Author Recognition in Discussion Boards

Mohamed Faouzi Atig[a], Sofia Cassel[a], Amendra Shrestha[a] , Lisa Kaati[b]
[a]Uppsala University, Sweden
[b] FOI. Swedish Defence Research Agency, Stockholm, Sweden

Abstract- In this work we consider the problem of detecting users in a discussion board that make use of several aliases. The elementary idea is to create cluster of users who have the same activity patterns. Theses clusters are further analyzed using stylometric analysis. We have implemented our algorithm and tested it on the ICWSM dataset boards.ie.

I. Introduction And Background

With the transfer from Web 1.0 to Web 2.0 the amount of user generated content has increased. Various social media allows users to communicate and upload content. Discussion boards (also called Internet forums, web forums, message boards etc.) are one way of communication on the Internet. A discussion board is a Web application that is used to publish user generated content in the form of a discussion. Discussion boards have an important social aspect and are active for a long period of time. There are discussion boards dedicated to almost every possible human activity and therefore Internet users can find a discussion board that suits theirs interests and needs. Discussion boards can be a source of information about consumers' opinion and this is one of the reasons why the need for mining and analyzing discussion boards has increased over the years.

One important aspect when analyzing discussion boards is to identify users that make use of multiple aliases. Identifying users that make use of multiple aliases can be of interest for forensic investigations, for example in the case of suspected grooming and identify frauds.

II. Author recognition

One way of identifying user with multiple aliases is to use author recognition techniques. Author recognition is usually done by measuring textual features and use these features to distinguish between texts written by different authors known as stylometric analysis. Previous research has shown that stylometric analysis can be used to create a "writeprint" [1] that captures the unique writing style of a person. The writeprint is similar to a fingerprint since it can be used to identify authors. The basic idea that people have distinctive writing styles is well-known and well-understood, and there are lot of research done on this topic. In a large scale setting such as the Internet, it has recently been shown [1] that it is possible to detect authors based on their writeprint in up to 80% in some cases. That is a remarkable result that might have serious implications on online anonymity.

In this work we focus on recognizing user with multiple aliases on discussion boards. We consider not only the writeprint of a user, but also the activity patterns of a user. The activity pattern is calculated using the time periods when a user is most active and when the user is inactive in his/her communication on the discussion board. Using this information we create inactivity and activity clusters where we group users that have a similar behavior.

III. Results

We use 6 different activity clusters, where each cluster represents 4 hours of the day. Each such activity cluster is then divided into 6 different inactivity clusters. For each user we calculate when the user is most active and similarly when the user is inactive. We use the results to and cluster the users.
For each cluster we can calculate the writeprint of each user that is included in the cluster. The writeprints can be compared with users that belong to the same cluster. The assumption behind this is that an Internet user has the same activity and inactivity periods even if he/she communicates using different alias names or identities. This assumption arises from the idea that we tend to sleep during the same hours every day and we tend to communicate during the same hours.

We have done some experiments with the on the ICWSM dataset boards.ie. We have used a set of 1000 users. Our experiments shows that by using clustering the probability for detecting a user that make use of multiple aliases increases with more than 15 % compared to only considering stylometric analysis. Previous work [3] has also shown that analysis of the time when messages are posted can be a useful feature to recognize authors in digital environments.

References

[1] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 7:1–7:29, 2008

[2] N. Narayan, H. Paskov, N. Gong, J. Bethencourt,, E. Stefanov, E. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *IEEE Symposium on Security and Privacy*, 2012 pp.300-314.

[3] F. Johansson, L. Kaati, A. Shrestha, "Detecting Multiple Aliases in Social Media," in *International Symposium on Foundations of Open Source Intelligence and Security Informatics*, 2013.