THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Scrutinize Source Code Using Metric and Suffix Array Based Token Technique To Unmask The Code Clones In Multiple Languages.

Manjit Kaur
Assistant Professor
Computer Science and Engineering
Lovely Professional University.
Phagwara,Punjab, India

Abstract— **In software industries, a technique called software cloning has come into existence. Software cloning has various broad aspects, out of them; the shadow of light is thrown on one of the aspect called code cloning. In code cloning, some significant quantity of code as desired by the user is taken from some pre-existing code and copied into some another code. In short, it is a kind of copying or pasting of code where some desired code is copied from one source and pasted into another source. The code in which pasting is done is called the replica of original code. In other words, the code which contains the replicated code is called the clone. [30]This research work proposes a Hybrid techique which is the goodmix of Metric based technique & Token based technique and also works on multiple languages to unmask the clones. In the Token based technique, Suffix array based string matching algorithm has been used to unmask Type1, Type2 and Type 3 clones.**

Keywords—clone, code smells, hybrid, match detection, plagiarism

I. INTRODUCTION

Modern world is the era of science and technology due to which many new technologies have been introduced at different times. Internet is one of the results of this. On one side, the introduction of internet and advancements related to it, opens new opportunities for people but on the other side, these new advancements of internet make people tedious and weary[30]. People now are too much dependent on internet that they start copying and pasting the things to accomplish their tasks instead of learn them or grasp them in mind. They are not brainstorming their minds. According to software and technology terms, this duplicity achieved by making copying or pasting of things which could lead to lack of originality is referred to as 'cloning' [30]. This copying of codes will lead to copyright infringements of original work of the authorized persons. This has been a question from many years in the mind of researchers who dedicated their research in this field of cloning that whether cloning is a legal or an illegal exercise. Then the answer to this is cloning is not illegal if it has been done with the permission of authorized person. For instance, reusing the code in the software development is an efficient method to reuse the design or requirements; which will save lot of time and cut costs in developing large software products. So to reuse the required things, the owner must be asked about copying his code [30]. In another case also, cloning can not be illegal if the content present on some website or on some journal is for free, but that too leads to the absence of one's originality and creativity. Cloning imparts many cons despite of having many pros, which can be easily justifiable to the fact that "Every rose has its thorns". [30]

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

II. RELATED WORK

There has been various proposed and review papers which are advocated by various researchers who worked in this area of cloning and implemented their proposed methodologies using various techniques and tools. [30]

Starting from the year 2006, there was a technique proposed by Rachel Edita Roxas [1] which automatically detects the cloning in students program with the help of the tool named Jplag. In 2007, Stefan Bellon [2] laid out a comparison and evaluation on clone detection tools which says that token and text based techniques works similarly and Marcelo & Baxter's [27] AST tool has higher precision but in parallel, it exhibits higher costs. In year 2008, Cory J. kapser [3] laid emphasis on negative characteristics of clone which are often called code smells. In 2009, a popular review paper stated by James R. Cordy & Chanchal K. Roy [4] was solely based on clone detection techniques and tools. Another technique was given by Tung Thanh Nguyens [5] in the same year whose research work was based on scalable and incremental clone detection with the assistance of Clemanx tool. [30]

In 2010, a technique was proposed by Perumal.A [6]. This technique made use of metrics for extracting similarity in the software clones which had been already detected. In 2011, a technique was proposed by G.Anil Kumar, who used light weight clone detection technique, which is also called hybrid technique as it is the combination of two techniques, metrics and text based along with refactoring support.[30]

In 2012, a technique was proposed by the Saif Ur Rehman & Kamran Khan [7] which constituted of LSC-Miner tool to detect clones. In the same year, another technique was semantic clone detection which was based on JSC Tracker [8] tool and it worked only on java code. Clone detection was further enhanced by Deepak Sethi [9] in the same year. He used a technique of data sets of data mining. In this proposed method, Solid SDD tool has been used that provides a better visualization of clone detection. Other attempt was made by the Rupinder Kaur & Prabhjot Kaur [12] in the same year which focuses on comparison of two token based techniques known by the name, CC-Finder[21] [24] and PMD to improve performance and efficiency. They concluded that there exists no such tool which detects all the clones efficiently; in fact, each tool has its own strengths and weaknesses which become the factor of its usefulness in detecting clones. [30]

In 2013, Dhavleesh Rattan [13] came up with review paper which was purely related to the various clone detection techniques and methodology which has to be followed while locating clones, pros & cons related to it, etc. In the same year, the other technique was proposed by Kanika Raheja [14]. The technique used here was metric based technique which worked only for Java language.[30]

During 2014, another technique proposed by Harpreet Kaur & Rupinder Kaur [15] was based on metrics. This technique detects clone not only in programming languages but also in web applications. In one of the other clone detection, clones were detected by neural networks and SIMCAD. As data mining seems to be a new emerging area for clone detection another technique was proposed by D. Gayathri Devi [16]. This technique follows an algorithm which detects clones for control structures such as for, while, do statements.[30]

In 2015, one of the latest proposals was given by the Manpreet Kaur and Madan Lal [17], which was a hybrid technique by the combination of the metric based & text based technique. In the present year 2016, Shashank Prabhakar [18] proposed a technique to detect clones which emphasis on detecting Type-1 and Type-2 clones.

In 2017, a technique proposed by Ryota Ami et.al who proposes the Tree based and Token based approach to detect Type1, Type2 and Type3 clones. Another technique was proposed by Ashish N.Runwal et.al in the same year who proposed the Code Clone Detection based on Logical Similarity. Their Proposed system presents an algorithm for clone detection based on comparing parts of abstract syntax tree(AST)of

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

programs and finding semantic coding styles.[34]In the same year Hui-Hui Wei et.al proposed a paper that address the software functional clone detection problem by learning supervised deep features.[35].A technique proposed by Seulbae Kim et called VUDDY, an approach for the scalable detection of vulnerable code clones, which is capable of detecting security vulnerabilities in large software programs efficiently and accurately.[36]One of the Design Code Clone Detection System technique proposed by Jasmandeep Kaur uses Optimal and Intelligence Technique based on Software Engineering.[38]

In 2018, a technique was advocated by Tijana Vislavski et.al who uses LICCA tool which is a cross-language code clone detection tool. Another technique proposed in the same year by Hongfa Xue et.al.They present an novel framework, Clone-Slicer, foridentifying domain-specific binary code clones (e.g., pointer-relatedcode) through program slicing.[37]

In 2019, a technique proposed by Debajyoti Mondal et.al, an visual analytics system, *Clone-World*, which leverages big data visualization approach to manage code clones in large software systems.[39]

Hence, from the literature survey, it can be analyzed and concluded that there are still various active areas in the cloning that can be further worked upon in the future by the new researchers [30] which includes code maintenance, clone removal,etc.

III. BACKGROUND

A. Types of clone

There are four types of clones namely: Exact clone, Renamed or Parameterized clone, Near-miss clone and Semantic clone. These clones are called Type 1 clone, Type 2 clone, Type 3 clone and Type 4 clone respectively.

a) Type 1 clone (Exact clone)

In Type 1 clone, the cloned code is exactly same as that of the original code. In this type of clone, the only difference comes in the form of comments, blank spaces etc; which makes the cloned code different from original code.Text based approach is appropriate to detect this type of clone.

Table 1. Type1 clone

CODE	**CLONE**
int mul(int c[],int p) {	int mul(int c[],int p){
int e=0; //mul	int e=0; //multiplication
for(int u=0; u<p; u++) {	for(int u=0; u<p; u++){
e=e*c[u];	e=e*c [u];
}	}
return e;	return e;
}	}

b) Type 2 clone(Renamed/parameterized clone)

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

In Type 2 clone,the cloned code contains the renaming of variables, literals, etc w.r.t the original code. This kind of clone is detected by Text based technique & Token based technique.

Table 2. Type 2 clone

CODE	CLONE

int mul(int c[],int p){
int e=0; //mul
for(int j=0; j<p; j++){
e=e* c [j];
}
return e;
}
	int multiply(int q[],int f){
	int i=0; //multiply
	for(int g=0; g<f; g++){
	i=i*q [g];
	}
	return i;
	}

c) Type 3(Near-miss clone)

In Type 3 clone, there is a insertion, deletion and modification of statements in the cloned code w.r.t the original code.Text based, Token based,Hybrid, Tree based and Metric based approach is best for unmasking such clones.

Table 3. Type 3 clone

CODE	CLONE

int multiply(int c[], int r){
int mul=0; //mul
for(int h=0; h<r; h++) {
mul=mul* c [h];
}
return mul; }
	int multiply(int x[], int e)
	{
	int y=0; //mul
	for(int z=0; z<e) {
	y=y* x [z];
	z++; }
	return y; }

c) Type 4 clone(Semantic clone)

In Type 4 clone, the cloned code is same as that of the original code in terms of functionality or behaviour but differ in the syntax and implementation.

Table 4. Type 4 clone

CODE	CLONE

int add(int n[],int m){
int c=0;
for(int d=0;d<m; d++){
c=c+ n [d];
}
return c;
}
	int add(int n[],int p){
	if(p==1)
	return n[p-1];
	else
	return n[p-1]+add[n,p-1];
	}

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

IV.CLONE DETECTION

A. Definition

It is the process of finding or detecting clones in code. It is used to find clone pairs in programs based on similarity. There are various advantages of finding the clones so as to detect the bugs at the earlier stage, plagiarism detection, etc. [30]

B. Steps in clone detection

There are various steps involved in the detection of clones.[30]

- Pre-processing
- Transform
- Match Detection
- Formatting
- Post Processing
- Aggregation

C. Techniques in clone detection

The various techniques involved in the clone detection are:[30]

- Text Based- This technique compares two code fragments line by line [4][13][14][30]. This technique is only for Type 1 clones. It doesn't consider any renaming of variables and any syntactical or semantically changes. It provides high accuracy. But it is not highly efficient to detect any other kinds of clones.

- Graph Based-This technique uses program dependency graph (PDG). It is good for detecting semantically similar clones. Semantically similar clones are those clones which are syntactically different but show similar behavior or perform same function. In other words, it can detect Type 3 and Type 4 clones efficiently. PDG are directed graph which determines two types of dependencies namely data dependency and control dependency which exists between statements of the source code [30].

- Metric Based-It is a straight forward technique. There are four types of metrics namely class, layout, method and control [4][13][14]. All these types follow a different metrics. Metric based approach is more scalable technique and gives accurate results for large software systems. It contains structural information only. So it is good for finding syntactic clones i.e. Type 1, Type 2 and Type 3 clones [30].

- Token Based-In this technique, there is a formation of lexical tokens [4][13][14]. It is good for detecting Type 1 and Type 2 clones. It gives fast response and is considered to be more efficient as compared to text based but also gives many false positives. It extracts tokens out of the source code with the help of lexical analysis and based on this token sequence is formed. There is a method called "functor" that maintains the order of tokens [30].

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- Tree Based-This technique is based on Abstract Syntax Tree (AST) which is obtained after converting the source fragments into some intermediate form. It is efficient for detecting Type 1, Type 2 and Type 3 clones [4][13][14]. It is a heavy weight technique and requires a sub tree comparison [30].

- Hybrid-This technique is the combination or the union of various other techniques like Tree, Text, Token, Metric and Graph [4][13][14][30] for the detection of code clones .

V.**PROPOSED WORK**

A proposed technique is a Hybrid technique which is the combination of Metric based and Token based technique. A proposed technique detects clones in multiple languages such as Java and Asp.net language. Under the Metric based technique a total of 20 Metrics has been computed from the source code and Threshold value has been set to be 3 in order to calculate Potential clones. This implies that if the Threshold value is equal to or greater than 3, only then Potential clones are exists in the code, otherwise not. For calculating Actual clones, Suffix Array Matching Algorithm has been applied on Tokenized code.

Suffix array computes repeated token sequences with the time complexity of O (N)[31]. Moreover it is a simpler algorithm. The space consumption while constructing the suffix array is five times lesser than other algorithms such as Suffix tree, etc [32].

VI. IMPLEMENTATION

The implementation of the tool for proposed technique has been carried out in Java language. The software being used for carried out the implementation of tool is Netbeans8.1. The tool detects the clones in multiple languages such as Java language and Asp.net language. This technique can detect Type1, Type2 and Type 3 clones.

In Figure 1, Metrics are calculated from 2 source files. Further these Metrics values are compared with each other in order to find Potential clones. Once Potential clones are detected, Tokenization of Source code has been carried on, where source code is converted into Tokenized strings as in Figure 2. After that, Suffix Array matching algorithm has applied on these Token Strings as in Figure 3 in order to find repeated Token subsequences. These repeated Token subsequences are actually the Actual clones depicted in Figure 4.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

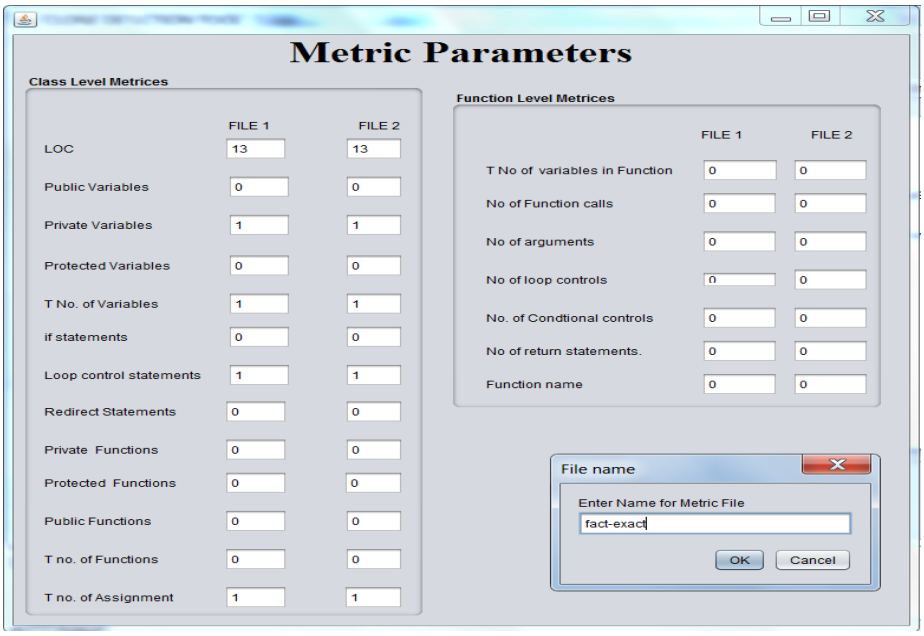

Figure 1. Metric calculation

The above figure declares that how Metric based approach works on the group of 20 Metrics. The Metrics used in the above figure belongs to the class of Class level Metrics and the Function level Metrics. This step is executed for the detection of Potential clones. After this Token approach has been applied to detect Actual clones.

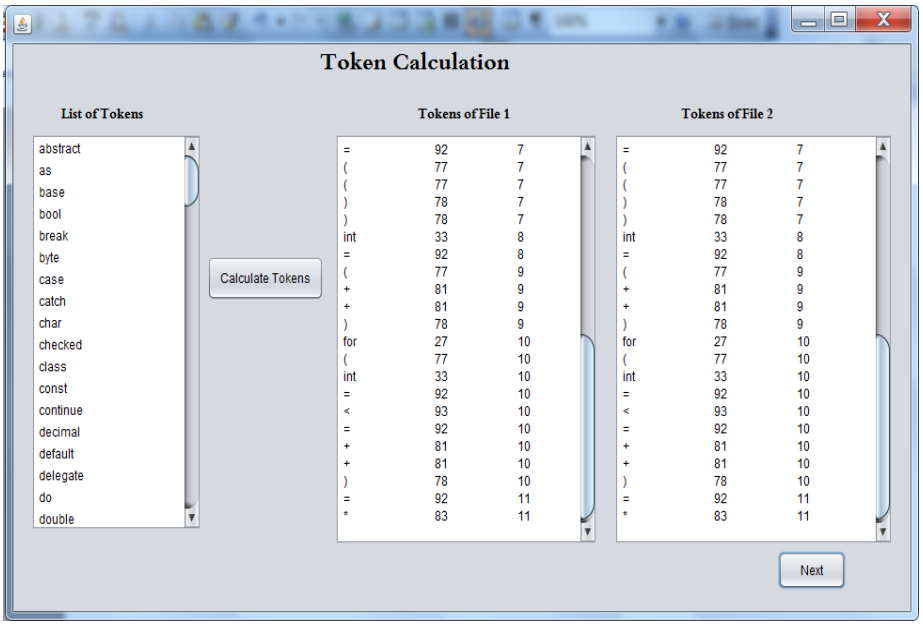

Figure 2. Token calculation

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

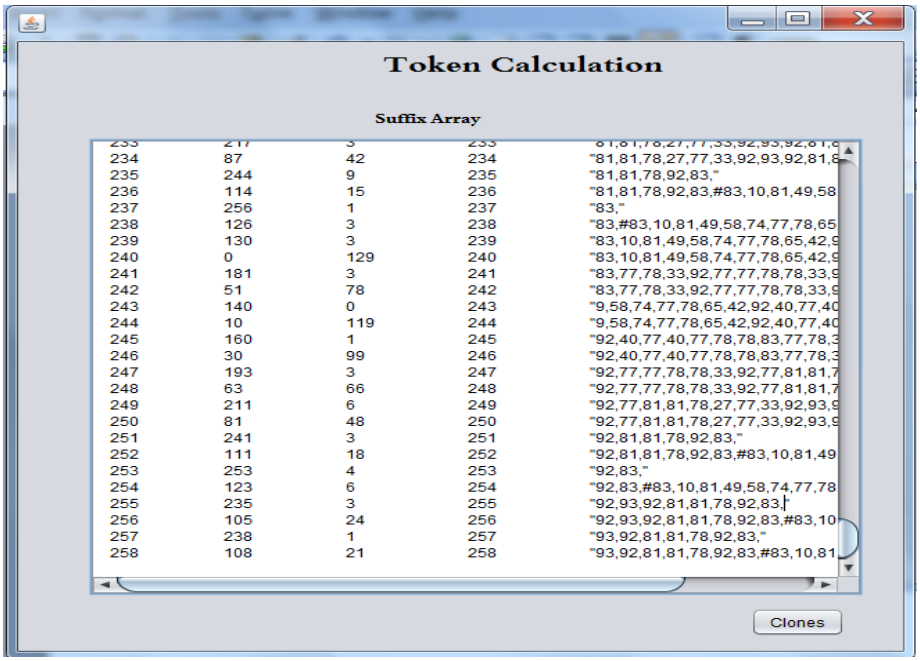

Figure 3. Application of Suffix Array on Tokens

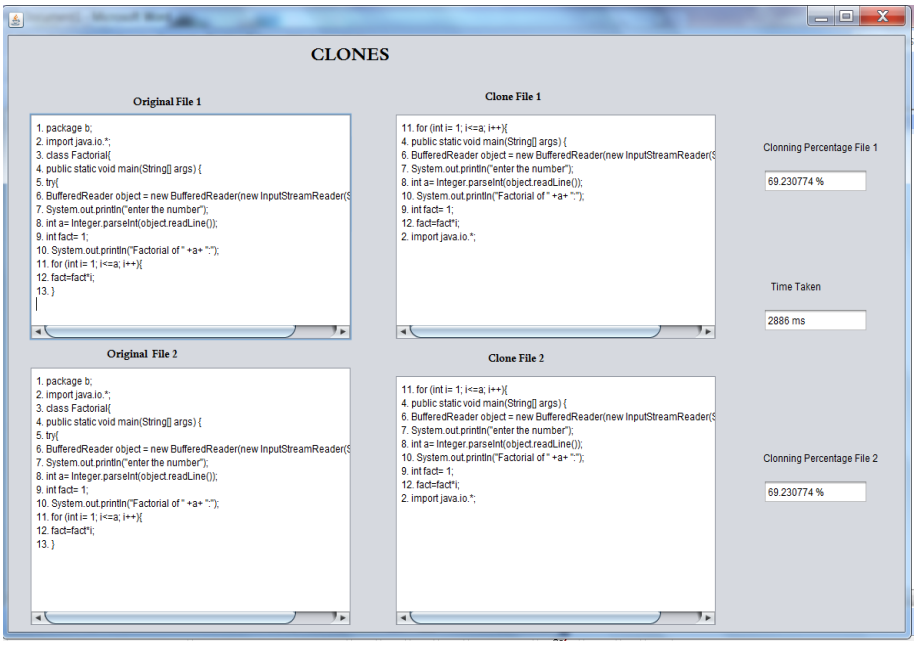

Figure 4. Actual clones

VII. RESULTS & DISCUSSIONS

The proposed hybrid technique can efficiently detects Type1,Type 2 and Type3 clones with the help of Metric based and Token based Technique which uses Suffix Array matching Algorithm to detect Actual clones. The output has been shown in terms of Cloning Percentage and Total Time Taken in finding clones. The advantage of using Suffix array algorithm over Suffix Tree Algorithm is, it directly form clone pairs

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

based on the output whereas Suffix Tree requires extra preprocessing step to form clone pair and clone classes[32][33]. Also Suffix array is more memory or space efficient than Suffix Tree. It consumes 5 times less memory than that of suffix tree.[32][33]

VIII. CONCLUSION & FUTURE SCOPE

From the study of various research papers, it can be concluded that cloning is in great demand today apart from its various shortcomings. It has proven to be an advantageous process in fast development of the software systems to meet the deadlines or to complete the work on time, etc [30]. It is considered as a great boon to industries. Also, various tools and techniques have been proposed to detect the clones, wherever required, to overcome the various pitfalls released by cloning like bug propagation, maintenance costs, etc [30].The proposed technique is the hybrid combination of Metric based and Token based Technique where Suffix array matching algorithm has applied on Tokenized string to detect clones in the Java and Asp.Net languages. Moreover the proposed technique can detect Type1, Type2 and Type3 clones. For future work, this technique will be further enhanced to detect Type4 clones. It can also be enhanced in such a way that it can detect clones for languages such as C++, C, PHP, etc. Clone removal techniques can also be added with this technique in order to further enhance it.

References

[1] Rachel Edita Roxas, “Automation generation of Plagiarism Detection among students Plagiarism”, In IEEE Transactions on Software Engineering, September 2006.
[2] Stefan Bellon, Rainer Koschke, Giuliano Antoniol, Jens Krinke, and Ettore Merlo.Comparison and Evaluation of Clone Detection Tools. In IEEE Transactions on Software Engineering, Vol. 33(9): 577-591, September 2007.
[3] C.Kapser and M. Godfrey“Cloning Considered Harmful” Considered Harmfu”. In WCRE, pp. 19 -28, 2006.
[4] Chanchal K. Roya, James R. Cordy, Rainer Koschke, “Comparison and evaluation of code clone detection techniques and tools: A qualitative approach,” Science of Computer programming, ELSEVIER, pp 470-495, 2009.
[5] Tung Thanh Nguyen, “ClemanX: Incremental Clone Detection Tool for evolving Software", ICSE’09, May 16-24, 2009, Vancouver, Canada 978-1-4244-3494-7/09@ 2009 IEEE.
[6] Kodhai.E, Perumal.A, and Kanmani.S, "Clone Detection using Textual and Metric Analysis to figure out all Types of Clones", Proceedings of the International Joint Journal Conference on Engineering and Technology, pp. 99-103, 2010
[7] Saif Ur Rehman, Kamran Khan, “An Efficient New Multi-Language Clone Detection Approach from Large Source Code,” International Conference on Systems, Man, and Cybernetics, IEEE, pp 937-940, 2012.
[8] Rochella Elva,Gary T. Leavens, “A Semantic Clone Detection Tool for Java Code,” March 2012.
[9] Deepak Sethi, Manisha Sehrawat, “Detection of code clones using Datasets,” IJARCSSE, pp 263-268, July 2012.
[10] Tahira Khatoon, Priyansha Singh, Shikha Shukla, “Abstract Syntax Tree Based Clone Detection for Java Projects,” Journal of Engineering, IOSR, pp 45-47, Dec 2012.
[11] Priyanka Bhatta, “HYBRID TECHNIQUE FOR SOFTWARE CODE CLONE DETECTION,” International Journal of Computers and Technology, pp 97-102, April 2012.
[12] Rupinder Kaur, H. K, “Evaluation of Token Based Tools On The Basis Of Clone Metrics” International Journal of Advanced Research in Computer Science and Electronics Engineering, 2012.
[13] Dhavleesh Rattan, Rajesh Bhatia, Maninder Singh, “Software Clone Detection: Systematic Review,” Information and Software Technology, ELSERVIER, pp 1165-1199, 2013.
[14] Kanika Raheja, Rajkumar Tekchandani, “An Emerging Approach towards Code Clone Detection: Metric Based Approach on Byte Code,” IJARCSSE, Vol.3, May 2013.
[15] Rupinder Kaur, Harpreet Kaur, “Clone Detection in Web Application using Clone Metrices” International Journal of Advanced Research in Computer Science and Software Engineering, July 2014.
[16] D.Gayathri Deviet al. “Comparison and evaluation on metrics based approach for detecting code clone” Vol. 2 No. 5 Oct-Nov 2011.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[17] Manpreet Kaur, Madan Lal, "Review on various code clone detection Techniques", Computer Science and Software Department, Punjabi University, May 2015
[18] Shashank Prabhakar, Sonam Gupta "A review on Code Clone Detection and implementation", Computer and Communication Engineering, February 2016
[19] Yogita Sharma "Hybrid Technique for Object Oriented Software Clone Detection", Master Thesis, Computer Science and Engineering Department, Thapar University, June 2011.
[20] Lingxiao Jiang, "DECKARD: Scalable and Accurate Tree-based Detection of Code Clones", pp 2-10.
[21] Mohammed Abdul Bari "Code Cloning: The Analysis, Detection and Removal" International Journal of Computer Applications (0975 – 8887) Volume 20– No.7, April 2011..
[22] Y. Higo, U. Yasushi, M. Nishino, and S. Kusumoto "Incremental code clone detection: A PDG-based approach". In WCRE, pages 3 –12, 2011.
[23] Anna Corazza, Sergio Di Martino, Valerio Maggio, Giuseppe Scanniello, "A Tree Kernel Based Approach for Clone Detection" 26th IEEE International Conference on Software Maintenance (2010).
[24] T. Kamiya, S. Kusumoto, "CCFinder: a multilinguistic token-based code clone detection system for large-scale source code". IEEE Trans. Software. Eng., 28(7):654–670, 2002.
[25] R. Komondoor and S. Horwitz, "Using Slicing to Identify duplication in Source Code", in: Proceedings of the 8th Static Analysis, (2001).
[26] J. Krinke, "Identifying Similar Code with Program dependence Graphs", in Proceedings of the 8th Working conference on Reverse Engineering, (2001).
[27] Ira Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant Anna, "Clone Detection Using Abstract Syntax Trees", In Proceedings of the 14th International Conference on Software Maintenance (ICSM'98), pp. 368-377, Bethesda, Maryland, November 1998.
[28] Baker, B.S, "On finding duplication and near-duplication in large software systems", Proc. of the 2nd IEEE Working Conference on Reverse Engineering, 1995, pp. 86-95.
[29] Sonam Gupta and P. C Gupta, "Literature survey of clone detection techniques." International Journal of Computer Applications (2014):41-44.
[30] M. Kaur, "Review on Software Cloning and Clone Detection," in Interenational Conference on Intelligent Circuits and Systems(ICICS 2016), Phagwara, 2016.
[31] L. P. Z. F. l. M. a. D. S. L. Qing Qing Shi, "A Novel Detection Approach for Statement Clones," IEEE, pp. 27-30, 2013.
[32] H. A. Basit, "Efficient Token Based Clone Detection with Flexible Tokenization," in ACM, 2004.
[33] G. Y. Munina Yusufu, "Efficient Algorithm for Extracting Complete Repeats from Biological Sequences," International Journal of Computer Applications, vol. 128, pp. 33-37, 2015.