Detecting Clone Profiles in Social Media Networks

C. R. Liyanage

169318J

Master of Science in Information Technology

Faculty of Information Technology

University of Moratuwa

February 2019

Detecting Clone Profiles in Social Media Networks

C. R. Liyanage

169318J

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology

February 2019

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

…….

Signature of Student

……..……

Date:

Supervised by

Name of Supervisor

…….

Signature of Supervisor

……..……

Date:

Acknowledgement

I wish to express sincere appreciation to,

My research supervisor, Mr. S. C. Premarathne for dedicating his valuable time for guiding me throughout this.

My husband and parents for providing me with continuous encouragement throughout my years of study.

Finally, for my colleagues for their cooperation and support.

Thank you.

Abstract

With the popularity of Online Social Networks (OSN), the number of different types of digital attacks has been increased causing lots of damages to their users. Identity Clone Attack (ICA) is one of the leading among them which illegally uses the information of a genuine user by duplicating them in another fake profile. These attacks severely affect a true and innocent identity since it can be misused by another malicious profile. Hence these clone profiles need to be identified and removed in order to increase the protection of users. Many researchers have tried to solve the problem of clone profiles in OSN, however more robust solutions are still to be taken. This study introduces a model to detect clone profiles on Facebook by clustering based on weighted categorical attributes and estimating the strength of friend relationship among friends. The list of possible clones with the amount of clone percentages to a given victim profile was presented as the output of the model. With the use of Agglomerative hierarchical clustering algorithm and Jaccard similarity measurement, a low average within cluster distance and a precision of 88.75% has shown in the results

Table of Content

CHAPTER 1 1

INTRODUCTION TO FAKE AND CLONE PROFILES 1

1.1. Introduction to the background 1

1.2. Profile Cloning Environment 1

1.3. A brief from previous work 3

1.4. Problem Statement 3

1.5. Aim and Objectives 3

1.6. Research Process, Inputs, Outputs and Beneficiaries in brief 4

1.7. Summary of the first chapter and heading to next chapters 4

CHAPTER 2 5

CURRENT STATE OF DETECTING DUPLICATE PROFILES IN OSN 5

2.1. Introduction to the background 5

2.2. Current Situation 6

2.3. Platform Selection 6

2.4. Data Collection 6

2.5. Existing Approaches 7

2.5.1. Using Classification Algorithms 7

2.5.2. Social Graph based approach 8

2.5.3. Matching similarity attributes 9

2.5.4. Analyzing user behavior changes 10

2.5.5. Validation 11

CHAPTER 3 12

TECHNOLOGIES USED FOR CLONE DETECTION PROCESS 12

3.1. MySQL 12

3.2. Java 12

3.3. RapidMiner Studio9.1 12

3.4. Datamining through Clustering 13

3.5. Similarity measurement algorithms 14

3.6. Hardware environment 14

3.7. Advanced Data Generator 14

CHAPTER 4 15

A NOVEL APPROACH FOR DUPLICATE PROFILE DETECTION 15

4.1. Introduction 15

4.2. Inputs to the model 15

4.3. Outputs from the 15

4.4. Process in brief 15

4.5. Platform Selection 16

4.6. Data Collection 16

4.8. Performing clustering 17

CHAPTER 5 18

CLONE PROFILE DETECTION - ANALYSIS & DESIGN 18

5.1. Introduction 18

5.2. Overview of the Proposed Model Design 18

5.3. Attribute feature selection 19

5.4. Artificial Clone Profile Set Generation 21

5.5. Filter candidates by name 21

5.6. Clustering based on attribute features 21

5.7. Network Similarity Calculation 23

5.7.1 Friend Network 25

5.7.2 Recommended friend network 26

5.7.3 Aggregate Friend Network Similarity 27

5.8. Profile Similarity Threshold Calculation and Clone Detection 27

5.9. Result Validation 28

CHAPTER 6 30

IMPLEMENTATION OF THE CLONE PROFILE DETECTION MODEL 30

6.1. Introduction 30

6.2. Data Preprocessing 30

6.2.1. Friend Network dataset 30

6.2.2. Attribute feature dataset 31

6.3. Recommended friend list generation 32

6.4. The Model 33

6.5. Detection Phase 1 – Filter by Names 33

6.6. Detection Phase 2 – Clustering on categorical data 34
6.7. Detection Phase 3 – Friend Network Similarity Calculation 36
6.8. Calculating the similarity threshold value 37
6.9. Testing Phase 37

CHAPTER 7 38

RESULTS EVALUATION & DISCUSSION ON WORK 38
7.1. Introduction 38
7.2. Clustering as a supervise method 38
7.3. Network Similarity performance evaluation 42
7.4. The Results Validation 44

CHAPTER 8 46

CONCLUSION & FUTURE WORK 46
8.1. Introduction 46
8.2. Challenges to the field of research 46
8.3. Challenges to the proposed model 47
8.4. How this study can be modified in future 48
8.5. Conclusion 49
REFERENCES 50

List of Tables

Table 5.1: Profile attributes/ features considering 20
Table 5.2: Attribute Weight Calculation 22
Table 5.3: Similarity Threshold Calculation 28

List of Figures

Figure 3.1: RapidMiner Studio 9.1 13
Figure 3.2: Advanced Data Generator tool 14
Figure 5.1: Proposed steps of detection stage 19
Figure 5:2 Example attribute set of victim and clone profile (victim top, clone bottom) 21
Figure 5.3: Clustering Algorithms with number of clusters 23
Figure 5.4: Clustering Algorithms with their distribution performances 23
Figure 5.5: Friends of two users 25
Figure 6.1: Subset of the list of friends for each user in OSN 30
Figure 6.2: Subset of the list of friends for each user in OSN 31
Figure 6.3: Part of the Binary Vectors corresponding to the availability of features for a particular user 32
Figure 6.4: Part of the annonymized Feature list corresponding to a particular user in a particular circle 32
Figure 6.5: Attribute adjacency matrix (4039*11) 32
Figure 6.6: Detection phase 1 and 2- system model 33
Figure 6.7 : Join by frist_names 34
Figure 6.8: weight values estimated for each attribute 34
Figure 6.9 Clusterin criteria in RapidMiner 35
Figure 6.10: clustering output for a given victim 35
Figure 6.11: detection phase 3 - network similarity 36
Figure 6.12: Calculated similarity and percentage 37
Figure7.1: Cluster centroid table 1 39
Figure 7.2: Cluster allocation plot 1 39

Figure 7.3: Cluster centroid table 2.....40
Figure 7.4 : Cluster allocation plot 2 40
Figure 7.5 Cluster centroid table 3.....41
Figure 7.6: Cluster allocation plot 3 41
Figure 7.7: clone similarity and percentage 1.....42
Figure 7.8: clone similarity and percentage plot 4.....42
Figure 7.9 : clone similarity and percentage 2.....43
Figure 7.10: clone similarity and percentage plot 5.....43
Figure 7.11: clone similarity and percentage 3.....44
Figure 7.12 : clone similarity and percentage plot 6.....44
Figure 7.13: Cluster density performance.....45

List of Equations

Equation 5.1: Jaccard Similarity Measurement 24
Equation 5.2: Friend Network Similarity based on friend lists of two profiles.....25
Equation 5.3: Friend Network Similarity between Friend list and recommended friend list of two profiles 26
Equation 5.4: Aggregate Network Similarity between two profiles.....27
Equation 6.1: Clone Percentage Calculation 36