A Behavioural biometrics based on stylometrics in cloud computing domain for cybercrime detection

Literature Review – Rough Draft

Neo Thokoa

COS 700 - Research

Cloud Computing statistically benefits companies and individuals by saving time, money and resources compared to traditional on-premise computers. However, with this burgeoning amount of benefits in cloud computing, there comes several critical issues that have been seen as drawbacks to cloud computing and growing popularity of virtualization among companies could lead them to being the next possible target of cyber criminals. Therefore, if prevention mechanisms, such as employing a solid cyber security systems to protect vendors from intrusion and also help protect user privacy, fail then we need to look at detection mechanisms to help identify perpetrators which can assist in the prosecution of these cyber criminals. Since there already exists a method utilising the analysis of behavioural biometrics to profile a suspect based on the manner in which they executed the crime in a physical world context, this paper will attempt to utilise the same philosophy to assist profile and identify a suspect that is performing cybercrime on the public cloud. This paper will look to propose a stylometric attribution model for user-identification in public cloud environment.

In a research article by Ahmed Abbasi (2008), Stylometry is introduced and a shortcoming of online stylometry is examined (ABBASI & CHEN, 2008). The shortcoming experienced is the lack of scalability in terms of number of authors and across application domains such as email, forums and chats. The paper also highlights that prior to the time it was written, techniques highly focussed on the identification rather than the profiling of these individuals. Research gaps that were identified and can be still relevant are Similarity detection. This is because most studies have been focused on the identification task rather than looking at similarity detection and so there is a need for models that can perform similarity detection as in the cyberspace, class definitions are not known prior to the application of the model. Furthermore, scalability across domains is in question as most work was done on a single domain which could to prove to be limited. Henceforth, analysis across domains is crucial in order to evaluate the stylometry’s across various modes of domains.

The research article proposed a writing technique by the name of Writeprints technique which is an unsupervised method. Previous methods utilised the stylometric analysis technique which require prior knowledge of author class labels (Narayanan, et al., 2012). Thus this is utilized more for identification tasks rather than the similarity tasks. This is where Unsupervised Stylometrics comes in as it makes categorizations with no prior knowledge of author classes. The shortcoming that had not been evaluated in this research paper was the researchers had not considered a very important case where the impact of intentional stylistic alterations that a cybercriminal might imploy would have on their model. The paper also highlights that a need to evaluate feature sets on individual-author-level that could gradually change after some time as a user’s writing style evolves.

A similar research article that was conducted in 2012 by Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, focused on the feasibility of Internet-Scale Author Identification, whereby with prior knowledge of writing domains, better results were reached. The problem found that on a large scale, results produced are qualitatively different. This paper however looked at identifying an anonymous author given a set of texts from 100000 authors (Brocardo, Traore, Saad, & Woungang, 2013). Thus even if it was to show a proof of concept, the results may differ on an even larger pool of users. Thus a lessoned learnt is to create a model that uses more abstract techniques that do not require a known number of users.

Moving forward, a research done by Marcelo Luiz Brocardo, Issa Traore, Sherif Saad in 2013, exploited the previous research, to narrowly focus on short messages, particularly focus on the authorship verification for short messages. The research aimed at highlighting steps to achieving supervised stylometry learning techniques, combined with a model called n-gram analysis to identify anonymous author. The paper states though that there needs further improvement on accuracy at a suitable level of authorship verification. The paper also highlights a need to address the underlying problem of increasing the robustness of the scheme.

# Bibliography

ABBASI, A., & CHEN, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*.

Brocardo, M. L., Traore, I., Saad, S., & Woungang, I. (2013). Authorship Verification for Short Messages using Stylometry. *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference*.

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C., & Song, D. (2012). On the Feasibility of Internet-Scale Author Identification. *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, 300-314 .