Digital Investigation 14 (2015) S116–S126

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2015 USA

E-mail authorship attribution using customized associative classification

Michael R. Schmid [a], Farkhund Iqbal [b, *], Benjamin C.M. Fung [c]

[a] Concordia Institute for Information Systems Engineering, Concordia University, QC, Canada
[b] College of Technological Innovation, Zayed University, United Arab Emirates
[c] School of Information Studies, McGill University, QC, Canada

Keywords:
Authorship
Crime investigation
Anonymity
Data mining
Associative classification
Writeprint
Rule mining

A B S T R A C T

E-mail communication is often abused for conducting social engineering attacks including spamming, phishing, identity theft and for distributing malware. This is largely attributed to the problem of anonymity inherent in the standard electronic mail protocol. In the literature, authorship attribution is studied as a text categorization problem where the writing styles of individuals are modeled based on their previously written sample documents. The developed model is employed to identify the most plausible writer of the text. Unfortunately, most existing studies focus solely on improving predictive accuracy and not on the inherent value of the evidence collected. In this study, we propose a customized associative classification technique, a popular data mining method, to address the authorship attribution problem. Our approach models the unique writing style features of a person, measures the associativity of these features and produces an intuitive classifier. The results obtained by conducting experiments on a real dataset reveal that the presented method is very effective.

Introduction

E-mail has emerged as one of the most popular means of online communication. Unfortunately, it is often used for sending unsolicited e-mails, conducting phishing scams, and for spreading malware due to the lack of standard security and privacy mechanisms. In many misuse cases, an offender either masks his/her actual identity or impersonates someone of high authority to trick a user into disclosing valuable personal information such as credit card or social insurance numbers. According to the annual report published by the Internet Crime Complaint Center,[1] 16.6% of the total reported 336,655 cybercrimes were e-mail scams called “FBI scams”, in which the attackers pretended to be an FBI official in order to defraud victims.

Most published methods are used as a postmortem panacea and there exists no concrete proactive mechanism for securing e-mail communication (Iqbal et al., 2010a). It has been shown Iqbal et al. (2010a) that analyzing e-mail content for the purpose of authorship analysis can help prosecute an offender by precisely linking him/her to a malicious e-mail with tangible supporting evidence. Most existing authorship techniques (de Vel et al., 2001a,b; Teng et al., August 2004; Zheng et al., 2003) study different stylometric features (e.g., lexical, structural, syntactical, content-specific and idiosyncratic) separately but very few of them have studied the collective effect of these features.

Building a writeprint by combining lexical, syntactical, structural, semantic, and content-specific attributes produces more promising results than when individual features are compared separately. This reveals the importance of

* Corresponding author.
E-mail addresses: michael.schmid@concordia.ca (M.R. Schmid), farkhund.iqbal@zu.ac.ae (F. Iqbal), ben.fung@mcgill.ca (B.C.M. Fung).

[1] http://www.ic3.gov/media/annualreport/2011_IC3Report.pdf.

http://dx.doi.org/10.1016/j.diin.2015.05.012

interdependence, correlation, and associativity of stylometric features on the accuracy of methods. *Frequent pattern mining* (Agrawal et al., 1993), *sequential pattern mining* (Agrawal and Srikant, 1995), and *association rule mining* are studied for analyzing associativity of features ((Fachkha et al., 2012), (Han et al., 2006)). In this paper, we employ *Associative Classification* (*AC*) (Agrawal et al., 1993), based on association rule discovery techniques, for authorship identification. The developed classification model consists of patterns that represent the respective author's most prominent combinations of writing style features.

There are many different implementations of AC, namely Classification based on Associations (CBA) (Liu et al., August 1998), Classification based on Predictive Association Rules (CPAR) (Han and Yin, 2003), Classification-based on Multiple Association Rules (CMAR) (Li et al., 2001), and Multi-class Classification based on Association Rules (MCAR) (Thabtah et al., 2005). Given the need to accurately quantify the match between the various author's writing styles and the anonymous e-mail, we have concentrated our research on *CMAR*. This variation on AC uses a subset of rules as opposed to a single best rule, to determine which class, or author in our case, is the best match.

Below are some of the pertinent contributions of this paper.

- To our knowledge, this is the first application of *AC* to the authorship attribution problem; the experimental results on real-life data endorse the suitability of the presented approach.
- Association rule mining in AC is different than traditional association rule mining; the former investigates the associativity of features with one another as well as with the target predetermined classes, whereas the later is limited to the analysis of the interdependence between features and do not associate them all to a target class. Therefore, extracted association rules reveal feature combinations that are relevant in distinguishing one author from another in authorship identification.
- Each instance in a classification model shows the features that are related, not only to each other but to the class label as well. As a result, the proposed method builds a concise and representative classifier that can serve as admissible evidence to support the identification of the true author of a disputed e-mail.

The rest of this paper is organized as follows: Section 1 provides a literature review on authorship analysis and classification analysis. Section 2 formally defines the authorship attribution problem and the notion of writeprint by *class association rule* (*CAR*) list. Section 3 describes our new data mining approach for modeling a writeprint from transformed semantic content. Section 4 evaluates the accuracy and efficiency of our suggested method on the Enron e-mail dataset.[2] Section 5 brings the paper to a conclusion.

[2] http://www.cs.cmu.edu/~enron/.

Related work

Authorship attribution is studied as a text categorization and classification problem in the literature (de Vel, August 2000). Generally, a classification model is built using the previously written documents of the suspected authors. The author names are used as class labels in the training and testing processes of model development. Unlike authorship verification, which is studied as one-class (Koppel & Schler) and two-class (Iqbal et al., March 2010b) classification problem, modern authorship attribution, which can be better understood by reading Stamatato's survey Stamatatos (March 2009), can be approached as a multi-class classification problem.

There is no single standard predefined set of features that best differentiates the writing style of individual writers, but some studies Grieve (July 2007) have identified the most representative features in terms of accurately classifying anonymous or disputed texts. Punctuation and n-gram features have proven to be highly representative on their own, but the combination of these features was discovered to be even more characteristic. The relative preference for using certain words over others along with their associations is another highly representative feature. Vocabulary richness, fluency in the language and grammatical and structural preferences of individuals are among these important writing style manifestations. Finally, spelling and grammar mistakes and rare word sequences are also quite characteristic of an authors writing style. One comprehensive study on stylistic features presented by Abbasi and Chen (2008) discusses these with sufficient detail.

Most methods require feature selection as an important step towards maximizing accuracy; our algorithm does not require feature selection because unimportant features will not meet the minimum support threshold. In other words, the algorithm itself performs feature selection, simplifying one of the more complex aspects of authorship attribution.

Authorship analysis has been quite successful in resolving authorship attribution disputes over various types of writings (Mendenhall, 1887). However, e-mail authorship attribution poses special challenges due to its characteristics of size, vocabulary and composition when compared to literary works (de Vel et al., 2001a,b). Literary documents are usually large in size, comprising of at least several paragraphs; they have a definite syntactic and semantic structure. In contrast, e-mails are short and usually do not follow well defined syntax or grammar rules. Thus, it is harder to model the writing patterns of their author. Ledger and Merriam (1994) established that authorship attribution would not be very accurate for texts containing less than 500 words, creating the need for better models Iqbal et al. (2010a) able to handle the characteristics inherent in e-mails. Moreover, e-mails are more informal in style and people are not as conscious about spelling or grammar mistakes particularly in these types of communications. Therefore, techniques that are appropriate for literary and traditional works are not always well suited for e-mail authorship attribution problems.

Iqbal et al. (May 2013) have shown that the e-mail authorship attribution problem can be solved by designing