international standards [4]: waterfall, incremental, and spiral.

During the design phase, establishing a set of quality requisites is vital: structure requirements for the software system (PS); air navigation specifications; user interface design prerequisites; multimedia component requisites for aircraft; usability demands; and technical prerequisites. The design stage formulates the response to the question, "How will the software system realize the imposed requirements?" Information flows during the software design stage [9] encompass software requirements portrayed through informational, functional, and behavioral analysis models. The information model outlines the data the software must process as per the customer's specifications. The functional model delineates a roster of information processing functions and software system modules. The behavioral model captures the desired system dynamics (operational modes). Concluding the design phase entails data

*Author: Independent Researcher, Riga, Latvia.*
*e-mail: kerimovsoftdev@gmail.com*

development, architecture formulation, and procedural software development.

Various approaches are employed for evaluating quality indicators, as outlined in standard [6]: measurement, recording, computation, expert assessment, and their combinations. Measurement involves specialized software tools to gather data on software characteristics such as volume, lines of code, operators, branches, entry/exit points, and more. Recording tracks factors like execution time, failures, and software start/end instances. Computation relies on statistical data collected during testing, operation, and maintenance to estimate indicators like reliability, accuracy, and stability. Expert assessment involves a panel of experienced evaluators who rely on intuition and experience rather than direct calculations or experiments. This method is used for reviewing programs, codes, documentation, and software requirements to assess factors like analyzability, documentation quality, and structured design [11].

In this context, the spiral life cycle model allows for the early assessment of software quality using a combination of calculation and expert evaluation techniques during the design phase.

1. *The Purpose of the Article:* Is to develop an adequate tool for determining the quality of software using the methods of metric analysis, which will make it possible to calculate the appropriate metrics with the help of quality indicators and determine the value of a complex indicator of the quality of a software product.
2. *Presentation of the Main Material:* The valuation of software can take the form of its monetary cost or be expressed through alternative means. Typically, clients hold their own notions regarding the maximum investment they're willing to make and the subsequent returns they expect, contingent on the software achieving its core objectives. The client's perspective might also encompass the software's functionality and specific expectations concerning its quality.

Typically, a client's initial focus revolves around the functional capacities of the software, often overlooking quality considerations, let alone the associated development costs. Consequently, during the initial phases of a software project, the focus may shift towards ensuring the client comprehends both the benefits of software utilization and the developmental expenses tied to attaining a particular level of software quality. Ideally, these crucial determinations should primarily occur when establishing user requirements for the software. Nonetheless, these considerations remain pertinent throughout the entirety of the software's development process. While standardized decision-making protocols might not exist, systems engineers must possess a clear understanding of the diverse avenues leading to specific levels of software quality and the corresponding developmental costs. This clarity aids in the anticipation of the overall expenditure associated with executing the software project.

To visually illustrate the correlation between the implementation costs of a software project and the level of software quality, we delve into the particulars of an information protection system's (ISI) development. Specifically, we analyze its functional model while bearing in mind its inherent intricacies. This model omits the depiction of information's inherent value- the object of confidentiality (e.g., bank deposit accounts or access codes), as such information retains its value over time. To facilitate understanding, the diagram introduces specific notations:

- *P:* Probability level indicating the extent of information protection (approximately $0.6 \leq P < 1.0$).
- *Z (P):* Permissible costs associated with safeguarding information as a function of the required level of protection. These costs rise as the demands for higher levels of information protection increase.

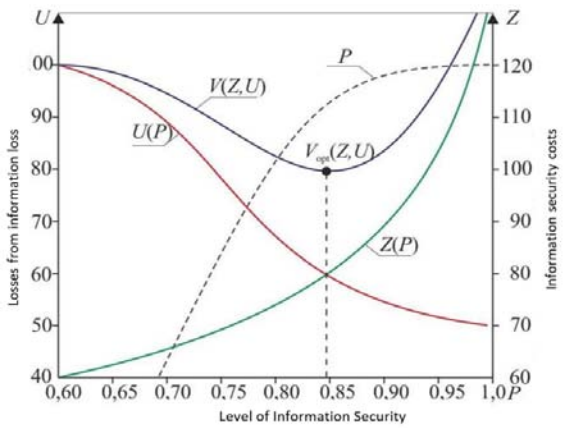



*Fig. 1:* The Main Features of the Process of Evaluating the Quality of SHI

The aspiration to achieve an exceedingly high level of information protection often ushers in a substantial escalation in expenses, potentially surpassing the intrinsic value of the information being safeguarded. The conceivable losses, or damages, borne by the information owner U(P), stemming from an insufficient level of protection, form a direct correlation with the extant level of protection, denoted as P. The diagram illustrates how the sum of Z(P) and U(P) collectively shapes the overall costs V(Z, U) associated with ensuring information security. Within this context, the optimal threshold for safeguarding, marked as $V_{opt(Z, U)}$, corresponds to the point where the combined costs of protection (Z(P)) and potential losses (U(P)) are minimized. This equilibrium signifies the balance between investing in protection measures and the potential losses due to inadequacies in protection, effectively preventing both excessive expenditures and heightened risks.

Striving to surpass this equilibrium point inevitably triggers a sharp escalation in Z(P), the expenses tied to information protection. Conversely, lowering the level of protection would lead to an escalation in potential losses, U(P), stemming from the compromised functionality of the system handling the safeguarding of information.

Consequently, the notion of software quality is intrinsically relative, gaining true comprehension within the context of real-world application scenarios. Therefore, the quality requirements established by relevant standards must be carefully aligned with the circumstances of the software's use and its specific domain of application.

Software quality embodies several critical components, notably:

1. *Quality of Software Development Processes:* This pertains to the efficacy, efficiency, and adherence to best practices during the creation of the software.
2. *Quality of Software Project Products:* Referring to the final software products themselves, encompassing attributes like functionality, reliability, and performance.
3. *Quality of Software Support or Implementation:* Addressing the competence and effectiveness of the software's implementation, utilization, and ongoing support.

This multi-faceted perspective illustrates how software quality is a nuanced and multifarious concept, emerging as a result of intricate interplays between development processes, product attributes, and the operational support environment.

The element concerning software development processes plays a pivotal role in gauging the extent of formalization and the inherent reliability of these processes across every stage of software evolution. This facet is intricately interwoven with the critical activities of verification and validation (abbreviated as V & V), which entail scrutinizing and endorsing the interim outcomes generated during these processes. The diligent pursuit of error detection and eradication within the finalized software is facilitated through rigorous testing methodologies. These approaches serve to diminish the occurrence of errors, thereby elevating the overall quality of the forthcoming software product.

Fostering excellence in the software project's products is underpinned by the meticulous application of procedures that govern the oversight of intermediate project deliverables at all developmental stages. These steps encompass meticulous checks to ascertain the attainment of the requisite quality standards. Furthermore, modern methodologies and resources dedicated to supporting the software product are harnessed to bolster this quality pursuit. The efficacy of software implementation hinges upon a symbiotic combination of factors, including the expertise of service personnel, the functional prowess of the software product, and the meticulous adherence to well-defined implementation protocols.

The framework for software quality is structured across four distinct levels of representation, as expounded by [7].

1. *First Level:* This pertains to the delineation of software quality's inherent attributes or indicators. Each of these indicators offers a unique vantage point from an end-user's perspective, encapsulating diverse facets of software quality. Established standards such as ISO/IEC 9126, DSTU 2844-1994, DSTU 2850-1994, and DSTU 3230-1995 elucidate a comprehensive quality model comprising six key characteristics or quality indicators for software: functionality, reliability, usability, maintainability, efficiency, and portability.
2. *Second Level:* Subsequent to the first tier, the focus shifts to expounding software quality attributes germane to each distinctive characteristic. This intricate articulation delves into the finer nuances and multifaceted features that contribute to each attribute. This assemblage of attributes subsequently underpins the metric analysis of software quality, enabling a comprehensive assessment across a spectrum of dimensions.

Therefore, a comprehensive understanding of the intricacies involved in assessing software quality has been elucidated. This endeavor encompasses a meticulous exploration of the very essence of software product quality, a subject subjected to the tenets of standardization. Concurrently, an in-depth investigation into the strata of the software quality model's representation has taken place. This discerning analysis has not only unveiled latent dimensions for refinement but also paved the way for the construction of judicious requisites tailored to the assessment of quality criteria.

Furthermore, it has facilitated the enhancement of the metric models used for the analysis of software quality and the calibration of quantitative measurement methods across every juncture of software project implementation.

The empirical landscape reveals a significant proportion of software errors manifesting during the critical phase of requirement formulation, accounting for 10-23% of the entire spectrum. A conspicuous trend emerges whereby the magnitude of software intricacy is positively correlated with the prevalence of conceptual errors within this stage (Hrytsiuk, 2018). It is noteworthy that as the complexity of software augments, the propensity for conceptual discrepancies becomes more pronounced. This phenomenon often arises due to the nherent challenges of grappling with extensive and multifaceted requirements.

Moreover, the formulation of software requirements engenders a vulnerability to information losses, primarily stemming from the interplay of incomplete articulation and variances in comprehending customer needs and the contextual milieu within the requirements specification. This predicament is particularly acute within software projects traversing the intersections of diverse domains of knowledge. It is unequivocally established that software endeavors marred by incomplete requirements and ill-prepared specifications invariably confront hurdles impeding successful realization.

Consequently, in light of such circumstances, the judicious recourse of subjecting the software requirements specification to rigorous analysis by impartial experts assumes paramount significance. This proactive measure serves as a pivotal bulwark against errors cascading through successive stages, encompassing requirement formulation, software architecture design, and subsequent construction phases [3].

Informed by the data presented in Table 1, a salient revelation surfaces wherein errors originating from requirement formulation and architectural design precipitate as a substantial portion, accounting for 25-55% of the overall error spectrum. It is compelling to note that this proportion is notably exacerbated as the magnitude of software complexity escalates, signifying a heightened susceptibility to errors during the nascent stages of development.

*Table 1:* Distribution of Errors Assumed at Different Stages of Software Development [2]

| Software development stage | Volume of Software/Share of Errors,% | | | | |
|---|---|---|---|---|---|
| | 2K | 8K | 32K | 128K | 512K |
| Formulation of requirements | 10 | 15 | 20 | 22 | 23 |
| Architecture design | 15 | 19 | 25 | 28 | 32 |
| Designing | 75 | 66 | 55 | 50 | 45 |

Consequently, we hold the conviction that an imperative avenue for further exploration lies in investigating the potential of harnessing metric analysis to ascertain software quality through insights gleaned from software requirements specifications. As a decisive stride towards this objective, we have conceived a bespoke software tool (depicted in Figure 2) meticulously architected to evaluate software quality via metric analysis. More specifically, it capitalizes on the utilization of quality metrics replete with both precise and prognostic values. A salient distinction of our tool, differentiating it from established counterparts, resides in its adeptness to dissect software based on ascertained metric values, prognosticating the trajectory of its developmental trajectory. Furthermore, the tool orchestrates a sequence of computations culminating in the generation of a comprehensive dataset, which in turn enables an extrapolation of metric outcomes. This inductive methodology endows the capacity for a quantitative assessment of the project's product quality and engenders the anticipation of developmental software quality attributes.

To orchestrate a systematic software development risk management paradigm, a project manager assumes the pivotal role of foretelling the precursors to potential predicaments, the emergence of adversities, or the occurrence of unfavorable events. This endeavor unfolds as an art of forecasting, grounded in empirically substantiated inferences regarding plausible trajectories of software project management execution, juxtaposing alternative courses and temporal dynamics. The interplay of forecasting management decisions intersects intimately with strategic and tactical contours delineating the risk landscape of program project implementation.

The development of the aforementioned software tool was steered within the contours of Microsoft Visual Studio. NET 2017 development environment. Significantly, this tool operates autonomously, devoid of any tether to internet connectivity. The commencement of the task hinged upon an intricate process of prototyping the user interface, progressively infusing augmentative functionalities into the software tool's architecture. The outcome of this endeavor, culminating in the software tool's user interface, is prominently featured in Figure 5.

A cornerstone of the software's architecture is encapsulated within the MetricsQualitySoftware.cs

class, an abstract entity that encapsulates pivotal functionalities essential for metric evaluation. This class is equipped with a suite of cardinal methods that underpin its operational dynamics. These include functions such as modifying metric parameter values (ChangeValue_OfParameter), accessing parameter names (GetNameOfParameter), furnishing fundamental metric information (SetInformation_OfMetric), illuminating metric definitions (ShowDescription_Of Metric), establishing metric parameter value functionality (SetAllParameters), ascertaining metric values (FindMetric), and facilitating metric parameter reference information display (ClearAllParameters_OfMetric).

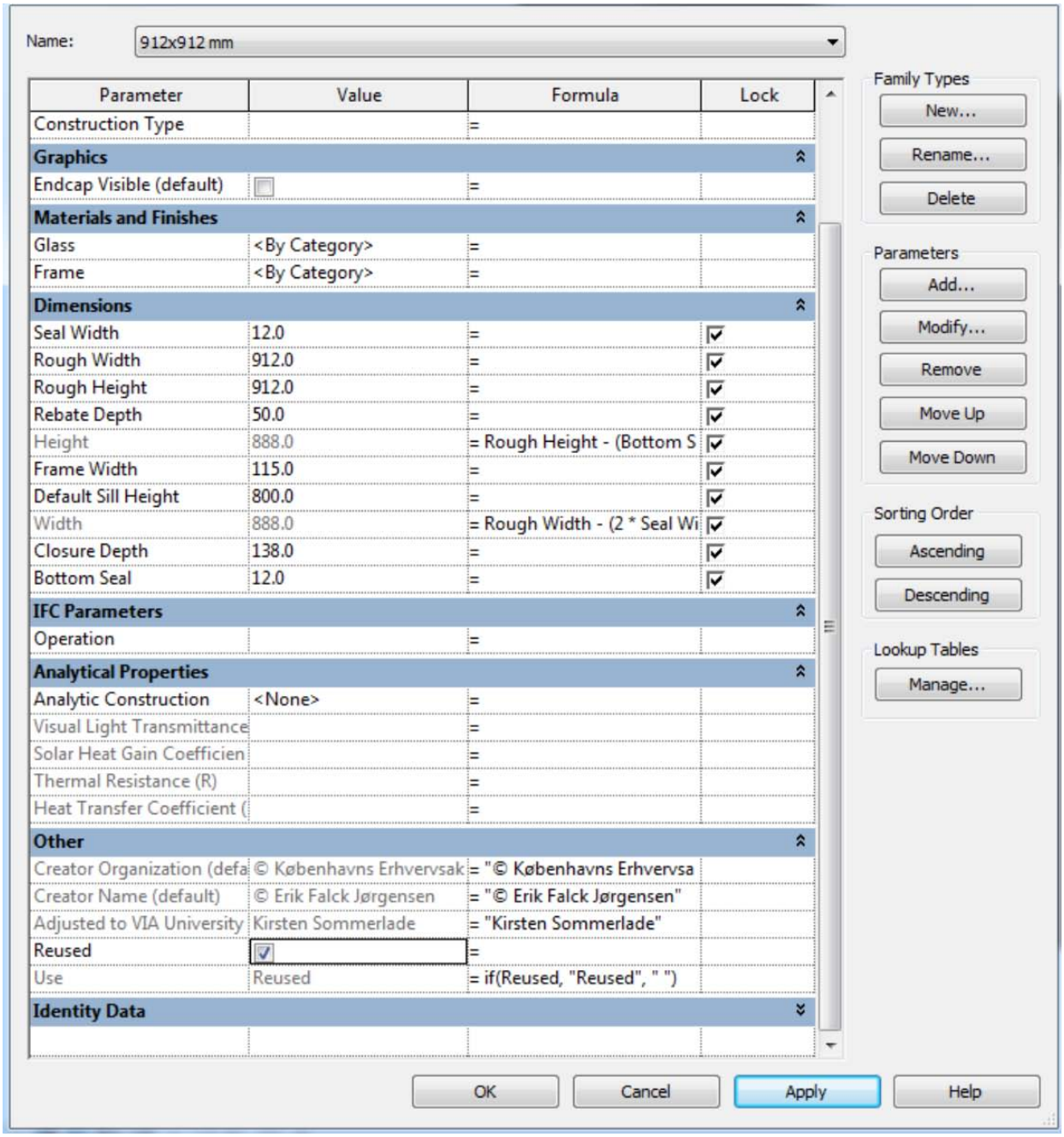



*Fig. 2:* Windows of the Software Tool for Determining Software Quality by Metric Analysis Methods

To facilitate the seamless manipulation, input, and retrieval of data within specific cells of the DataGrid table, the software employs the DataGridHelper.cs class. This crucial class encompasses key methods that empower efficient data handling: first, the ability to retrieve the value of a designated cell by specifying the row and column indices (GetCell), and second, the capability to retrieve data based solely on the row index (GetRow).

The architecture encompasses a series of distinct metrics classes, namely CHPmetric.cs, CPP metric.cs, MBQmetric.cs, MMTmetric.cs, RUPmetric.cs,

CCCmetric.cs, CPTmetric.cs, SCCmetric.cs, SCTmetric.cs, SDTmetric.cs, SQCmetric.cs, FPmetric.cs, LCmetric.cs, DPmetric.cs. Each of these classes is crafted to inherit from the abstract MetricsQualitySoftware.cs class, thereby inheriting its foundational structure, while also seamlessly overriding its methods to align with the specific requisites of their respective contexts.

The design also embraces auxiliary model classes such as MyTableInfo_OfAllMetrics.cs, MyTable Info_OfAllParameters.cs, and MyTableInfoCharacteristic _forMetricFp.cs. These model classes are meticulously sculpted to serve as repositories for recording the data harvested from distinct DataGrid tables. They also boast the capacity to efficiently dispense the synthesized tabular information.

Illustrating the software tool in action, let's delve into an illustrative scenario that underscores its operational prowess. In a bid to engender a comprehensive understanding of the tool's underlying mechanics, a meticulous examination is undertaken to ascertain both the quality and overarching forecasted assessment of the developmental process. This exploratory analysis culminates in the extraction of essential input data pertinent to the metrics, as delineated in Table 2. Following the meticulous input of all pertinent metrics' parameters and their subsequent calculation utilizing the software tool, a comprehensive dataset is curated, pivotal for constructing an informed forecast concerning the software's quality attributes.

The software tool instantiates the delivery of diverse representations of the culled information. Foremost, it furnishes an all-encompassing tabular display of metric values (Figure 4), thereby proffering a succinct overview of the analytical outcome. Furthermore, it leverages graphical illustrations to visually convey the insights, employing both pie charts and histograms (Figure 3) to distill the intricacies of the analysis. This holistic visualization augments the clarity and interpretability of the results. Conclusively, the software tool culminates in the holistic assessment of the software's quality, synthesizing the intricate array of metrics and their concomitant implications.

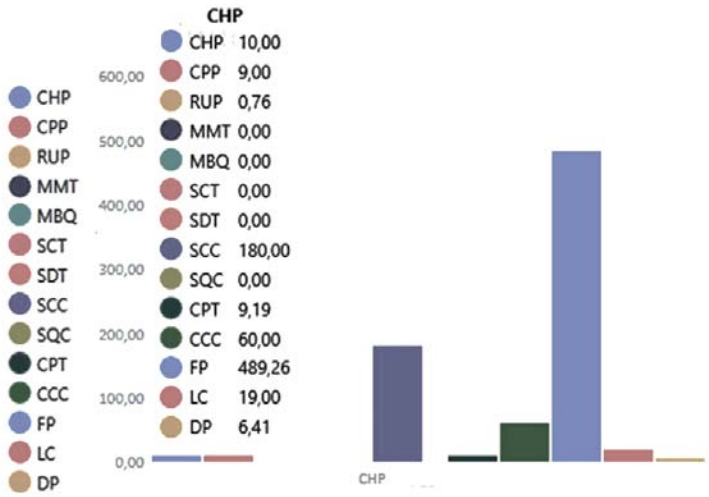
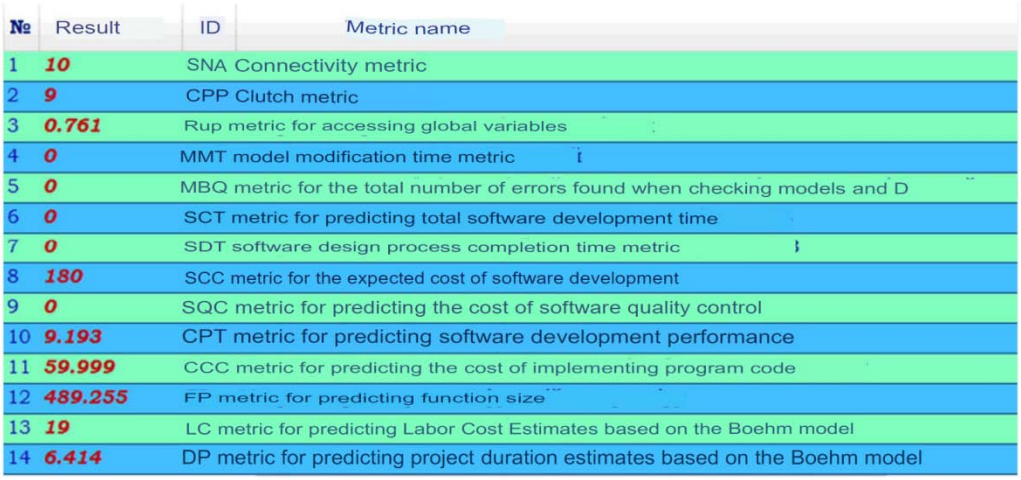



*Fig. 3:* Graphic Presentation of Results in the form of a Histogram

*Table 2:* Input Data for the Software Tool

| No. for/p | Parameter name | Value |
|---|---|---|
| 1 | How many times will the module actually access the global variable | 265 |
| 2 | How many times a module could access a global variable | 348 |
| 3 | The number of lines of program code | 4670 |
| 4 | The duration of the implementation of the software project | 126 |
| 5 | Part of the software architecture design stage in the process of its development | 2 |
| 6 | Number of module errors | 108 |
| 7 | Number of modules | 345 |
| 8 | Expected number of lines of function source code | 54, 34, 28, 58, 6 |
| 9 | Estimated cost to develop a feature line | 1 |
| 10 | Part of the stage of verification , validation and testing of software in the process of its development | 1 |
| 11 | Part of the product quality control stage of the project at the verification , validation and testing stages | 2 |
| 12 | The expected number of lines of source code in a similar function | 45, 30, 25, 50, 5 |

| | | |
|---|---|---|
| 13 | Productivity of the process of developing a similar function | 2 |
| 14 | Predicted performance of the software development process | 3 |
| 15 | The number of external inputs to the function that affect the executed function differently | 5, 11, 6, 5, 34 |
| 16 | The number of external outputs of the function for significantly different algorithms and non-trivial functionality | 8, 56, 7, 7, 12 |
| 17 | Number of external requests | 3, 3, 10, 2, 4 |
| 18 | Number of internal logical files or unique logical groups of user data | 1, 1, 53, 5, 7 |
| 19 | Number of external logical files or unique logical groups of user data | 4, 1, 1, 8, 2 |
| 20 | Connectivity level | functional |
| 21 | Clutch type | by content |
| 22 | Number of functions | 5 |



| № | Result | ID | Metric name |
|---|---|---|---|
| 1 | 10 | | SNA Connectivity metric |
| 2 | 9 | | CPP Clutch metric |
| 3 | 0.761 | | Rup metric for accessing global variables |
| 4 | 0 | | MMT model modification time metric |
| 5 | 0 | | MBQ metric for the total number of errors found when checking models and D |
| 6 | 0 | | SCT metric for predicting total software development time |
| 7 | 0 | | SDT software design process completion time metric |
| 8 | 180 | | SCC metric for the expected cost of software development |
| 9 | 0 | | SQC metric for predicting the cost of software quality control |
| 10 | 9.193 | | CPT metric for predicting software development performance |
| 11 | 59.999 | | CCC metric for predicting the cost of implementing program code |
| 12 | 489.255 | | FP metric for predicting function size |
| 13 | 19 | | LC metric for predicting Labor Cost Estimates based on the Boehm model |
| 14 | 6.414 | | DP metric for predicting project duration estimates based on the Boehm model |

*Fig. 4:* Obtained Results of Metrics

In the realm of software engineering, a sophisticated and advanced software tool has been meticulously crafted with the explicit purpose of ascertaining the quality of software through the adept utilization of metric analysis methodologies. This innovative tool transcends mere analysis, extending its reach into the realm of forecasting the prospective efficacy of the software development process. A notable feature of this software is its intrinsic capability to curate a comprehensive dataset that plays a pivotal role in the determination of a multifaceted indicator encapsulating the quality of the resultant software product. To concretize the tool's operational essence, an illuminating example elucidating its function is presented. Moreover, a comprehensive research endeavor has been undertaken to scrutinize and discern the quality of select software entities, culminating in a holistic prognostic-cation concerning the triumphant trajectory of their developmental journey.

This contemporary software marvel, meticulously fashioned to evaluate software quality, harnesses the power of metric analysis paradigms, enabling the seamless translation of quality indicators into precise metrics. Through this harmonious synergy, the intricate fabric of software quality is meticulously woven, ultimately manifesting in the articulation of a multifaceted metric indicative of software excellence. An in-depth examination of the research findings precipitates several salient conclusions, shedding luminous insight into the complex tapestry of software quality assessment.

The labyrinthine path of software quality assessment is unveiled, wherein the fundamental tenets of this process are dissected with precision. The concept of software product quality, assuming a central role in standardization, undergoes profound analysis. Simultaneously, the stratification of the software quality model is scrutinized, thereby establishing a robust framework conducive to iterative enhancements. This involves the meticulous refinement of quality assessment criteria, augmentation of metric analysis models, and the development of methods for quantitative measurement. Consequently, this holistic approach encompasses all facets of software project realization.

## III. Conclusion

For gauging software quality during the design phase, the spiral model of the software life cycle emerges as the most fitting approach. Examining the methods of assessing quality indicators (metrics) reveals that solely calculation and expert measurement techniques are viable at this stage. This is due to the inability to measure characteristics of software that hasn't been developed and the impracticality of recording execution moments for non-existent software.

1. The bedrock of successful software project implementation is unveiled through meticulous exploration. The crux of this revelation lies in the ardent aspiration of project managers to engender software solutions that bear inherent value. This value is both a catalytic agent in solving intricate challenges and a cornerstone in accomplishing tactical and strategic objectives. A nuanced understanding of this value leads to the discernment that it can be encapsulated either in monetary terms or via alternative metrics. This profound insight is fortified by the recognition that customers harbour their distinct perception of maximum investment thresholds, intertwined with the anticipated returns rooted in the attainment of overarching objectives through software deployment. Moreover, this discernment extends to the articulation of software functionality and the quality paradigm, encapsulating the customer's discerning expectations.
2. The unique contours of metric analysis as a conduit for assessing software quality come to the fore. A pivotal observation is the absence of homogenous standards for metrics, resulting in diverse methodologies proposed by individual system providers to gauge software quality. The enigmatic interpretation of metric values surfaces as an additional challenge, as these values often elude the comprehensive grasp of the majority of users. The interplay of these facets underscores the complexity inherent in selecting a software implementation route. As a corollary, pivotal determinants in this selection process include financial viability, temporal dynamics, and the reputation of the design company. Notably, however, these determinants do not inexorably guarantee the desired software quality outcome.
3. A groundbreaking feat materializes in the form of a bespoke software tool architected to gauge software quality by harnessing the potential of metric analysis methodologies. This innovative tool ingeniously extends its functionality beyond analysis, adroitly projecting the future efficacy of the developmental process. The hallmark of this innovation is its adeptness in formulating a dataset of paramount importance, intricately intertwined with the determination of a comprehensive quality indicator encompassing the software product's inherent excellence.
4. A culmination of insightful observations culminates in the crystallization of pertinent recommendations, offering guidance in the employment of the developed information visualization technique. This technique augments the interpretability and efficacy of software quality assessment, paving the way for enhanced decision-making and informed trajectories in software development endeavors.

## References Références Referencias


1. Bozic, Velibor. (2023). Methods and Techniques of Software Development. 10.13140/RG.2.2.27516.00645.
2. Bozkurt, Erkam. (2022). The usage of cybernetic in complex software systems and its application to the deterministic multithreading. Concurrency and Computation: Practice and Experience. 34.10.1002/cpe.7375.
3. Cheng, Kwok Sun & Huang, Pei-Chi & Ahn, Tae-Hyuk & Song, Myoungkyu. (2023). Tool Support for Improving Software Quality in Machine Learning Programs. Information. 14. 53. 10.3390/info14010053.
4. Hong, Sirui & Zheng, Xiawu & Chen, Jonathan & Cheng, Yuheng & Zhang, Ceyao & Wang, Zili & Yau, Steven & Lin, Zijuan & Zhou, Liyang & Ran, Chenyu & Xiao, Lingfeng & Wu, Chenglin. (2023). MetaGPT: Meta Programming for Multi-Agent Collaborative Framework.
5. Kovari, Attila & Katona, Jozsef. (2023). Effect of software development course on programming self-efficacy. Education and Information Technologies. 1-27. 10.1007/s10639-023-11617-8.
6. Luo, Ke & Deng, Wei. (2023). Software engineering database programming control system based on embedded system. Applied Mathematics and Nonlinear Sciences. 10.2478/amns.2023.1.00473.
7. Nagalakshmi, S. (2023). Software Development Techniques In Current Scenario. Data Analytics and Artificial Intelligence. 3. 50-53. 10.46632/cllrm/3/2/10.
8. Romli, Rohaida & Nordin, Noorazreen & Omar, Mazni & Mahmod, Musyrifah. (2018). A Review on Meta-Heuristic Search Techniques for Automated Test Data Generation: Applicability Towards Improving Automatic Programming Assessment. 896-906. 10.1007/978-3-319-59427-9_92.
9. Shafiq, Muhammad & Alghamedy, Fatemah & Jamal, Nasir & Kamal, Tahir & Daradkeh PhD., P. Eng, Dr. Yousef & Shabaz, Dr. Mohammad. (2023). Scientific programming using optimized machine learning techniques for software fault prediction to improve software quality. IET Software. 17. n/a-n/a. 10.1049/sfw2.12091.
10. Stuikys, Vytautas & Damasevicius, Robertas. (2013). A Background of Meta-Programming Techniques. 10.1007/978-1-4471-4126-6_3.
11. Tietz, Vanessa. (2021). Development of a Meta-language and its Qualifiable Implementation for the Use in Safety-critical Software.

# Global Journals Guidelines Handbook 2023

www.GlobalJournals.org

# Memberships

## Fellows/Associates of Computer Science Research Council

### FCSRC/ACSRC Memberships

## Introduction

FCSRC/ACSRC is the most prestigious membership of Global Journals accredited by Open Association of Research Society, U.S.A (OARS). The credentials of Fellow and Associate designations signify that the researcher has gained the knowledge of the fundamental and high-level concepts, and is a subject matter expert, proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice. The credentials are designated only to the researchers, scientists, and professionals that have been selected by a rigorous process by our Editorial Board and Management Board.

Associates of FCSRC/ACSRC are scientists and researchers from around the world are working on projects/researches that have huge potentials. Members support Global Journals' mission to advance technology for humanity and the profession.

# FCSRC

### Fellow of Computer Science Research Council

FELLOW OF COMPUTER SCIENCE RESEARCH COUNCIL is the most prestigious membership of Global Journals. It is an award and membership granted to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Fellows are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Fellow Members.

# Benefit

## To the institution

### Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

## Exclusive Network

### Get access to a closed network

A FCSRC member gets access to a closed network of Tier 1 researchers and scientists with direct communication channel through our website. Fellows can reach out to other members or researchers directly.They should also be open to reaching out by other.

Career | Credibility | Exclusive | Reputation


## Certificate

### Certificate, LoR and Laser-Momento

Fellows receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career | Credibility | Exclusive | Reputation


## Designation

### Get honored title of membership

Fellows can use the honored title of membership. The “FCSRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FCSRC or William Walldroff, M.S., FCSRC.

Career | Credibility | Exclusive | Reputation

## Recognition on the Platform

### Better visibility and citation

All the Fellow members of FCSRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation. All fellows get a dedicated page on the website with their biography.

Career | Credibility | Reputation

## Future Work

### Get discounts on the future publications

Fellows receive discounts on future publications with Global Journals up to 60%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career | Financial

## GJ Account

### Unlimited forward of Emails

Fellows get secure and fast GJ work emails with unlimited forward of emails that they may use them as their primary email. For example, john [AT] globaljournals [DOT] org.

Career | Credibility | Reputation

## Premium Tools

### Access to all the premium tools

To take future researches to the zenith, fellows receive access to all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

## Conferences & Events

### Organize seminar/conference

Fellows are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same 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. Additionally, they get free research conferences (and others) alerts.

Career | Credibility | Financial

## Early Invitations

### Early invitations to all the symposiums, seminars, conferences

All fellows receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive


## Publishing Articles & Books

### Earn 60% of sales proceeds

Fellows can publish articles (limited) without any fees. Also, they can earn up to 70% of sales proceeds from the sale of reference/review books/literature/publishing of research paper. The FCSRC member can decide its price and we can help in making the right decision.

Exclusive | Financial

## Reviewers

### Get a remuneration of 15% of author fees

Fellow members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

## Access to Editorial Board

### Become a member of the Editorial Board

Fellows may join as a member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. Additionally, Fellows get a chance to nominate other members for Editorial Board.

Career | Credibility | Exclusive | Reputation

## And Much More

### Get access to scientific museums and observatories across the globe

All members get access to 5 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 10 GB free secure cloud access for storing research files.

## ASSOCIATE OF COMPUTER SCIENCE RESEARCH COUNCIL

ASSOCIATE OF COMPUTER SCIENCE RESEARCH COUNCIL is the membership of Global Journals awarded to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Associate membership can later be promoted to Fellow Membership. Associates are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Associate Members.



# Benefit

## To the institution

### Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

## Exclusive Network

### Get access to a closed network

A ACSRC member gets access to a closed network of Tier 2 researchers and scientists with direct communication channel through our website. Associates can reach out to other members or researchers directly.They should also be open to reaching out by other.

Career | Credibility | Exclusive | Reputation


## Certificate

### Certificate, LoR and Laser-Momento

Associates receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career | Credibility | Exclusive | Reputation


## Designation

### Get honored title of membership

Associates can use the honored title of membership. The “ACSRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., ACSRC or William Walldroff, M.S., ACSRC.

Career | Credibility | Exclusive | Reputation

## Recognition on the Platform

### Better visibility and citation

All the Associate members of ACSRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation.

Career | Credibility | Reputation

## Future Work

### Get discounts on the future publications

Associates receive discounts on future publications with Global Journals up to 30%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career | Financial

## GJ Account

### Unlimited forward of Emails

Associates get secure and fast GJ work emails with 5GB forward of emails that they may use them as their primary email. For example, john [AT] globaljournals [DOT] org.

Career | Credibility | Reputation

## Premium Tools

### Access to all the premium tools

To take future researches to the zenith, associates receive access to all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

## Conferences & Events

### Organize seminar/conference

Associates are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same 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. Additionally, they get free research conferences (and others) alerts.

Career | Credibility | Financial

## Early Invitations

### Early invitations to all the symposiums, seminars, conferences

All associates receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive


## Publishing Articles & Books

### Earn 30-40% of sales proceeds

Associates can publish articles (limited) without any fees. Also, they can earn up to 30-40% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.

Exclusive | Financial

## Reviewers

### Get a remuneration of 15% of author fees

Associate members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

## And Much More

### Get access to scientific museums and observatories across the globe

All members get access to 2 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 5 GB free secure cloud access for storing research files.

| Associate | Fellow | Research Group | Basic |
|---|---|---|---|
| $4800<br>**lifetime designation** | $6800<br>**lifetime designation** | $12500.00<br>**organizational** | APC<br>**per article** |
| **Certificate,** LoR and Momento<br>**2** discounted publishing/year<br>**Gradation** of Research<br>**10** research contacts/day<br>**1 GB** Cloud Storage<br>**GJ** Community Access | **Certificate,** LoR and Momento<br>**Unlimited** discounted publishing/year<br>**Gradation** of Research<br>**Unlimited** research contacts/day<br>**5 GB** Cloud Storage<br>**Online Presense** Assistance<br>**GJ** Community Access | **Certificates,** LoRs and Momentos<br>**Unlimited** free publishing/year<br>**Gradation** of Research<br>**Unlimited** research contacts/day<br>**Unlimited** Cloud Storage<br>**Online Presense** Assistance<br>**GJ** Community Access | **GJ** Community Access |

# 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<br><br>Above 200 words | No specific data with ambiguous information<br><br>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