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

Global Journal of Computer Science and Technology : C

Software and Data Engineering

DISCOVERING THOUGHTS AND INVENTING FUTURE

HIGHLIGHTS

- Failure Using Software Metrics
- Approach to Quality Testing
- Intelligent Information Retrieval
- Large Numeric Data Sets

Datacentre

Volume 12 | Issue 12 | Version 1.0 ENG

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 12 Issue 12 (Ver. 1.0)

Open Association of Research Society

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

Global Journals Inc.

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

Sponsors:Global Association of Research

Open Scientific Standards

Publisher's Headquarters office

Global Journals Inc., Headquarters Corporate Office, Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141 United States

USA Toll Free: +001-888-839-7392

USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

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

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

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

eContacts

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

Pricing (Including by Air Parcel Charges):

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

Editorial Board Members (HON.)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contents of the Volume

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Quantitative Analysis of Fault and Failure Using Software Metrics

By Shital V. Tate & S. Z. Gawali
Bharati Vidyapeeth Deemed University, College of Engineering, Pune

Abstract - It is very complex to write programs that behave accurately in the program verification tools. Automatic mining techniques suffer from 90–99% false positive rates, because manual specification writing is not easy. Because they can help with program testing, optimization, refactoring, documentation, and most importantly, debugging and repair. To concentrate on this problem, we propose to augment a temporal-property miner by incorporating code quality metrics. We measure code quality by extracting additional information from the software engineering process, and using information from code that is more probable to be correct as well as code that is less probable to be correct. When used as a preprocessing step for an existing specification miner, our technique identifies which input is most suggestive of correct program behaviour, which allows off-the-shelf techniques to learn the same number of specifications using only 45% of their original input.

GJCST-C Classification: *D.2.8*

Strictly as per the compliance and regulations of:

Quantitative Analysis of Fault and Failure Using Software Metrics

Shital V. Tate [α] & S. Z. Gawali [α]

Abstract - It is very complex to write programs that behave accurately in the program verification tools. Automatic mining techniques suffer from 90–99% false positive rates, because manual specification writing is not easy. Because they can help with program testing, optimization, refactoring, documentation, and most importantly, debugging and repair. To concentrate on this problem, we propose to augment a temporal-property miner by incorporating code quality metrics. We measure code quality by extracting additional information from the software engineering process, and using information from code that is more probable to be correct as well as code that is less probable to be correct. When used as a pre-processing step for an existing specification miner, our technique identifies which input is most suggestive of correct program behaviour, which allows off-the-shelf techniques to learn the same number of specifications using only 45% of their original input.

I. Introduction

Software remains buggy and testing is still the leading approach for detecting software errors. Incorrect and buggy behaviour in deployed software costs up to $70 billion each year in the US[1]. Thus debugging, testing, maintaining, optimizing, refactoring, and documenting software, while time-consuming, remain significantly important. Such maintenance is reported to consume up to 90% of the total cost of software projects[2]. Maximum maintenance time is spent studying existing software since maintenance concern is incomplete documentation.

Consistently, however, verification tools require specifications that describe some aspect of program accuracy. Creating accurate specifications is difficult, time-consuming and error-prone. Verification tools can only point out disagreements between the program and the specification. Even assuming a sound and complete tool, an defective specification can still yield false positives by pointing out non-bugs as bugs or false negatives by failing to point out real bugs. Crafting specifications typically requires program-specific knowledge.

Specification mining can be compared to learning the rules of English grammar by reading essays written by high school students; we propose to focus on the essays of passing students and be doubtful of the essays of failing students. We claim that existing miners have high false positive rates in large part because they treat all code equally, even though not all code is created equal. For example, consider an execution trace through a recently modified, rarely-executed piece of code that was copied-and-pasted by an inexperienced developer. We argue that such a trace is a poor guide to correct behaviour when compared with a well-tested, infrequently-changed, and commonly-executed trace.

Various pre-existing software projects are not yet formally specified[3]. Formal program specifications are difficult for humans to construct[4], and incorrect specifications are difficult for humans to debug and modify[5]. Accordingly, researchers have developed techniques to automatically infer specifications from program source code or execution traces[6],[7],[8],[9]. These techniques typically produce specifications in the form of finite state machines that describe legal sequences of program behaviours.

Unfortunately, these existing mining techniques are insufficiently precise in practice. Some miners produce large but approximate specifications that must be corrected manually [5]. As these large specifications are indefinite and difficult to debug, this article focuses on a second class of techniques that produce a larger set of smaller and more precise candidate specifications that may be easier to evaluate for correctness. These specifications typically take the form of two-state finite state machines that describe temporal properties, e.g. “if event a happens during program execution, event b must eventually happen during that execution.” Two-state specifications are limited in their expressive power; comprehensive API specifications cannot always be expressed as a collection of smaller machines[8].

Recognize and illustrate lightweight, automatically collected software features that fairly accurate source code quality for the purpose of mining specifications. In this approach explain how to lift code quality metrics to metrics on traces, and empirically measure the utility of our lifted quality metrics when applied to previous static specification mining techniques. To avoid false positives recommend two novel specification mining techniques that use our automated quality metrics to learn temporal safety specifications.

II. On Going Methodology

a) *Specification Mining With Few False Positive*

This methodology presents a new automatic specification miner that uses artifacts from software

Author α : Department of Information Technology, Bharati Vidyapeeth Deemed University, College of Engineering, Pune-46.

engineering processes to capture the reliability of its input traces.

The main contributions of this project are:

– A set of source-level features related to software engineering processes that capture the trustworthiness of code for specification mining. We analyze the relative analytical power of each of these features.
– Experimental evidence that our notions of trustworthy code serve as a basis for evaluating the trustworthiness of traces. We provide a characterization for such traces and show that off-the-shelf specification miners can learn just as many specifications using only 60% of traces.
– A novel automatic mining technique that uses our trust-capturing features to learn temporal safety specifications with few false positives in practice. We evaluate it on over 800,000 lines of code and explicitly compare it to two previous approaches. Our basic mining technique learns specifications that locate more safety-policy violations than previous miners (740 vs. 426) while presenting far fewer false positive specifications (107 vs. 567). When focused on precision, our technique obtains a low 5% false positive rate, an order-of-magnitude improvement on previous work, while still finding specifications that locate 265 violations. To our knowledge, this is the first specification miner that produces multiple candidate specifications and has a false positive rate under 90%.

i. *Approach*

In this approach present a specification miner that works in three stages:

1. Statically estimate the trustworthiness of each code fragment.
2. Lift that judgment to traces by considering the code visited along a trace.
3. Weight the contribution of each trace by its trustworthiness when counting event frequencies for specification mining.

The code is most trustworthy when it has been written by experienced Programmers who are familiar with the project at hand, when it has been well-tested, and when it has been mindfully written.

b) Mining Temporal Specification for Error Detection

If we use implicit language-based specifications (e.g., null pointers should not be dereferenced) or to reuse standard library specifications then it can reduce the cost of writing specifications. More recently, however, a variety of attempts have been made to conclude program-specific temporal specifications and API usage rules automatically. These specification mining techniques take programs (and possibly dynamic traces, or other hints) as input and produce candidate specifications as output. Basically specifications could also be used for documenting, refactoring, testing, debugging, maintaining, and optimizing a program. Centre of attention is that finding and evaluating specifications in a particular context: given a program and a generic verification tool, what specification mining technique should be used to find bugs in the program and thereby improve software quality? Thus we are concerned both with the number of "real" and "false positive" specifications produced by the miner and with the number of "real" and "false positive" bugs found using those "real" specifications.

In this methodology propose a novel technique for temporal specification mining that uses information about program error handling. Our miner assumes that programs will generally adhere to specifications along normal execution paths, but that programs will likely violate specifications in the presence of some run-time errors or exceptional situations. Intuitively, error-handling code may not be tested as often or the programmer may be unaware of sources of run-time errors. Taking advantage of this information is more important than ranking candidate policies.

i. *Contributions*

– Propose a novel specification mining technique based on the observation that programmers often make mistakes in exceptional circumstances or along uncommon code paths.
– Present a qualitative comparison of five miners and show how some miner assumptions are not well-supported in practice.
– Finally, we give a quantitative comparison of our technique's bug-finding powers to generic "library" policies. For our domain of interest, mining finds 250 more bugs. We also show the relative unimportance of ranking candidate policies. In all, we find 69 specifications that lead to the discovery over 430 bugs in 1 million lines of code.

III. Proposed System For Quantitative Analysis of Fault and Failure

In proposed system, aim to develop a system which can be used to measure the quality of the code considering different aspects affecting the quality of the code. The term quality of the code can be explained using different factors such as code clone, author rank, code churn, code readability, path feasibility etc.

To Present a new specification miner that works in three stages. First, it statically estimates the quality of source code fragments. Second, it lifts those quality judgments to traces by considering all code visited along a trace. Finally, it weights each trace by its quality when counting event frequencies for specification mining.

This system develops an automatic specification miner that balances true positives – as required behaviours –with false positives – non-required behaviours. We claim that one important reason that previous miners have high false positive rates is that they falsely assume that all code is equally likely to be correct. For example, consider an execution trace through a recently modified, rarely-executed piece of code that was copied and-pasted by an inexperienced developer. We believe that such a trace is a poor guide to correct behaviour, especially when compared with a well-tested, stable, and commonly-executed piece of code. Patterns of specification adherence may also be useful to a miner: a candidate that is violated in the high quality code but adhered to in the low quality code is less likely to represent required behaviour than one that is adhered to on the high quality code but violated in the low quality code. We assert that a combination of lightweight, automatically collected quality metrics over source code can usefully provide both positive and negative feedback to a miner attempting to distinguish between true and false specification candidates.

Code quality information may be gathered either from the source code itself or from related artefacts, such as version control history. By augmenting the trace language to include information from the software engineering process, we can evaluate the quality of every piece of information supporting a candidate specification (traces that adhere to a candidate as well as those that violate it and both high and low quality code) on which it is followed and more accurately evaluate the likelihood that it is valid.

The system architecture of the system is as in following figure, which explains the modules to be generated.

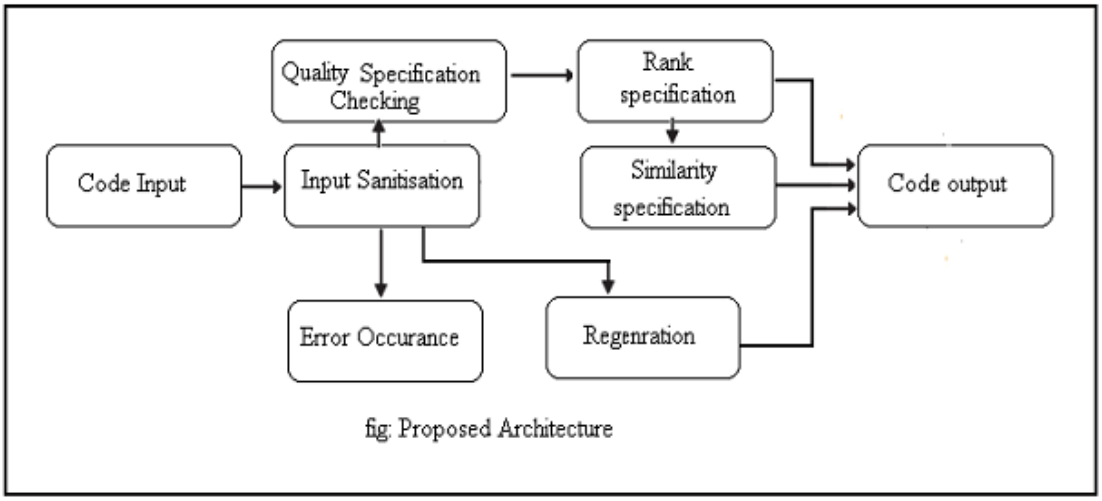


Figure 1 :

a) Description of proposed system

Proposed system for quantitative analysis of and fault and failure using software metrics uses the following stages-

1. Accept input in the form of computer program code.
2. Perform input sanitization.
3. Check for error occurrence in the code.
4. Check for the quality specification regarding the given code.
5. Specify the rank for the different condition, using calculated result.
6. Generate output in the form of quality report.

IV. Conclusion

Testing, maintenance, optimization, refactoring, documentation, and program repair these are the various applications of formal specification. Though human programmers should not produce and verify such specification manually. These technique is also problematic since it treat all parts of program as equally indicative as correct behaviour. We encode this intuition using dependability metrics such as analytical execution frequency, copy paste code measurements, code duplication software readability or path feasibility. We compare the bug finding power of various miners. This technique improves the performance of existing trace based miners by focusing on high quality traces. Our technique is also useful to improve the quality of code through specification mining.

References Références Referencias

1. National Institute of Standards and Technology, "The economic impact of inadequate infrastructure for software testing," Tech. Rep. 02-3, may 2002.
2. R. C. Seacord, D. Plakosh, and G. A. Lewis, Modernizing Legacy Practices, 2003.
3. M. Das, "Formal specifications on industrial-strength code from myth to reality," in Computer-Aided Verification, 2006, p. 1.

4. H. Chen, D. Wagner, and D. Dean, "Setuid demystified," in USENIX Security Symposium, 2002, pp. 171–190.
5. G. Ammons, D. Mandelin, R. Bod´ık, and J. R. Larus, "Debugging temporal specifications with concept analysis," in Programming Language Design and Implementation, 2003, pp. 182–195.
6. G. Ammons, R. Bodik, and J. R. Larus, "Mining specifications," in Principles of Programming Languages, 2002, pp. 4–16.
7. D. R. Engler, D. Y. Chen, and A. Chou, "Bugs as inconsistent behaviour: A general approach to inferring errors in systems code," in Symposium on Operating System Principles, 2001, pp. 57–72.
8. M. Gabel and Z. Su, "Symbolic mining of temporal specifications," in ICSE, 2008, pp. 51–60.
9. J. Whaley, M. C. Martin, and M.S. Lam, "Automatic extraction of object-oriented component interfaces," in ISSTA, 2002.
10. Claire Le Goues, Westely Weimer "Measuring code quality to improve specification mining" IEEE Trans. Software Eng.
11. Mohammed Kayed and Chia-Hui Chang," FiVaTech: Page-Level Web Data Extraction from Template Pages" IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 2, February 2010
12. S. R. Chidamber and C. F. Kemerer, "A metrics suite for object oriented design," IEEE Trans. Softw. Eng., vol. 20, no. 6, pp. 476–493, 1994.
13. D. Detlefs, G. Nelson, and J. B. Saxe, "Simplify: a theorem prover for program checking," J. ACM, vol. 52, no. 3, pp. 365–473, 2005.
14. M. Di Penta and D. M. German, "Who are source code contributors and how do they change?" in Working Conference on Reverse Engineering. IEEE Computer Society, 2009, pp. 11–20.
15. C. Kapser and M. W. Godfrey, ""Cloning Considered Harmful" in WCRE, 2006, pp. 19–28.
16. J. Krinke, "A study of consistent and inconsistent changes to code clones," in WCRE. IEEE Computer Society, 2007, pp. 170–178.
17. C. Le Goues and W. Weimer, "Specification mining with few false positives." In TACAS, 2009, pp. 292–306.
18. T. J. McCabe, "A complexity measure," IEEE Trans. Software Eng., vol. 2, no. 4, pp. 308–320, 1976.
19. N. Nagappan and T. Ball, "Using software dependencies and churn metrics to predict field failures: An empirical case study,"in ESEM, 2007, pp. 364–373.
20. J. C. Sanchez, L. Williams, and E. M. Maximilien, "On the Sustained Use of a Test Driven Development Practice at IBM," in Agile 2007. IEEE Computer Society, August 2007, pp. 5–14.
21. W.Weimer and N. Mishra, "Privately finding specifications," IEEE Trans. Software Eng., vol. 34, no. 1, pp. 21–32, 2008.
22. W. Weimer and G. C. Necula, "Mining temporal specifications for error detection," in TACAS, 2005, pp. 461– 476.
23. D. R. Engler, D. Y. Chen, and A. Chou. "Bugs as inconsistent behavior: A general approach to inferring errors in systems code". In Symposium on Operating Systems Principles, pages 57–72, 2001.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data Mining in Clinical Practices Guidelines

By Mayura Kinikar, Harish Chawria, Pradeep Chauhan & Abhijeet Nashte
Maharashtra Academy of Engineering Alandi, Pune

Abstract - This paper proposes text mining of clinical practices to extract decision-making steps. These steps should be formed in- logical functions capable of branching on different plan set on some deciding variables. The probable action sequence will be notified on the data of patient given to the conditions of clinical guideline and this will also give critical conditions that need immediate attention. In this project medical grammar rules are applied to extract key decision making steps from the clinical guidelines. In the first step lexical analysis is performed to key- words like 'if this then perform this, all the medical terms will be identified and this extracted rule set will be used to create a XSLT file. The patient data in form of an XML file will be then applied to the XSLT transformations or rule sets to derive final result of action plan specific to that patient.

Keywords : *Clinical Data Repository(CDR), Virtual Medical Record(VMR), Abstract Syntax Notation(ASN), Electronic Medical Records(EMR), Medical Logic Modules(MLM), HealthCare Data Dictionary(HDD).*

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Data Mining in Clinical Practices Guidelines

Mayura Kinikar [α], Harish Chawria [α], Pradeep Chauhan [α] & Abhijeet Nashte [α]

Abstract - This paper proposes text mining of clinical practices to extract decision-making steps. These steps should be formed in- logical functions capable of branching on different plan set on some deciding variables. The probable action sequence will be notified on the data of patient given to the conditions of clinical guideline and this will also give critical conditions that need immediate attention. In this project medical grammar rules are applied to extract key decision making steps from the clinical guidelines. In the first step lexical analysis is performed to key- words like 'if this then perform this, all the medical terms will be identified and this extracted rule set will be used to create a XSLT file. The patient data in form of an XML file will be then applied to the XSLT transformations or rule sets to derive final result of action plan specific to that patient.

Keywords : Clinical Data Repository(CDR), Virtual Medical Record(VMR), Abstract Syntax Notation(ASN), Electronic Medical Records(EMR), Medical Logic Modules(MLM), HealthCare Data Dictionary(HDD).

I. Introduction

Data Mining is the science of finding patterns in huge reserves of data, in order to generate useful information from it. Data Mining has potential applications in several fields, one of which is Health Care. The myriad possibilities of improvement in Health Care through Data Mining only further justify the need to apply data mining principles to clinical data. However, prior to applying data mining techniques to garner information from data, the data has to be 'prepared' to ensure the veracity of the information obtained. 'Preparing' the data involves removal of incorrect information or 'noise' from the data and ensuring that the data mining principles are applied on real data. This paper gives a detailed description of the purpose, design and implementation of the Data Mining Framework. The primary purpose of the Data Mining Framework is to help determine trends in patient records to improve Health Care.

Health Care Preparing the data, prior to applying data mining techniques, is a critical step of data mining. It primarily consists of four steps. The first step is the selection of data. This is following by data cleaning, forming the new data and finally formatting the data.

Data has to be selected, prior to applying data mining principles. The data is chosen based on its completeness and correctness. Constraints on the data, such as the data type, could also be factors in the selection of the data.

Author α : Department of Computer Engineering Maharashtra Academy of Engineering Alandi, Pune.

Most of the medical data for the use of doctors are achieved/ available in text form in medical reference books and on Internet in web pages. To derive decision-making information from bulk of data, careful observation of the guideline is required, which tampers quick decision.

Data Mining can automate the process of finding trends and patterns in databases. Large Databases can be analyzed quickly and effectively, using high performance systems and time effective algorithms. As the speed of processing improves, the size of the database can be increased. Also, as the size of the database increases, the accuracy of the predictions also improves. Co-relation among data can be identified.

The data-mining tool, based on the type of task that has to be performed, creates the data model. The modeling phase is an important aspect of the data-mining process. There are a vast number of data-mining techniques that can be applied to the data model. The data modeling technique, which is applied depends on the type of the model. The trends and patterns that are found in the data are based on the data mining technique applied. These patterns can be fed to decision support systems, which can make informed predictions on the data. Some of the more popular data modeling techniques are discussed in the following section

1. Decision tree is an attribute classifier, which takes the form of a tree. It can consist of leaf nodes or internal nodes. The former signifies the value of an attribute while the latter are decision nodes. These nodes, as the name signifies, specify a test that can be performed on the attribute, so that the attributes can be classified as sub-trees. Decision trees algorithms basically use the divide-and-conquer approach to classify. They thus work in a top-down manner, using an attribute at each level split on, that best separates the classes. Decision trees signify rules and are easily interpretable by humans.
2. In Rule Induction methods, rules are created from the data based on statistical values. In this approach, all values in a class are considered. At each stage a rule is generated for these values. Testing the rule under construction with new values creates a rule with maximum accuracy. The value is chosen such that it maximizes the probability of the desired classification.

II. Detailed Clinical Models

While coded vocabularies provide much of the raw material needed to describe clinical information, they are not sufficient alone. If we want to state that a patient has a diagnoisis of breast cancer, we could store the concept, <254837009 | SNOMED-CT | breast cancer>, in her electronic medical record. If we wanted to state that the patient had a family history of breast cancer, we could store the concept, <275862002 | SNOMED-CT | family history of breast cancer>, in her record. Suppose now that we wanted to store the fact that it was the patient's sister that had breast cancer. Currently SNOMED-CT does not have a concept for this. Although a coded vocabulary like SNOMED-CT could add concepts like this, it is not a practical solution. It would require the maintainers of the vocabulary to create concepts for most combinations of disease and family members.

III. Clinical Data Repository

The CDR is a robust electronic medical record system which makes extensive use of coded vocabularies and detailed clinical models.

The detailed clinical models used by the CDR are defined using Abstract Syntax Notation One (ASN.1). ASN.1 is an ISO standard for describing electronic messages [ASN]. As its name implies, ASN.1 provides a syntax for describing messages that is abstract from any specific encoding. However, in addition to the abstract specifica-tion, ASN.1 also defines multiple specifications for specific encodings, including binary and XML encodings. Thanks to its flexibility and efficiency, ASN.1 isused in many dif-ferent areas ranging from telecommunications to genome databases.

The best analogies for understanding what ASN.1 is and how it works are nested structs in the C programming language and XML. All three are tools for defining nested data structures where each field in the structure can have a name and a type. All three tools have distinct concepts for definitions and instances. This means that while instances of C structs are always regions of memory and instances of XML are always text documents, instances of ASN.1 can be represented in many different forms depending on the chosen encoding rules. Since all of the encodings are representationally complete with respect to the abstract model, they are interchangable. Figure 10 gives an example of an ASN.1 definition for a simple detailed clinical model. This figure also illustrates that the type of each item in an ASN.1 definition may be a primitive (e.g. REAL) or it may be the result of another definition (e.g. a CodedConcept).

All coded concepts in the CDR are drawn from IHC's Healthcare Data Diction-ary (HDD), another technology jointly developed by IHC and 3M. The HDD is effectively a large coded vocabulary (over 800,000 concepts with over 4 million synonyms) containing both locally defined concepts and concepts from other coded vocabularies. The names of all the detailed clinical models used in the CDR and the fields they contain are defined as concepts in the HDD. The result is that all data stored in the CDR can be viewed as name-value pairs. The name portion of the pairs is always coded concepts.

The CDR is made up of a database and a set of services that operate on the database. For the most part, data in the database is only accessed through the services. The services perform a couple of functions. First, they provide a common access mechanism to ensure consistent security, auditing, and error handling. Equally important is the way Detailed Clinical Model Table Relational the services handle detailed clinical models. To applications built on the services, the CDR behaves more like an object- oriented database than a relational database. The applications pass instances of detailed clinical models to the services and get other instances of detailed clinical models back. Internally, the data is actually stored in a relational database, but this fact is almost completely hidden from any application.

Although the underlying database is relational, the data is not stored in a traditionally normalized relational manner. Instead the CDR has one table where the services store each instance of a detailed clinical model, formatted as an ASN.1 BER string. Every row in this table has a binary field that holds a BER string. Other fields in each row provide information for indexing purposes such as a patient identifier.

In addition, the services shred the BER strings into another small set of tables. These tables are used for indexing purposes. In effect, all of the data in the CDR is stored twice, once in the BER string and once in the relational tables.

This allows the services to do fast indexed searches in the relational tables to identify the detailed clinical models of interest. They can then read back the entire instance with a single row read instead of the large number of joins it would take to reconstitute the models if they were stored only in a normalized relational format. This is advantageous because applications commonly need the entire detailed clinical models rather than just the information present in a single row of a relational table.

IV. Arden Syntax

The Output from the above step is executable logic in Arden Syntax. Arden Syntax was developed in 1990 as a language for encoding medical knowledge. It was developed in an attempt to address the need to share medical knowledge between hospitals and

other medical institutions. Arden Syntax is currently maintained by the Health Level Seven (HL7) Arden Syntax Special Interest Group and is an ANSI 20 standard. Many vendors of electronic medical records have implemented Arden compilers in their systems.

Arden Syntax is written in units called medical logic modules (MLMs). Each MLM contains the logic necessary for making one medical decision. Portions of an MLM. An Arden Syntax MLM is made up of categories and slots. The three categories in Arden Syntax are the maintenance category, the library category, and the knowledge category. Each category contains a list of slots. The slots in the mainenance category contain information related to knowledge base maintenance and change control.

The maintenance category does not contain any clinical information. The slots in the library category describe the sources of information used in creating the MLM, keywords, and related information. The knowledge category of an MLM is where the clinical logic is represented.

The most significant slots in this category are the data slot and the logic slot. The data slot contains mappings of symbols used in an MLM to data in the target electronic medical record. The logic slot, as its name implies, contains the logic that operates on the data. While Arden Syntax is the best option currently available for sharing medical logic across institutions, it suffers from what is known as the "curly braces problem." Arden Syntax does not specify a notation for referencing data elements in the target electronic medical record (EMR). Rather, such references are written in a form that is understood by the native EMR and placed inside curly braces (e.g. the curly braces may contain a SQL statement specific to a given EMR). This means that while the logic of the module should be portable from one EMR to the next, the references to the data in the EMR are not portable. One proposed solution to the "curly braces problem" is to use and abstraction call the virtual medical record (VMR).

V. Virtual Medical Record

VMR is an abstraction of a data model for a medical record [PRC+04]. It is intended that decision logic can be written against a VMR and then distributed to any number of healthcare organizations, each possibly using a different EMR. Each EMR would have a mapping to the VMR and would therefore be able to translate VMR logic into native queries. In the MLM in Figure 4, curly braces follow the "READ" keyword. In this case the curly braces contain an abbreviated snippet of XML representing a VMR query. The specification of a standard VMR is a current effort of the Clinical Decision Support Technical Committee of HL7 and established it.

The VMR that we use in this project is based on some early work from HL7.This VMR consists of a small set of classes that describe clinically relevant information. These classes include Observation, Substance Administration, and Encounter. Each class has a number of attributes. For example the Observation class has a "code" attribute that specifies the type of the observation, a "value" attribute, and other attributes for capturing information such as the timing and status of the observation. In this project we limit our VMR queries to queries on the code and value attributes of the Observation class. This small subset of the VMR captures a large majority of the information needed to determine clinical trial eligibility.

VI. Extraction and Formula Generation

The first part, Criterion Extraction, takes a web page describing a clinical trial as input. For this thesis we used clinical trials from ClinicalTrials.gov, an internet site sponsored by the National Institutes of Health and the National Library of Medicine. We created a Python script that reads the web page describing a trial and extracts the eligibility criteria as well as available context information. The context information consists of items such as whether a criterion is an inclusion criterion or an exclusion criterion. The output of this part is an XML document containing the criteria and context information. Adapting the system to work with trials from 3 - System Design 24 other sources would involve modifying the Python script to understand the format of the new source.

The second part of Step 1, Formula Generation, takes the XML document with the extracted criteria as an input. This process parses each criterion using a link grammar parser [ST91]. From this it then creates a first order predicate calculus formula representing each criterion as Figure 3 illustrates. This process relies partiallyon recognizable sentence or phrase structure. Since the authors of clinical trials sometimes use telegraphic or ungrammatical phrasing, and since the link grammar parser we are using in this work is not familiar with many medical terms and syntactic constructs, the system is not able to correctly parse some criteria into predicate formulas. 1) The number of constructed rules is equal to or greater than the user-specified threshold. .

The root element of this XML document is labeled "criteria". This element contains a "trial" attribute whose value is the URL of the clinical trial. The "criteria" element contains a sequence of "criterion" elements. Each of these "criterion" elements contains

a sequence of "text" elements followed by a "formula" element. The "text" elements contain text that the system extracts from the trial document. The last "text" element in a sequence contains the eligibility criterion of interest. The preceding "text" elements contain available context information system.

VII. Concept Mapping

The process further takes the XML file described above as input. It attempts to map each criterion to concepts and data structures in the target electronic medical record. For each criterion that is successfully mapped we generate executable code for determining if a patient meets the criterion.CDR (Clinical Data Repository) stores clinical data as instances of detailed clinical models that can be viewed as a series of nested name-value pairs. Recall also that all of the pair names are coded concepts, as are some of the pair values. Since all of these coded concepts are in the HDD (Healthcare Data Dictionary), the mapping task consists largely of trying to match words and phrases from consisting of concepts that are either names or values in the detailed clinical models that are stored in the CDR. The content of the HDD is stored in a normalized relational fashion and we kept the same relational structure in our working subset. This way our system could easily use the live HDD or our subset of the HDD by merely changing a configuration parameter. In addition, we created an abstraction of the HDD for our system with a Terminology Server interface that defines a set of methods for making vocabulary related queries. We then created an implementation of this interface against the HDD.

VIII. Code Generation

Although we have chosen to use Arden Syntax as the language of our executable code, we constructed the code generation subsystem using the same separation of interface and implementation that we used in other areas. Therefore generating code in a different language would only require the interested party to supply an appropriate implementation of the generator interface.

Using metadata from the target EMR, the system determines that "heart disease" is valid in the value part of a name-value pair. Thus, for instance - A sample Arden Syntax read statement containing a VMR query. "VMR Query" element contains a "value" element. If the mapped concept serves as the name part of a pair, then a "code" element replaces a "value" element in the query. The valid values of this attribute depend on the type of the element that is contained within the "value" element. In this case the "value" element contains a "cd" element representing a coded concept. The comparison operations that are valid for a coded concept include "equals" and "isa." If the contents of the "value" element represented a numeric value, then numeric comparison operators such as "equals," "less than," and "greater than" would be applicable.

The second step in generating code to determine eligibility takes place after all of the criteria have been considered for mapping. In this step we generate the Arden Syntax MLM. The MLM we generate is focused on the executable logic. Even though the vast majority of slots in an MLM are required by the specification, only a handful are useful for machine execution. Most of the remaining slots are intended for human perusal. Therefore, for this project we populate only the small number of slots that are useful for automated rocessing. We do not generate any slots in the maintenance category. In the library category we populate the inks slot with the URL of original clinical trial.

IX. Conclusion

Clinical medicine is one of the most interesting areas in which data mining may have an important practical impact. The widespread availability of large clinical data collections enables thorough retrospective analysis, which may give healthcare institutions an unprecedented opportunity to better understand the nature and peculiarity of the undergoing clinical processes.

Future Scope

Combine the whole process of clinical process with computer generated treatment recommendations.

Fast and Efficient implementation of clinical guideline reference for the medical practitioners.

Dynamic updating of clinical guidelines according to trial results.

References Références Referencias

1. Rafael S. Parpinelli, Heitor S. Lopes, Member, IEEE, and Alex A. Freitas.
2. Predictive data mining in clinical medicine: a focus on selected methods and Applications Riccardo Bellazzi, Fulvia, Ferrazzi and Lucia Sacchi.
3. Predictive data mining in clinical medicine: Current issues and guidelines by Riccardo Bellazzia,, Blaz Zupanb
4. Data Mining Framework by Hemambika Payyappillil, College of Engineering and Mineral Resources at West Virginia University.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Intelligent Information Retrieval

By Ayushi Gupta, Divya Verma & Kajal Gambhir
Dronacharya College of Engineering

Abstract - The World Wide Web has become an invaluable information resource but the explosion of information available via the web has made web search a time consuming and complex process. Index-based search engines, such as AltaVista, Google or Info seek help, but they are not enough. This paper describes the rationale, architecture, and implementation of a next generation information gathering system – a system that integrates several areas of Artificial Intelligence (AI) research under a single umbrella. Our solution to the information explosion is an information gathering agent, IIR, that plans to gather information to support a decision process, reasons about the resource trade-offs of different possible gathering approaches, extracts information from both unstructured and structured documents, and uses the extracted information to refine its search and processing activities.

GJCST-C Classification: *H.3.3*

Strictly as per the compliance and regulations of:

Intelligent Information Retrieval

Ayushi Gupta [α], Divya Verma [σ] & Kajal Gambhir [ρ]

Abstract - The World Wide Web has become an invaluable information resource but the explosion of information available via the web has made web search a time consuming and complex process. Index-based search engines, such as AltaVista, Google or Info seek help, but they are not enough. This paper describes the rationale, architecture, and implementation of a next generation information gathering system – a system that integrates several areas of Artificial Intelligence (AI) research under a single umbrella.

Our solution to the information explosion is an information gathering agent, IIR , that plans to gather information to support a decision process, reasons about the resource trade-offs of different possible gathering approaches, extracts information from both unstructured and structured documents, and uses the extracted information to refine its search and processing activities.

I. Introduction

The World Wide Web has given the researchers, businessmen, corporate, students, hobbyists and technical groups a medium by which they can share the information they have, with others. The ease of HTML and platform independence of the web documents has lead to a tremendous growth of the web, that has outstripped the technologies that are used to effectively search in these pages, as well as proper navigation and interpretation.

With the aim of inception of AI (Artificial Intelligence) in the searching techniques, the first step we have decided is to find out those limitations in the current searching methodologies, which make the result unsatisfactory and not up to the expectations. Some of the key features of today's search engines are:

- **Meta Searching:** The scope of each search engine is limited and no search engine has the database that covers all the web pages. This problem was noted long ago and was solved with the help of Meta search sites that make use of multiple search engines to search for the "Query String. The common names of such search engines are 37.com (which searches 37 search sites simultaneously), metacrawler.com and many others. Another advantage of these Meta search sites is that they incorporate advanced features which are absent in some of the member search sites (Member search sites are those sites which return the search result to Meta Search engines). But the basic methods used in these Meta search sites are more or less same as those used in any other search engines.
- **URL Clustering:** URL clustering was a basic problem from which most of the earlier search sites were affected. Suppose we search for 'GRE' and we intend to get the link to all those sites that have information on GRE exam. But a search engine without URL clustering will give results.
- **Shopping Agent:** It is an intelligent enhancement over the other searching techniques, which *tries* to give the most appropriate site, not just any site that has high frequency of the Key Phrase. For example, if a person with the intention of shopping searches the web for 'Printer', then the normal search engines will return the pages which have high frequency of the word - 'Printers'. There can be a case when one of the results contain the information irrelevant from the point of shopping, but still that page has very high frequency of Query String. Say, a person on his personal home page writes : 'My **printer** HPDeskJet 640C is a 1997 model', then the search engines will return this page as well (Note that the user has not used any Boolean Operator in the search string). While a shopping agent gives the details in a commercial format like price range, model, other options, second hand options and just many results. The implementation of Shopping Agent was first serious step forward in the direction of making *Intelligent Information Retrievers.* We will be using its powers in the design of our new Information Gathering Agent.
- **Personal Information Agent (PIA):** This is the most important step forward in the direction of incorporating AI in searching. The PIAs try to retrieve your personal interests and give the result accordingly. The information is gathered either from the previous search results and mostly by a questionnaire. But the current day PIAs are very slow and less adapting. In this paper, we will try to confer PIA with the power to give satisfying and fast search results.

II. Intelligent Information Retrieval

The solution to the problem of *Intelligent Information Retrieval* is to integrate different Artificial Intelligence (AI) technologies, namely scheduling, planning, text processing, information extraction and interpretation into a single information gathering agent, which we christen as **Intelligent Information Retriever (IIR).** IIR locates, retrieves and processes information to support a human decision process. During thinking, we human adopt a top-down and a down-top structured

E-mail α : sweetuaayushi@gmail.com
E-mail σ : divyaverma417@gmail.com
E-mail ρ : kjlgambhir7@gmail.com

analysis. Before discussing how this can be implemented through AI, first let's have a glimpse at how human is able to do this. For this, we create a scenario in which a person wants to buy a book that is not available at the local book stores. The person now has two options: Order the book from the publisher and second option is to go to a nearby town and have the book from there, provided that the person has the information that the book is available at the book stores of that city. To complicate the situation, further assume that the book is by a foreign publisher and that publisher has no branch in the country of the person, so ordering a book from the publisher will result in a time consuming process. Let us further assume that the overhead expenses involved in visiting the neighboring town is more than the actual cost of the book. Now the person will subconsciously list all the parameters in his mind that may affect the decision of buying the book. The typical, probably minimum list of questions that will come in his mind are:

1. Whether the book really worth buying?
2. Whether the book is required urgently?
3. Is there any alternative to that book?
4. Do I have enough money to buy that book?
5. Do I have enough money to bear the overhead expenses involved in visiting neighboring town/city?
6. How will I get to the neighboring city / How will I order the book from the publisher?

So, in any such decision making, humans make use of following :

- Interpretation [derived from pt. 1 and 2 above]
- Comparison [pt. 3]
- Monetary Factors [pt. 4 & 5]
- Planning and scheduling [pt.6].

Our aim is to incorporate above decision making constructs in searching making proper use of AI (Artificial Intelligence). We will be implementing all this through a new information-gathering agent, that we have already christened as Intelligent Information Retriever (IIR).

"The IIR is a data-driven as well as expectation-driven and adaptable information gathering agent that does information interpretation through decision making constructs properly adjusted and incorporated with the existing powerful points of today's search engines, most prominent of which being Personal Information Agent and Shopping Agent." After having formally designed the definition of IIR, we are in a position to be equipped with the tools and techniques that will be used in the design of IIR.

III. Criticism

The integration of different components in IIR - The Task Assessor, Decision Maker, CORE, Object database, Information Extractor is itself a major accomplishment in its own kind. Despite the integration issues, the combination of the different AI components in IIR and the view of information gathering as an interpretation task have given IIR some very strong abilities.

In terms of limitations, the following points should be noted:

- Initially, due to smaller Object Database, the results will be lesser efficient (but still more efficient than current technology). This problem can be overcome by having a large database before the start of the service.
- The form fields to be filled by the user *may* increase, if precise results are desired.
- The cost of implementation will be very high.

Despite these limitations, this Intelligent Information Retriever is a major enhancement over the current search engines and is a serious step forward in the direction of incorporating Artificial Intelligence in searching for more efficient results.

References Références Referencias

1. Intelligent information retrieval. en.wikipedia. org/wiki/intelligent_information_retrieval
2. Intelligent techniques for IIR www.sigir.org/forum/2006D/2006d_sigirforum_siddiqui.pdf
3. Intelligent information retrieval laboratory iir.csie.ncku.edu.tw

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data Mining Based on Semantic Similarity to Mine New Association Rules

By Sandeep Jain & Aakanksha Mahajan
Doon Valley Institute of Engineering And Technology

Abstract - The problem of mining association rules in a database are introduced. Most of association rule mining approaches aim to mine association rules considering exact matches between items in transactions. A new algorithm called "Improved Data Mining Based on Semantic Similarity to mine new Association Rules" which considers not only exact matches between items, but also the semantic similarity between them. Improved Data Mining (IDM) Based on Semantic Similarity to mine new Association Rules uses the concepts of an expert to represent the similarity degree between items, and proposes a new way of obtaining support and confidence for the association rules containing these items. An association rule is for ex: i.e. for a grocery store say "30% of transactions that contain bread also contain milk; 2% of all transactions contain both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule and this rule is represented as Bread →Milk. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. This paper then results that new rules bring more information about the database.

Keywords : *Data mining, Semantic similarity, Association Rules, Support, Confidence, Fuzzy logic.*

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Data Mining Based on Semantic Similarity to Mine New Association Rules

Sandeep Jain [α] & Aakanksha Mahajan [σ]

Abstract - The problem of mining association rules in a database are introduced. Most of association rule mining approaches aim to mine association rules considering exact matches between items in transactions. A new algorithm called "Improved Data Mining Based on Semantic Similarity to mine new Association Rules" which considers not only exact matches between items, but also the semantic similarity between them. Improved Data Mining (IDM) Based on Semantic Similarity to mine new Association Rules uses the concepts of an expert to represent the similarity degree between items, and proposes a new way of obtaining support and confidence for the association rules containing these items. An association rule is for ex: i.e. for a grocery store say "30% of transactions that contain bread also contain milk; 2% of all transactions contain both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule and this rule is represented as Bread → Milk. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. This paper then results that new rules bring more information about the database.

Keywords : Data mining, Semantic similarity, Association Rules, Support, Confidence, Fuzzy logic.

I. Introduction to Data Mining

Data mining (DM), also known as knowledge discovery in databases (KDD), has been recognized as a new area for database research. This positive and evolutionary cycle is now occurring in area named data mining or knowledge discovery in database for efficiently discovering interesting rules from large collections of data. Informative knowledge discovering and new valuable data finding in database are very attractive in various business scenes.

Data mining (DM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns such as association rules. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data "and "the science of extracting useful information from large data sets or databases"[1]. It involves sorting through large amounts of data and picking out relevant information. It is usually used by businesses and other organizations, but is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimentation. Although data mining is a relatively new term, the technology is not. Companies for a long time have used powerful computers to sift through volumes of data such as supermarket scanner data, and produce market research reports. Continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy and usefulness of analysis.

II. A Conceptual Model of Data Mining

Many useful studies have been done in data mining and knowledge discovery in database. By basing on the concept that features the process aspects of data mining, we gives attention to the interaction between a human and a machine and the purpose clarification. Figure 1. Shows the conceptual model of data mining:

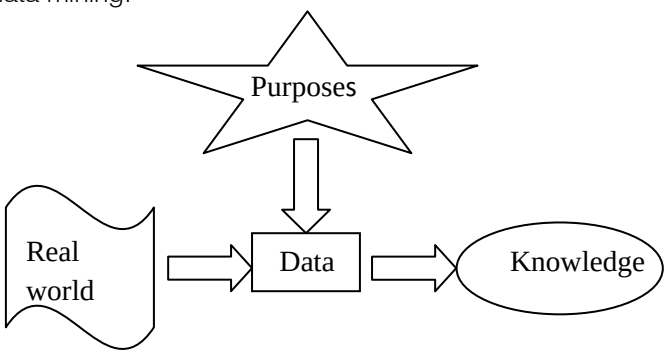


Figure 1 : Conceptual of data mining

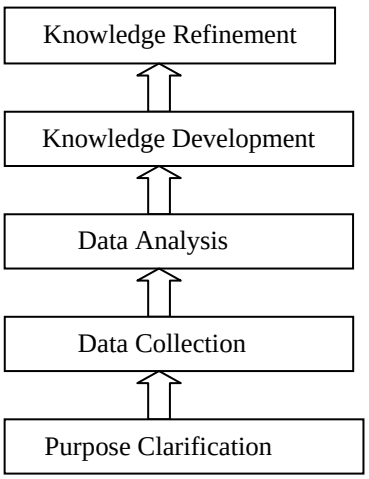


Figure 2 : Data Mining Process

Author α σ : Department of Computer Science & Engineering, Doon Valley Institute of Engineering And Technology.
E-mail α :1sandeepjain4891@gmail.com
E-mail σ : er.aakanksha@yahoo.co.in

A conceptual model of data mining is proposed by generalizing the actual application development process. Data mining is the process which extracts knowledge from real world environment according to a certain purpose. In this process, top-down and bottom-up approaches are performed as problem solving methods. The top-down approach clarifies purpose, defines problems to be solved, then breaks down the problems into elements until solvable level.

On the other hand, the bottom-up approach collects data from the real world, analyzes them, and then integrates the findings. Both approaches are combined into data mining to find solvable goal, to select a suitable method for the goal and then to develop knowledge based on the method. The data mining process is shown in Figure 2. The steps below are the generalized data mining process. Before applying the process, we should define the benefits of developing target applications clearly to give the purpose.

1) Purpose Clarification: Clarifying the purpose, the problems to be solved, and the hypothetical goal of solution through the top-down approach.
2) Data Collection: Collecting data from the real world and visualizing them through the bottom-up approach.
3) Data Analysis: Analyzing the data collection to verify the hypothetical goal of solution through the combination of the top-down and the bottom-up approach.
4) Knowledge Development: Selecting a suitable method for the goal and developing knowledge based on the method.
5) Knowledge Refinement: Testing and refining the knowledge. If necessary, back to the previous steps.

III. Improved Data Mining Based on Fuzzy Weighted Association Rules

Data Mining has been researched a lot due to its utility in many applications, and one of its most used tasks is Association Rule Mining. Given a set of transactions, where each transaction is a set of items, an association rule is an expression X => Y, where X and Y are sets of items (or item sets). The meaning of such a rule is that transactions which contain items in X tend to also contain items in Y. The support of the rule X => Y is the percentage of transactions that contain both X and Y. The confidence of the rule X => Y is the percentage of transactions containing X that also contain Y.an example of an association rule is "90% of transactions that contain bread also contain butter; 3% of all transactions contain both of these items."The 90% is referred to as confidence and the 3%, the support of the rule. The problem of mining association rules is to find rules having minimum support and confidence. Many algorithms were developed to solve the problem of mining association rules. In general, new approaches were motivated by finding new ways of dealing with different attributes types or increasing computational performance. However, new approaches could address other issues. In this paper, we concern about semantics on mined data. Known algorithms only consider exact matches when mining frequent item sets, not generating some association rules which could bring important information.

In our approach, besides exact matches, the semantic similarity between items is also taken on account. For example, consider the set of transactions shown in Table 1.

TID	Attribute1	Attribute2
1	Chair	Table
2	Sofa	Desk
3	Chair	Desk
4	Chair	Table

Table 1 : A set of transaction examples

If this set of transactions were mined by a traditional association rule mining algorithm, the following association rules would be obtained:

Chair => table (support 50%, confidence 67%)
Sofa => desk (support 25%, confidence 100%)
Chair => desk (support 25%, confidence 33%)

Thus, if a minimum support of 50% and a minimum confidence of 60% were established, the only rule generated would be chair=> table. In this situation, only strings of characters are being considered, and as they have the same characters, with the same order and the same length, the mining algorithm will recognize a match. Table and desk, for example, are totally different words, but it does not mean they are totally different items. If we semantically analyze the words table and desk, we can consider them similar (both are furniture and have similar utilities, for example). In this case, there is not an exact match, but there is a kind of "similarity match", which can be also useful to find relevant association rules and therefore important information. That is what traditional approaches can not reveal: association rules including semantically similar items. To make it possible, in this paper we present an algorithm called IDM.

IV. Algorithm

a) *Semantic Similarity*

The objective of data mining is to discover knowledge, and that is why so many approaches try to make rules more understandable. Analyzing the meaning of mined data (i.e., the data semantics) naturally contributes to increase the quality of information obtained through the mining process, and consequently better the decisions guided by this

information will be. Database transactions have different attribute types. The attributes can be quantitative or categorical[6] i.e. during the mining process, quantitative attributes cannot be semantically analyzed, but some categorical attributes can be. Known algorithms usually deal with categorical attributes as if they were mere character strings. These strings are recognized and counted along the transactions, and associations between them are found. In this case, matches occur only when strings have exactly the same characters, in the same order, with the same length. However, different strings can represent similar meanings. Consider, for instance, the words cupboard and wardrobe. Although character strings are totally different, they represent semantically similar words. Cupboard and wardrobe are different objects, but they both have shelves and doors and are used for storing things. They are not identical, but they are similar. This semantic similarity between items is ignored by traditional algorithms, what can make them lose important information. This additional analysis considering associations between similar items may reveal other association rules, which can be also relevant. We call semantically similar data mining the mining process which also considers the semantic similarity between data items, extending the usual way of mining association rules.

In this paper, we present a new algorithm called IDM. In IDM, the semantic similarity between data is expressed by a similarity degree between items. Thus, if the value of similarity degree between items is 1 (one), this means that compared items have maximum similarity. According to the reflexive property of binary fuzzy relations, it can only occur if an item is compared to itself. Therefore, when comparing two non-identical items, the similarity degree (sim) between them must be a value greater or equal to zero and less than one ($0 \leq sim < 1$). During the mining process, if the similarity degree between items is greater than a user-defined parameter, a semantic similarity association is detected, meaning that items contained in this association are similar enough (and therefore interesting to the user). Next section shows how IDM detects these semantic similarity associations and uses them to get important association rules.

b) Algorithm Structure

IDM is based on Apriori and, as an association rule mining algorithm, it needs user-provided minimum support and minimum confidence parameters to run. Moreover, by using fuzzy logic concepts, IDM also needs a user provided parameter which indicates the minimum similarity degree desired, called minsim. Thus, there are the following parameters:

- minsup, which indicates the minimum support;
- minconf, which represents the mininum confidence;
- minsim, which is the minimum similarity degree necessary to consider two items similar enough, and then associate them during mining.

All of these parameters are expressed by a real value in the interval [0, 1]. The steps performed by IDM are shown below

1. Data Scanning: Identifying items and their domains
2. Determining similarity degrees between items for each domain
3. Identifying similar items
4. Generating candidates
5. Calculating the weight of candidates
6. Evaluating candidates
7. Generating rules

Now, consider as an example a table containing transactions of buys from a furniture store (Table 2), where Tid is an identifier for each transaction, whereas Dom1, Dom2 and Dom3 contain items bought by the furniture store customers.

Moreover, suppose henceforth that we have the following parameter values:

- minimum support (minsup) = 0.45
- minimum confidence (minconf) = 0.3
- minimum similarity (minsim) = 0.8

Tid	Dom1	Dom2	Dom3
10	Chair	Table	wardrobe
20	Sofa	Desk	cupboard
30	Seat	Table	wardrobe
40	Sofa	Desk	cupboard
50	Chair	Board	wardrobe
60	Chair	Board	cupboard
70	Chair	Desk	cupboard
80	Seat	Board	cabinet
90	Chair	desk	Cabinet
100	Sofa	desk	cupboard

Table 2 : Transactions of buys from a furniture store

c) Data Scanning

The first step is a data scanning that identifies items in the database. IDM identifies each item, generating 1-itemsets (itemsets with size one). Moreover, in this step each item is associated to a domain, which is important because they make possible to relate items according to their similarity only when is convenient — that is, if they belong to the same domain. When mining relational tables, domains can be defined by the column where the item is. Thus, considering the furniture store example, after data scanning we have items and domains identified, as shown in Table 3.

Items	Domain
sofa, chair, seat	Dom1
board, desk, table	Dom2
cabinet, cupboard, wardrobe	Dom3

Table 3 : Items and domains identified by data scanning

In this example, domain Dom1 contains items of furniture where one can sit, domain Dom2 contains items of furniture where one can place things on them, and domain Dom3 contains items of furniture where one can store things. Each domain contains items used in similar situations, what makes domains identification semantically coherent. The number of items belonging to domain determines its size. Thus, all domains in Table 3 have size 3.

d) Determining Similarity Degrees

After having items and their domains identified, it is time to determine the values of similarity relations within each domain. These values must be supplied by a domain specialist (usually the user himself). This task corresponds to one of the steps of KDD [3], prior to the step of data mining. Alternatively, it would be possible to obtain these values automatically, through a rule or method. However, to determine the similarity values between items so that the semantics is considered, it is necessary to adopt a way of reproducing, with high fidelity, the capacity of the human mind of doing this. Any rule chosen to determine these values automatically will consider non-semantic factors, decreasing the quality of the analysis realized and this way going against the objective of the semantically similar data mining, which is to enrich the analysis and consequently enrich the information obtained from the rules. In each domain, the similarity degree values are stored in a similarity matrix. In the furniture store example, 3 domains were identified, and the correspondent similarity matrices can be seen in Table 4. The values in the matrices inform the similarity degree between the items of the domain. For example, chair is 70% similar to sofa. Next subsection shows how each similarity matrix is consulted to identify similar items.

e) Identifying Similar Items

In this step, the similarity matrix of each domain is analyzed, thus identifying pairs of items whose similarity degree is greater than or equal to minsim. These pairs of items compose fuzzy associations of size 2. In IDM, these associations are expressed through fuzzy items,[2] which are representations where the ~ symbol is used to indicate the relation between items. Thus, supposing that the sufficiently similar items are item1 and item2, for example, a fuzzy item on the form item1~item2 represents the fuzzy association between them.

Dom 1	sofa	seat	chair
sofa	1	0.75	0.7
seat	0.75	1	0.6
chair	0.7	0.6	1

Dom2	desk	table	Board
desk	1	0.9	0.75
table	0.9	1	0.7
board	0.75	0.7	1

Dom3	cabinet	wardrobe	cupboard
cabinet	1	0.9	0.85
wardrobe	0.9	1	0.8
cupboard	0.85	0.8	1

Table 4 : Domains and their respective similarity matrices

In the furniture store example, the similarity matrices in Table 4are analyzed and, considering the minsim value (0.8), the associations shown in Table 5 are obtained.

Domain	Value	Similarity relation1	Equivalent fuzzy item
Dom2	0.9	sim(desk, table)	desk~table
Dom3	0.9	sim(cabinet, wardrobe)	cabinet~wardrobe
Dom3	0.85	sim(cabinet, cupboard)	cabinet~cupboard
Dom3	0.8	sim(cupboard, wardrobe)	cupboard~wardrobe

Table 5 : Similarity relations that satisfy minsim

After obtaining the set of fuzzy associations of size 2, the existence of similarity cycles is verified. A similarity cycle is a fuzzy association of size greater than 2 that only exists if all of its items are, in pairs, sufficiently similar. That is, according to the intersection operation in fuzzy set theory, the minimum value among the similarity degrees involved in the cycle must be greater than or equal to minsim. It is what occurs, in the furniture store example, with the cycle cabinet~wardrobe~cupboard, shown in Figure 3. In this figure, arrows represent the similarity relations between items, and near them are the values that express the relation values. Thus, it is possible to see that in this example all the items are, in pairs, sufficiently similar ($0.9 \geq 0.8$, $0.85 \geq 0.8$ and $0.8 \geq 0.8$). Or else, it can be verified that the minimum value among the similarity degrees involved is greater than or equal to minsim ($\min(0.9, 0.85, 0.8) \geq 0.8$). Whereas the minimum size of a similarity cycle is 3, the maximum size is equal to the size of the analyzed domain.

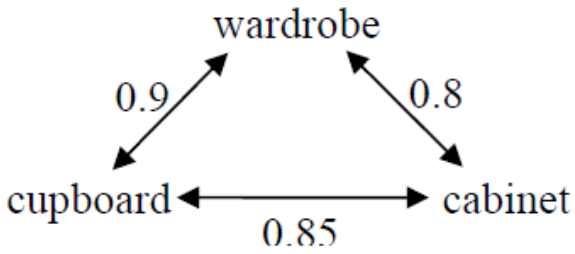


Figure 3 : Similarity cycle

```
1   A2 = { Set of fuzzy associations of size 2}
2   for (k = 3; k < size(Dn) ; k++)
3   compare each pair of fuzzy associations in Ak-
4   1
5   if (prefix(ai) = prefix(aj), i ≠ j)
6   //if suffixes union is sufficiently similar
7   if ((suffix(ai) U suffix(aj)) ∈A2)
8   ak = ai U suffix(aj)
9   end if
10  end if
11  end compare
12  Ak = {set of all ak}
13  end for
    Sn = Group of all Ak
```

Figure 4 : Algorithm to find similarity cycles

Ak	Set of fuzzy associations of size *k*
Dn	Each one of the *n* domains analyzed
Ak	Each fuzzy association in *Ak* set
Sn	Set of similar items on domain *n*
size(*Dn*)	*Dn* size
prefix(*ak*)	*ak* fuzzy association prefix
suffix(*ak*)	*ak* fuzzy association suffix

Table 6 : Notation used in figure 4

In the furniture store example, *Dom3* size is 3, and that is why no fuzzy association of size greater than 3 can be obtained. However, for bigger domains, there can be cycles of bigger sizes. That is why IDM checks for the existence of similarity cycles iteratively on each domain, where the fuzzy association sets of size k-1 are analyzed on each step k (k ∈ N, k ≥ 3) to obtain fuzzy associations of size k. The notation sim(item1,item2) represents the similarity relation between item1 and item2.

A fuzzy association *ak* is of the form {*s1*, *s2*, ..., *sk-1*, *sk*}, where *s1*, *s2*, ..., *sk-1*, *sk* are the *k* items which composes it. Its suffix is on the form {*sk*}, whereas its prefix is on the form {*s1*, *s2*, ...,*sk-1*}. Every obtained *A k* in this step are grouped in *Sn*. This is how the step of identifying similar items ends, and then another iterative part of the algorithm begins. In this part, for each step *k* (*k* ∈ N), the *k*-item set candidates are generated from the frequent item sets obtained on previous step (*k*-1). Also, weights of *k* item set candidates are calculated and candidates are evaluated.

f) Generating Candidates

The way candidates are generated is very similar to the way it is done in Apriori. However, in IDM, besides items identified during the data scanning step, fuzzy items — which represent fuzzy associations obtained in the step of identifying similar items — also integrate the generated candidates. At the end of this step, we have the set of *k*-item set candidates, which is submitted to the step of calculating the weight of candidates.

g) Calculating The Weight Of Candidates

In this step, the weight of each item set candidate is calculated. The weight of an item set corresponds to the number of its occurrences in the database. In IDM, differently from what happens in A priori, an item set can have fuzzy items, hence called fuzzy item set. The notation item1∼item2, has the following meaning: if item1 and item2 are very similar, they can be considered as being practically identical; thus, if occurrences of item1 or item2 are found in the database, they will be associated and, together with the similarity degree between items, they will compose a fuzzy occurrence of item1∼item2. Therefore, we need to know if the item set is fuzzy or not, before calculating its weight: if the item set is not fuzzy, we calculate its weight in the conventional way, counting its exact occurrences; if the item set is fuzzy, we shall consider its fuzzy occurrences to obtain its weight. To understand how fuzzy occurrences happen, suppose that the similarity degree between item1 and item2 is 0.8. In this case, each occurrence of item2 in the database can be considered equal to 80% of item1 occurrence. Consequently, for each item1 occurrence we sum one item1 occurrence (of course), and for each item2 occurrence we sum 0.8 item1 occurrence (Table 7– situation A).

The problem can also be seen in the contrary manner, summing one item2 occurrence for each item2 occurrence and 0.8 for each item1 occurrence (Table 7– situation B). Notice that, for situation A, the fuzzy occurrences totalize the value of 2.8 (1.0 + 1.0 + 0.8), whereas for situation B fuzzy occurrences totalize the value of 2.6 (0.8 + 0.8 + 1.0).

Tid	Dom1	
10	item1	1.0
20	item1	1.0
30	item2	0.8

Situation A

Tid	Dom1	
10	item1	0.8
20	item1	0.8
30	item2	1.0

Situation B

Table 7 : Fuzzy Occurrences

Hence, depending on situation, the result obtained for the same similar items could be different. To avoid this distortion, it is necessary to balance this counting. To do that, consider weight (item1) as the number of item1 occurrences, weight (item2) as the number of item2 occurrences; and sim (item1, item2) as the similarity degree between item1 and item2. Thus, for situation a in Table 7, the number of occurrences is given by the expression.

$$\text{weight}(item_1) + \text{weight}(item_2) \times \text{sim}(item_1, item_2)$$

In the same way, for situation B in Table 7, the number of occurrences is given by the expression.

$$\text{weight}(item_1) \times \text{sim}(item_1, item_2) + \text{weight}(item_2)$$

We adopt the arithmetic average between situations A and B to balance the two situations, getting the fuzzy weight of item1∼item2 through the Equation 1.

$$\textbf{Fuzzy Weig}\square\textbf{t} = \frac{[\textbf{weig}\square\textbf{t(item1)} + \textbf{weig}\square\textbf{t(item2)}][1 + \textbf{sim1(item1, item2)}]}{2}$$

Equation1. Fuzzy weight for two similar items

Equation 1 is useful to calculate the weight of fuzzy items formed by an association of only two similar items. After this, itemset candidates are evaluated in the next step of IDM.

h) Evaluating Candidates

This is the step of IDM where the support of itemset candidates is evaluated, similar to what is done in Apriori. The support corresponds to the weight divided by the number of rows (or total of transactions) in the database (Equation 2). If the itemset candidate is fuzzy, its weight is also fuzzy, and then when its weight is divided by the number of rows, the result is its fuzzy support. Thus, generically, the support of each item set is calculated from its weight, and it is verified if its support is greater than or equal to minsup. In negative case, the item set is considered not frequent, and is therefore discarded. In positive case, the item set is stored in the set of frequent item sets.

$$\textbf{Support} = \frac{\textbf{weight(itemsets)}}{\textbf{number of rows in the database}}$$

Equation2. Support of the item set

The end of this step is also the end of the iterative part of IDM. At this time, all frequent item sets are grouped in a set, from which it is possible to start the step of generating rules.

i) Generating Rules

Association rules have antecedents (items left of arrow) and consequents (items right of arrow), as shown in Figure 5.

Antecedent →Consequent

Figure 5 : Antecedent and consequent of the rule

If confidence, given by Equation 3, is greater than or equal to minconf, then rule is valid.

$$\textbf{Confidence} = \frac{\text{Support(rule)}}{\text{Support(anteecedent)}}$$

Equation 3. Rule confidence

Regardless of supports being fuzzy or not, confidence is obtained in the same way. When IDM is concluded, all valid rules are exhibited, showing antecedent, consequent, support and confidence of each rule, in the format shown in Figure 6.

Anteecedent→ Consequent sup = < support value > conf = <confident value>

Figure 6 : Association Rule Format

In IDM, antecedents and consequents of the rule can contain fuzzy items, and the values of support and confidence reflect the influence of the similarity degree between items in their calculations.

V. Tests

We realized some tests to compare the results obtained with IDM and Apriori, using real data about furniture store. We started testing our first set of data, named FURNITURE STORE, containing transactions with the following attributes. There are semantic similarities in the domain and the similarity degrees between its items are shown in Table 8.These similarity values are manually decided.

Item1	Item2	Similarity
Chair	Sofa	70
Sofa	Seat	75
Desk	Table	90
Desk	Board	75
Table	Board	70
Cupboard	Wardrobe	90
Cupboard	Cabinet	85
Wardrobe	cabinet	80

Table 8 : Similarity degrees for furniture store

We mined FURNITURE STORE using Apriori with parameters minsup = 40 and minconf = 40, obtaining the rules shown in Figure 8. We also mined FURNITURE STORE using IDM with the parameters minsup = 40, minconf = 40 and minsim = 80, obtaining the rules shown in Table 9.

Test with Apriori over the set FURNITURE STORE, with minsup = 40 and minconf = 40,Itemsets pair above minimum support and minimum confidence rule:
Rules generated Chair → Sofa sup= 50% conf= 66.6% Sofa → Chair sup= 50% conf= 66.6%

Table 9 : Test With Apriori Over the Set Furniture Store

In Table 10, the underlined rules are those ones which are obtained by IDM, but are not obtained by Apriori. The additional rules bring more information, which can be useful for decision making. When the association rule contains fuzzy

Test with Apriori over the set FURNITURE STORE, with minsup = 40 and minconf = 40,Itemsets pair above minimum support and minimum confidence rule:
Rules generated Chair~sofa → table sup= 50% conf= 100% Table→ chair~sofa sup= 50% conf= 100%
Chair → Sofa sup= 50% conf= 66.6% Sofa → Chair sup= 50% conf= 66.6% Chair~sofa → table sup= 50% conf= 100% Table→ chair ~ sofa sup= 50% conf= 100%

Table10 : Test With Idm Over the Set Furniture Store

items, its support and confidence values are calculated considering the semantic similarity between items. Association rules obtained by IDM contain fuzzy items like chair~sofa (chair and sofa can be considered similar) and which represents interesting semantic similarities not revealed by Apriori. Analyzing the additional rules obtained by IDM, we can show that IDM generates more association rules than Apriori does, with the same support and confidence parameters. As expected, the computation performance of Apriori is better than the computational performance of IDM, because IDM has a more complex structure to find semantically similarity items.

VI. Conclusion and future Work

We have discussed the data mining algorithms and techniques, which have been used by the researchers to implement the data mining for very large data. With the creation and application of IDM, it has become possible to discover association rules that reflect the semantic similarity among data. The use of fuzzy logic concepts in IDM contributed to make information representation and manipulation closer to the human language, making them more understandable. The better the comprehension of the obtained knowledge, the bigger the knowledge utility. We have also discussed the data mining challenges, in which the researches are required for developing efficient and uniform data mining algorithms, software tools and techniques for very large, high dimensional and complex data.

As future work, here in this paper because a human expert knowledge is used reason that it is easy for human to recognize objects which are existing in a database or to understand the meanings from just short conversion with using their background knowledge. Thus in near future we are thinking to enhance our system in such a way that their should not be a requirement to have an expert for finding similarity between items.

References références referencias

1. W.H. Au, K.C.C. Chan, An effective algorithm for discovering fuzzy rules in relational databases, in: Proc. IEEE Internat. Conf. Fuzzy Systems,vol. II, 1998, pp. 1314–1319.
2. W.H. Au, K.C.C. Chan, FARM: a data mining system for discovering fuzzy association rules, in: Proc. FUZZ-IEEE'99, vol. 3, 1999, pp. 22–25.
3. Han, J. and Kamber, M. (2001) "Data Mining - Concepts and Techniques", 1st Edition.Nova York: Morgan Kaufmann.
4. Chen, G. and Wei, Q. (2002) "Fuzzy association rules and the extended mining algorithms",
5. Fuzzy Sets and Systems, v. 147, n. 1-4, p. 201-228 X.Wu, C.Zhang, and S.Zhang, Mining both Positive and Negative Association Rules, Proc. Of 19th Int. Conf. on Data Machine Learning,pp.658-665,2002.
6. T. P. Hong, K. Y. Lin and S. L. Wang, "Mining fuzzy association rules from quantitative transactions", Soft Computing, Vol. 10, No.10, pp. 925-932, 2006.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Query Join Processing Over Uncertain Data for Decision Tree Classifiers

By V. Yaswanth Kumar & G. Kalyani
JNTU, Kakinada

Abstract - Traditional decision tree classifiers work with the data whose values are known and precise. We can also extend those classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Example sources of uncertainty measurement/quantization errors, data staleness, and multiple repeated measurements. Rather than abstracting uncertain data by statistical derivatives, such as mean and median, the accuracy of a decision tree classifier can be improved much if the complete information of a data item is used by utilizing the Probability Density Function (PDF). In particular, an attribute value can be modelled as a range of possible values, associated with a PDF. The PDF function has only addressed simple queries such as range and nearestneighbour queries. Queries that join multiple relations have not been addressed with PDF. Despite the significance of joins in databases, we address join queries over uncertain data. We propose semantics for the join operation, define probabilistic operators over uncertain data, and propose join algorithms that provide efficient execution of probabilistic joins especially threshold. In which we avoid the semantic complexities that deals with uncertain data. For this class of joins we develop three sets of optimization techniques: item-level, page-level, and index-level pruning. We will compare the performance of these techniques experimentally.

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Query Join Processing Over Uncertain Data for Decision Tree Classifiers

V. Yaswanth Kumar [α] & G. Kalyani [σ]

Abstract - Traditional decision tree classifiers work with the data whose values are known and precise. We can also extend those classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Example sources of uncertainty measurement/quantization errors, data staleness, and multiple repeated measurements. Rather than abstracting uncertain data by statistical derivatives, such as mean and median, the accuracy of a decision tree classifier can be improved much if the complete information of a data item is used by utilizing the Probability Density Function (PDF). In particular, an attribute value can be modelled as a range of possible values, associated with a PDF. The PDF function has only addressed simple queries such as range and nearest-neighbour queries. Queries that join multiple relations have not been addressed with PDF. Despite the significance of joins in databases, we address join queries over uncertain data. We propose semantics for the join operation, define probabilistic operators over uncertain data, and propose join algorithms that provide efficient execution of probabilistic joins especially threshold. In which we avoid the semantic complexities that deals with uncertain data. For this class of joins we develop three sets of optimization techniques: item-level, page-level, and index-level pruning. We will compare the performance of these techniques experimentally.

I. Introduction

Classification rules can be represented as below. Consider the information about the insurance company information.

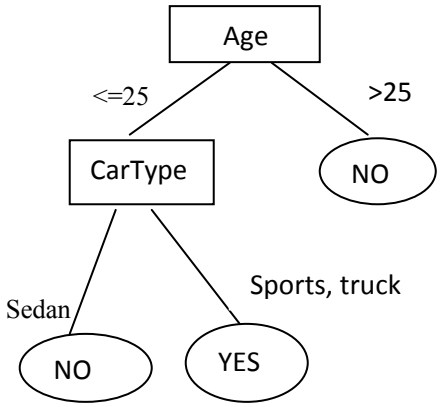
Insurance info (age: **integer**, cartype: **string**, highrisk: **boolean**) if age is between 16 and 25 and cartype is either sports or truck, then the risk is high.

Consider the information about the insurance customers.

Age	Cartype	Highrisk
23	Sedan	False
30	Sports	False
36	Sedan	False
25	Truck	True
30	Sedan	False
23	Truck	True
30	Truck	False
25	Sports	True
18	Sedan	False

Author α : Student, DVR & Dr HS MIC College of Technology, Kanchikacherla, Krishna(dt).
Author σ : Assoc. professor, DVR & Dr HS MIC College of Technology, -Kanchikacherla, Krishna(dt).

Trees that represent classification rules are called classification trees or decision trees.



Data uncertainty arises naturally in many applications due to various reasons. We briefly discuss three categories here: dimension errors, data mustiness, and repeated dimensions.

a) **Dimension Errors**: Data obtained from measurements by physical devices are often imprecise due to dimension errors.
b) **Data mustiness**: In some applications, data values are continuously changing and recorded information is always out of date.
c) **Repeated dimensions**: Perhaps the most common source of uncertainty comes from repeated dimensions. For example, a patient's body temperature could be taken multiple times during a day.

Type-1 Probabilistic Relations

Type-1 uncertainty refers to confidence if a tuple belongs to a relation or not. Consider the table represents a part of my personal address book. It is not really likely that my address book contains the phone number of the Dutch Queen, where it
is very likely that the address book contains the phone number of one of my fellow students, Ruud van Kessel.

Type-2 Probabilistic Relations

With Type-2 uncertainty, the value of the key-attribute is deterministic but values of other attributes in the relation may be uncertain. Table shows a relation which depicts where Kings and Queens of different countries around the world live. There is no uncertainty about the country where the King or Queen lives. Since the attribute values of the field "town" for Queen Beatrix

and King Carl XVI Gustaf represents uncertainty, it is not possible to tell in which village they live with complete certainty, based on this list.

If the probability that the join pair meets the join condition exceeds the threshold, it is included in the result, otherwise the pair is not included. This threshold can either be user specified or a system parameter. The tuple pairs when their probabilities exceed a certain threshold as Probabilistic Threshold Join Queries (PTJQ) we focus on threshold joins and develop various techniques for the efficient (in terms of I/O and CPU cost) algorithms for PTJQ. In particular, we develop three pruning techniques:

Name	Country	Town
BeatrixvanOranje	The Netherlands	The Hague/0.9 Amsterdam/0.1
Carl XVI Gustaf	Sweden	Stockholm/0.5 Malmo/0.5

(a) Type-1 Probabilistic Relations

Name	Phone number	Probability
Beatrixvan Oranje	+31701234567	0.01
Ruud van Kessel	+316 12345678	0.99

(b) Type-2 Probabilistic Relations

1) **item-level pruning**, where two uncertain values are pruned without evaluating the probability.
2) **page-level pruning**, where two pages are pruned without probing into the data stored in each page.
3) **index-level pruning**, where all the data stored under a subtree is pruned. Two useful types of join operations specific to uncertain attributes: value join (v-join) and distribution join (d-join). V-join is a natural extension of the join operation on deterministic data. The PDF (probability Density Function) can be used to calculate the Range of values to an attribute which contains attribute uncertainity. PDF also calculates probability of matching uncertain tuples present in different relations while performing join operation. Each join-pair is associated with a probability to indicate the likelihood that the two tuples are matched. We use the term **Probabilistic Join Queries** (PJQ). For join conditions over uncertain data, the result is generally not boolean, but probabilistic.

II. Related Work

The model for managing uncertain data is proposed in moving-object environments and in sensor networks. Recently, the Trio System has been proposed to handle such uncertainty. Another representation of data uncertainty is a "probabilistic database", where each tuple is associated with a probability value to indicate the confidence of its presence. Probabilistic databases have also been recently extended to semi-structured data and XML. Probabilistic queries are classified as value-based (return a single value) and entity-based (return a set of objects). Probabilistic join queries belong to the entity-based query class.

Aggregate value queries and nearest neighbor evaluation algorithms are presented. To our best knowledge, probabilistic join queries have not been addressed before. Also these works did not focus on the efficiency issues of probabilistic queries. Although examine the issues of query efficiency, their discussions are limited to range queries. There is a rich vein of work on interval joins, which are usually used to handle temporal and one-dimensional spatial data. Different efficient algorithms have been proposed, such as nested-loop join, partition-based join, and index-based join. Recently the idea of implementing interval joins on top of a relational database. All these algorithms do not utilize probability distributions within the bounds during the pruning process, and thus potentially retrieve many false candidates. We demonstrated how our ideas can be applied easily to enhance these existing interval join techniques.

III. Implementation

The INLJ (IndexedNestedLoopJoin) algorithm can recover I/O performance by organising the pages in a Tree structure. Let R and S denote the two relations that are being joined, and assume that R has fewer tuples than S. If neither join input has an index on the joining attribute, the indexed nested loops join algorithm first builds an index on the smaller input R.

The index is built by extracts the *key–pointerinformation* for each tuple. The key–pointer information is then spatially sorted based on the MBR. We can develop the efficient query join processing technique by the following sequence of operations.

a) Data Refinement

Take any Real-world Data which is possible to containing Uncertainity. Clean the data i.e. removing unnecessary data for Our project and Represent the most appropriate Data.

b) Formulating Range values using PDF function (Probability Density Function)

PDF summarizes how odds/probabilities are distributed among the events that can arise from a series of trials.By using PDF function we can replace the uncertainity values as Ranges.

c) Similarity matching between Uncertainty tuples

(By using probabilistic Joining Queries) Calculate the probability of joining the two uncertainity

tuples. Each join-pair is associated with a probability to indicate the likelihood that the two tuples are matched.

d) Removing Uncertainity by using INLJ

Although uncertainty tables can be used to improve the performance of page-based joins, they do not improve I/O performance, simply because the pages still have to be loaded in order to read the uncertainity tables.In INLJ we can use Interval Index.

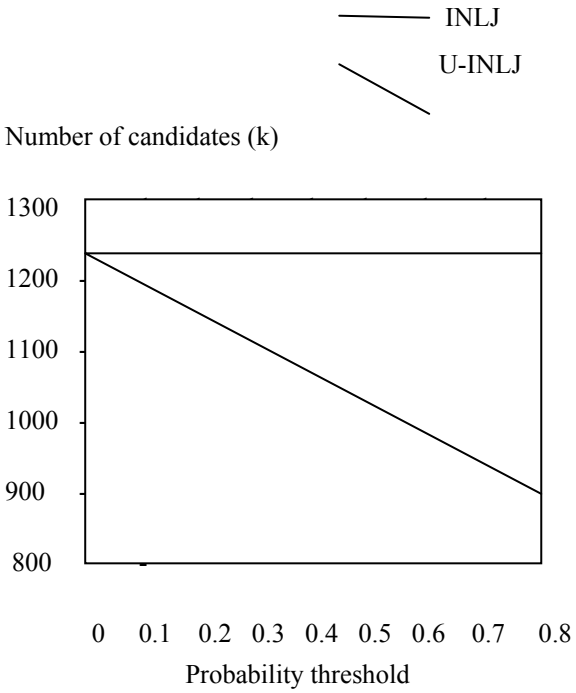
Conceptually, each tree node still has an uncertainty table, but now each uncertainty interval in a tree node becomes a Minimum Bounding Rectangle (MBR) that encloses all the uncertainty intervals stored in that MBR. Page-level pruning now operates on MBRs instead of uncertainty intervals.

e) Construct the Decision Tree for Query processing

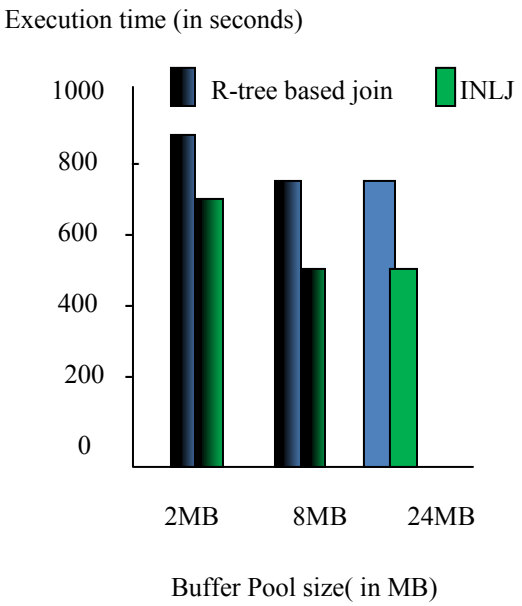
Splitting of an attribute depends on the attribute selection measures (Information gain, Gain ratio, Gini Index).Higher value of an attribute can be selected as splitting one. In this way the output Can be represented in the Decision Tree form by classifying the result into different classes.

IV. Performance

Index-Level Pruning The above problem can be alleviated by organizing intervals with an index. shows that both **INLJ** and **U-INLJ** have a much better performance in *Npair* than **BNLJ** and **U-BNLJ.**



In the above Graph, the comparison between page-level join algorithms with the Index-level join algorithms (INLJ and U-INLJ). In the Index-level join algorithms whenever the Threshold increases the output candidate pairs are Reduced. So, we can join the tables based on most similarity tuples. This leads to high performance in the Results, Next we can compare the Execution time between R-tree based join algorithm and INLJ(Indexed-Nested Loop Join) algorithm are compared in the below graph. The Horizontal row specifies the size of the Datasets and vertical row specifies the Execution time in seconds. In all different type of Datasets the Execution time of INLJ is Better than R-tree based join algorithm.



Finally I can prefer the Indexed-Nested Loop join algorithm as a Probability Threshold Joining Algorithm for Removing the Uncertainity while Joining of multiple table where the joining attribute has uncertain values. so, the Result of joining is efficient and we get the close to Exact Results.

V. Conclusion

Uncertainity management is the mounting topic in Data mining in recent times. In this paper we identify the situation of maintaining uncertain attributes present in the database relations. We suggest a method for getting better join processing of relations in requisites of I/O cost which are having uncertainity attributes present. In this paper we propose the implementation of INLJ, which is capably handle the uncertain values when compared to the earlier uncertainty handlings.

References références referencias

1. J. R. Quinlan, “Induction of decision trees,” Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
2. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993, ISBN 1-55860-238-0.
3. J. Chen and R. Cheng, “Efficient evaluation of imprecise location dependent queries,” in ICDE. Istanbul, Turkey: IEEE, 15-20 Apr. 2007, pp. 586–595.

4. M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in PAKDD, ser. Lecture Notes in Computer Science, vol. 3918. Singapore: Springer, 9–12 Apr. 2006,pp. 199–204.
5. R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter, "Efficient indexing methods for probabilistic threshold queries over uncertain data," in VLDB. Toronto, Canada: Morgan Kaufmann, 31 Aug.–3 Sept. 2004, pp. 876–887.
6. R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Querying imprecise data in moving object environments," IEEE Trans. Knowl. Data Eng.,vol. 16, no. 9, pp. 1112–1127, 2004.
7. T. M. Mitchell, Machine Learning. McGraw-Hill, 1997, ISBN 0070428077 .
8. R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *Proc. SIGMOD*,2003.
9. R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J.Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *Proc. VLDB*, 2004.
10. D. Zhang, V. Tsotras, andB.Seeger.Efficient temporal join processing using indicies. In *Proc. ICDE*, 2002.

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Approach to Quality Testing

By Drakshaveni G
MCA, BMSIT, Bangalore

Abstract - Time is the most important resource for any activity. Software development is no exception. With an increase in the competition in the market, it is very essential for the companies to release their products into the market at the earliest with good quality to earn profit. In order to develop the products early, companies implement various techniques like Rapid Application development and implement Agile methodologies like Scrum & Xtreme Programming to obtain tangible and useful features of the software at the earliest. It is just not sufficient to develop the product, it is more important to develop a quality product. But how does one know if the product is of good quality or not? This question is best answered by the Quality Assurance team which Tests the product end to end against requirements and standards. They conduct various kinds of tests and check the behavior of the system/ product in various conditions. If it passes all kinds of tests (which is dependent on the kind of the product or the system), then the QA team assures that the product is fit for use. Hence it is very important for the QA time to be given enough time to test the product/system before it is released into market. But it is difficult to judge as to how much time is required to test a product / system completely. An ideal answer would be "Years". Time is a major constraint in any software activity. Many software projects are time bound. Then how does one ensure that within a given time, the product/system is tested to the level where a confidence can be achieved? The answer is one needs to adopt faster means of testing. Many Testing techniques are available which in help achieving this objective. The major drawback of some of the proved techniques is these techniques are dependent on the kind of the system that is being tested. So what are the other ways available to save time for testing? The answer is in the question itself. One can save some time in optimizing the way of testing itself. If we carefully look into each testing activity, we realize that we can improve these to a great extend to provide maximum output. In this paper, it is intended to showcase one such approach.

Keywords : *Test Management, Defect tracking, ,SDLC MODEL, Test planning.*

GJCST-C Classification: *B.8*

Strictly as per the compliance and regulations of:

Approach to Quality Testing

Drakshaveni G

Abstract - Time is the most important resource for any activity. Software development is no exception. With an increase in the competition in the market, it is very essential for the companies to release their products into the market at the earliest with good quality to earn profit. In order to develop the products early, companies implement various techniques like Rapid Application development and implement Agile methodologies like Scrum & Xtreme Programming to obtain tangible and useful features of the software at the earliest. It is just not sufficient to develop the product, it is more important to develop a quality product. But how does one know if the product is of good quality or not? This question is best answered by the Quality Assurance team which Tests the product end to end against requirements and standards. They conduct various kinds of tests and check the behavior of the system/ product in various conditions. If it passes all kinds of tests (which is dependent on the kind of the product or the system), then the QA team assures that the product is fit for use. Hence it is very important for the QA time to be given enough time to test the product/system before it is released into market. But it is difficult to judge as to how much time is required to test a product / system completely. An ideal answer would be "Years". Time is a major constraint in any software activity. Many software projects are time bound. Then how does one ensure that within a given time, the product/system is tested to the level where a confidence can be achieved? The answer is one needs to adopt faster means of testing. Many Testing techniques are available which in help achieving this objective. The major drawback of some of the proved techniques is these techniques are dependent on the kind of the system that is being tested. So what are the other ways available to save time for testing? The answer is in the question itself. One can save some time in optimizing the way of testing itself. If we carefully look into each testing activity, we realize that we can improve these to a great extend to provide maximum output. In this paper, it is intended to showcase one such approach.

Keywords : Test Management, Defect tracking, ,SDLC MODEL, Test planning.

I. Introduction

Testing is an important part of SDLC[1]. It is in this phase that the quality of the product is assured. The product is tested in detail against the requirements and standards. This is not as simple as it sounds. There are many phases in a Test life cycle like Test planning, Design, execution etc. Fig 1.0 depicts a typical Testing Life cycle implemented in many projects.

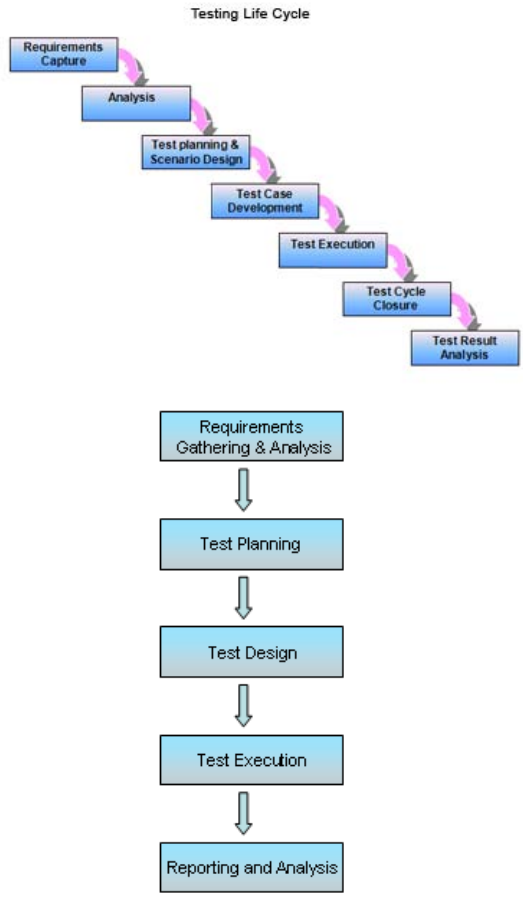


Fig. 1.0 :

When the Requirements are provided, the requirements are captured in a repository using a suitable convention which is easily understood by the stakeholders. Then it is analyzed for implementation. In this phase static testing techniques are implemented. Once the Analysis is complete, Planning comes into picture. This is a complex phases. It is in this phase that the testing approach is designed based on various constraints defined by the project. The success of Testing is largely dependent on the Test Planning & Design phase. Once the planning & Design phase is crossed, then comes the execution i.e. execution of test cases against the requirements and standards. After test execution, the results are analyzed and the quality is measured accordingly. If more test cases / Test scenarios (depending on which one is used) fail, then the product is sent back to the development team to fix the issues. After the issues are fixed, the product is tested again. This process continues till the there are very minimum or no

Author : Asst Professor, Dept. of MCA, BMSIT, Bangalore.

issues detected. If the product is tested and no issues are found, then the product is certified as fit for use.

In the above stated model, each phase has its on set of Inputs, processes and outputs. Any testing activity will be incomplete without supporting artifacts. Hence proper documentation is very important. It is not just sufficient that the product is tested for what it has to do. It is equally important to test a product for what it should not do as well. This is the reason that there will be many input conditions that need to be checked for. Hence testing is a time consuming activity. With Time being a constraint, how to complete so many activities in minimum time? What are the ways in which one can execute the testing activities faster? Let's see few of them.

II. Description

There are many ways through which the testing activities can be optimized. By implementation of these, lot of time can be saved which can be used to execute other activities which yield more result. Few of the most common and useful approaches in this regards are:

1. Transition from Manual to Automation Testing.
2. Usage of good Tools.
3. Forming a team of highly skilled persons.

Let's see how the above approaches can help save time and improve the testing quality.

a) Transition from Manual to Automation Testing

Transition from Manual to Automation is beneficial for the features of Automation. Few of them are Reliability, Repeatability, Fast, accuracy and Cost Reduction [2]. The Automation tools use Test Script language to develop scripts that imitate user actions.

The Test Script Language (like VB Script) is easy to learn and implement. These scripts can be designed to test the applications as a human would do. But the difference is they are faster. They can execute a task as many number of times as instructed without any errors.

As an example, if it takes 1 hour to execute 5 test cases, the Automation scripts can do it in 30 min (Depending on the way the script is developed and the application behaviour).Hence those features of the application which are very stable and require repeated testing, can be automated. The Manual testers can use the same time to test other parts of the application where more defects are likely t be found. Hence more testing can be done within available time. Hence the product can be tested better.

b) Usage of good Tools

Usage of tools like Test Management Tools [3], Defect Tracking Tools and Configuration tools, helps in achieving testing objective faster. These tools have templates/ formats which can be used readily. For example, Quality Center (a Test Management Tool) provides a single application for management of requirements, Test cases, execution and defect management. The reports generated from these tools give a single window data of the status of Testing. This can help the project manager decide the quality of the product. Without this, a subjective decision would be required to be taken regarding the quality of the product, which is a risk by itself. If data is required then it would take a considerable time to collect the same and put it in a format which is easily understood by the stakeholders. Many important Testing artifacts can be stored in such tools which provide the inputs at the execution time. In absence of such tools, the testers would have to refer documents from various sources and when the execution is at its peak, this takes lot of time.

The defect tracking tools help save some time by intimating the right persons at right time irrespective of their location. Example, if a defect is logged, the developer is intimated instantaneously through a mail or an alert, with all the required details as provided by the tester. In such scenarios, most of the times, the developer need not get in a discussion with the tester to know what the defect is all about. Also he can communicate with other teams instantaneously with the all required information. This saves lot of time.

c) Forming a Team of Highly Skilled Persons

This is a time tested approach [4]. Having team with people who know the application/product/system better are always beneficial. They can help the team members by guiding them in the following areas:

a. Implementation of special testing techniques like Orthogonal Array Technique, branch Testing etc
b. Usage of Tools in an optimum way.
c. Usage of right inputs to test the application
d. Understanding the domain and the purpose of each testing phase.
e. Understanding the requirements, Test processes, Test Reports etc.

In absence of a proper guide, one needs to spend more time in understanding the things. With time constraint, if a tester can't understand the full functionality, the functionality can't be tested properly. This is a risk. And if the defect is found at later stage, it would cost more to the project to fix the same. At the same time, there should be new people in the team who can test the product as an new user would do. They might have a new way to seeing a functionality which might help in testing it better and making it better. Testing is a creative activity which requires a good understanding of the application, domain, Testing principles and new ways to thinking. Hence the team should be a combination of people with such mentality so that optimum testing can be done and more defects can be detected at early state. This way rework can be reduced to some extent and hence the time.

III. Challenges

Till now we saw how to save time using Automation, Tools and right team, let us look at some of the challenges in implementing these approaches.

a) Automation

a. Availability of skilled professionals - the newer the technology, tools, methods, and domain, the smaller the pool of skilled professionals
b. Stability of implementation technology - the newer the technology, the lower the stability and the greater the need to balance the technology with other technologies and manual procedures
c. Stability and power of tools - the newer and more powerful the development tool, the smaller the pool of skilled professionals and the more unstable the tool functionality
d. Effectiveness of methods - what modeling, testing, version control, and design methods are going to be used, and how effective, efficient, and proven are they
e. Domain expertise - are skilled professionals available in the various domains, including business and technology
f. Good Automation Tools are costly. It might not be possible for all projects to afford Automation as per the budget defined for the project.
g. Availability of resources trained in usage & implementation of Automation tools. It is not easy to get people who are well versed in usage of these tools. If they are, then they need to be paid more. This again puts constraint on the project budget.
h. The new resources involved in automation, need to be trained in usage of the tools which itself consumes time & money.
i. Automation can be implemented for those applications which are stable. If the application is not stable, then automaton can't be implemented effectively.

b) Tools

a. Good Test Management, Defect Tracking & Configuration Management Tools are costly.
b. Availability of resources trained in usage & implementation of Automation tools. It is not easy to get people who are well versed in usage of these tools. If they are, then they need to be paid more. This again puts constraint on the project budget.
c. Team members need to be trained in usage of the tools which itself consumes time & money.

c) Forming a Highly Skilled Team

a. Difficult to judge the expertise in the required area. Complete knowledge cannot be checked in just a couple of interviews.
b. Availability of experts employment of such people is bit difficult and as they look for high salaries which puts a constraint on the budget of the project.

IV. Proposed Solution

Every problem has a solution. Most of the times, the solution to a problem is very much available but with some search. Having seen the difficulties in implementing the time saving techniques, let us see if we can save some time in the testing process itself. For this, we need to understand what a tester does. Here are some of the steps which most of the testers follow:

1. Understand the requirements and develop the test scenario.
2. Derive useful test cases out of these test scenarios.
3. Execute the test cases and verify & validate the features of the system.
4. If there is any discrepancy observed, confirm it and log a defect.
5. Once the defect is fixed, retest and closed the defect if it is fixed.
6. Capture the results of test case execution and use it further for metrics and analysis.

It is possible to help the testers save some time in step 4 stated above. Normally when a discrepancy is observed, a good tester captures the basic information like how the defect was detected, what was the defect, screen shot of the system (if possible) and associated details. Along with execution of the test case, the tester needs to spend some time recreating the defect to confirm the behavior. Lets take an example.

Suppose a tester is executing a test case for a web based application. That test case has 12 steps and it takes 20 minutes to execute the test case. When the tester finds that step 6 is failing, he logs a defect in the Defect Tracking Tool. He first needs to capture the screenshot of the application to let the developer know as to what is the defect. Then he needs to write the steps to reproduce the defect. It should be elaborate enough for the developer to understand the steps of re-creation the defect. Then he needs to provide some information like Version of the application tested, date & time at which the defect was detected, actual result of the step and expected result etc. On an average depending on the expertise of usage f the defect management tool and the application knowledge, it can take somewhere around 10 to 15 minutes to log the defect with all the required details. This is a considerable amount of time when compared to the test case execution time. This does not stop here. As a good practice, all steps of the test case should be executed to be sure that there are no defect goes detected. If there is another defect detected in step 9 of the same test case, again some considerable time and effort should be spent for defect logging. In such a process, the time spent on defect logging is more than the test case execution time itself.

Here is a method which can help reduce the defect logging time so that the tester can use the same time in execution of other important test cases. This is a solution that should be incorporated in the tool form which the test case is being executed.

1. Make sure that a test case has only on e scenario.
2. If a step in the test case fails, the tool should capture the following details and store it in a repository

a. Screenshot of the application.
b. Step description and expected result of all the previous steps in the 'steps to reproduce the defect section' along with expected result of the step that failed so that the developer can fix the code accordingly.
c. Time stamp information. (I.e. Date and time on which the defect was observed)
d. Version of the application that is being tested (Build number). This information can b entered in some file from where the tool can pick this up.
e. The test data that was used while execution of the test case. This information can be captured by making sure that each test case has a unique test data available at the time of execution. This information can be stored in a separate file with Test Data & Test case mapping so that the tool picks up this information when required.
f. Severity of the issue. This should be defined during the test case design for each step. The test case designer should know the impact of failure of the step that is being executed, on the application.

The above points are just few of the information that can be captured. Other mandatory information can also be captured.

3. Once all the above information is available, the tool should format the information in a way which is easily understood by all stakeholders and should present it to the tester who wishes to log the defect (can be in form of a pop-up window).
4. Depending on the time availability, the tester can review the captured details before submitting the defect or can add extra information, if required.

This way the process of defect logging which takes 10 to 15 minutes, can be reduced to just couple of minutes. If a tester executes 15 such test cases, and on an average he detects 4 defects, he can easily save 30 minutes (approximately) which can be used for adhoc testing or execution of some other important test case. If there are 5 testers, who perform a similar job, in the above scenario, 2.5 hours can be saved.

V. Advantages of the Above Method

1. Time saving: The process of defect logging becomes faster. Instead of taking 10 minutes, the defect can be logged in a couple of minutes.
2. Since the information that needs to be captured becomes standardized, chances of missing out details are reduced to a great extent.
3. Since the information is captured immediately the timestamp information is more accurate which helps the developers to check the logs as to what exactly happened at that instance of time.
4. This method necessitates that the test case be understandable, logical and complete. This helps the new resources of the team to execute the test case easily (which otherwise would have required more time to execute the same). So considerable time is saved and the tester would feel more comfortable.
5. Along with this, it also necessitates defining the impact f failure of each step, which prevents subjective decision on the assigning the severity to the defect which is to be logged for the failure of that step.
6. Depending on the number of High or Medium or Low severity steps, the severity of the test case can be defined. Depending on the number of High or Medium or Low test cases, the Severity of the Test Set can be determined. This is helpful when the testers do not get sufficient time to execute the complete regression cycle and need to prioritize the test case execution based on the severity of the Test Set / Test Case

The advantages are not limited to the above. The availability of extra time itself defined the usefulness of this method.

VI. Challenges in Implanting This Method

Although there are many advantages of this method, there are a couple of drawbacks too.

1. This method calls for use of a tool which can capture all the required information and translate it into useful information as contained in a defect log. This involves time, effort and cost. The project might not have so much time or money to spend on development of such a specialized tool.
2. This method requires defining more details like severity of each step of a test case, capturing unique test data for each test case etc which requires more time and effort.

VII. Where Can This Method Be Useful?

The approach proposed in this paper can be added in any Test Management Tool which has Test case execution and Defect Management features. This approach will be helpful in testing projects where Defect Finding Rate is high and more test cases need to be executed in a shorter time.

VIII. Conclusion

Testers should be given sufficient time to test the system/ product/application in order to get a good quality product. When there is a time crunch, projects should optimize the way testing is carried out. The method stated in this paper can help the Testing project irrespective of the SDLC model followed, to save reasonable time which can be used for performing other important testing activities.

Acknowledgement

We wish to thank all the authors for the providing the informational support

References Références Referencias

1. Testing Common Body of Knowledge.
2. Software Engineering-Ian Somerville.
3. A very comprehensive book on the testing techniques.
4. Johnson, M., Ho, C-w, Maximilien, M., and Williams, L., Incorporating Performance Testing in Test-Driven Development, IEEE Software.

This page is intentionally left blank

Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 12 Issue 12 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch

By D.Pramodh Krishna, A.Senguttuvan & T.Swarna Latha
Sree Vidyanikthan Engg.Coll., A.Rangempet, Tirupati, A.P, India

Abstract - The paper is about the clustering on large numeric data sets using hierarchical method. In this BIRCH approach is used, to reduce the amount of data, for this a hierarchical clustering method was applied to pre-process the dataset. Now a day's web information plays a prominent role in the web technology, large amount of data is consumed to communicate, but some with intruders there is loss of data or may changes occur in the interaction, so to recognize intruders they detect to build an intrusion detection system for this a hierarchical approach is used to classify network traffic data accurately. Hierarchical clustering is performed By taking network as an example. The clustering method could produce high quality dataset with far less instances that sufficiently represent all of the instances in the original dataset.

Keywords : *Hierarchical clustering, support vector machine,data mining,KDD cup.*

GJCST-C Classification: *H.3.3*

Strictly as per the compliance and regulations of:

Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch

D.Pramodh Krishna [α], A.Senguttuvan [σ] & T.Swarna Latha [ρ]

Abstract - The paper is about the clustering on large numeric data sets using hierarchical method. In this BIRCH approach is used, to reduce the amount of data, for this a hierarchical clustering method was applied to pre-process the dataset. Now a day's web information plays a prominent role in the web technology, large amount of data is consumed to communicate, but some with intruders there is loss of data or may changes occur in the interaction, so to recognize intruders they detect to build an intrusion detection system for this a hierarchical approach is used to classify network traffic data accurately. Hierarchical clustering is performed By taking network as an example. The clustering method could produce high quality dataset with far less instances that sufficiently represent all of the instances in the original dataset.

Keywords : Hierarchical clustering, support vector machine,data mining,KDD cup.

I. Introduction

Data clustering is an important technique for exploratory data analysis and has been studied for several years. It has been shown to be Useful in many practical domains such as data classification. There has been a growing emphasis on analysis of very large data sets to discover the useful patterns. This is called data mining and clustering is regarded as a particular branch. So, as the data set size increases they do not scale up well in terms of memory requirement. Hence an efficient and scalable data clustering method is proposed based on a new in memory data structure called CF Tree which serve as in in-memory summary of the data distribution.

We have implemented in a system called BIRCH (Balanced Iterative Reducing And Clustering Using Hierarchies) and studied its performance extensively in terms of memory requirement and scalability.

The SVM technique is unable to operate at such a large dataset due to system failures caused by insufficient memory, or may take too long to finish the training. Since this study used the KDD Cup 1999 dataset, to reduce the amount of data, a hierarchical clustering method was applied to pre-process the dataset before SVM training. The clustering method could produce high quality dataset with far less instances that sufficiently represent all of the instances in the original dataset.

This study proposed an SVM-based intrusion detection system based on a hierarchical clustering algorithm to pre-process the KDD Cup 1999 dataset before SVM training. The hierarchical clustering algorithm was used to provide a high quality, abstracted, and reduced dataset for the SVM training, instead of the originally enormous dataset. Thus, the system could greatly shorten the training time, and also achieve better detection performance in the resultant SVM classifier.

This study proposed an SVM-based intrusion detection system, which combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique. The hierarchical clustering algorithm provided the SVM with fewer, abstracted, and higher-qualified training instances that are derived from the KDD Cup 1999 training set. It was able to greatly shorten the training time, but also improve the performance of resultant SVM. The simple feature selection procedure was applied to eliminate unimportant features from the training set so the obtained SVM model could classify the network traffic data more accurately. The famous KDD Cup 1999 dataset was used to evaluate the proposed system. Compared with other intrusion detection systems that are based on the same dataset, this system showed better performance in the detection of DoS and Probe attacks, and the be set performance in overall accuracy.

II. Background

a) Data transformation and scaling

SVM requires each data point to be represented as a vector of real numbers. Hence, every non-numerical attribute has to be transformed into numerical data first. The method is simply by replacing the values of the categorical attributes with numeric values. For example, the protocol type attribute in KDD Cup 1999, thus, the value tcp is changed with 0, udp with 1, and icmp with 2. The important step after transformation is scaling. Data scaling can avoid attributes with greater values dominating those attributes with smaller values, and also avoid numerical problems in computation. In this paper, each attribute is called with linear scaling to

Author α : Assistant Professor, Dept. of CSE, Sree Vidyanikthan Engg. Coll., A.Rangempet, Tirupati, A.P, INDIA-517 102.
E-mail : pramodhkrishna.d@gmail.com
Author σ : Professor, Dept. of CSE, Sree Vidyanikthan Engg. Coll., A.Rangempet, Tirupati, A.P, INDIA-517 102.
E-mail : asenguttuvan@rediffmail.com
Author ρ : M.Tech Student, Dept of CSE, Sree Vidyanikethan Engg. Coll., A.Rangempet , Tirupati , A.P, INDIA-517 102.
E-mail : swarnalatha514@gmail.com

the range of [0, 1] by dividing every attribute value by its own maximum value.

b) Clustering feature (CF)

The concept of a clustering feature (CF) tree is at the core of BIRCH's incremental clustering algorithm. Nodes in the CF tree are composed of clustering features. A CF is a triplet, which summarizes the information of a cluster.

c) Defining the CF trees

Given n d- dimensional data points in a cluster {xi}, where i = 1, 2, . . ., n, the clustering feature (CF) of the cluster is a 3-tuple, denoted as CF = (n, LS, SS), where n is the number of data points in the cluster, LS is the linear sum of the data points, i.e., $\sum_{i=1}^{n} x_i$, and SS is the square sum of the data points, i.e., $\sum_{i=1}^{n} x_i^2$.

III. Related Work

a) Theorem (CF addition theorem)

Assume that CF1 = (n1, LS1, SS1) and CF2 = (n2, LS2, SS2) are the CFs of two disjoint clusters. Then the CF of the new cluster, as formed by merging the two disjoint clusters is

$CF_1 + CF_2 = (n1+n2, LS1+ LS_2, SS1+SS_2)$

For example, suppose there are three points (2, 3), (4, 5), (5, 6) in cluster C1, then the CF of C1 is

$CF_1 = \{3, (2+4+5, 3+5+6), (2^2+4^2+5^2, 3^2+5^2+6^2)\}$

$= \{3,(11,14),(45,70)\}$

Suppose that there is another cluster C2 with CF2 = {4, (40, 42), (100, 101)}. Then the CF of the new cluster formed by merging cluster C1 and C2 are

$CF_3 = \{3+4, (11+40, 14+42), (45+100, 70+101)\}$

By Definition and Theorem, the CFs of clusters can be stored and calculated incrementally and accurately as clusters are merged. Based on the information stored in CF, the centroid C and radius R of a cluster can be easily computed. The definitions of C and R of a cluster are given as follows. Given n d-dimensional data points, say {xi} and i = 1, 2. . . n, in a cluster:

$$\text{the centroid } C = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \text{and}$$

$$\text{the radius } R = \frac{\sum_{i=1}^{n} \|x_i - C\|^2}{n}.$$

Where, R denotes the average distance of all member points to the centroid. As mentioned earlier, CF stores only the abstracted data point, i.e., statistically summary of data points that belong to the same cluster. After a data point is added into a cluster, the detail information of the data point itself is missing. Therefore, this approach can save space significantly for densely packed data.

b) CF tree

A CF tree is a height-balanced tree with two parameters, branching factor B and radius threshold T. Each non-leaf node in a CF tree contains the most B entries of the form (CF_i, $child_i$), where1≤ i≤ B and $child_i$ is a pointer to its i^{th} child node, and CF_i is the CF of a cluster pointed by the child i. An example of a CF tree with height h = 3 is shown in Fig. 1. Each node represents a cluster made up of sub-clusters, which represents its entries. It is different from a non-leaf node is that a leaf node has no pointer to link to other nodes, and contains at most B entries. The CF at any particular node contains information for all data points in that node's sub-trees. A leaf node must satisfy the threshold requirement, that is, every entry in every leaf node must have its radiuses less than threshold T.

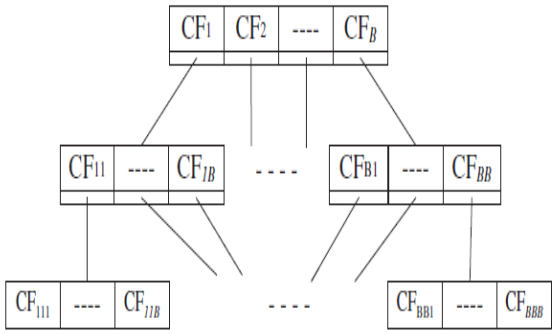


Fig. 1 : A CF Tree with height h=3

A CF tree is a compact representation of a dataset, each entry in a leaf node represents a cluster that absorbs many data points within its radius of T or less.

A CF tree can be built dynamically as new data points are inserted. The insertion procedure is similar to that of a B+-tree to insert a new data to its correct position in the sorting algorithm. The insertion procedure of CF tree has the following steps.

1. Identify the appropriate leaf
2. Modify the leaf
3. Modify entries on the path to the leaf

S no.	String Value	Equivalent Numerical Value
1	Tcp	0
2	Udp	1
3	Icmp	2
4	http	0
5	domain_u	1
6	Auth	2
7	Smtp	3
8	finger	4
9	telnet	5
10	ecr_i	6

Table 1 : Replacement of the string values

IV. Experimental Results

Here the scenario is mainly responsible in the task of clustering. The process of data clustering is performed using the WEKA 3.7 tool's Hierarchical clustered. The testing in this test case can be done in the stage of mentioning the parameters of the clustering such as distance parameter i.e. Euclidean distance, link type such as centroid and number of clusters value. The testing strategies here mainly the setting of those parameters regarding the hierarchical clustered in the WEKA tool. The clustered output can be visualized on the results for the test results.

This test case deals with the training of the clustered dataset. Both the train set and the test set are givenas input for classification and regression/prediction of the accuracy values with the following option lists.

1	2	3	4	5	6
0	0	0	0	0.320175	0.188728
0	0	0	0	0.087719	0.018894
0	0	0	0	0.412281	0.005072
0	0	0	0	0.399123	0.00844
0	0	0	0	0.425439	0.001964

Table 2 : Dataset values in .csv format

This is the first step in the process of data transformation. Here the KDD data set which is in the text format is changed in to the comma separated values i.e., from .txt to .csv. The reason for the format change is the incompatibility of the tool or the language that we use in the further processing steps.

After the format conversion the data set can be view in the office excel as it support csv format viewing in windows environment. Here the data present in the KDD data set before transformation is seen in the excel sheet i.e. in the csv format.

The java code has been written for the data transformation, all the non integer values should be transformed to integer values in the data set. So that program is executed in this screen shot.

The data set has been transformed i.e. all the variables present in the data set has been transformed to the integer values which is in the text format. The data set which is transformed is present in the text format we have to convert that in to .csv format.

After the data transformation the data scaling has to be performed. So in order to perform the data scaling we use the WEKA tool here and load the transformed file into WEKA pre- processing task.

a) Data Clustering

WEKA provides a wide range of clusterers for the process of data clustering. Here we use the hierarchical clusterer for our clustering process which has been clearly mentioned in the screen shot. The testing regarding the clustering has been discussed in the test case 3 in the system testing chapter.

Once the type of clustering has been selected now the query tab is double clicked for the setting the hierarchical clustering parameters. The paramaters of hierarchical clustering that we see here are,

1. Distance Function
2. Link Type
3. NumClusters

The process of clustering is continued once these parameters are correctly set by the user.

The clustered output shown in WEKA mainly show the percentage values of the clusters made by it and the cluster values. The percentage mainly show the amount of dataset those are regarding a single property i.e. based on a single attack present in the considered test set. The output of the clustered is also compared to this test set values in the further procedures such as data training etc.

This is an additional option that is being provided in WEKA tool for visualizing the clustered output in the form of graphs with different color representations. To select this option right click on the result list that is being displayed on the left of the WEKA Explorer sorted with their timing status in hh:mm:ss

This is how you visualize the clustered output in the form of a graph. Here we come across a slider labelled Jitter, which is a random displacement given to all points in the plot. Dragging it to the right increases the amount of jitter, this is useful for spotting concentrations of points. Without jitter, a million instances at the same point would look no different to just a single lonely instance. Here we have the option save to save our clustered output in the external memory to use that in our training procedure.

Once the clustering has been visualized and saved, the clustered output is saved in the external memory with a new column known as cluster of the type nominal which clearly mentions the complete dependency of the particular row with the respective cluster.

This is treated as the clustering output in dataset format which is in turn given as the input for the process of data training.

This is the visualization graphs for each of the column values that is being seen in the KDD dataset. The graphs are drawn according to the level of values present in the complete column. This helps the user in accessing the column present in the particular column in the dataset.

```
=== Run information ===
Scheme:
weka.clusterers.HierarchicalClusterer -N
2 -L SINGLE -P -A
"weka.core.EuclideanDistance -R first-
last"
Relation:     testsample-1_clustered
Instances:    1000
Attributes:   44
```

```
                Instance_number
                1
                2
                3
                4
                5
                6
Cluster
Test mode:    evaluate on training data
=== Clustering model (full training set)
===
Cluster 1
Time taken to build model (full training
data) : 2.81 seconds
=== Model and evaluation on training set
===
Clustered Instances
0         1 (  0%)
1       999 (100%)
```

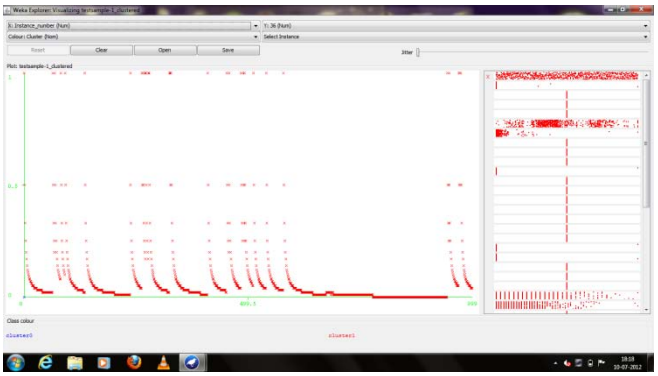


Fig. A : Visualization of the clustered data

V. Conclusion

In this paper, we have proposed an hierarchical clustering approach using BIRCH algorithm it proposed an SVM-based network intrusion detection system with BIRCH hierarchical clustering for data pre-processing. The BIRCH hierarchical clustering could provide highly qualified, abstracted and reduced datasets, instead of original large dataset, to the SVM training. Thus, in addition to a significant reduction of the training time, the resultant SVM classifiers showed better performance than the SVM classifiers using the originally redundant dataset.

However, in terms of accuracy, the proposed system could obtain the best performance. Some new attack instances in the test dataset, which never appeared in training, could also be detected by this system.

References Références Referencias

1. Abraham, A., Grosan, C., & Martin-Vide, C. (2007).Evolutionary design of intrusion detection programs. International Journal of Network Security, 4(3), 328–339.
2. Bouzida, Y., & Cuppens, F. (2006). Neural networks vs. decision trees for intrusion detection. <http://www.rennes.enstbretagne.fr/~fcuppens/articles/monam06. pdf>.
3. Guha, S., Rastogi, R., & Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In Proceedings of the international conference on data engineering (ICDE'99) (pp. 512–521).
4. Guha, S., Rastogi, R., & Shim, K. (1998). Cure: An efficient clustering algorithm for large databases. In Proceedings of the ACM SIGMOD (SIGMOD'98) (pp. 73–84).
5. Hsu, C. -W., Chang, C. -C., & Lin, C. -J., (xxxx). A practical guide to support vector classification. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
6. Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: A hierarchical clustering algorithm using dynamic modelling. Computer, 32, 68–75.
7. KDDCup, (1999). Intrusion detection data set. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
8. Khan, L., Awad, M., & Thuraisingham, B. (2007). A new intrusion detection system using support vector machines and hierarchical clustering. The International Journal on Very Large Data Bases, 16(4), 507–521.

Global Journals Inc. (US) Guidelines Handbook 2012

www.GlobalJournals.org

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

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

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

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

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

AUXILIARY MEMBERSHIPS

ANNUAL MEMBER

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

PAPER PUBLICATION

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

Process of submission of Research Paper

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

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

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

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

(II) Choose corresponding Journal.

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

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

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

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

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

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

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

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

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

Scope

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

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

2. ETHICAL GUIDELINES

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

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

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

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

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

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

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

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

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

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

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

Please mention proper reference and appropriate acknowledgements wherever expected.

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

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

3. SUBMISSION OF MANUSCRIPTS

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

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

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

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

4. MANUSCRIPT'S CATEGORY

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

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

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

Research articles: These are handled with small investigation and applications

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

5.STRUCTURE AND FORMAT OF MANUSCRIPT

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

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

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

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

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

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

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

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

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

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

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

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

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

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

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

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

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

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

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. **Pick a good study spot:** To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

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

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

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

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. **Think technically:** Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be

sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

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

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to

shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic

principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

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

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

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

Index

T

U

V

W

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