Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 640 – 645

Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

Authorship Verification of Online Messages for Forensic Investigation

Smita Nirkhi [1], Dr.R.V.Dharaskar [2], Dr.V.M.Thakare[3]

[1]*Smita Nirkhi,Research Scholar,G.H.Raisoni College Of Engineering,Nagppur,India*
[2]*Former Director, Disha Group of Institutions, Raipur, Chattisgargh, India*

Abstract

Online messaging provides a convenient and effective means of fast communication. Along with personal communication it is being used by organizations for official communication. Many organizations make use of online messaging for exchanging sensitive and secret information. Although online messaging is used for legitimate purpose, it can be misused by various means. An attacker may masquerade as legitimate user by hijacking the connection or by conducting man-in-the-middle attack or by obtaining physical access to a user's computer. The reason behind masquerade as someone else could be spying, snooping and other malicious intentions. Unsupervised Techniques for Forensic Analysis of online messages are relatively unexplored area in Authorship Analysis research. This paper explores the application of unsupervised techniques for authorship Verification problem. The approach is based on comparing the similarity between a given unknown documents against the known documents using various features so that an unknown document can be classified as having been written by the same author.

Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: Authorship Identification ; cluster Analysis; Multidimensional Scaling; Masquerade;Cyber Forensics

1. Introduction

The use of patterns of vocabulary and grammar by an author unconsciously could be an effective discriminator to verify the authorship. Stylometry method is used to extract the linguistic patterns from available text of author to

*Smita Nirkhi. Tel.:+91-8551898648
E-mail address: smita811@gmail.com

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.111

identify the authorship[1,2]. Authorship verification problem is based on the assumption that even though the writing style of an author many change a bit over a period of time [3], each author has a inimitable writing style tendency. Forensic authorship analysis consists of attribution of authorship of a document by analyzing the writing styles or stylometric features from the textual content.

Authorship analysis is classified into three different research areas which include authorship identification, authorship verification, and authorship characterization[15]. Authorship attribution determines the most probable author of an anonymous document by comparing it with the known available documents. Authorship verification deals with examination to check whether an unknown document was written or not written by a specific individual. Authorship characterization is helpful to determine the characteristics like gender, age etc. of the author of an anonymous document. Authorship analysis of physical and electronic documents is an imminent research area and currently many researchers are working in this research area[4,7]. As mentioned in research paper[6], the verification of documents for Authorship purpose is remarkably complex than basic authorship identification and less work has been done on it. Most of the previous work focused on plagiarism detection and authorship verification did not focus on online text documents. However, authorship verification for online documents is useful to solve various criminal cases such as blackmailing and terrorist activities etc[13]. The challenges for applying the authorship identification to online messages are as follows.

- Length of Online document is short.
- An online document has less formal writing style and the vocabulary pattern is not stable.
- Online documents are different than normal text documents in composition style and in format of structure.
- Due to the internationalization of cybercrime, multilingual problems become a new challenge for authorship Analysis.

Along with above mentioned challenges, forensics analysis of online messages using authorship analysis can be potentially used for various applications. Evaluating the linguistic features of online messages & comparing them to known writing styles offers the intelligence community a tool for identifying patterns of terrorist. Authorship attribution has been applied to various cybercrimes. Examples include webpage spam[12,15], malware[11,9], pornography [14], online terrorism postings[1] web forum postings[2] and malware code[7, 6], phishing, Identity theft and masquerade.

The authorship analysis problem can be solved by two types of machine learning algorithm, the supervised and unsupervised learning. The supervised learning algorithms are also suitable when the information of both training and testing data that is the data's label is known beforehand. Thus in the classification process of the authorship identification task, the supervised learning algorithms are mainly used. The unsupervised learning algorithms do not need the information of the data beforehand. Unsupervised learning algorithms use the extracted characteristics of the samples in order to group them into different groups that sharing the similar properties. This paper proposed the use of unsupervised techniques for solving the authorship verification problem.

2. Literature Survey

Authorship Verification problem has been studied by few researchers. This section described the works by M. Koppel et al.[7] in 2004, F. Iqbal et al.[10] in 2008, X. Chen et al.[8] in 2011, and O. Canales et al.[5] in 2011. Koppel et al. proposed a method to check the dissimilarity between the anonymous document produced by the suspect and available writing samples[7]. This approach has been produced 95.70% of accuracy for documents containing at least 500 words .But when comparing it with the online messages this approach is not realistic because the length of online messages could be less than 500 words.

Iqbal et al.[10] worked on authorship verification of email messages and used 292 different features to perform

analysis of these features by using various classifiers and regression model. Enron e-mail corpus was used for conducting experimentation. It has produced Equal Error Rate in the range of 17.1% to 22.4%.
X.Chen and Hao identified 150 stylistic features for authorship verification of e-mail messages[8].He has used Enron dataset with number of authors= 40 authors .They have obtained classification accuracy of 84% for 10 emails and accuracy of 89% for 15 short e-mails.
Canales et al. performed verification on sample online test documents to verify online test takers. Keystroke dynamics and 82 stylistic features were used as feature set[5]. These features set were analyzed using a K-Nearest neighbor (KNN) classifier. The dataset contains documents of 40 students. Document size is ranging from 1710 and 70,300 characters with Equal Error Rate of 30%.

3. Applied Unsupervised Techniques

Unsupervised methods used for experimentation in this paper are, multidimensional scaling and Hierarchical clustering. The results are represented on a scatter plot and a tree-like diagram (dendrogram).The results obtained using this analysis techniques speak for themselves, which gives a practitioner an opportunity to notice with the naked eye any peculiarities or unexpected behaviour in the analyzed corpus. Also, given a tree-like graphical representation of similarities between particular samples, one can easily interpret the results in terms of finding out to which group of texts a disputable sample belongs to. The last stage of the analysis involves a human interpretation of the generated plots.

3.1 Hierarchical Clustering

The hierarchical clustering algorithm builds a hierarchy of clusters. The cluster formation process is of two types, the agglomerative way and the divisive way. In the agglomerative clustering is bottom up approach. Every time the hierarchy goes up for one level, pairs of clusters are merged. Divisive hierarchical clustering is based on top-down approach. All of the samples in the dataset will stay in a same cluster together. When the hierarchy moves down for one level, the current clusters will be splitted into the smaller clusters. Both of the agglomerative and divisive approaches make the merging or splitting decisions base on the measurement value of the dissimilarity between the sets of observations. This can be achieved by using an appropriate metric and a linkage criterion. In the agglomerative approach, the clustering process is completed when all samples stays in the same cluster, while, in the divisive approach, each sample will stay in its cluster. Below figure demonstrates a simple hierarchy result after the clustering process with the divisive approach.

3.2 Multidimensional Scaling

If we could see how the various texts are arranged in feature space, we could get an idea of how well the different types of style markers group the texts by author. One way to do this is to use multi-dimensional scaling to map the texts onto a two-dimensional space. Multidimensional scaling takes as its input a matrix of distances between points, and attempts to fit those points into an n-dimensional space while preserving distances as much as possible. A two-dimensional MDS plot provides a convenient visual representation of distances between texts, since works that are close together on the plot are close together in feature space.

4. Proposed Methodology

The methodology used for solving the problem of authorship verification is given in the following steps.

Step I : It calculates word frequency lists for all messages in the corpus. Then we obtain N × M matrix
Step II : For each word M calculate the mean μ_m and standard deviation σ_m of frequencies for the whole corpus
Step III : In this step frequencies are standardized by converting them into z-scores. Each fwb, is replaced by their Z-scores. z-scores are calculated by using the mean and standard deviation of whole corpus and therefore it is Corpus dependent.
Step IV : In this step, the distance between the texts in the corpus is measured. Distance between message a and b

is calculated using delta measure. Which is the distance between messages a and b.

All such distances can once again be stored in the form of a symmetric matrix and such matrix can be treated as input data for some clustering algorithms

5. Experimental Results

5.1 corpus used

Enron corpus is used for experimentation purpose. In this paper we have considered only 4 authors to visualize the results. Enron corpus was made public during the legal investigation concerning the Enron Corporation. The current version contains 619,446 messages belonging to 158 users. This dataset was collected and prepared by the CALO.

5.2 Cluster Analysis using Hierarchical Clustering

Fig-1 shows result for cluster analysis using hierarchical clustering. Three separate clusters for 3 authors are shown in here. Fig-2 shows Multidimensional scaling for three authors. Both techniques are useful to visualize the results.

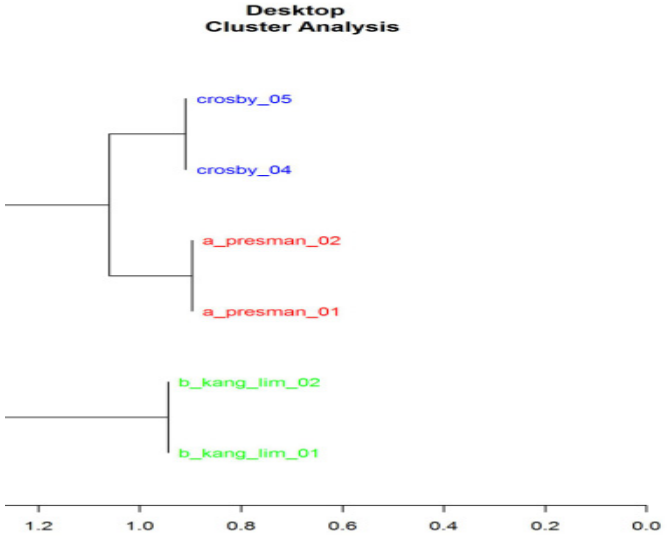

Fig.1. Result for Hierarchical Clustering

Clustering is used to show the similarity between two documents. Clustering algorithm used here is hierarchical agglomerative clustering. As shown in figure small pairs of closely related documents are combined and form groups. After that these small groups are combined into larger group till all the documents are connected into single large cluster. Fig shows dendrogram for three authors. The text documents written by one author are similar and are placed on neighbouring branches.

5.3Multidimensional Scaling

This visualization technique is based on distance matrix.MDS represents the original matrix by two dimensional map where values in the vector become co-ordinates of the document. This representation is as shown in fig –2 where more similar documents appear closer together**.**

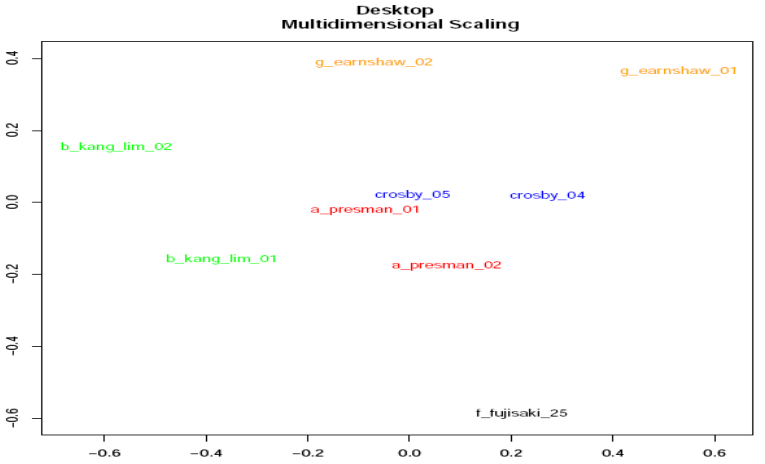

Fig. 2.Multidimensional Scaling

6. Conclusion

Authorship verification of online messages is a current research area. In this paper, we have explored authorship verification as a clustering problem and used unsupervised machine learning methods for the authorship verification problem. The proposed solution used cluster analysis and multidimensional scaling techniques, which provides visualization of clusters, helpful to investigator to visualize the results. According to the available literature on law enforcement and forensics linguistics, accuracy from 70-90% is acceptable in initial phase of investigation when investigator has very few clues to start the investigation .The proposed method helps investigator to speed up the analysis process and can find out the hidden stylometric patterns from writing style. These extracted results can be verified with other available evidences for consistency by the experts. After analyzing the entire evidences expert has to apply his /her knowledge to solve the case.

References

1. J. Li, R. Zheng, and H. Chen, “From fingerprint to writeprint,” Communication ACM, vol. 49, pp. 76–82, April 2006.
2. J. L. Hilton, On verifying wordprint studies: Book of Mormon authorship, ser. Reprint (Foundation for Ancient Research and Mormon Studies). F.A.R.M.S., 1991.
3. F. Can and J. M. Patton, Change of Writing Style With Time. Kluwer Academic Publishers, 2004, vol. 38.
4. A. Abbasi and H. Chen, “Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace,” ACM Trans. Inf. Syst., vol. 26, pp. 1–29, April 2008.
5. O. Canales, V. Monaco, T. Murphy, E. Zych, J. Stewart, C. T. A. Castro, O. Sotoye, L. Torres, and G. Truley, “A stylometry system for authenticating students taking online tests,” P. of Student Faculty Research Day, Ed., CSIS. Pace University, May 6 2011.
6.C. Sanderson and S. Guenter, “Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation,” in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, ser. EMNLP ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 482–491.
7. M. Koppel and J. Schler, “Authorship verification as a one-class classification problem,” in Proceedings of the 21st international conference on Machine learning, ser. ICML ’04. Banff, Alberta, Canada: ACM, 2004, pp. 62–69.
8. X. Chen, P. Hao, R. Chandramouli, and K. P. Subbalakshmi. Authorship similarity detection from email messages. In Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition, MLDM’11, pages 375–386, Berlin, Heidelberg, 2011. Springer-Verlag.
9. A. Abbasi and H. Chen, “Applying authorship analysis to extremist-group web forum messages,” IEEE Intelligent Systems, vol. 20, pp. 67–75, September 2005.
10. F. Iqbal, R. Hadjidj, B. C. Fung, and M. Debbabi, “A novel approach of mining write-prints for authorship attribution in email forensics,” Digital Investigation, vol. 5, pp. S42–S51, 2008.
11. F. Iqbal, L. A. Khan, B. C. M. Fung, and M. Debbabi, “E-mail authorship verification for forensic investigation,” in Proceedings of the 2010 ACM Symposium on Applied Computing, ser. SAC ’10. New York, NY, USA: ACM, 2010, pp. 1591–1598.
12. F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, “A unified data mining solution for authorship analysis in anonymous textual communications,” Information Sciences, vol. 231, pp. 98–112, 2013.
13. C. E. Chaski, “Who’s at the keyboard: Authorship attribution in digital evidence investigations,” International Journal of Digital Evidence,

vol. 4, no. 1, pp. 1–13, Spring 2005.
14. F. Mosteller and D. L. Wallace, Inference and Disputed Authorship: The Federalist. Addison-Wesley, 1964.
15. Smita Nirkhi and R.V.Dharaskar," Comparative study of authorship identification techniques for cyber forensics analysis", International Journal of advanced computer science and application, vol. 4, no. 5, 2013