International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A SURVEY ON CLONE DETECTION AND CLONE ANALYSIS

[1]Er. Rajnish Bala, [2]Er. Navpreet Rupal
[1]MTech CSE
Shaheed Udham Singh College of Engg. & Technology, Tangori
[2]Asst. Prof
Shaheed Udham Singh College of Engg. & Technology, Tangori
[1]Rajnijindal27@gmail.com, [2] er.nrupal@gmail.com

Abstract: *Code replication is ordinary difficulty, and a well recognized sign of terrible design. But Code replication is one of the nearly all well liked forms of software use again amongst developers. Clone discovery or code repetition discovery is the method troubled with the detection of code rubble that fundamentally calculate the same consequences .The most important aim of clone discovery is to recognize clone code and put back them with a single function call where the purpose would mimic the performance of a single example from the set of clones. As consequences of that, in the last decade, the issue of detect code replication led to a variety of tools that can mechanically find duplicate blocks of code. In this document dissimilar methods for code clone discovery, dissimilar tools and method used for that and the code examination will be discuss.*

Keywords: *clone analysis, clone detection, code clone, Evaluation and Maintenance.*

I. Introduction

In software expansion, it is ordinary to reuse a number of code remains by copying with or with no small modification. These kinds of code remains are called code clone. Cloning or repetition of codes can be harmful or beneficial. A basis code cloning occurs when a developer reuses accessible code in a new background by making a copy that is distorted to provide new functionality. There are other reason like copy & paste; lack of industrial data in developers and sometimes inadvertently clones are introduced. Clones are sector of code that are parallel according to various designation of similarity –Ira, Cloned code can be awkward for the reasons: multiple duplicate of code add to size of source code, preservation costs and, not in agreement changes to cloned code can create faults and which lead to wrong program performance .So a code clone discovery method is wanted. Code clone discovery is the procedure of locating segments of alike source code, according to the meaning of similarity, within a software system. Code clone psychoanalysis uses those consequences to look at code cloning in a software system. The goal of clone analysis is to understand the use of code cloning and study the individual code clones [1].

Code clone discovery and psychotherapy can be done as a three step process:

a) Generate a list of candidate clones,
b) Post processing the results, and
c) Analyzing the clones.

WHY DO CLONES OCCUR?

A Software clones emerge for a diversity of reasons:

- Code reclaim by copying pre existing idiom
- Code styles
- Instantiations of definitional computation
- Breakdown to identify and use theoretical data types
- Presentation enhancement
- An Accident [2].

DATA MINING

We consider that in order leads to authority and achievement, and gratitude to complicated technology such as computer, satellites, etc., we have been collect marvelous amounts of information. An originally, with the advent of computer and means for mass digital storage, we started collect and storing all sorts of data, including on the power of computer to help sort through this mixture of information [6]. Unluckily, these massive collection of data stored on different structures very speedily became irresistible. This original chaos has led to the creation of ordered databases and database management systems. The capable database management systems have been very important assets

for management of a large quantity of data and particularly for effectual and efficient retrieval of exacting information from a large collection when needed. The spread of database organization systems has also contribute to recent massive meeting of all sorts of in order.

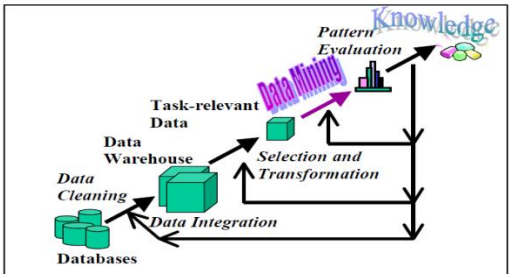

Figure 1.1: Data Mining is the core of Knowledge Discovery Process.

WHAT IS (NOT) DATA MINING?

What is not Data Mining?	**What is Data Mining?**
Look up phone number in phone directory	Certain names are more prevalent in certain US locations
Query a Web search engine for information about "Amazon"	Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

DATA MINING TASKS:
Categorization (Predictive)
- cluster (Descriptive)
- Connection Rule Discovery (Descriptive)
- Chronological Pattern Discovery (Descriptive)
- Failure (Predictive)
- Departure Detection[6] (Predictive)

II. RELATED WORK

Chanchal K. Roy[3], 2009 In this document, we provide a qualitative comparison and evaluation of the existing state of the art in clone detection technique and tools, and systematize the large amount of in order into a coherent conceptual framework. We begin with background concepts, a generic clone detection process and an overall taxonomy of in progress technique and tools. We then pigeonhole, weigh against and evaluate the techniques and tools in two different dimensions. First, we pigeonhole and compare approach based on a number of facets, each of which has a set of attribute. Second, we qualitatively evaluate the classified techniques and tools with respect to nomenclature of editing scenarios designed to model the creation of Type-1, Type-2, Type-3 and Type-4 clones.

Thomas LaToza[4], 2005 In this document ,Code clones, piece of code alike enough to be careful duplicate or clones of the same functionality, are a difficulty. In spite of attitude that code should never be copied and paste and agile dictum that all repetition be detached, there are often good reasons to copy and paste code. Developers wish to do be able to do amazing implement elsewhere that they can't call straight. This might be since the code makes assumption or design decisions that the developers need to modify.

Randy Smith[5],2009 In this paper described as, most preceding work on code clone discovery has focused on finding the same clones, or clones that are the same up to identifiers and literal values. Though, it is often significant to find alike clones, too. One confront is that the meaning of similarity depends on the context in which clones are being found. Therefore, we suggest new method for finding alike code blocks and for quantify their similarity. Our method can be used to find clone clusters, sets of code block all within a user full similarity threshold of each other. Also, given one policy block, we can find all parallel blocks and present them rank ordered by parallel. Our method has been used in a clone recognition tool for C program. The thoughts could also be included in many existing clone exposure tools to afford more give in their definition of related clones.

III. DETECTION TECHNIQUES

The detection of code clone is a two stage process which consists of a transformation and a comparison phase. In the first stage, the spring text is distorted into an external format which allows the use of a more professional association algorithm. During the ensuing comparison segment the actual matches are detected. Due to its middle role, it is reasonable to classify discovery technique according to their internal format. This part gives an overview of the dissimilar technique available for each category while selecting an envoy for each category.

a) **String Based**: String based a technique use basic cord transformation and comparison algorithm which makes them self governing of indoctrination languages. Techniques in this group differ in fundamental string contrast algorithm. A compare calculated signature per line, is one possibility to recognize for matching substrings [7]. Line corresponding which comes in two variants, is an

option which is selected as envoy for this category because it uses general string manipulations.

b) **Token Based:** Token based technique use an additional complicated alteration algorithm by construct a token stream from the source code, hence need a laxer. The attendance of such tokens makes it probable to use better contrast algorithms. Next to parameterized matching with suffix trees, which acts as hand over, we include [7] in this group since it also transform the source code in a token structure which is afterwards matched. The later tries to remove much more detail by abbreviation non exciting code fragments.

.

IV.BACK PROPAGATION NEURAL NETWORK

The Back Propagation neural network is artificial neural network based on error back propagation algorithm. The Back Propagation (BP) neural network model consists of an input layer, some hidden layers and an output layer. Each connection connecting neurons has a distinctive weighting value. In training the network, the nodes in the BP neural network obtain input information from exterior sources, and then go by to hidden layer which is an interior information processing layer and is answerable for the information conversion, and then the nodes in the output layer supply the required output information. After that, the anti propagation of error is transported by distinct the actual output with wanted output. Each weight is revised and back propagated layer by layer from output layer to hidden layer and input layer. This process will be continued until the output error of network is reduced to an acceptable level or the predetermined time of learning is achieved. The processing results of information are exported by output layers to the outside.

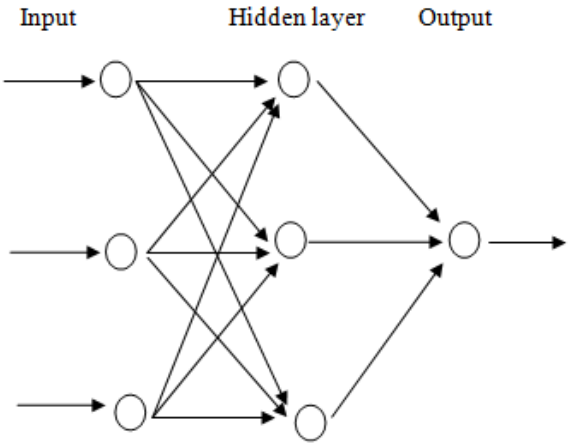

Figure: Back Propagation Neural Network

BP neural network consists of many neurons that are arranged in a form of three layers: input, hidden and output. The neurons are linked by weights W y In training the network with a given architecture, the back propagation approach, finds a single best set weight values by minimization of suitable error function .[10]

In a multi layer feed forward neural network, the processing elements are arranged in layers and only the elements in adjacent layers are connected. It has a minimum of three layers of elements (i.e., input layer, the middle or hidden layer, and the output layer). The name "back propagation" (BP) derives from the fact that computations are passed feed forward from the input layer to the output layer, following which calculated errors are propagated back in other direction to change the weights to obtain a better performance. BP algorithm is an extension of the least mean square algorithm that can be used to train multi-layer networks

The three-layered free forward neural network is displayed in Figure which is comprised by input layer, hidden layer and output layer. Once the network weights and biases are initialized, the network is ready for training. The training process requires a set of examples for proper network behavior, such as network inputs p and target outputs t. During training the weights and biases of the network are iteratively adjusted to minimize the network performance function. The number of hidden layer is always difficult to determine in ANN creation. It is generally agreed that one hidden layer is sufficient for most of purposes [8]. In the present study, only 1 hidden layer will, thus, be used in the BP network for simplicity.

V. NEURAL NETWORK

Machine learning algorithms help a lot in decision making and neural network has performed well in classification purpose in medical field. Most popular techniques among them is neural network. Neural networks are those networks that are composed of simple elements which operate parallel. A neural network [9] can be trained to perform a particular function by adjusting the values of the weights between elements. Network function is determined by the connections between elements. There are several activation functions that are used to produce relevant output.

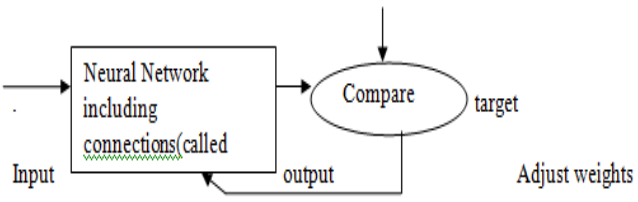

Figure 5 : Neural Net Block Diagram

Training can be either supervised or unsupervised. In supervised training system learns by trying to predict outcomes for known examples. System compares its predications with the known results and learns from is mistakes. In unsupervised training system no output or result is shown as part of training process. With the delta rule, as with other types of back propagation, learning is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of output, and the backward error propagation of weight adjustments. Simply, when a neural network is initially presented with a pattern it makes a random 'guess' as to what it may be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights. Within each hidden layer node is a sigmoid activation function which polarizes network activity and helps it to be stable in nature. Back propagation performs a gradient descent within the solution's vector space towards a 'global minimum' along the steepest vector of the error surface. The global minimum is that imaginary solution with the lowest possible error. Back propagation is a method of training artificial neural network. It requires a desired output for each value in order for calculation of loss function gradient. Following algorithm will show how BPNN works in classification in medical imaging.

VI. CONCLUSION

Code clone is a significant difficulty. As a form of use again, it is more often than not caused by programmers' copy and paste behavior. Though it seems to be a simple and effectual method, these repetition behavior are usually not recognized, which cause a lots of unenthusiastic effects on the excellence of the software, growing the amount of the code which needs to be maintain, and replication also increase the flaw likelihood and resource supplies. In this manuscript I largely listen vigilantly on exposure techniques and clone analysis methods. This will help for indulgent code clones and the different techniques used. The textual approach gives a crude overview of the duplicate code that is fairly easy to obtain, so it is most fitting during problem detection and problem appraisal. The token based move toward provides a precise picture of a given piece of duplicated code and is robust against rename operations. Therefore it works best in mixture with fine-grained refactoring tools that work on the level of statement. Syntactic technique are very good at enlightening duplicate subroutines, irrespective of small difference, so it works best in mixture with refactoring tools that labor on the technique level.

REFERENCES

[1] Prajila Prem," A Review on Code Clone Analysis and Code Clone Detection", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 12, June 2013.

[2] Baxter, Ira D., et al. "Clone detection using abstract syntax trees." Software Maintenance, 1998. Proceedings., International Conference on. IEEE, 1998.

[3] Roy, Chanchal K., James R. Cordy, and Rainer Koschke. "Comparison and evaluation of code clone detection techniques and tools: A qualitative approach." Science of Computer Programming 74.7 (2009): 470-495.

[4] LaToza, Thomas. "A literature review of clone detection analysis." (2005).

[5] Smith, Randy, and Susan Horwitz. "Detecting and measuring similarity in code clones." Proceedings of the International Workshop on Software Clones (IWSC). 2009.

[6] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Vol. 1. Boston: Pearson Addison Wesley, 2006.

[7]. Rysselberghe, Filip Van, and Serge Demeyer. "Evaluating clone detection techniques from a refactoring perspective." *Proceedings of the 19th IEEE international conference on Automated software engineering*. IEEE Computer Society, 2004.

[8] Mitchell, Melanie. *An introduction to genetic algorithms*. MIT press, 1998.

[9] N.B. Karayiannis, A.N. Venetsanopoulos,"Efficient Learning Algorithms for Neural Networks, " IEEE , vol. **23**, 1993, pp. 1372 -1383.

[10] Wang, Yuanfei, Wei Zhang, and Wen Fu. "Back Propogation (BP)-neural network for tropical cyclone track forecast." *Geoinformatics, 2011 19th International Conference on*. IEEE, 2011.