ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 10, October 2013

Study of Behavioural Patterns for Author Identification

Kunal Borkar, Abhishek Chandorkar, Dr. R. S. Prasad

Abstract* - *An author's distinct writing style can be studied by analyzing written documents. Hence, identification of an unknown author is an important task from the point of view of security. This task will help save a lot of efforts which are otherwise taken to verify the identity. An author's writing style is analysed by studying some of the features in a document written by him. In this paper, we will study the features which are useful in author identification and the ways by which this process is performed.

Keywords-author identification, stylometry, features, stylistics, feature extraction

I. INTRODUCTION

Lately, our life has been boosted and speeded up due to the frequent use of the Internet. The Internet provides a medium for everything from getting the latest news, to creating a movie, playing games and so on, the list is endless. Due to the emergence of blogging and social networking (online community) sites like Facebook, communication between people can take place irrespective of any geographical boundaries and beyond time zones. This has not only reduced the degree of separation between people but also allowed them to interact with a larger audience. However, malicious and rogue users too frequently use this online community for performing malicious acts. Blogs too induce online communities by allowing users to share their views and opinions to a large audience base. Consequently, they too have been seen as a medium for propagation and also as a breeding place for malicious activities. This massive reach and impact of online media has provided a thriving ground for internet-based terrorism, where sheer size of the community and its interactions are taken undue advantage of. Identifying such individuals and hidden groups of users is an important task for checking cyber-crime and preserving online privacy. The text data however contains plenty of clues which we will call as features, which can help to reveal the identity of such individuals.

However, author identification was first used to determine which author wrote a specific passage or a book. Its use for identification of individuals on the Internet came about much later. Author identification divides text into several parts(features) and uses them as the tool to identify the writing style. A branch of this is the stylometric research in which linguistic characteristics are used to identify the author of a text. Actually, most of the features used for author identification are stylometric, especially in literary authorship. In stylometry research it is generally accepted that each author has a distinct writing habit which is mostly unconscious. These habits are revealed from the use of words and the construction of the sentences using the words and grammar. The more unconscious a process is, the more difficult it is to control. Therefore words and grammar can a reliable indicator of the author. These individual differences in use of language are known as idiolect. Due to the unconscious use of grammar and syntax, we can perform the task of author identification based on stylometric features.

II. RELATED WORK

Most of the previous work done in the field of author identification was done to identify the authors of unknown scripts of books like the Bible. In 1877, Mendenhall, a meteorologist, proposed the first quantitative approach. He proposed that word-length distribution is a feature which varies from author to author. In 1964, Mosteller and Wallace used function words and Bayesian analysis to identify authors. It was a statistical approach and not a quantitative one. Then, McCombe carried out various experiments to find out which features can be used for author identification. She performed tests using word unigrams as a classification feature. She showed that the results using this method are promising. But no method she used was successful in classification based on word bigrams. Lately, Carole Chaski implemented a method which used both syntactic and statistical analysis for identifying authors, which proved to be very useful.

III. MACHINE LEARNING ALGORITHMS

Training data samples are used to create a model. Machine learning algorithms are used to learn such training date samples. This is a classification model which is used to generalize over unseen data. Characteristics of the unseen data are used by the classification model to predict the class label. This class label is used to label the unseen data sample. Different types of machine learning algorithms have been used for this process in different ways. For author identification, methods like decision tree and Support Vector Machine (SVM), neural networks etc. are used. Each method gives different classification results. Some of the methods and the authors who have done research using these methods are given in the table below.

TABLE I
An overview of which machine learning algorithms have been used previously in the author identification research [6]

	Diederich	Hoorn	Zheng et.al	Van der Knaap & Grootjen	Joachims
FCA				Yes	
Decision Tree			Yes		
Nearest Neighbour		Yes			
Neural Network		Yes	Yes		
Support Vector Machine	Yes	Yes	Yes		Yes

III. FORENSIC STYLISTICS

Forensic stylistics is the task of applying stylistics to the process of author identification. The stylistics is based on two main rules:

1) Two authors having a common mother tongue do not write in the same manner.
2) The writer does not write in the same way all the time.

The stylistics can be classified into two approaches: qualitative and quantitative.

In the qualitative approach, the errors and personal behaviour of the authors is assessed. They are also known as idiosyncrasies. According to Chaski, this approach could be quantified through databasing. But the databases which would be required for this approach have not yet been developed completely. Without such types of databases to use the significance of stylistic features, the checker's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias. The best result reported was about 72% of recognition rate which is comparatively on the lesser side.

The second approach is known as stylometry. This approach is quantitative and computational, focusing on readily computable and countable language features e.g. word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths. Experimental results have shown that usually this approach provides better results than the qualitative one. Hence, we have chosen to focus more on this approach in our study.

IV. LINGUISTIC FEATURES

Scientific methods to identify authors are based on empirical hypotheses. They do not require any special talent for the person who wishes to use them. Nine important hypotheses have been used till now to identify authors. They are: Vocabulary Richness, Hapax Legomena, Readability Measures, Content Analysis, Spelling Errors, Grammatical Errors, Syntactically Classified Punctuation, Sentential Complexity, and Abstract Syntactic Structures [5].

Vocabulary Richness is given by the ratio of the number of distinct words (type) to the number of total words (token). Hapax Legomena is the ratio of the number of words occurring once to the total number of words. Readability Measures compute the supposed complexity of a document, and are calculations based on sentence length and word length. Content Analysis classifies each word in the document by semantic category, and statistically analyze the distance between documents. Spelling Errors quantifies the misspelled words. Prescriptive Grammatical Errors test errors such as sentence fragment, run-on sentence, subject-verb mismatch, tense shift, wrong verb form and missing verb. Syntactically Classified Punctuation takes into account end-of-sentence period, comma separating main and dependent clauses, comma in list etc. Finally, Abstract Syntactic Structures computationally analyzes syntactic patterns. It uses verb phrase structure as a differentiating feature [5].

V. STRUCTURAL FEATURES

Structural features also provide the means to analyze the features and help in author identification. Some of the structural features are paragraph count, sentence count, number of tabs between paragraphs etc. It helps to analyze the documents which have a large number of paragraphs or are enormous in size. In these cases, word count like feature would prove very tedious; hence structural features prove to be more useful.

VI. GENDER SPECIFIC FEATURES

These features can help identify the gender of the author. This can prove beneficial in case two authors have much writing style in common but their gender is different. Following are some of the ways by which we can analyse the gender of the author. Let M be the total number of distinct words. Then the words ending in able, ive, ible etc are an indication of the author's gender. Thus, in case there is almost same similarity between two authors after analysing their linguistic and structural features, then a difference could be obtained by analysing the gender-specific features.

VII. APPLICATIONS

Besides identifying the writers of books and tracing the individuals who perform malicious activities on the internet, author identification has a number of other applications in the real world. Recently, online examinations have increased in number and one problem faced by the organisers of the exam is that of the identity of the candidate giving the exam. The candidate fills the given form online, and he is asked to write a few lines about himself. Then by analysing his writing, and the writing of the person who shows up at the time of the exam, identity can be ascertained. Secondly, historians and archaeologists find a number of ancient scriptures and sometimes struggle to tell from which era they originated or who wrote them. So author identification can help a great deal in solving this problem. Lastly, it can protect the privacy of a person who requests for it.

VIII. PROGRAM CODE AUTHORSHIP

This is one of the most important applications in which we can use the concept of author analysis. It can be used in the context of software theft and plagiarism, software author tracking and intrusion detection. For e.g. consider a team of many people working on a software project. Then software author tracking helps in the identification of a particular code fragment in the large code. This can prove to be of enormous help during the process of debugging when detecting bugs and finding who was responsible for that piece of code can be a tedious task. It can also be useful during software upgrades and maintenance.

A computer virus or a Trojan horse can be identified in the same manner. Like we identify the author of free text by using linguistic evidence, we can identify the author of a piece of program code by examining peculiar characteristics or metrics of programming style. The typical program metrics which can be used for this task are the use of uppercase and lowercase characters, multiplicity of program statements per line. These features are known as typographical features. Stylistic metrics are the features like length of variable names, preference for 'for' or 'while' loops etc. Programming structure metrics like placement of symbols, use of debugging symbols can also be used. Like programs, e-mails can also be analysed by the use of these features to determine whether an e-mail is

genuine or it is sent by an offender who has enclosed a virus in it.

VIII. CONCLUSIONS

We have studied the need of author identification, how it is the need of the day in this Internet age where forging is easy and identification needs to be foolproof. We have studied the two types of analysis of text. We focused on the quantitative approach since it is more efficient than the qualitative one. Also, we have studied the features which can give an idea of an author's distinct identity and how these features are classified. Finally, we learnt where this task of author identification can be used in the real world.

REFERENCES

[1] Joachim Diederich." Computational methods to detect plagiarism in assessment "ITHET 2006 Paper No. 145

[2] Vineet Chaoji, Apirak Hoonlor and Boleslaw K. Szymanski. "Recursive Data Mining for Author and Role Identification"

[3] Luuk van der Knaap and Franc Grootjen." Author Identification in Chatlogs using Formal Concept Analysis"

[4] Arvind Narayanan Hristo Paskov Neil Zhenqiang Gong John Bethencourt Emil Stefanov Eui Chul Richard Shin And Dawn Song "On the Feasibility of Internet-Scale Author Identification"

[5] Daniel Pavelec, Edson Justino and Luiz.S.Oliveir "Author Identification Using Stylometric Features" Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial , Vol 11

[6] Marcia Fissette, Dr.F.A.Grootjen " Author Identification In Short Texts" 2010

Kunal Borkar B.E. Computer Engineering, Vishwakarma Institute of Information Technology, Pune University.
Abhishek Chandorkar B.E. Computer Engineering, Vishwakarma Institute of Information Technology, Pune University.
Dr. R.S.Prasad Ph.D. (Computer Sc. & Engg.) M.E. (Computer Engg.), MBA(Mktg)