Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004

E-MAIL AUTHORSHIP MINING BASED ON SVM FOR COMPUTER FORENSIC

GUI-FA TENG[1,2], MAO-SHENG LAI[1], JIAN-BIN MA[2], YING LI[2]

[1] Department of Information Management, Peking University, Beijing 100091, China
[2] School of Information Science and Technology, Agricultural University of Hebei, Baoding 071001,China
E-MAIL: tguifa@pku.edu.cn, search@vip.163.com

Abstract:

In this paper, we describe our work which attempts to mine e-mail authorship for the purpose of computer forensic. We extract various e-mail document features including linguistic features, header features and structural characteristics. These features together are used with the Support Vector Machine learning algorithm to classify or attribute authorship of e-mail messages to an author. The primary experiments on a number of e-mail documents have given ideal results, which indicate that the project has laid a firm groundwork for the future work.

Keywords:

Authorship mining; E-mail; Support Vector Machine; Computer Forensic

1. Introduction

Nowadays many industries and governments have become dependent on the use of e-mail as an expedient and economical form of communication over the Internet and Intranet. E-mail is used for many purpose, for example, many companies and institutions rely on e-mail for transaction business and individuals for communication. Thus the amount of e-mail traffic has increased markedly particularly since the inception of the World Wide Web.

Unfortunately with the increase in e-mail traffic comes an undesired increase in the use of e-mail for illegitimate reasons. Examples of misuse include: sending Spam or unsolicited commercial e-mail, which is the widespread distribution of junk e-mail; sending threats; sending hoaxes or forbidden propaganda; the distribution of computer viruses and worms and unauthorized conveyance of sensitive information. Furthermore, criminal activities such as traffic in drugs or child pornography can easily be aided and abetted by sending simple communications in e-mail messages. As a result, these e-mail misuse phenomena do a lot of harm to people's benefit, even influence social stability.

Now there are not effective methods for preventing these phenomena. The current methods are merely some passive defending measures such as e-mail filtering, installing firewall, etc. But they have inhered limitations and cannot put an end to the e-mail misuse phenomena. So the ability to provide empirical evidence and identify the original author of e-mail misuse is an important factor in the successful prosecution of an offending user by means of law.

Computer forensic technique can be thought of investigation of computer-based evidence of criminal activity, using scientifically developed methods that attempt to discover and reconstruct event sequences from such activity. The principal objectives are to collect sufficient and accurate evidence for courtroom. Undoubtedly, it would be useful to have a computer forensic technique that can be used to mine the illegitimate e-mail's real identity and collect evidence for computer forensic professionals and law enforcement agencies.

Identifying the author's identity encounters difficulties since the sender will attempt to hide his/her identity in order to avoid detection. However, humans are creatures of habit and have certain personal traits which tend to be persisted. That is why, for example, we have a handwriting style that is consistent during periods of our life, although the style may vary as we grow older. Likewise for writing, an author is sure to have certain writing habits, and these habits are unconscious and deeply ingrained. It means that even if one were to make a conscious effort to disguise one's style, this would be difficult to achieve. The writing habits usually display in certain characteristics that pertain to language, composition and writing, such as particular syntactic and structural layout traits, pattern of vocabulary usage, unusual language usage, stylistic and sub-stylistic features. So it is feasible to grasp the documents' characteristics and identify their real authorship based on text categorization.

A closely related area of authorship mining is text categorization which attempts to categorize a set of text documents based on its content-type. Text categorization

0-7803-8403-2/04/$20.00 ©2004 IEEE

provides support for a wide variety of activities in information mining and information management. The main techniques are to extract text characteristic and then use a learning algorithm such as decision trees [1], neural network, Bayesian probabilistic approaches, or support vector machines [2] to classify the text document. Compared to longer, formal text documents, e-mail documents have several characteristics which make authorship categorization challenging. Firstly, e-mail documents are generally brief and to the point, and can be punctuated with a larger number of grammatical errors etc. Secondly, the author's composition style used in e-mails can evolve quite rapidly over time. Finally, e-mail documents have generally few sentences and paragraphs. These provide difficulties compared to text classification.

The rest of the paper is organized as follows: Section 2 describes our choice of the feature selection and extraction methods. Section 3 briefly outlines the support vector machines learning algorithm. Section 4 provides our research methods. And conclude the paper with a discussion on ongoing works in section 5.

2. Feature selection and extraction

We adopt Vector Space Model (VSM) to represent the document information. In VSM, text-based documents are represented as vectors in a high-dimensional vector space where the value of dimensions is based on the words occurring in that document. Each document is represented as a vector of term and weight pairs. So a document d will be represented by a vector $V=((t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n))$. We calculate the weight of the vector by the common technique $tf \bullet idf$ (term frequency-inverse document frequency) value:

$$W(t,\vec{d})=\frac{tf(t,\vec{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{d}}\left[tf(t,\vec{d})\times\log(N/n_t+0.01)\right]^2}} \quad (1)$$

where $w(t,\vec{d})$ is the weight of term t in document d, $tf(t,\vec{d})$ is the frequency of term t in document d, N is the total number of documents, n_t is the number of documents that contain term t.

A technique for managing the computational costs associated with text classification is feature selection. The most popular approach to feature selection is to select a subset of the available features using methods like document frequency (DF), information gain (IG), mutual information (MI), term strength (TS), the χ^2-test (CHI). A prior study has showed that CHI statistics generally outperforms other feature selection techniques. So we adopted χ^2 (CHI) as the feature selection criteria.

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (2)$$

where A is the number of times t and c co-occur, B is the number of time the t occurs without c, C is the number of time the c occurs without t, D is the number of times neither c nor t occurs, and N is the total number of documents.

Unlike some formal text documents, e-mail documents are more than free-form text. The body of a document is unstructured text. A document also contains a structured header. The header fields include the sender (the From field), a list of recipients (the To field), short description (the Title field), reply message (the Reply field) and attachment message. The body of e-mail documents sometimes is short. So the structural characteristics and header features cannot be ignored.

The From and To field always contain e-mail address. Because the word of From or To field appears in the document only once, the weight of vector equals 1 or 0. If the particular word appears in the document, then the value is 1, or the value is 0. The header and some structural features are listed in table1.

Table 1. E-mail document's header and structural features

Attribute	Attribute type
1	The From message
2	The To message
3	Whether or not have title
4	Whether or not have attachments
5	Whether or not have reply
6	Uses a greeting acknowledgement
7	Uses a farewell acknowledgement
8	Contain signature text
9	Mean sentence length
10	Mean paragraph length
11	Number of blank lines/total number of lines

3. Support vector machine classifier

Support Vector Machine (SVM) is a relatively new class of machine learning techniques first introduced by Vapnik [14]. Based on the structural risk minimization principle of the computational theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements from the training set.

Given a set of linearly separable points

$S=\{x_i|i=1,2,\cdots,N\}$, each point x_i belongs to one of two classes labeled as $y_i\in\{-1,+1\}$, a separating hyper-plane divides S into two sides, each containing points with the same class label only. The separating hyper-plane can be identified by the pair (w,b) that satisfies

$$w\bullet x+b=0 \quad (3)$$

for any training sample $x_i\in S$, the goal of the SVM learning is to find the optimal separating hyper-plane (OSH) that has the maximal margin to both sides. This can be formularized as:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 \quad (4)$$

$$\text{subject to} \quad y_i(w\bullet x_i+b)\geq 1 \quad (5)$$

The points that are closest to the OSH are termed support vectors (Figure 1).

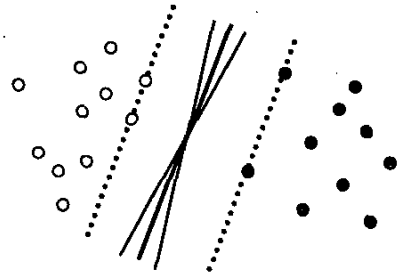

Figure 1. Separating hyper-plane (the set of solid lines), optimal separating hyper-plane (the bold solid line), and support vectors (data points on the dashed lines)

During classification, SVM makes decision based on the globally optimized separating hyper-plane. It simply finds out on which side of the OSH the best pattern is located. This property makes SVM highly competitive with other traditional pattern recognition methods in terms of predictive accuracy and efficiency.

The SVM problem can be extended to non-linear models by mapping the input space into a very high-dimensional feature space based on kernel function. Examples of common kernel functions are:

Polynomial $K(x_i,x_j)=(c+(x\bullet y))^d$ (6)

Sigmoid $K(x_i,x_j)=\tanh(k(x\bullet y)+\theta)$ (7)

RBF $K(x_i,x_j)=\exp\left(-\gamma\|x-y\|^2\right)$ (8)

The idea of the Support Vector Machine is to find a model for which we can guarantee the lowest true error by controlling the model complexity (VC-dimension) based on the structural risk minimization principle. This avoids over-fitting, which is the main problem for other learning algorithm. So the distinctive advantage of the SVM is its ability to process many high-dimensional applications, such as text classification and authorship categorization.

4. Research methods

In preprocessing, citations (marked with >, |, etc.) removal and sentence segmentation of the original E-mail documents should be performed using pre-specified rules. Some header information such as mail address, reply text, titles and attachments should not be removed because the header of e-mail contains some important information that we can't conceive.

When the classification problem has only a small set of data to work with, it can be difficult to provide enough data for disjoint training and testing data. This is likely to be the case in a forensic study, where there may only be a small amount of data to produce a model of authorship. So we want to measure the performance of each feature by cross-validation evaluation methods to provide a more meaningful result.

There are some free Support Vector Machine softwares available. We adopt Libsvm as the machine learning algorithm which is a simple, easy-to-use, and efficient software for SVM classification and regression.

As Support Vector Machine only compute two-way categorization, multi-class classification models should be used. These models are 'one against all' classification and 'one against one' classification. To avoid computation cost, the approach used in this research was 'one against all'.

To evaluate the categorization performance on the e-mail document corpus, we calculate the accuracy, recall (R), precision (P), and combined F_1 performance measures commonly employed in the text categorization, where:

$$F_1=\frac{2PR}{(P+R)} \quad (9)$$

To obtain an overall performance figure over all binary categorization tasks, a macro-averaged F_1 statistic $F_1^{(M)}$ is calculated, where:

$$F_1^{(M)}=\frac{\sum_{i=1}^{N_{AC}}F_{1,AC_i}}{N_{AC}} \quad (10)$$

where N_{AC} is the number of author category and F_{1,AC_i} is the per-author-category F_1 statistic for author category $AC_i(i=1,2,\cdots,N_{AC})$.

$$F_{1,AC_i}=\frac{2P_{AC_i}R_{AC_i}}{(P_{AC_i}+R_{AC_i})} \quad (11)$$

5. Conclusions and outlook

We attempt to investigate methods for e-mail authorship mining for the purpose of computer forensic. We resort to the techniques of text categorization and try to extract various e-mail document features including linguistic features, header features and structural characteristics and use the Support Vector Machine as the learning algorithm to classify or attribute authorship of e-mail. Primary experiments on a limited number of documents have given promising results, which indicate that these methods are feasible. And the research provides a new approach to computer forensic.

Further work must be done in the future. Firstly, we consider that some performance will be gained if we combine SVM with other machine learning algorithms. Secondly, some additional features should be investigated in order to extract the optimal feature sets of e-mail documents. Thirdly, authorship characterization or cohort profiling for authorship, such as gender, language background, and education level should be investigated further in order to reduce the size of the list of possible suspects for forensic purpose. Fourthly, some citation text in the e-mail documents should be identified in order to increase the classification accuracy.

References

[1] C.Apte, F.Damerau, S.Weiss, "Text mining with decision rules and decision tree", In workshop on Learning from text and the Web, Conference on Automated Learning and Discovery, 1998.

[2] T.Joachims, "Text categorization with support vector machines: Learning with many relevant features", In Proceedings of the European Conference on Machine Learning, Springer, 1998.

[3] I.Krsul, "Authorship analysis: Identifying the author of a program", Technical report, Department of Computer Science, Purdue University, 1994.

[4] I.Krsul and E.Spafford, "Authorship analysis: Identifying the author of a program", Computers and Security, No.16, 1997,pp.248-259.

[5] E.Spafford and S.Weeber, "Software forensics: tracking code to its authors", Computers and Security, No.12, 1993,pp.585-595.

[6] Malcolm Walter Corney, "Analysing E-mail Text Authorship for Forensic Purpose", Master thesis, University of Software Engineering and Data Communications.

[7] E. Stamatatos, N. Fakotakis, G. Kokkinakis, "Automatic Authorship Attribution", Dept. of Electrical and Computer Engineering University of Patras.

[8] Malcolm Corney, Olivier de Vel, Alison Anderson, George Mohay, "Gender-Preferential Text Mining of E-mail Discourse".

[9] Olivier de Vel, "Mining E-mail Authorship", KDD-2000 Workshop on Text Mining, ACM International conference on knowledge Discovery and Data Mining, Boston, MA, USA, 2000.

[10] O. de Vel, A. Anderson, M. Corney, G. Mohay, "Multi-Topic E-mail Authorship Attribution Forensics". ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, November 8, 2001, Philadelphia, PA.

[11] Yuta Tsuboi, "Authorship Identification for Heterogeneous Documents", Master's Thesis, Nara Institute of Science and Technology, University of Information Science, 2002.

[12] Ji He, Ah-Hwee Tan, Chew-Lim Tan, "On Machine Learning Methods for Chinese document categorization", Applied Intelligence, No.18, 2003,pp.311-322.

[13] C.Cortes and V.Vapnik, "Support vector networks", Machine Learning, vol.20, pp.273-297, 1995.

[14] V.Vapnik, "The Nature of Statistical Learning Theory", Wiley, New York, 1998.

[15] A.Anderson, M.Corney, O.de Vel, G.Mohay, "Identifying the authors of Suspect E-mail", communications of the ACM, 2001.

[16] R.Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", In International Joint Conference on Artificial Intelligence, 1995.

[17] B.Kjell, "Authorship attribution of text samples using neural networks and Bayesian classifiers", In IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, 1994.

[18] Y.Yang, "An evaluation of statistical approaches to text categorization", Journal of Information Retrieval, vol.1, pp.67-88, 1999.