International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

Study of Different Methods for Author Identification

Vishal Chandani, Ninad Deshmane, Kshitij Buva, Suvrat Apte, Dr. R.S. Prasad (Head of Department)
NBN Sinhagad School of Engineering,SavitribaiPhule Pune University: Computer Engineering
NBNSSOE
Pune, India

***Abstract*—The documents written by the author exhibit a distinctive style and identification of an anonymous document of any author is one of the most important and ambitious task in computer system. Evaluation of an individual's work is an important characteristic in distance education. It is difficult for institutions to verify whether the individual participated is the person enrolled. Author identification has played a crucial role to prevent plagiarism as well as verifying an individual's identity. In our paper we have focused on methods like "support vector machines", that are useful for author recognition to prevent infringement of copyright. It also includes various methods such as 'Type-Token ratio', 'CUSUM technique', 'Readability Measures', 'E-mail content analysis' and 'Chi-Square method'. These methods are used in text classification and in various other applications.**

Keywords*— *plagiarism, infringement, support vector machine, type-token ratio, cusum, chi-square.

I.INTRODUCTION

In the age of information revolution, electronic documents are becoming principle media of business and academic information. Thousands and thousands of documents are produced and made available on the internet every day. In order to determine the actual author of each of these articles or documents, we need a system that could do that by itself in a self learning way. In addition to this we need a system to determine the richness of each of these e-documents so as to separate the relevant ones from the less relevant ones. Thus, a system using text analysis would effectively be serving this purpose.

Text mining is an emerging field that attempts to extract relevant information from natural language text. It may be generalized as the process of analyzing text to extract information that is useful for particular purposes. Unlike the data stored in databases, textual data is unstructured and ambiguous. Therefore, it is algorithmically complex to deal with. Text has always been the most common way of exchanging the information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial. [6]

In the age of E-learning, face to face interaction is impractical due to distance. Therefore, institutions are providing online courses as a part of distance education. It is difficult for the institutions to detect plagiarism in assessment of answers submitted by students. Thus, there is need of Author Identification. [1] It also plays an important role in E-mails, short text messages and in many other domains for avoiding duplication of author identity. This paper summarises different methods which have proved to be beneficial for authorship identification.

II. RELATED WORK

The attempts for identifying authors were made since mid 18th century. The first significant attempt was made by Mendenhall in the year of 1887. It was based on the plays of world famous English writer Shakespeare. Mendenhall applied his technique to the Bacon-Shakespeare controversy in the year of 1901. This work was followed by Zipf in 1932 and by Yule in 1938. This effort made by Zipf and Yule highlights the statistical analysis of text. In the later half of 20th century, the research in author identification was continued by Mosteller and Wallace. Their work was based on the authorship of 'The Federalist Papers'. [7][4] This paper consisted of 146 essays based on politics written by number of authors like Alexander Hamilton, James Madison, John Jay. This was undisputedly the best contribution in Author Identification. Their approach for Author Identification was based on Bayesian statistical analysis of the occurrences of prepositions and conjunctions like 'or', 'to', 'but', 'and', etc. Thus, it was helpful in classifying the authors. The research in Author Identification then saw a tremendous attention and sped up rapidly. Many papers were published in the following years. Decades of research was made. These research papers consisted of a small set of authors due to the scarcity of textual data. For last few years, the research in Author Identification has changed its direction towards advanced machine learning techniques. Internet today has enormous amount of documents available with it and thus it is very challenging to carry out proper Author Identification process over entire Internet. In the period of 2006 to 2011, only Koppel et al have tried successful Author Identification on Internet scale.

III. METHODS

There are a number of methods for carrying out Author Identification. Some of the commonly used methods are listed below:-

A. The Chi-Square Method / Distribution

The Chi-Square Method is a statistical method for evaluating the relationship between expected values and observed values. It is a tentative test in which the sampling distribution of the test is a chi-square distribution when the null hypothesis is true. The Null Hypothesis is the assumption that the hypothesis in question is false. In this method the truth value of the null hypothesis is analysed. It should be evaluated as false. This measure is known as "statistical significance". There may be a possibility where the results derived from the conclusion may suggest that the null hypothesis is false. A significant threshold needs to be set by the analyser to rectify the mistake. The results above the threshold are considered significant. A result below 5% is often considered as significant because the approximate threshold used is 0.05.

B. Type / Token Ratio

This method is commonly used for small sized documents. The number of tokens, say 'n' in a text is said to be equal to the number of words in that text. But even if the text has 'n' words, all the words are not unique and many of them are repeated. Analysts have calculated the average repetition of words to be around 40% of the original word count. The term 'types' represents the number of unique words in a text and the term 'token' represents the total word count of the text. This method is based on the ratio of 'types' and 'tokens'. The average ratio for every author is stored in the database. When a new document is to be identified, the ratio for that document is calculated and matched with the ratio values in the database. This is not one of the accurate methods of Author Identification. [1][4]

C. E-mail Content Method

The e-mail content analysis is limited by sensitive and ethical values. The open email content includes newsgroups, mailing lists etc. However, in such type of public e-mail databases, it is generally very complex to detect large enough 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.[9] 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.

D. CUSUM Technique

Every author has his/her own and unique writing style. The author tends to use some of the words frequently. These words are helpful in determining the author. This technique starts by measuring the length of the sentences in terms of number of words.

The proper nouns used in the sentences are treated as single symbols. The average sentence length for a particular document is calculated and then eachsentence is compared with the average length and is marked with a '+' or a '-' according to its length. This allows good identification even if the feature set is large. In CUSUM technique the text is analysed by the use of function words by an author and shorter words like vowel words (words which begins with a vowel) and the combinations of both i.e. the short + vowel word. There are nine test samples included in this technique for identifying the authors. Thus, this method plays a very important role in author identification. [1][4]

E. Readability Measures

All the documents are not readable for a general domain of people. This method is based on the complexity of documents. The complexity of documents usually is based on the technical terms used in the writing. The readability of the document is calculated by counting number of words in a sentence or counting number of technical terms or average number of syllables per word. Some of the important aspects of readability are:

1. Readability index formulae are language dependent.
2. Readability does not reject understandability.
3. Readability does not consider actual content.

This method is said to be one of the accurate methods of Author Identification. [1][4][5]

IV. CONCLUSION

In the process of Author Identification the important aspects of a document remains unknown and it is difficult to analyze the document completely. This has been one of the major issues faced in identifying the authors. The above specified methods in this paper provide a solution to overcome this problem. Of all the above methods mentioned, Cusum is the most efficient technique for identifying an author. All these methods are based on Machine Learning. Thus, Author Identification can be useful for detection of plagiarism in the context of e-learning.

REFERENCES

1. Joachim Diederich, "Computational methods to detect plagiarism in assessment" Paper No. 145: DiederichJ.:Computational methods to detect plagiarism in assessment 2006ITHET.
2. Todd K. Moon, Peg Howland, Jacob H. Gunther, "Document Author Classification using Generalized Discriminant Analysis", Utah State University.
3. D. Holmes, "A Stylometric Analysis of Mormon Scriptures and Related Texts," Journal of the RoyalStatistical Society, A, 1992.
4. Akhil Sanjeev Gokhale, RajendraKrishnatDalbhanjan, Dr.Rajesh.S.Prasad, "Review and Study of Different Methods for Author Identification", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 6, August 2012.
5. S. Theodoridis and K. Koutrombas "PatternRecognition." New York: Academic Press, 1999.
6. EfstathiosStamatatos, "A Survey of Modern Authorship Attribution Methods", Dept. of Information and Communication Systems Eng.University of the Aegean Karlovassi, Samos – 83200, Greece.
7. Abdur Rahman, Haroon A. Babri, Mehreen Saeed, "Feature Extraction Algorithms for Classification of Text Documents", ICCIT 2012, pp. 231-236.
8. Berry M (ed.) (2003). Survey of Text Mining: Clustering, Classi_cation, and Retrieval. Springer-Verlag.ISBN 0387955631.
9. Nihar Ranjan, Dr. Rajesh. S. Prasad, "Author Identification in Text Mining Used in Forensics", International Journal of Research in Advent Technology Volume 1, Issue 5, Dec 2013.
10. Uplavikar Nitish Milind, Wakhare Sanket Shantilalsa, Dr. Rajesh. S. Prasad, " Feature Based Text Summarization", International Journal of Advances in Computing and Information Researches Volume 1– No.2, April 2012.