ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

ANALYTICAL REVIEW ON THE TECHNIQUES OPTED FOR DETECTION OF CLONING IN SPREAD SHEETS

MITALI, Dr.SUSHIL GARG

M-TECH Scholar, RIMT ,Mandi Gobindgarh , Punjab, India

Principal, RIMT ,Mandi Gobindgarh , Punjab, India

ABSTRACT: Spreadsheets are widely used in industry. However, spreadsheets are error-prone, many companies have lost money because of these spreadsheet errors. One of the causes for spreadsheet problems is the prevalence of copy-pasting. This paper focuses on the methods of identifying data clone in spread sheets and their efficiency. The paper also presents suitable methods for visualizing the detected data clones.

Keywords: Spread Sheets, Data Clone, Data Mining, Fatal Errors

I. INTRODUCTION

Spreadsheets are widely used in industry [1]. However, spreadsheets are error-prone, numerous companies have lost money because of spreadsheet errors. One of the causes for spreadsheet problems is of copy-pasting. Based on existing text-based clone detection algorithms, different algorithms have been designed to detect data clones in spreadsheets: formulas whose values are copied as plain text in a different location.

Data cloning is the practice of duplication of data and in the case of spreadsheets the values of different formulas and same results are copied in same locations in the spreadsheets. Most of the companies relies on the spreadsheets especially for their accounts section and spread sheets being prone to errors can result in loss to the companies. This survey paper describes the techniques and algorithms being used for the detection of data clones in the spreadsheets. It is seen that when the formula results are copied as data for the other parts of spreadsheets it is risky and it is easy coping in excel with Paste Special ‘where a user can only copy values. The results of the evaluation of cloning in the spreadsheets clearly indicate that

1) Data clones are common.
2) Data clones pose threats to spreadsheet quality.
3) Duplicated fragments can increase the maintenance efforts.
4) Data clones can increase the probability of introducing a bug.

II. TYPES OF DATA CLOING

a) REGULAR
b) IRREGULAR

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

In a regular cloning the entire data from one spread sheet can be copied to another data format or to another spread sheet itself. This cloning method is done if the data is getting updated in the industry and the concerned person wants to keep the previous and the new record all together.

Irregular data cloning is done very often in the industry as to avoid work and to save time. Industry people copy the database from one organization to another to build up links with the clients already associated with the first organization.[3].

On the basis of textual similarities, clone can be categorized into:

2) Exact clone: When the Fragment of data that is identical or similar to another is known as exact clone.
3) Renamed clone: When the two fragments of data are same except for the name by which they are represented or saved.
4) Near miss clone: Near miss clone clusters can occur when user updates a copied cell, but does not update the original cell.

Now figure1 shows an example of cluster cloning between two spreadsheets

Variance Matrix

	A	B	C	D
A	0.1156	0.0320	0.384	0.448
B	0.0320	0.1300	0.0480	0.056
C	0.384	0.0480	0.1476	0.0672
D	0.448	0.660	0.672	0.1684

Covariance Matrix (figure 1)

According to figure 1 the shaded region shows the duplicated clusters in two spreadsheets and the values under column steady represents the original values which are not being duplicated.

III. REASONS OF CLONING IN THE SPREADSHEETS

Why would spreadsheet users resort to copy-pasting data from one spreadsheet file to the other, if most spreadsheet systems have a 'better' way to do this? In most of the experiences, this practice can have several reasons.

Firstly, sometimes users are not aware of a ways to link spreadsheets; they do not know how to use the link-formulas.

Secondly, users are often unaware of the risks of copy-pasting data and it seems to be the easiest way and the aim is not at changing the spreadsheet user's behavior, since that would, most likely, involve changing the process around a spreadsheet and that would be hard to implement in a company rather it is better to follow an approach that allows users to proceed as they normally would, but reducing the risks by detecting and visualizing the copy-paste relationships.

IV. PARAMETERS OF EVALUATION FOR CLONE DETECTION

This survey paper indentifies the main four parameters to be considered for the detection of data clones which are as follows:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

Step Size: This indicates the search radius in terms of numbers of cells. Setting it to 1 means we are only looking for direct neighbors, with step size 2, a 'gap' of 1 cells is allowed in a cluster.[5]

Match Percentage: This is used when clusters are matched. This percentage indicates what percentage of the cells has to match. Setting it to 100% means the values have to match exactly, lower percentages allow for the detection of near-miss clones.

Minimal Cluster Size: This sets the minimal number of cells that a cluster has to consist of. Very small clusters might not be very interesting, hence there has to be a minimal threshold.

Minimal Different Values: This represents the minimal number of different values that have to occur in a clone cluster. Similar to small clusters, those clusters consisting of a few different values will be of less interest. [6]The user can indicate whether clones are found within worksheets, between worksheets, between spreadsheets or a combination of those.

V. METHODS AND ALGORITHM FOR CLONE DETECTION:

There are several algorithms for the identification of the data clone and the percentage in which they have been copied. Here is a review of some of the finest algorithms.

A) ICA (Increment Component Analysis): They employ a generalized suffix-tree that can be updated efficiently when the source changes [4]. The amount of effort required for the update only depends on the size of the change, not the size of the code base. Unfortunately, generalized suffix-trees require substantially more memory than read-only suffix-trees, since they require additional links that are traversed during the update operations. Since generalized suffix-trees are not easily distributed across different machines and the memory requirements represent the bottleneck with respect to scalability. Consequently the improvement in incremental detection comes at the cost of substantially reduced scalability.

B) AST Based Incremental Method: Nguyen et al. presented [5] an AST-based incremental approach that computes characteristic vectors for all sub trees of the AST for a file. Clones are detected by searching for similar vectors. If the analyzed data changes, vectors for modified files are simply recomputed. As the algorithm is not distributed, its scalability is limited by the amount of memory available on a single machine. A related approach that also employs AST sub tree hashing is proposed by Chilowicz et al. [5]. However, such systems often contain substantial amounts of cloning [3] making clone management for them especially relevant. Instead, his approach does not require a parser.

C) Neural Logistics for Clone Detection: The neural networks are non-linear statistical data modeling tools that are inspired by the functionality of the human brain using a set of interconnected nodes [6] networks are widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are chiefly applied to credit card spread sheet data, automobile insurance spread sheet data and corporate fraud. Literature describes that neural networks can be used as a financial fraud detection tool. The neural network fraud classification model employing endogenous financial data created from the learned behavior pattern can be applied to a test sample. The neural networks can be used to predict the occurrence of corporate fraud at the management level [7].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

A neural network (NN) is a feed-forward, artificial neural network that has more than one layer of hidden units [7] between its inputs and its outputs. Each hidden unit, j, typically uses the logistic function1 to map its total input from the layer below, xj , to the scalar state, yj that it sends to the layer above.

y_j= logistic(xj) =1/1 + e^{-xj} , $x_j = bj + \sum y_i w_{ij}$

Where bj is the bias of unit j, i is an index over units in the layer below, and wij is a weight on a connection to unit j from unit i in the layer below. For multiclass classification, output unit j converts its total input, xj , into a class probability, pj.

D) Text Based Techniques: Text based techniques perform little or no transformation to the raw source data of spread sheet before attempting to detect identical or similar (sequences of) data. [8]

E)Token Based Technique: Token-based techniques apply a lexical analysis (tokenization) to the source code and, subsequently, use the tokens as a basis for clone detection. [9]

F)PDG-based Approach: PDG approaches go one step further in obtaining a source code representation of high abstraction. Program dependence graphs (PDGs) contain information of a semantic nature, such as control and data flow which look for similar sub graphs in PDGs in order to detect similar data. It first augments a PDG with additional details on expressions and dependencies, and similarly applies an algorithm to look for similar sub graphs [10]

G)Root Cause Detection: It is a challenging issue to identify a defective system in a large scale network. The data collected from the distributed systems for trouble shooting [12] is very huge and may contain noisy data, so manual checking and detecting of the abnormal node is time consuming and error prone. In order to solve this, RCD is an automated system that detects the anomaly. It works on two principles.

a) **OUTLIER DETECTION**
b) **MAIN CAUSE**

OUTLIER DETECTION determines the nodes that are "far away" from the majority as potential anomalities by analyzing the low dimensional matrix produced by feature extraction, a cell-based algorithm is used to quickly identify the outliers.
MAIN CAUSE determines whether a deadlock occurrence or memory leakage problem that caused the node to act like a anomaly.

VI. VISUALIZATION OF DETECTED CLONES

After the detection of data clones there are two ways the data clones can be visualized.

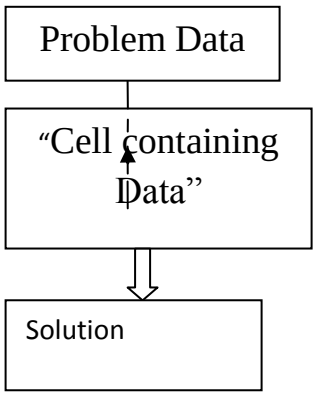
A)DFD'S: First is generation of data flow diagram which will show how data is cloned between two worksheets, which is done by drawing an arrow between the worksheets. DFD's provides the basic understanding of relationship between spreadsheets that are having clones. According to this worksheets are represented as rectangles and the arrows are used to indicate a formula dependency between two worksheets. Figure 2 shows the kind of DFD's being used to visualize the detected data clones

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013



Figure(2) representing general format

Figure 2 shows the dependency between the problem data and the particular cell in the spreadsheet whose matching is being done .problem data is the data which is suspected to be cloned the dashed line shows the data clone dependency and the solid line shows the formula dependency. Here data clone dependency means copied input value and formula dependency means the result of the formula is being copied.

B) POP-UP'S : The second way is the addition of pop-up boxes into the spreadsheets to show where data is copied and in some special cases for example near miss clones the popup boxes will show which cell is differing. Pop-ups are added to the spreadsheets to both the source and the copied clone fragment .this will in turn give the warning to the user that the data has been copied so he can update the changes and can reduce the probability of bug. By links the dependencies are shown explicitly to make changes.

VII. CONCLUSION

This survey paper focuses on the need of study of data clone detection in the spreadsheets as spreadsheets are one of the most important functioning blocks in an organization and errors in the spreadsheets can result in big loses to the companies. The different kinds of data clones that can encounter and can degrade the quality of spreadsheets and increases the maintenance cost. In accordance with this, the paper presents various techniques and algorithms to detect these data clones in the spreadsheets to make them less error prone and implementing ways to visualize these detected data clones which shows the dependency between the fragments and provides warning to the user about the cloned status of spreadsheets. This paper can lead to various directions for the further study to improvise the techniques and algorithms used for the detection of data clones to increase the quality of spreadsheets.

REFERENCES

[1] H. A. Basit, D. C. Rajapakse, and S. Jarzabek. "A study of clones in the STL and some general implications. In Proc. of the Int'l Conf. on Software Engineering, " pages 451{459, 2005.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

[2] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier. "Clone detection using abstract syntax trees. In Proc. of the Int'l Conf. on Software Maintenance " pages 368{377, 1998.
[3] K. Beck. "Extreme Programming explained, embrace change" Addison-Wesley, 2000.

[4] Hang Dai and Jingshi He Dongguan "China Research Journal of Applied Sciences, Engineering and Technology" 6(5): 895-899, 2013 ISSN: 2040-7459; e-ISSN: 2040-7467 2013
[5] T. T. Nguyen, H. A. Nguyen, J. M. Al-Kofahi, N. H. Pham, and T. N. Nguyen, "Scalable and incremental clone detection for evolving software," ICSM'09, 2009.
[6] Ghosh, S., & Reilly, D. L. (1994). "Credit card fraud detection with a neural-network", 27th Annual Hawaii International, Conference on System Science 3 (1994) 621–630.
[7] Beasley, M. (1996)." An empirical analysis of the relation between board of director composition and financial statement fraud. The Accounting Review", 71(4), 443–466.
[8] J. H. Johnson, "Identifying redundancy in source code using fingerprints,"in Proc. of CASCON '93, 1993, pp. 171–183.
[9] M. Fisher and G. Rothermel, "The EUSES spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms," ACM SIGSOFT Software Engineering Notes, vol. 30, no. 4, pp. 1–5, 2005.
[10] I. D. Baxter, A. Yahin, L. M. de Moura, M. Sant'Anna, and L. Bier, "Clone detection using abstract syntax trees," in Proc. of ICSM '98, 1998, pp. 368–377.
[11] R. Komondoor and S. Horwitz, "Using slicing to identify duplication in source code," in Proc. of SAS '01, 2001, pp. 40–56.
[12] G.D.K.Kishore1, Maddali Sravanthi Automated Anomaly and Root Cause
Detection in Distributed Systems. International Journal of Engineering Trends and Technology- Volume3Issue1- 2012