IDENTIFICATION AND ANALYSIS OF INDIVIDUAL USERS ON NATURAL LANGUAGE WITH N-GRAM ANALYSIS METHODOLOGY

Imamulhakim Syahid Putra, Bochari Rachman, Afriyudi

Universitas Bina Darma
e-mail: putratevez10@gmail.com

Abstrak

The user profile is one approach to identify intruders on a computer system. User profiles are very important in all applications, especially to identify specific information about the user itself. Basically profiling build or information about the user or the user experience. So, User Profile be used to collect information related to user activity. This study focuses on the user psychometric writing style that identifies the user based on natural language. The purpose of this study was to facilitate the identification of the user based on the style of writing. So it can detect people who want to attack the computer security system. The methodology used in this study are N-Gram. N-Gram analysis can identify the user to accurately seen from the results achieved from any type of N-grams are used. In this study the N-Gram will compare each writing activity so that the user can determine the type of user authorization.

Kata kunci: *N-Gram, User Profil, Identification, Information, Writing*

1 INTRODUCTION

One of the most popular internet application nowadays are social media sites [1] or Social Networking. The birth of social networking sites has changed the way users communicate. These internet sites attract billions of cyber users [2]. Social networking media, is a media that is widely used to access the information [3]. The use of the website not only focus on the distribution of information, building an online business but this time the website has been used for interaction between users. Currently the social networking site has been popularly used as the interaction between user and interactive web user who lead a more active developing of websites such as Facebook and MSN is the revolution of Web 2.0 [4]. Social networking is required for some activities such as profiles and identification of user habits [5] , in addition to the computer security is a branch of technology known as information security as applied to computer. One of the goals of computer security, among others, as a protection against theft of information, data errors, maintenance and availability of information.

Computer security today is not only limited to the scope of the hardware, software and users but it has grown on the side of the user interaction between these three components.

Such as the security of the collection of information about a user or users on social networking sites that might be expected to obtain accurate information about the user or simply called User Profiling. User profiling is a way to gather personal information or individual to category-specific category based on the characteristics of such a situation, the appearance and properties [6].

This study continues previous research studies conducted by Darussalam, entitled "Profiling and Individual Identifying Users by Their Command Line Usage and Writing Style", at the University of South Australia, the contribution of this study is to analyze the writing style on Natural language and Formal Language with the aim to allow to detect intruders posing as real users.

The aim of this study is also to apply to the user profile anomaly detection, anomaly detection is the main ingredient of the intrusion detection system. In this study, will see some journals have been conducting research on anomaly detection associated with the user profile [7].

From the explanation above, the writer uses the N-gram analysis as a method to be used in this study. N-gram analysis method is used to retrieve pieces of n pieces of character letters of a word that continuity is read from the source text to the end of the document. N-gram analysis is one method that will be used in this study. N-gram method is' the language model based collinear relationship [8]. Collinear relationship is a piece of n-characters from a string. N-gram method is used to retrieve pieces of n pieces of character letters of a word that continuity is read from the source text to the end of the document [9].

One-on-one advantages of using the N-gram and not a whole word as a whole is that the N-gram is not overly sensitive to writing errors contained in a document [10]. With the N-gram analysis method is also expected to be able to distinguish whether the user is the author of A has a different writing styles with user B is called Negative Identification. Or conversely the same writing style user, user A can be identified (same or similar) by comparing other writings of the user A or called Positive Identification. Mapping posts of users will use two methods, namely graphs and t-test. From the graph can be seen if the same user or a different user has kesamaaan or differences in their writing style. Then to get measurable results, the authors compare the writing style of each user by using t-test tools.

Based on the above background, the author outlines are still some problems that arise, especially on social networking users such as blogs, twitter, and facebook as the difficulty of proving the safety of the personal data and messages published on social networking user. Where information or messages published by users certainly not altered or changed. The benefits gained from this research area: 1) It is expected to be useful for identifying characteristics psycometrik user profile on the intrusion detection system, and 2) Distinguish between positive and negative identification.

The title of this research is "Identification and Analysis of Individual Users On Natural Language With N-Gram Analysis Methodology". Here the researcher defines a problem that this discussion does not widen too far off the mark that will facilitate the discussion and preparation of this thesis. By using the methods of N-Gram analysis, to identify the user and the style of writing in the papers of William Shakespeare and blogs Oprah Winfrey in two ways: positive and negative identification.

2 RESEARCH METHODOLOGY

2.1 Design and Research

In this study, the authors used the help of software applications using Java programming language. There are two classes in this application is "Ngram.java" and "Data.java". This study author takes from January to August 2015.

2.2 Populasi dan Sampel

Populasi dari penelitian ini adalah seluruh pegawai dan dosen serta mahasiswa Fakultas Teknik Universitas Sriwijaya yang diperkirakan Jumlah tenaga Dosen 192 orang, Jumlah pegawai. Agar tercipta efisiensi, maka digunakan metode sampling. Dan mengenai hal ini ditetapkan sebesar 10% sebagai sampel yaitu 280 orang dan seluruhnya dijadikan responden yang penetapannya digunakan metode simple random sampling.

2.3 Data Research

The data used as a sample in this study is Oprah Winfrey Blog and Novel of William Shakespeare. Data were taken through a common website then the author compares the data after processing through the N-Gram analysis. The author will investigate whether these blogs belong to the same person or different that people posing as Opera Winfrey or William Shakespeare.

2.4 Concepts and Methods

This study aims to identify the user profile, especially on social networking sites like blogs or tweets (Natural Language). In this study, the authors use the method of N-Gram analysis to identify the user (author identification). In addition this study will compare the results of identification. Furthermore, this study also evaluates how accurately the N-Gram analysis can identify the characteristics of the user. The author uses java applications as one of the software used to calculate the N-Gram.

2.5 Data Collection Technique

There are several research methods that I use to complete the stages of activities in the framework of the research. The quality of the results will be known by what method is used to obtain data and information relating to research to be conducted. In this case the author uses several methods of data collection and data analysis as follows: 1) Primary Data. Ie data collected comes from the research field studies, collect data directly (from Novel and Blog) to the problem of what is going on to become the object of meticulous research in particular by observation method (observation). Namely to obtain more precise data and in accordance with the information obtained, then I make observations directly to the field to find out the real situation in order to get an accurate assessment, and 2) Secondary Data. Ie data collected by studying literature, that is by using qualitative data.

Table 1: Positive Identification 2-gram

Percentage 2-gram		
Week1	Week2	1
Week1	Week3	1
Week2	Week3	1

Table 2: Positive Identification 2-gram

Z-score 2-gram		
week1	week2	1
week1	week3	1
week2	week3	1

3 RESULTS AND DISCUSSION

3.1 Results

In this research, we investigate user writing styles which aims to be able to identify users positively and negatively. We investigate formal language and natural language by use n-gram methodology. There are five participants in formal language and two famous writers for natural language. 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. The result shows that formal language can identify users in term of positive and negative identification. However, for natural language, the n-gram analysis is successful for positive identification but not for negative identification. Thus, formal language is shown to be more generally accurate. In this research the author uses a sample on Oprah Winfrey Blog.

In Table 1 and 2 show the results of a comparison between week1 week vs. 2 vs. week3 week1 and week2 vs week3. Based on the style of writing after the data is normalized, using two methods, namely normalization Percentage and Z score normalization. The data has been normalized later, the similarity test using T-Test Pair two sample for means. The result of the comparison states that the data are 100% the same.

This main objective of this research is to identify a users writing style in term of formal language and natural language especially for user reauthentication. This research used n-gram methods to assess the user in two writing styles. The investigation used software that created from previous researcher for counting n-gram. The software will store the counting gram from the history file of the participants into excel file. After that we use excel to create n-gram spectrum and from n-gram spectrum we compare each participants to see the similarity and different their writing style. Further experiment have to be made for continue the investigation, as follows.

Firstly, for formal language we can investigate by divided to period of time for instance per month or week, rather than compare different machine in different work place. It is because working in office or home could be different in term of psychologies.

Secondly, we should try another gram, such as 1,2,4,6,7,8,9,10,12,13, since every gram length appears to show a different result, and another gram length could give a more accurate

result for formal and natural language. Thirdly, we can assume one person could have more than one writing style. Try to collect their writing style and match it to each other users writing style.

Finally, another analysis should be to compare each of the participants in both directions. For example, after compare A person writing style to B person writing style, we should swap the order, so B must compared to A as well. Thus, from both result we can compare it and see how the result.

3.2 Discussion

In a study conducted by taking samples of the novel of William Shakespeare and Blog Oprah Winfrey. Data were taken through a common website then the author compares the data after processing through the N-Gram analysis, with discussion as follows:

1. Analysis of the N-gram method using the help of software applications using Java programming language. There are two classes in this application is "Ngram.java" and "Data.java". This program can be run with the command "java Ngram [n]" where n is the value or amount of N-Gram. This software will generate the N-gram frequencies will be placed in "csv" folder in Microsoft Excel and "csv" folder is present in the form of txt history data.

2. The use of the N-gram analysis offers the efficiency and accuracy in the analysis and is not too sensitive to writing errors contained in a document.

4 CONCLUSION

This research method using N-gram analysis taking samples on the novel of William Shakespeare and Blog Oprah Winfrey, the use of N-gram analysis is less sensitive to error in writing contained in a document.This study was to analyze the writing style and Formal Language Natural language with the aim to allow to detect intruders posing as real users.

Referensi

[2] L. A. Abdillah, 2014, Social media as political party campaign in Indonesia, *Jurnal Ilmiah MATRIK*, vol. *16*, pp. 1-10, 2014.

[2] L. A. Abdillah, 2014, Indonesian's presidential social media campaigns, in *Seminar Nasional Sistem Informasi Indonesia (SESINDO2014)*, ITS, Surabaya.

[3] D. R. Rahadi and L. A. Abdillah, 2013, The utilization of social networking as promotion media (Case study: Handicraft business in Palembang),in *Seminar Nasional Sistem Informasi Indonesia (SESINDO2013)*, Inna Grand Bali Beach Sanur & STIKOM Bali, Bali, pp. 671-676.

[4] Vosecky, J., DAN, H., & Shen, V. Y., 2009, User identification across multiple social networks. Networked Digital Technologies. *First International Conference*, (pp. 360-365).

[5] Raad, C., & Dipanda., 2010, *User profile matching in social networks.* 297-304.

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

[7] Grzech, A., 2006, Anomaly detection in distributed computer communication systems. *Cybernetics and Systems.* 635-652.

[8] Luo, F., OU, Q., & Wei, G., 2010, Research on n-gram-based malicious code feature extraction algorithm. *Computer Application and System Modeling (ICCASM),International Conference* , V6-89-V6-92.

[9] Sugianto, S. A., & Liliana, R. S., 2013, Pembuatan Aplikasi Predictive Text Menggunakan Metode N-Gram-Based. *Jurnal Infra* , Volume *1* No 2, 125-130.

[10] Mustafa, S., 2004, *Character Contiguity in N-gram-based Word Matching: the Case for Arabic Text Searching.* Jordan: Department of Computer Information System Faculty of Information Technology Yarmouk University Jordan.