Knowledge Engineering and Data Science (KEDS)
Vol 1, No 2, September 2018, pp. 55–63

pISSN 2597-4602
eISSN 2597-4637

Profiling and Identifying Individual Users by Their Command Line Usage and Writing Style

Darusalam [a, 1, *], Helen Ashman [b, 2],

[a] *Department of Technology, Policy and Management, Delft University of Technology*
Building 31 Jaffalaan 5, Delft and 2628 BX Delft, Netherlands
[b] *School of Information Technology and Mathematical Sciences, University of South Australia*
Mawson Lakes Campus(D3-13), Adelaide, South Australia 5095, Australia
[1] *d.darusalam@tudelft.nl*;* [2] *helen.ashman@unisa.edu.au*
* *corresponding author*

ARTICLE INFO

Article history:
Received 31 May 2018
Revised 10 July 2018
Accepted 10 July 2018
Published online 31 August 2018

Keywords:
Profiling
User Identifying
Intrusions
detection
Identification
N-Gram

ABSTRACT

Profiling and identifying individual users is an approach for intrusion detection 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. This research extends previous research on re-authenticating users with their user profiles. This research focuses on the potential to add psychometric user characteristics into the user model so as to be able to detect unauthorized users who may be masquerading as a genuine user. There are five participants involved in the investigation for formal language user identification. Additionally, we analyze the natural language of two famous writers, Jane Austen & William Shakespeare, in their written works to determine if the same principles can be applied to natural language use. This research used the n-gram analysis method for characterizing 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 formal language and natural language as a user profile characteristic that could enable intrusion detection where intruders masquerade as real users.

I. Introduction

Profiling is a way of grouping things or individuals into categories or groups based on characteristics such as situation, appearance, traits [1]. The term *profiling* means to get information about user's activities, and it is possible to perform anomaly detection over a user profile to allow user identification.

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* [2]. According to Vosecky *et al*. [3] 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. Other work involves the social networking technologies used for identifying user behavior [2]. Pannell and Ashman [4] evaluate an intrusion detection system (IDS) to analyze user's activities, and propose to help an administrator to identify and quickly respond to the intrusion.

Other work discusses the use of behavioral biometrics for intrusion detection applications [5]. Another interesting research by Pepyne *et al.* [6], analyses user profiling for computer security on particular users such as insurance adjusters and bank tellers Another work investigates user profiling based on tag based in social media recommendation [7]. A further work outlines the purpose of user profile as gather related information based on user interests [8]. The ontology-based semantic similarity method is used to extend and sustain a user profile based on web access behaviour of user

https://doi.org/10.17977/um018v1i22018p55-63

in music domain. However the lack of data support to justify whether the method is effective and no future work is discussed to develop the research.

One of the researches is to identify individual user by process profiling [9]. Another research from [1] examined the UNIX operating system to identify the user based on the login host, the login time, the command set and command set execution time of the profiled user. Another piece of related research in regarding identifying user in social network that concern in trust and privacy [10].

Another interesting research focuses on the connection of network topology and semantic similarity of user keywords [11]. Categories of keyword and the notion of the distance among multiple categories trees and keyword across were used in a forest model. Work by Takeda *et al* [12] outline characteristic expression in literary work. Their problem is, take literary work as positive examples (first writer) and negative examples of works by another writer especially in Japanese Poems (Waka poems and prose texts). The method is to create a sequence list of substring of writer goodness.

There is also research about the misuse of social networks for Automated User Profiling [13]. They analyze users' weakness when registered to the social networking such as Facebook by used their email address. Other research investigates how to identify a user based on similar profile [3]. They used social networking such as MSN and Facebook to collect user profiles. User's profiles are used to create tools especially for a profile comparison, to decide if similar profile is belonging to the same person or not. The use of vector-based comparison algorithm is a method to compare each user profile.

This research will focus on evaluating the potential of two psychometric user characteristics, namely writing style in both Natural Language (Jane Austin & William Shakespeare) and Formal language (command line histories). 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 [14], 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 by profiling program and user behaviors [15]. Anomaly detection over 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 [14]. However, this research will focus on one potential psychometric user characteristic and will consider whether user's writing styles in two 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 determine which is the more effective of the two.

This research will analyse 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 a negative identification. The second way is to detect whether the current user can unquestionably be verified as the true user – this is positive identification. 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.

II. Methods

This research aims to identify a user especially in Natural language (Jane Austen and William Shakespeare writing style) and Formal Language (command line history). 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. We use the software to perform four types of n-gram analysis, namely 3-gram, 5-gram, 11-gram and 15-gram.

A. 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 111001000010101010010000 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 project 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 visualize their n-gram patterns graphically to view their frequency pattern.

B. T-Test

The t-test is a method that can be performed to decide whether two data sets (samples) are similar or dissimilar and to conclude whether they could have come from the same population. It assesses whether the means of two groups are statistically different from each other. This analysis is useful when we want to determine whether the means of two groups are similar or different. We will use t-tests to assess both natural language and formal language samples, between two samples from the same user (for positive identification purposes) and between two samples from different users (for negative identification purposes).

We next consider which form of t-test is appropriate to this research.

- One sample t-test
  The one-sample t-test is used to decide whether a specific sample comes from a specific population. For example, when we want to know about a specific sample of university students is similar to or different from university students in general. In the current research we are comparing series of words or commands, and while it may later be feasible to identify a user from a single n-gram value, at this early stage, it is more appropriate to decide whether individuals can be identified from larger quantities of their writings.
- Independent t-test
  The independent t-test, or two sample t-test, is used to determine whether two samples are statistically similar or different to each other between the means in two unrelated groups. For example, when we want to know between university students female and male are different or similar to some psychological characteristics. In this research, the samples may not be unrelated, especially when comparing two samples from the same user.
- Dependent t-test
  The dependent t-test, also called the paired-group t-test, correlated-group t-test, matched-groups t-test or dependent-group t-test. This t-test is used to compare two related samples (matched or related in the same way) that are both measured once or the same sample measured on two separate occasions. For example, when we want to know how the effect of using a particular drug for insomnia, for the patient is similar or different after consuming the drug. In this case, we will see what the effects of the drug on the patient are before and after consuming the drug. This is highly suitable to this research as we need to positively identify a user by comparing a current sample of the user's writing to an older sample of their writing.

From the explanation above we conclude that the dependent t-test or paired group t-test is the most suitable method to test our investigation. We use the t-test by proposing the following competing hypotheses:

- The test hypothesis is the means of population behind the different of two samples.
- The null hypothesis is the means of population behind the similarity of two samples.

A probability value p is output by the t-test. The result of probability value is a comparison to the chosen level of significance α to conclude the test result. A common default is $\alpha = 0.05$:

- If the probability value is equal to or less than the level of significance, we can reject the null hypothesis and conclude that the two samples of writing are different
- If the probability value is more than the level of significance, we can accept the hypothesis and conclude that the two samples of writing style are the same.
- Normalization of samples

Before performing any t-test, we will see the distribution of data collection from each gram whether the distribution of data is normal or non-normal. If the data non-normal we will transform the data to the normal data. This is because the samples we are analyzing are of radically different size. While it would be possible to choose subsamples from each sample so that each subsample is an identical size, we elected to normalize each whole sample instead, as command lines users, in particular, may have different tasks at different times, and the subsamples may not accurately reflect the user's command line habits in a subsample. By normalizing the samples, we most accurately preserve each user's writing styles, but at the same time cast them into the same numerical range so that different sample sizes do not confound the results.

Fig. 1 shows three types of normalization, we use to make normal distribution: a percentage normalization, max-min and Z score. Firstly, percentage normalization is counting each value of the n-gram and divided by total value of all n-gram and time to one hundred. Secondly, max-min normalization is counting the total value all the n-gram and divided by a total number from the reduction of maximum number and minimum number of n-gram. Lastly, the z score normalization is counting each value of the gram reduction by average total all the n-gram and divided by standard deviation. We will assess all three normalization methods in this research to determine which is most appropriate for the task of identifying users.

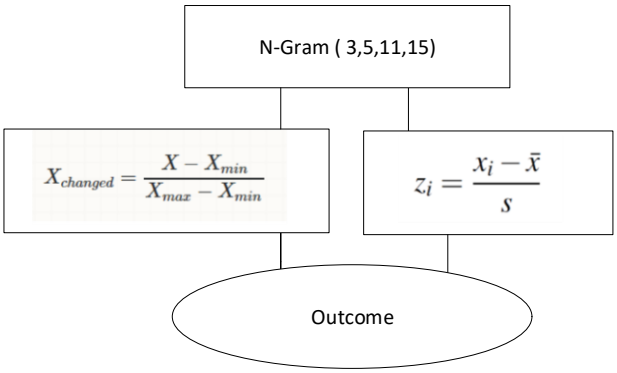

Fig. 1. Normalization process

III. Results and Discussions

A. Natural Language

Fig. 2 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.

B. Formal Language

There were five users involved in this experiment (Fig. 3). One example of formal language is command line history, where the user interacts with the computer through a command line. By use n-gram analysis we will identify those user's 'writing style', namely their command line usage habits. The figure below shows how we compare each of our formal language samples:

We will follow the same procedure for formal language as for natural language. Namely we will analyze each sample, and compare samples by the same user to see if we can achieve positive identification. We will then compare the samples from different users to see if we can achieve negative identification.

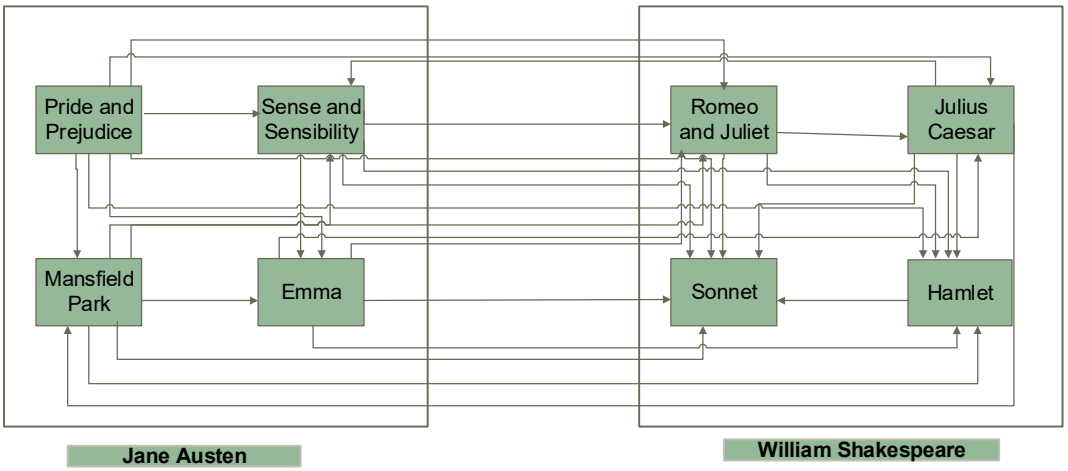

Fig. 2 method for comparison of natural language samples

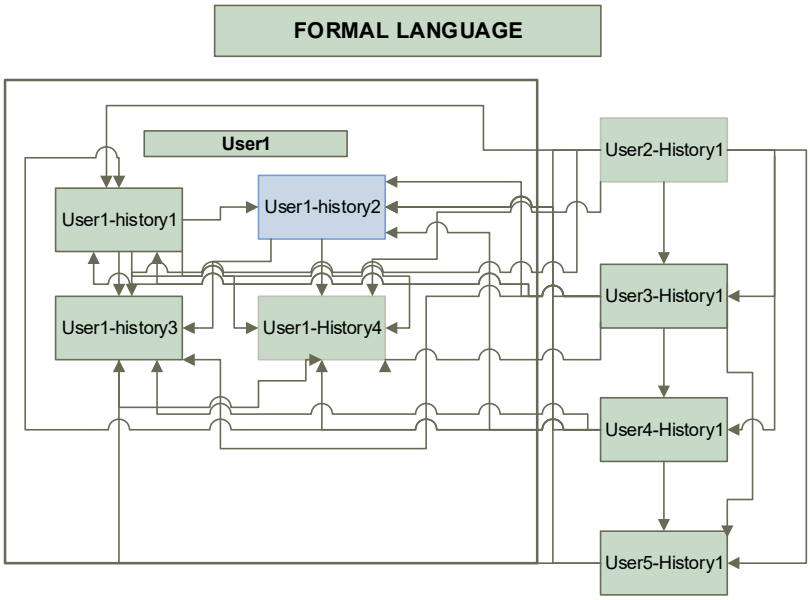

Fig. 3. Method for comparison of formal language samples

C. Summary of Formal Language

1) Positive Identification

Table 1 shows positive identification summary for formal language, for all six possible pairings of samples, and for the four different n-gram lengths, calculated for each of the three normalization methods. For all four n-gram lengths, we find that the Percentage and Z score normalization methods correctly identify that the user is the same in each case. However, the Max-Min normalization method fails to identify that the user is the same in all but one case for each n-gram length. These results suggest that positive identification can be reliably achieved using the n-gram analysis method for formal language, using either the Percentage or Z score normalization methods. However, it indicates also that the Max-Min normalization method is not useful for positive identification in formal language samples.

2) Negative Identification

Table 2 is a negative identification summary for formal language, for all possible pairings of samples, and for the four different n-gram lengths, calculated for each of the three normalization methods. The results are less clear than were observed in the positive identification table. The Max-

Table 1. Positive identification summary of formal language

n-gram	Normalization Type	Correct Identification	Incorrect Identification	Rate Percentage
3 Gram	Percentage	6/6	0/6	100 %
	Max-Min	1/6	5/6	16.6 %
	Z Score	6/6	0/6	100 %
5 Gram	Percentage	6/6	0/6	100 %
	Max-Min	1/6	5/6	16.6 %
	Z Score	6/6	0/6	100 %
11 Gram	Percentage	6/6	0/6	100 %
	Max-Min	1/6	5/6	16.6 %
	Z Score	6/6	0/6	100 %
15 Gram	Percentage	6/6	0/6	100 %
	Max-Min	1/6	5/6	16.6 %
	Z Score	6/6	0/6	100 %

Table 2. Negative identification summary of formal language

n-gram	Normalization Type	Correct Identification	Incorrect Identification	Rate Percentage
3 Gram	Percentage	23/28	5/28	82.14 %
	Max-Min	20/28	8/28	71.43 %
	Z Score	19/28	9/28	67.86 %
5 Gram	Percentage	13/28	15/28	46.43 %
	Max-Min	17/28	11/28	60.71 %
	Z Score	14/28	14/28	50.00 %
11 Gram	Percentage	16/28	12/28	57.14 %
	Max-Min	26/28	2/28	92.86 %
	Z Score	16/28	14/28	57.14 %
15 Gram	Percentage	23/28	5/28	82.14 %
	Max-Min	24/28	4/28	85.71 %
	Z Score	24/28	4/28	85.71 %

Min normalization method is correct between 60.71 % (out of 100 %) and 92.86 % (out of 100 %) of the time, showing an improvement over its use in positive identification. The other two normalization methods were not as reliable as in the positive identification tests

D. Summary of Natural Language

1) Positive Identification

Table 3 is a positive identification summary for natural language, for all possible pairings of same-author samples, and for the four different n-gram lengths, calculated for each of the three normalization methods. Firstly, for the 3-gram percentage normalization and Z score the percentage rate is 100 % success. However, Max-min's percentage rate only 11.11 % (out of 100 %) similarity for the paired comparison in User1's command line history for a different machine. It means that Max-Min normalization fails to identify positive Identification. Secondly, for 5-gram, 11-gram and 15-gram the percentage normalization and Z score are 100 % (out of 100 %) success for positive identification. On the other hand, the Max-min gives a different result for each gram. For instance 5-gram show 11.11 % (out of 100 %) same as 3-gram, 11-gram is 33.33 % (out of 100 %) and 15-gram is 50 % (out of 100 %).

2) Negative Identification

Table 4 is a negative identification summary for natural language, for all possible pairings of different-author samples, and for the four different n-gram lengths, calculated for each of the three normalization methods. The table above is the summary of negative identification shows the

Table 3. Positive identification summary of natural language

n-gram	Normalization Type	Positive Result (Correct Identification)	Negative Result (False Identification)	Rate Percentage
3 Gram	Percentage	18/18	0/18	100 %
	Max-Min	2/18	16/18	11.11 %
	Z Score	18/18	0/18	100 %
5 Gram	Percentage	18/18	0/18	100 %
	Max-Min	2/18	16/18	11.11 %
	Z Score	18/6	0/18	100 %
11 Gram	Percentage	18/18	0/18	100 %
	Max-Min	6/18	12/18	33.33 %
	Z Score	18/18	0/18	100 %
15 Gram	Percentage	18/18	0/18	100 %
	Max-Min	9/18	9/18	50 %
	Z Score	18/18	0/18	100 %

Table 4. Negative identification summary of natural language

n-gram	Normalization Type	Positive Result (Correct Identification)	Negative Result (False Identification)	Rate Percentage
3 Gram	Percentage	0/16	16/16	0 %
	Max-Min	16/16	0/16	100 %
	Z Score	0/16	16/16	0 %
5 Gram	Percentage	0/16	16/16	0 %
	Max-Min	16/16	0/16	100 %
	Z Score	0/16	16/16	0 %
11 Gram	Percentage	0/16	16/16	0 %
	Max-Min	16/16	0/16	100 %
	Z Score	0/16	16/16	0 %
15 Gram	Percentage	0/16	16/16	0 %
	Max-Min	2/16	14/16	12.5 %
	Z Score	0/16	16/16	0 %

unsatisfactory result for each gram. The table 4 negative identification summary shows how negative identification for natural language fails for user identification. However, for Max-min normalization especially for 3-gram, 5-gram, and 11-gram show that we success 100 % (out of 100 %) to identify the negative identification. However, we cannot trust max-min normalization since in both formal language and natural language max-min normalization show the result always different.

IV. Conclusion

In this research, we investigate user writing styles which aim 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 using a 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. Further experiment has to be made to continue the investigation, as follows. Firstly, for the formal language, we can investigate by dividing the period of time for instance per month or week, rather than compare on different machines in the different work place. 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.

References

[1] V. N.P.Dau, *et al.*, "profiling users in the UNIX OS Environment," 2000.

[2] M. Maia, *et al.*, "Identifying user behavior in online social networks," presented at the Proceedings of the 1st Workshop on Social Network Systems, Glasgow, Scotland, 2008.

[3] J. Vosecky, *et al.*, "User identification across multiple social networks," in *Networked Digital Technologies, 2009. NDT '09. First International Conference on*, 2009, pp. 360-365.

[4] G. Pannell and H. Ashman, "User Modelling for Exclusion and Anomaly Detection: A Behavioural Intrusion Detection System," Berlin, Heidelberg, 2010, pp. 207-218.

[5] A. A. E. Ahmed and I. Traore, "Detecting computer intrusions using behavioral biometrics," 2005.

[6] D. L. Pepyne, *et al.*, "User profiling for computer security," in *Proceedings of the 2004 American Control Conference*, 2004, pp. 982-987 vol.2.

[7] C. C. Hung, *et al.*, "Tag-Based user profiling for social media recommendation," 2008.

[8] M. Reformat and S. K. Golmohammadi, "Updating user profile using ontology-based semantic similarity," in *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*, 2009, pp. 1062-1067.

[9] S. McKinney and D. S. Reeves, "User identification via process profiling: extended abstract," presented at the Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies, Oak Ridge, Tennessee, 2009.

[10] C. Dwyer, *et al.*, "Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace," 2007.

[11] P. Bhattacharyya, *et al.*, "Social Network Model Based on Keyword Categorization," in *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, 2009, pp. 170-175.

[12] M. Takeda, *et al.*, "Discovering Characteristic Expressions from Literary Works: a New Text Analysis Method beyond N-Gram Statistics and KWIC," Berlin, Heidelberg, 2000, pp. 112-126.

[13] M. Balduzzi, *et al.*, "Abusing Social Networks for Automated User Profiling," in *Recent Advances in Intrusion Detection*. vol. 6307, S. Jha, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 422-441.

[14] G. Pannell and H. Ashman, "User Modelling for Exclusion and Anomaly Detection: A Behavioural Intrusion Detection System," in *User Modeling, Adaptation, and Personalization*. vol. 6075, P. De Bra, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 207-218.

[15] W. Wei, *et al.*, "Profiling program and user behaviors for anomaly intrusion detection based on non-negative matrix factorization," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, 2004, pp. 99-104 Vol.1.