E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

AUTHOR IDENTIFICATION IN TEXT MINING FOR USED IN FORENSICS

Prof. Nihar Ranjan[1] ,
Department of Computer Engineerin[1],
SITS Narhe,University of Pune,Pune [1]
nihar.pune@gmail.com[1]

Dr. R.S. Prasad[2]
Department of Computer Engineering[2]
NBNSSOE,University of Pune,Pune[2]
rsprasad_viit@yahoo.com[2]

ABSTARCT:

We have reviewed into text data and e-mail content mining for author identification, or authorship categorization, for the purpose of forensic investigation. We have focused our discussion on the ability to discriminate between authors for the case of both e-mails as well as across the simple text data. An set of e-mail document features including structural characteristics and linguistic patterns which can be derived with a Support Vector Machine learning algorithm, which can be used for mining the e-mail content. The experiments using a number of e-mail documents and the text data generated by different authors on a set of topics gave good results for both author identification & categorization.

KEYWORDS: *AUTHOR, FORENSIC, LINGUISTIC, SVM.*

1. INTRODUCTION

An Author identification & categorization is an important problem in many areas including information retrieval and computational linguistics, but also in the areas such as law and journalism where knowing the author of a document may be able to save human lives.

The most common methods for testing candidate algorithms is a text classification problem[3]: given known sample documents from a small, finite set of candidate authors, which if any one wrote a questioned document of unknown authorship? It has been remarked, however, this may be an unreasonably easy task. A more challenging problem is author verification where given a set of documents written by a single author and the questioned documents, the problem is to determine whether the document under consideration has been written by that specific author or not.

Computer forensics undertakes the reconstruction of the sequence of events arising from an intrusion carried out by an external agent or as a result of illegal activities performed by an authorized user.

The field of computer forensics covers a wide set of applications, it uses a variety of evidence and is well supported by a number of different techniques. Application areas of this includes forensic accounting, law enforcement, commodity analysis, threat analysis, tracing illegal activities ,Finding unauthorized users etc. Evidence are made available to computer forensics investigators may varied and can be sourced from different entities, for example, storage devices, networks telecommunication traffic, cloud stored data. Computer forensics investigations can involve a wide variety of techniques or methods which includes information hiding analysis, data mining, text mining & analysis, causal analysis, timeline analysis and so on.

In the context of computer forensics, the mining of e-mail & text data authorship has a couple of characteristics. First, the identification of an author is usually attempted from a small set of known candidates, rather than from a large set of unknown authors. Second, the text body of the e-mail is not the only source of authorship identification. Other evidence can be in the form of e-mail headers, unique email tags, e-mail attachments, time stamps, etc. and can be used in conjunction with the analysis of the email text body

2. AUTHOR CATEGORIZATION

Authorship categorization is the task of determining the actual author of a piece of work. T be specific, we are interested to categorize textual work given other text samples produced by the same author [4]. Here we assume that only one author is responsible for producing the text generation by, or text modified by, multiple authors are not considered in this case. Authorship categorization or identification can be done using various approaches.

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

2.1. *Domain Expert*

This is the simplest method to identify new e-mail documents and allocate them to well-defined author categories. This approach can be time-consuming and very expensive and, have more limitations. It does not provide continuous measure of the degree of confidence by which the allocation has been made. The domain expert method can establish a set of fixed rules which can be used to classify new e-mail documents. Unfortunately, in several cases, the set of rules can be large and unwieldy, which is difficult to update, and unable to adapt the changes in document content or author characteristics.

2.2 *Text Categorization*

It is the method which attempts to categories set of text documents on the basis of its contents Text categorization provides support for a wide variety of activities in information retrieval and information management. It has broad applications in document filtering and can be also used to support document retrieval by generating the categories required in document retrieval. Many methods are proposed that automatically learn the rules that have been proposed for text categorization.

3. AUTHORSHIP ANALYSIS

Authorship analysis includes other distinct fields such as author characterization and similarity detection among the documents. Authorship characterization determines the author profile or characteristics of the author that produced a original piece of work. Whereas example characteristics include gender, educational and cultural backgrounds, language familiarity etc. Similarity detection analysis calculates the degree of similarity between two or more pieces of work without identifying the authors. A similarity criterion is used extensively in the context of plagiarism detection which involves the complete or partial replication of a piece of work with or without permission of the original author. However, that authorship categorization and author characterization are different from plagiarism detection. Plagiarism detection attempts to detect the similarity between two sub different pieces of work but is unable to determine if the documents were produced by the same author. Authorship analysis[4] has been used in a small but diverse number of application areas. Examples include identifying authors in literature, in program code, and in forensic analysis for criminal cases. The most extensive and comprehensive application of authorship analysis is in literature and in published articles.

4. FORENSIC ANALYSIS

The forensic analysis of text is attempted to find text to authors for the purpose of criminal investigations. The forensic analysis of text generally includes techniques derived from linguistics and behavioral profiling. Linguistic techniques usually employ common features such as grammatical errors, spelling mistakes, and stylistic deviations. These techniques do not quantify linguistic patterns and fail to discriminate between authors with a high degree of precision. However, the use of language based author attribution testimony as admissible evidence in legal proceedings has been identified in many of the cases .The textual analysis of the unabomber manifesto is a well known example of the use of forensic linguistics. In this case, the manifesto and the suspect bomber used a set of similar characteristics, such as a distinctive vocabulary, irregular hyphenations etc. Techniques which based on the scientific evidence of language have not, to the authors' knowledge, been used in the court proceedings. Author profiling is based on the behavioral characteristics contained within an author's text. For example, educated guesses on the type of personality of an author based on particular sequences of words are employed in profiling studies.

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

5. CHALLENGES WITH E-MAIL ANALYSIS

E-mail content or documents have several characteristics which make authorship categorization challenging compared with the longer formal text documents such as literature works or published articles. First, e-mails are generally short in length indicating that certain language based metrics may not be appropriate .Second; the composition style used in drafting an e-mail document is normally different from normal text documents written by the same author. So, an author profile derived from normal text documents may not necessarily be the same as that obtained from an e-mail document. For example, e-mail documents are generally brief and to the point, can involve a dialogue between two or more authors can be punctuated with a larger number of grammatical errors etc. Also, e-mail interaction between authors can be frequent and rapid, similar to speech interactivity and rather dissimilar to normal text document interchange patterns. Indeed, the authoring composition style and interactivity characteristics attributed to e-mails shares some elements of both formal writing and speech. Thirdly, the author's composition style used in e-mails can vary depending upon the intended recipient and can evolve quite rapidly over time. Fourthly, the vocabulary used by authors in e-mails is not stable, facilitating imitation. Thus the possibility of being able to disguise authorship of an e-mail. Furthermore, similar vocabulary subsets may be used within author communities. Finally, e-mail documents have generally few sentences / paragraphs, thus making contents profiling based on traditional text document analysis techniques, such as the (bag of words)"[2] representation, more difficult. However, as stated previously, certain characteristics such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, stylistic and sub stylistic features will remain relatively constant for a given e-mail author. This provides the major motivation for the particular choice of attributes/features for the authorship categorization of e-mails.

6. METHOD

6.1 *Support vector machine (SVM)*

The basic concepts of Support Vector Machines (SVM)[2] is based on the idea of structural risk minimization which can minimize the generalization error which is bounded by the sum of the training data error and the term which will depends on the dimension of the classifier and on the number of training set examples. The purpose of a structural risk minimization performance measure in contrast with the empirical risk minimization approach can be used by conventional classifier. Conventional classifiers attempt to minimize the training data error which does not necessarily achieve a minimum generalization error. Therefore, SVMs has theoretically a greater ability to generalize. The number of free parameters used in the SVM depends on the margin that separates the data and does not depend on the number of input features. Thus the SVM does not require a reduction in the number of available features in order to avoid the problem of over-fitting .This feature is clearly an advantage in the context of high-dimensional applications, such as text document analysis and authorship categorization, as long as the data vectors are separable with a large margin. SVMs require the implementation of optimization algorithms for the minimization procedure which can be computationally expensive. A few research scholars have applied SVMs to the problem of text document analysis and categorization using approximately thousands of features in some cases, concluding that, in most of the cases, SVMs out-performs conventional classifier. SVMs also can be used for classifying e-mail text and documents as spam or non-spam and compared it to boosting decision trees.

The classifier that can be used in the experiments is the Support Vector Machines classifier it can scales to a large number of sparse instance vectors as well as efficiently handling a large number of support vectors. These experiments explored a number of different kernel functions for the SVM classifier namely, the linear, polynomial, radial basis and the sigmoid functions. We obtained maximal F1 classification results on our training data set with a polynomial kernel of degree 3. The “LOQO” optimizer is used for maximizing the margin. Support Vector Machines only compute two-way categorization, Q two-way classification models is generated, where Q is the number of authors categories ,Q = 3 for our e-mail document corpus, and each SVM categorization was applied Q times. This produced Q two-way confusion matrix.

6.2 *E-Mail Corpus*

E-ISSN: 2321–9637
Volume 1, Issue 5, December 2013
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

The choice of the e-mail corpus is limited by privacy and ethical considerations. The publicly available e-mail corpus includes newsgroups, mailing lists etc. However, in such type of public e-mail databases, it is generally very difficult to find a significant large and "clean" corpus of both multi-author and multi-topic e-mails. The resulting author-topic matrices of multiple authors discussing the same set of topics is generally sparse and often characterized by having some interdependent topics .Also, there is no control over the authors' characteristics or profile. One approach that avoids the problems of e-mails obtained from newsgroups etc. is to generate a controlled set of e-mails for each author and topic. The resulting author-topic Matrix is non-sparse with maximum independence between topics and minimal bias towards particular author characteristics. This approach was used in the experiment. The corpus of e-mail documents used in the experimental evaluation of author-topic categorization contained a total of 160 documents sourced from three different language authors, where each author contributing e-mails on three topics the topics chosen were movies, religion and research. The relatively small numbers of e-mail documents per topic category was not thought to be critical as it has been observed that as few as a total of 15 or 20 documents for each author should be sufficient for satisfactory analysis and categorization performance. The body of each e-mail document was parsed, based on an e-mail grammar that we designed, and the relevant e-mail body features were extracted. The body of the e-mail was pre-processed to remove any salutations, replied text and signatures. However, the existence, position within the e-mail body and type of some of these is retained as inputs to the categorizer. Attachments are excluded, though the e-mail body itself is used.

To evaluate the categorization performance on the e-mail document corpus, we calculate the accuracy, recall (R), precision (P) and combined F1 performance measures commonly employed in the information retrieval and text categorization literature ,where: F1= 2RP/ R+P

CONCLUSION

We have reviewed the authorship analysis and categorization for the case of both aggregated and multi-topic e-mail documents. We used an extended set of predominantly content free e-mail document features such as structural characteristics and linguistic patterns. The classifier used was the Support Vector Machine learning algorithm. Experiments on a number of e-mail documents generated by different authors on a set of topics gave encouraging results for both aggregated and multi-topic author categorization. However, one author category produced worse categorization performance results, probably due to the reduced number of documents for that author.

References

[1] A. Anderson, M. Corney, O. de Vel, and G. Mohay. "Identifying the Authors of Suspect E-mail".
[2] Ode vel,A.anderson,M.corney "Mining Email Content for Author Identification "
[3] C. Apte, F. Damerau, and S. Weiss. "Text miningwith decision rules and decision trees".
[4] O. de Vel. "Mining e-mail authorship"
[5] Abhishek Chandorkar, Kunal Borkar, Dr. Rajesh. S .Prasad "Implementation of an Author Identification System",
[6] W. Cohen. "Learning rules that classify e-mail"
[7] O.de Vel."Mining e-mail authorship". InProc. Workshop on Text Mining, ACM International. Conference on Knowledge Discovery and Data Mining.