A Proposed System for Author Identification Using Statistical Method

Akhil Gokhale [1], Kunal Borkar [2], Dr. Rajesh. S. Prasad [3]
1, 2 Department Of Computer Engineering, VIIT, Pune, India

3 Department Of Computer Engineering, DCOER, Pune, India

***Abstract* – It is important to protect the identity of the works of an author. The content can be misused in many ways possible. Author identification is the solution for this problem. It analyses an author's writing style by extracting and studying features of his writing and comparing it with other documents. This paper proposes the implementation method for an author identification system. The proposed method consists of extraction of semantic features from the text, scoring these features and comparing them with a set of scores already stored previously in a database. This method is based upon the principles and concepts of text analysis.**

I. INTRODUCTION

The use of the Internet in this era has been widespread. It has grown in gigantic proportions in the last decade or so. Newer sites and newer techniques have been discovered, used and also misused. Blogging and social networking sites have added to the problem. Identities of persons are replicated or misused and malicious acts are performed which amounts to cyber crime. Hence to control this cyber crime, we have to identify the author of a particular document or even a small text. Thus author identification has gained increasing importance.

In author identification process, the basic principle is that each and every author has a distinct writing style. Any manually generated material will reflect some characteristics of the person who generated it. These characteristics may be enough to determine whether two documents are written by the same person. These characteristics are analyzed by studying the writing of a particular author, extracting relevant features, and calculating a score which is unique and fixed for a given author.

Author identification can be formulated as a typical classification problem which is dependent on discriminant features which represents the writing style of an author. In this context, stylometric features which includes various measures of vocabulary richness, efficiency of using language, lexical repetition, using minimum words to convey maximum meaning etc play a very important role.

Stylometry is going to be the method to perform authorship identification. It is a kind of study by which a person can tell about another person by studying his/her writing style. Stylometry grew out of earlier techniques of analyzing texts for evidence of authenticity, authorial identity, and other questions. The modern practice of the discipline received major impetus from the study of authorship problems in English Renaissance drama. Modern stylometry requires lot of aid from computers for statistical analysis, artificial intelligence and access to the growing corpus of texts available via the Internet.

Authorship identification has direct applications in various fields and in various ways. It can be used to determine the author of legal documents, manuscripts or any important text. It can also be used in the detection of cyber crime cases on micro-blogging sites. In the era of growing competition between companies in the IT world, authorship identification can prove to be of tremendous help by helping to detect the author of any source code. Finally, piracy of books, poems and songs can be avoided by authorship identification. In this paper we will put forth a method to bring about authorship identification and its obtained results and observations.

II. LITERATURE SURVEY

A mathematician Augustus De Morgan first proposed the use of statistical tools to test the questions of authorship. He advocated the use of average word length to numerically characterize authorship style. Afterwards, Thomas Mendenhall a physicist proposed that an author has a characteristic curve of composition which is determined by how an author uses words of different lengths frequently. Since then most of the work done was similar which proposed the use of different features for authorship identification.

In the electronic era, the plagiarism issue came to the fore. This called for greater accuracy and more novel methods to perform authorship identification. Stylometry was one of the most popular and accurate methods adopted. Stylometry is a method in which a person can tell about another person by studying his writing style. Other methods are stylistics and computational linguistics which address specific issues of an author. Fields like text classification, machine learning and forensic linguistics also impact on the current study. Plagiarism detection can be considered to be complementary to stylometric authorship identification.

In authorship attribution there are three kinds of evidence which can be used to detect the authorship.

- External evidence: This includes the handwriting or signed manuscript of the author.
- Linguistic evidence: This concentrates on the patterns of the words and the actual words used in the documents.
- Interpretative evidence:- it is mainly concerned about the information which can be derived from the document i.e. when was it written, what the author meant by it etc.

By using statistical methods, accurate calculations can be performed and this has helped to successfully deduce author identity in the past.

Recently, interest has been growing in applying stylometry to the content generation where the content is checked and its authenticity is determined. Shane Bergsma, Matt Post and David Yarowsky are currently evaluating stylometric techniques to determine authorship identification in the field of scientific writing. Rajesh Prasad and Uday Kulkarni in their paper [6] have proposed the use of various innovative features and their extraction which plays an important role in the task of text analysis. The features discussed by them are related to

word occurrence, sentence occurrence and their similarity among various paragraphs.

III. PROPOSED SYSTEM

We propose a system which will help identify the author of any random text document by extracting features, performing some preprocessing tasks and comparing the results with the already stored results in the database. The system will compare the results with a small number of selected authors whose profile will be stored in the database. This is a limitation of the proposed system. All efforts will be put into trying to minimize this limitation. Also, the system will accept only the documents which will be well-formatted and will be written according to the rules of English language.

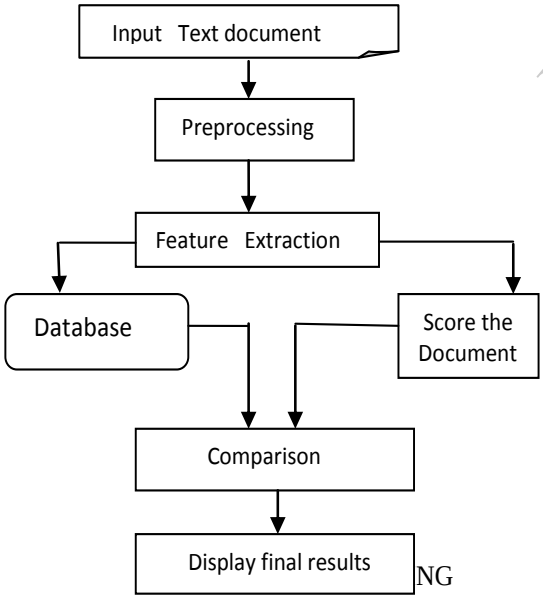
This system will consist of the following important modules:

Preprocessing: Some preprocessing is required to be performed on the document before it can be subjected to the process of authorship identification.

Feature Extraction: For identifying the author, features are extracted from the document, some calculations are performed upon them, a score is calculated and the results are stored against the corresponding author in the database.

Results: After the preprocessing, feature extraction and scoring is performed, the results will be compared to the already stored results in the database and the author to whose result the current results matches most will be determined. The percentage of similarity will be calculated and the final results will be displayed.

The system can be displayed in the form of a block diagram as follows:



NG

For the document to be ready for author identification process, some preprocessing is required to be done on the text in the document. This gives more efficient and accurate results.

The main preprocessing which we are going to perform in the proposed system is Stop-words Removal. Stop-words are the unnecessary words in a text which do not add any additional meaning to the text. They are simply the connectors or the basic building blocks without which a sentence cannot be constructed at all. Some examples of stop-words are conjunctions like 'and', 'or', articles like 'a', 'an, 'the', question words like 'what', 'which' and other words like 'of', 'for' etc. These words are useless and they add more complexity to the feature extraction process if not removed. Hence these stop-words are removed before performing any feature extraction or any other function on the input document.

V. PROPOSED FEATURES

An author's writing style is found out from the document written by him. There are several features which are unique and distinct to a particular author and these features cannot be changed much consciously. They remain more or less the same for a particular author. Hence extracting these features and studying them gives accurate information about the writing style of an author.

Features to be extracted are of various types. They can be based upon the words in a sentence, sentences in a text or upon individual words. Since documents for authors will vary in length we will calculate the average of each feature so comparison is fair. Some of the most useful features like title count and statistical count have been taken as stated in [6]. We are considering the following features for extraction and giving information about the writing style of an author:

Word count: This feature calculates the total number of words in the text of an author. Since preprocessing has been performed earlier, only the meaningful words will be counted which gives precision and accuracy to this feature and also saves a lot of unnecessary processing. Hence,

$$Word\ count = Number\ of\ words\ in\ the\ text \quad (1)$$

Phrase count: Phrases are parts of speech which convey a meaning or a moral in a fixed arrangement of words. For e.g. Tit for tat is a phrase which talks about revenge. Clearly, an author who uses more phrases in his text has a better knowledge on the English language and has good control over it. Hence phrase count can be used as a differentiating feature between authors.

$$Phrase\ Count = (No\ of\ phrases\ in\ the\ text \times 100) \div Word\ Count \quad (2)$$

Punctuation count: Punctuation symbols are essential to write any text in the English language. Hence this feature may seem worthless. However, even a punctuation symbol has information conveying properties. For e.g. we can tell the difference between a sentence ending with '?' and a sentence ending with '!'. Hence the punctuation count and the various types of punctuation symbols which the author uses is a very important distinguishing feature.

$$Punctuation\ Count\ = (No.of\ punctuations\ in\ the\ text\ \times\ 100) \div Word\ Count \quad (3)$$

Statistical count: These are the words which convey the information how much the author is willing to use statistical and precise data. It conveys whether the author is being precise or is making an estimate.

$$Statistical\ count = (No.of\ statistical\ term\ in\ the\ text\ \times\ 100) \div Word\ count \quad (4)$$

Title count: A document has a title or heading which conveys the information what the document is about. If the title is used by the author frequently in the text, we can say that the author is sticking to the topic and not deviating from it. On the other hand, if he is not referring to the title frequently in the document, then he is somewhat deviating from the main topic. For e.g. If the title is 'The Tiger' then the author should refer to the tiger more often and hence mentions the word 'tiger' frequently in the document This clearly shows that the author talks about the actual topic and not anything else. Hence the feature of title count is an important distinguishing feature as well.

$$Title\ count = (No.of\ times\ title\ occurs\ in\ text\ \times\ 100) \div Word\ count \quad (5)$$

Complexity of words: The English language is vast and has a large number of synonym/antonym pairs. Some words are simple and known to all while some words are complex and are known only to those people who have a good knowledge about the language. For e.g. a normal author can say ' He spends a lot of money'. However another author who has good knowledge about English can say 'He is extravagant'. Hence use of complex words conveys the author's study and mastery of the language. Hence this is feature is considered as a distinguishing feature.

$$Complex\ count = (No\ of\ complex\ words\ in\ the\ text\ \times\ 100) \div Word\ count \quad (6)$$

By calculating all these features, we will then calculate an average score of these features which will remain more or less constant for a given author. Then the score of the input document will be compared with the scores already in the database, and the final results will be generated. The author's similarity will be given according to individual features as well as according to the total average of the scores of the features.

VI. CONCLUSION

A lot of study has been done in the field of text mining and natural language processing. We thought of constructing the above proposed system for authorship identification and to gain maximum accuracy and efficiency. Although this system has some practical limitations, all possible efforts are being put to construct the system efficiently and minimize the limitations The system will give outputs in the format of a similarity percentage of the author. The author whose documents are scored and previously stored in the database, and who matches most closely to the current document will be displayed as the results. Thus we can determine by this system how similar an author is to another one by studying their features and analyzing their documents with the help of these features.

REFERENCES

[1] Vineet Chaoji, Apirak Honloor and Boleslaw K. Syzmanski, "Recursive Data Mining for Author and Role Identification", Proceedings of 3rd Annual Information Assurance Workshop ASIA'08, Albany, NY 4-5 June 2008, pp 53-62.
[2] Joachim Diederich, "Computational methods to detect plagiarism in assessment", ITHET 2006 Paper No. 145.
[3] Masahiro Terachi, Ryosuke Saga and Hiroshi Tsuji, "Trends Recognition in Journal Papers by Text Mining", 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.
[4] Lakshmi, Pushpendra Kumar Pateriya, "A Study on Author Identification through Stylometry", International Journal on Computer Science & Communication Networks, Vol 2(6), 653-657.
[5] Daniel Pavelec, Edson Justino, and Luis S.Oliveira, "Author Identification using Stylometric Features", Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol 11, No 36, Pages: 59-65, 2007.
[6] Rajesh Shardanand Prasad, Uday Kulkarni, "Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization", Journal of Computer Science 6(11): 1366-1376, 2010 ISSN 1549-3636
[7] Rajesh. S. Prasad, U. V. Kulkarni, J. R. Prasad, "Connectionist approach to Generic Text Summarization", World Academy of Science, Engineering and Technology 31 2009.
[8] Rajesh. S. Prasad, Dr. U. V. Kulkarni, "Two approaches to automatic text summarization: Extractive methods and Evaluation", ISSN 0974-4983, IJCECA January 2010- March 2010 Spring Edition 2010 PP[24-36] Volume 01, Issue No 01.