PROFILING AND IDENTIFYING INDIVIDUAL USERS BY THEIR NATURAL LANGUAGE USAGE FOR POSITIVE IDENTIFICATION

Darusalam[1], Widya Cholil[2]
Dosen Politeknik Sekayu
Dosen Universitas Bina Darma
Jalan Kol. Wahid Udin Lingkungan I kayuara, Sekayu[1]
Jalan Jenderal Ahmad Yani No.12, Palembang[2]
Pos-el: darusalam85@gmail.com[1], widya_cholil@mail.binadarma.ac.id[2]

Abstract: *Profiling and identifying individual users is an approach to help recognize intrusions in a computer system. User profiles are important in many applications since they record highly user-specific information - profiles are basically built to record information about users or for users to share experiences with each other. Thus, user profiles are used to aggregate relevant information about a user's activities, and to identify patterns in their behavior. This research uses the n-gram analysis method for characterizing each user's style, and can potentially provide accurate user identification. As a result, n-gram analysis of a user's typed inputs offers another method for intrusion detection as it may be able to both positively and negatively identify users. The contribution of this research is to assess the use of a user's writing styles in both natural language as a user profile characteristic that could enable intrusion detection where intruders masquerade as real users.*

Keywords: *User Profiling, Identification, Computer Security, Anomaly Detection*

Abstrak: *Profil dan identifikasi pengguna individu adalah suatu pendekatan untuk mengenali penyeranggan di sebuah system komputer. Profil penggunal sangat penting disetiap aplikasi, dikarenakan mencatat secara detail semua informasi pengguna untuk meberikan pengetahuan satu sama lain. Profil pengguna digunakan untuk mendapatkan informasi yang relevan tentang aktivitas pengguna personal komputer dan mengidentifikasi tingkah laku pengguna dalam menggunakan komputer. Riset ini mengunakan n-gram analisuntuk menanilis setiap karakter dari gaya penulisan pengguna method ini mempunyai potensi untuk memberikan identifikasi pengguna secara akurat. Hasilnya n-gram memberikan metodelogi yang lain untuk mendeteksi intrusión detection dan memungkinkan untuk menentukan positive identifikasi dan negative identifikasi. Kontribusi dari reset ini adalah untuk menilai setia gaya penulisan pengguna di natural language profil pengguna yang bisa memungkinan mendeteksi intrusion detection yang menyamar sebagai pengguna asli*

Kata kunci*: Profil Pengguna, Identifikasi, Keamanan komputer, anomaly deteksi*

1. INTRODUCTION

Profiling is a way of grouping things or individuals into categories or groups based on characteristics such as situation, appearance, traits (N.P.Dau, Rau & J.Templeton 2000). The term profiling means to get information about a user's activities, and it is possible to perform anomaly detection over a user profile to allow user identification. This research plans to investigate psychometric user characteristics and so as to be able to identify a user.

There are many examples of profiling in the area of information technology (N.P.Dau, Rau & J.Templeton 2000), such as profiling users to know about their computer usage patterns. As a result, users are able to use and distribute the system resource more efficiently and to offer best services. Furthermore, in other areas, profiling is used for user identification in Internet-based commerce.

Inter-social network functionalities and operations are necessary for several activities such as profiling and identifying user behavior

(Raad, Chbeir & Dipanda 2010). User profiles are used to collect relevant information about a user's activities and anomaly detection is performed over these user profiles. Furthermore, user profiles are important in a range of applications, due to being able to record and compare against user-specific information (Ashman & Holland). According to (Ashman & Holland) user profiles are a good method for reauthentication and intrusion detection.

This journal will focus on evaluating the potential of psychometric user characteristics, namely writing style in Natural Language (Jane Austin & William Shakespeare). To evaluate this particular characteristic this research will use the n-gram analysis method, and will aim to identify users in two ways, positive user identification and negative user identification. The work will however not implement the use of these user characteristics in an intrusion detection system, however it establishes whether they can be used in such a system.

Profiling and identifying can help recognize intrusions. According to Pannell and Ashman (Pannell & Ashman 2010), user profiling is already a necessary part of the personalization of information delivery, and they propose it as an approach for identifying attacks to a computer system is by profiling program and user behaviors (Wei, Xiaohong & Xiangliang 2004). Anomaly detection over a user profile can detect when an intruder is masquerading as a genuine user.

The research extends previous research that implemented an intrusion detection system based on biometric characteristics such as keystroke analysis and mouse use and psychometrics characteristic such as user prose style and favorite web pages (Pannell and Ashman 2010). However this research will focus on one potential psychometric user characteristic and will consider whether a user's writing styles in different scenarios can be assessed with n-gram analysis in order to identify the user. Users' writings may be in the form of text in a novel, books, blogs, tweets and emails, and this is a form of natural language. On the other hand, users' writings also occur in the way they interact with computers, issuing commands through a command line interface, and this is a form of formal language. This research will perform the same analysis on data of both types, using exactly the same analysis, and will determine firstly whether either form can be used successfully for user identification, and if so, the research will then will determine which is the most effective of the two.

This research analyze the two different forms of data in two ways, firstly to check whether it can detect when the current user does not match the user profile and is hence an intruder – this is negative identification. The second way is to detect whether the current user can unquestionably be verified as the true user – this is positive identification. This research provide positive identification result only. Most intrusion detection systems assume that the user is genuine until anomalies or broken rules show otherwise, that is, they only make use of negative identification. However it might be useful to constrain a user's activities until they positively identify themselves, perhaps not allowing the user to make significant changes until their current login session has been positively identified.

In this research, the default position will be that the system has no evidence about the user's identity, other than the fact that the user managed to log in. Analyzing their activity after logging in should either give positive information that correlates strongly with the user's profile and confirms their identity, or it should mismatch the profile and the user would then be rejected from the system.

The research will answer the following questions: Q1: Does the use of n-gram analysis to profile users' writing styles in their natural language allow accurate user identification? (a) if so, does it allow both positive identification?
Q2: If the profiling of both natural language writing styles and command usage allow accurate user profiling, which is the most accurate?
This research will contribute the following knowledge: (a) Proposing and assessing the usefulness of two psychometric characteristics for a user profile in an intrusion detection system (IDS) - specifically, comparing formal and natural language psychometric characteristics using n-gram analysis; (b) Distinguishing between positive and negative identification, and showing whether this distinction is practical.

2. METHODOLOGY

This research aims is to identify a user especially in Natural language (Jane Austen and William Shakespeare writing style). This research will use the n-gram analysis method for author identification – this is an established authorship attribution method. Furthermore the research also evaluates how quickly the system learns this characteristic of the user model. The structure of the method in the research is indicated below.

2.1 Experimental Set Up

The implementation part is explained about how the application produced the n-gram frequency. This software application was written in java programming language. There are two classes in this software, "n-gram.java" and "Data.java". The program running with the command "java n-gram [n]" n is the value of n-gram. This software will produce the n-gram frequency that placed in the Comma Separated Value "csv" folder and distributed to Microsoft Excel. In 'csv' folder contain the history of data which txt. Formatted and can be read by Microsoft Excel or other that equivalent and can be ready to use for n gram analysis.

We will use this software for counting the n-gram of history of data from the user writing style. This software was given from previous research (Ashman and Holland). We use the software for perform four types of n-gram analysis, namely 3-gram, 5-gram, 11-gram and 15-gram.

2.2 Literature Review

This section will discuss some previous research into several core aspects of this minor thesis. This literature review will investigate some similar research involving n-gram analysis and additional literature to support features of the project. n-gram analysis is one of the methods that will be used for the project. Reviewing some papers, which use n-gram analysis, demonstrates how n-gram analysis has successfully profiled both features and users in other application areas.

2.2.1 Profiling in online social networks

There are many researches on computer science that use social networks for user profiling. Social networking is one of the applications that engage the user to be more active and permit user to create and maintain their own web pages, Maia et al (2008). According to Vosecky et al. (2009) varieties of social networking have different manners to display and store information user profile on user's web profile. Social network has become one of the applications to identify user profile.

This research continues previous research by Ashman and Holland (in draft). They examined users to identifying anomaly detection over user model. They classified user model characteristics into two classes. Firstly there are behavioral biometrics, which represents a user's physical characteristics, for example ability to use mouse, habitual miskeying errors and typing speeds. Secondly there are psychometrics that represents a user's personal preferences or decision-based characteristics, for example prose style and favored web pages. In addition, n-gram analysis is used to identify users over their command line histories. This research will focus on psychometric user characteristics.

2.2.2 n-gram analysis method

There are many papers that use n-gram analysis method and it is implemented in many applications. n-gram analysis that use for many purposes, including computer virus detection, author profile and language independent authorship. n-gram analysis is one of the methods that will be used for this project. According to Luo et al (2010), n-gram method is 'language model based on collinear relation'. Some of the n-gram analysis is use in many purposes that we will explore here and evaluate for this project.

Luo et al (2010) outline the use of an n-gram-based malicious code feature extraction algorithm with statistical language model. By using trigram (3-gram) model they can characterise malicious code features and hence detect the virus. As a result they can reduce the time and space of computer rather than detect the virus from heuristics scheme that is costly and ineffective. The use of n-gram analysis offers efficiency and correctness in the analysis of malicious code.

Another approach that use n-gram analysis for virus detection (Reddy & Pujari 2006).

They merged some classifiers with the use of Dempster Shafer theory such as SVM, IBK and Decision Tree to get accurate classifiers rather than use one Theory. However using n-gram analysis for virus detection lacks semantic awareness. As a result they had difficulties to analysis the appropriate n-gram they find.

There is also an n-gram analysis method used for automatically detect malicious code (Zhang et al. 2007). Experiments were carried out by collecting 201 different windows executable files (109 benign codes and 92 malicious code). The result showed that by using n-gram method they could successfully distinguish between malicious code and benign code.

Other research is in area of anomaly detection that use n-gram modelling to create normal profile (Hubballi, Biswas & Nandi 2011). They outline an investigation to build a program based on anomaly detection by use of Occurrence Frequency model. The model is effective in short system call sequences. In addition an effect of the method is to build normal program model that can be applied in some level of infection in the training dataset. The authors also mentioned the advantage of the method is that the detection becomes immune to accidental 'infection' in the training dataset. n-gram Tree Method is an effective way to create a profile of normal behaviour. The benefits of n-gram tree analysis are that it is easy to use and fast operational.

2.2.3 Anomaly Detection

The purpose of the project is to apply anomaly detection over user profiles. According to Grzech (2006) anomaly detection refer to a fundamental of intrusion detection system. In addition, we look at other work to compare how profiling and identifying for anomaly detection use different methods, which we will explore here.

Grzech (2006) examines different architectures of anomaly detection - it could be as multiagent systems that support classification system to determine the activity as normal or abnormal detection. The author also provides the simple illustration hierarchical architecture of a spread anomaly detection system, which it is possible to implement in the structure of a multiagent decision supporting system. The author explains detail about Anomaly detection which divided in two categories are normal and abnormal. The example of hierarchical anomaly detection system provided to give brief example.

There is research that investigates the anomaly based intrusion detection in Linux Kernel (Almassian, Azmi & Berenji 2009). A sufficient feature list has been arranged to difference between normal and abnormal behaviour. The model used is to introduce new tools to the Linux kernel as protection module that function to log initial data to prepare features list. Recognize and classified input vector was used support vector machine (SVM). The evaluation was implemented on the research by use three experiments, including one-class SVM, Binary Classification and Sequence of delayed samples, however future work is not discussed for further study.

Work by Wang et al (2004) outline the used of non-negative matrix factorization (NMF) to profile user behaviours and normal program for anomaly detection. This new manner audit data flowed to system call and used UNIX commands as information source. This manner telling the normal program and user behaviour was build according to deviation and features, from user behaviour and normal program above a predetermined threshold is call anomaly detection. The authors also implemented methods to test with the system call data from AT& T research lab, Unix command data and University of New Mexico. As a result, the aim of the method is improved computational expense, detection accuracy and carrying out as real time intrusion detection. The advantages and disadvantages of NMF are also mentioned for a comparison to get effective result from the analysis.

2.3 N-Gram Analysis

An n-gram is a contiguous sequence of n letters, words or phonemes. For example size 1 of n-gram refer to unigram, size 2 of n-gram refer to bigram, size 3 of n-gram refer to trigram, size 4 refer to four-gram and in the general case is called an n-gram.

An n-gram analysis is able to count the frequency of n-grams in a given file. For example, in the binary string level 3-gram such as 1110010000101010010000 has the following character-level trigrams

```
111, 110, 100, 001, 010, 100, 000,
000, 001, 010, 101,
010,..................000
```

And in the sentences “in this work we aim to get the certain knowledge”, has the following word-level 3-grams:

```
In this work
This work we
Work we aim
We aim to
Aim to get
To get the
Get the certain
The certain knowledge,
```

And for the phrase “in this work”, has the following character-level 3-grams:

```
In ,n t,  th, thi, his, is , s w,
wo, wor, ork
```

This research uses varying sizes of n-gram such as 3-gram, 5-gram, 11-gram and 15-gram. Firstly, we will evaluate the use of n-gram analysis of user generated formal language such as their command line histories to profile users’ command usage in their command line histories. Secondly, we will evaluate the use of n-gram analysis of natural language to profile users and to ensure the accurate user identification. After that we will compare each writing style from each user and see how different or significance of their pattern in term of natural language and formal language. Next we will visualise their n-gram patterns graphically to view their frequency pattern.

2.3.1 User Samples

In this part we analyze two forms of writing style: natural language and formal language.

2.3.2 Natural Language

There were two famous authors used in the natural language identification. Firstly, William Shakespeare, we take tree famous novel and one poem from his writing. They are Romeo and Juliet, Julius Caesar, Sonnets and Hamlet. Secondly, we take Jane Austen’s writings Emma, Pride and Prejudice, Sense and Sensibility and Mansfield Park. The figure below shows how we compare the samples:

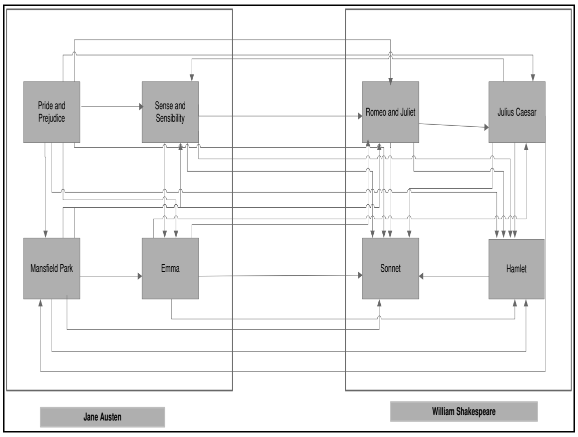

Figure 1. Method for Comparison of Natural Language Samples

Figure 1 shows how we compare both authors' writing styles. Firstly, we will see the result of one author. We compare each of Jane Austen's writings to each other, using 3-gram, 5-gram, 11-gram and 15-gram analyses. We then use the t-test to measure their similarity and if the t-test for both pairs in each comparison shows they are from the same author, we have successfully performed a positive identification. Secondly, we will do the same procedure for William Shakespeare's writings.

We will then compare the writing styles of each of Jane Austen's works with each of Shakespeare's and if the t-tests indicate they are different, then we will have successfully performed a negative identification.

3. RESULT

3.1 Positive Identification Using Natural Language Result of 3 Gram

Success and fail for positive identification of T-Test with the same participant for Jane Austin (1 person) Normalization a percentage of n-gram count

Table 1. Positive Identification Result of 3-Gr

Novel 1	Novel2	Z score norma -liza-tion	Max-min normaliza-tion	Z score normali-zation
Pride and Prejudice	Sense and Sensibility	1	2.52E-22	1
Pride and Prejudice	Mansfield Park	1	0.170681	1
Pride and Prejudice	Emma	1	1E-18	1
Mansfield Park	Sense and Sensibility	1	0.533889	1
Mansfield Park	Mansfield Park	1	1.49E-24	1
Sense and Sensibility	Mansfield Park	1	1.04E-39	1

As show on the table above the result of normalization a percentage of n-gram count and Z score show 100% they are the same person. However Max-min normalisation shows that Pride and Prejudice vs Mansfield Park, Pride and Prejudice vs Emma and Emma vs Mansfield Park are the same but the others are different.

3.2 Positive Identification Using Natural Language Result of 5 Gram

Success and fail for positive identification of T-Test with the same participant for Jane Austin (1 person)

Table 2. Positive Identification Result of 5-Gr

Novel 1	Novel2	Z score norma-lization	Max-min normali zation	Z score normali zation
Pride and Prejudice	Sense and Sensibility	1	0	1
Pride and Prejudice	Mansfield Park	1	4.4E-146	1
Pride and Prejudice	Emma	1	2.8E-108	1
Mansfield Park	Sense and Sensibility	1	1E-175	1
Mansfield Park	Mansfield Park	1	8.24E-05	1
Sense and Sensibility	Mansfield Park	1	1.2E-164	1

As show on the table above the result of normalization a percentage of n-gram count and Z score show 100% they are the same person. However Max-min normalisation shows that only Pride Prejudice vs Sense and Sensibility are the same but the others are different.

3.3 Positive Identification Using Natural Language Result of 11 Gram

Success and fail for positive identification of T-Test with the same participant for Jane Austen (1 person).

As show on the table above the result of normalization a percentage of n-gram count and Z score show 100% they are the same person. However Max-min normalisation shows that Pride and Prejudice vs sense and Sensibility, Pride and Prejudice vs Mansfield Park and Pride and Prejudice vs Emma are the same person but the other are different person.

Table 3. Positive Identification result of 11-gr

Novel 1	Novel2	Z score normali -zation	Max-min normali-zation	Z score norm alizati on
Pride and Prejudice	Sense and Sensibility	1	0	1
Pride and Prejudice	Mansfield Park	1	0	1
Pride and Prejudice	Emma	1	0	1
Mansfield Park	Sense and Sensibility	1	9.55E-71	1
Mansfield Park	Mansfield Park	1	2E-99	1
Sense and Sensibility	Mansfield Park	1	0.035547	1

3.4 Positive Identification Using Natural Language Result of 15 Gram

Success and fail for positive identification of T-Test with the same participant for Jane Austin (1 person)

Table 4. Positive Identification Result of 15-gr

Novel 1	Novel2	Z score norma-lization	Max-min normali-zation	Z score norma-lization
Pride and Prejudice	Sense and Sensibility	1	0	1
Pride and Prejudice	Mansfield Park	1	0	1
Pride and Prejudice	Emma	1	0	1
Mansfield Park	Sense and Sensibility	1	5.16E-29	1
Mansfield Park	Mansfield Park	1	1.75E-20	1
Sense and Sensibility	Mansfield Park	1	1.58E-119	1

4. CONCLUSION AND FUTURE

In this research, we investigate user writing styles which aims to be able to identify users positively. We investigate formal language language by use n-gram methodology. We compare the result of n-gram analyses from each participant and assess how successful this comparison by use t-test for paired two samples for means. For natural language, the n-gram analysis is successful for positive identification but not for negative identification.

Research question 1: does the user of n-gram analysis to profile users' command usage in their command line histories allow accurate user identification?

Table 5. Success Total

Normalization Type	Success Total
Percentage	100%
Max-min	16.66%
z Score	100%

The investigation of positive investigations is show successfully for identify the same user especially for a percentage and z score normalization both normalization show 100% (out of 100%) success but not for max-min normalization only 16.66% (out of 100%).

Research question 3: if the profiling of both writing styles and command usage allow accurate user profiling, which is the most accurate?

According to investigation result, natural language show accurate information in term of investigation user profiling for percentage and Z score normalization but not for Max-min normalization.

REFERENCES

Almassian, N, Azmi, R & Berenji, S. 2009. *'AIDSLK: An Anomaly Based Intrusion Detection System in Linux Kernel'*, Information Systems, Technology and Management: pp. 232-243.

Ashman, H & Holland, S. *'Profiling and identifying users with n-gram analysis on their command line histories'* (in draft).

Grzech, A. 2006. *Anomaly detection in distributed computer communication systems.* Cybernetics and Systems, 37: 635-652.

HUBBALLI, N., BISWAS, S. & NANDI, S. 2011. *Sequencegram: n-gram modeling of system calls for program based anomaly detection.* In: Communication Systems and Networks (COMSNETS), 2011 Third International Conference on, 4-8 Jan. 2011 2011: 1-10.

Maia, M, Almeida, J, Virg & Almeida, l. 2008, *'Identifying user behavior in online social networks'.* Proceedings of the 1st Workshop on Social Network Systems, Glasgow, Scotland.

N.P.Dau, V, Rau, V & J.Templeton, S. 2000. *'Profiling users in the UNIX OS Environment'.*

Pannell, G & Ashman, H. 2010, *'User Modelling for Exclusion and Anomaly Detection: A Behavioural Intrusion Detection System'.* in De Bra, P, Kobsa, A & Chin, D (eds), User Modeling, Adaptation, and Personalization, vol. 6075, Springer Berlin / Heidelberg: pp. 207-218.

Reddy, DKS & Pujari, AK. 2006, *'n-gram analysis for computer virus detection'.* Journal in Computer Virology, vol. 2, no. 3: pp. 231-239.

WEI, W., XIAOHONG, G. & XIANGLIANG, Z. 2004. *Profiling program and user behaviors for anomaly intrusion detection based on non-negative matrix factorization. Decision and Control*, 2004. CDC. 43rd IEEE Conference on: 14-17 Dec. 2004 2004. 99-104 Vol.1.