IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Authorship Attribution of Messages

Aneri R. Oza[1], Kajal K. Ladkat[2], Mrunalini V. Bhosale[3], Pranjali M. Marne[4], Dr. Rajesh S. Prasad[5]

Pursuing BE, Dept of Computer Engineering, NBN Sinhgad School of Engineering, Pune, India[1, 2, 3, 4]

Professor, NBN Sinhgad School of Engineering, Pune, India[5]

Abstract: Author identification can be implemented using n-gram method which exhibits a unique way of identification of any author of an anonymous document. It is indeed one of the important parts of the Computer System and author can be uniquely identified by its different textual features like punctuations, phrases, length of the sentence and others. To prevent plagiarism and verifying an individual's identity Author identification has played a crucial role. In this paper focus is on Author identification and verification for a set of provided text using n-gram method. Our future vision is to identify and verify an author with high performance and solution for feature extraction from various text documents.

Keywords: Author verification, author identification, n-gram method, false acceptance rate, false rejection rate.

I. INTRODUCTION

In this era of information revolution, electronic documents are becoming principle media of business and academic information. On internet large amount of documents are created and made available on daily basis. In the early 1990s, author identification research was dominated by various attempts to define features for evaluating writing style. One of the important things is to determine the actual author of the document or article. There is a necessity to handle large amount of electronic texts available through Internet System. Along with that system should find out richness of the document and compare the relevance of the document with other author based on some threshold parameter. Test analysis is one way to find out the author of the document based on text analysis. For avoiding the duplication of the author identity, authorship verification also plays an important role in short text messages and many other domains [11]. Authorship analysis can be carried by using different perspectives which are authorship identification, authorship verification and profiling [1]. The authorship attribution describes the way to determine the most similar author of a target document between lists of known authors [10]. Verification tasks consist in determines if a given text was written by a person given a small set of document. The textual plagiarism can be described by the unacknowledged use of an exact copy or a slightly modified version of the text written by another author. The main aim of authorship verification and attribution is the problem of answering the question whether or not a sample text was written by a specific person, given a few other documents known to be authored by them.

The n-gram is based on tokens (characters, words) that are most common in the considered document which are indeed helpful in detection of authorship.

II. RELATED WORK

The attempts for identifying authors were made since mid18th 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 latter 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 '[3][4].

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 [5]. 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 accordingly. The research in Author Identification then saw a tremendous attention and speed up rapidly as it was used to assess the document with its author.

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 the direction

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

towards advanced machine learning techniques of Author Identification. 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. These researchers have successfully implemented their work respectively.
Plagiarism detection used as a method which attempts to detect the similarity between two different pieces of work but is unable to determine if the documents were produced by the same author[6].

A document is represented by a feature vector that contains one Boolean attribute for each word that occurs in the training collection of documents. When generalized this method by using word sequences it forms a sequence, termed n-gram as a feature. For generation of n-gram features, consideration was made for small value of n, number n-gram features that can be discovered in the document, which increases in a way such that for every n-gram, there is at least one n+1gram that has n-gram as a starting sequence. Thus features are growing linearly and the number of features with minimum frequency grows more slowly. So, an efficient algorithm for generating these feature sets, should therefore avoid generating all n-grams [7].

III. N-GRAM METHOD

There are a number of methods for carrying out Author Identification such as text processing and token ratio based approach but here we have implemented N-Gram approach. Most important benefits of n-gram model are simplicity and scalability.

N grams –Authorship Verification:
A collection of profiles are generated separately for individual users possessed by N-Gram model. There are two modes of operations, namely, training and testing, where the user profiles are built and then checked, respectively. The training phase consists of two steps. During the first step, from sample documents, the user profile is derived by extracting n-grams.
During the second step, a user specific threshold is computed which is used in the verification phase later. There is a need to calculate the threshold value for each user, based on n grams algorithm.

Following steps will be used in N grams approach:

1. Find out the threshold value for each user;
2. Divide the given user U data into block of character of size p.
3. Find out the n grams for each block defined of size p. For another users also find out they also need to find out the n grams from the set of document.
4. If and only if the percentage of unique n-grams shared by block of user U is greater than threshold value specified for the user then a block p is said to be a genuine sample of user U [10].

In our work we have used, $\in_u$, the threshold calculated for the user from database and r which is considered as zero.
For any existing user data we have 25 records in db, False Rejection =0.

```
Loop 25 times
{
        If threshold ≤ record threshold from database;
        False rejection= false Rejection+1;
}// Loop complete
```

$$\text{False Rejection Rate} = \frac{\text{False Rejection}}{\text{Number of record(25 in our case)}}$$

Figure 1: FRR

For any user, we have 25 records of users in database.

```
Loop 25 times
{
        Threshold of new ≥ threshold+ r;
        False Acceptance = False Rejection + 1;
}// Loop complete
```

$$\text{False Acceptance Rate} = \frac{\text{False Acceptance}}{\text{Number of users}}$$

Figure 2: FAR

IV. EXPERIMENTS AND RESULTS

The document to be identified is provided as an input to the system. The result is calculated by applying n-gram method on the document provided. The evaluated result is compared with the average of the features extracted (according to threshold value) from the other documents of genuine authors stored in the database. The author of the document is the one whose average values are close enough with the values of input document.

Example:
We have considered a set of documents written by different authors and stored them at the back end. There is a threshold value calculated of the stored documents and then compared to the unknown document as an input. If the unknown document has value more than threshold then it is a genuine author else not. Based on n-grams of the user block and the user threshold value we can say that user is genuine author of block if n-gram value of block is greater than threshold value calculated for user using FRR (False Rejection Rate), FAR (False Acceptance Rate).

TABLE I: 4 GRAMS FOR 3 USERS

No. of users	FRR	FAR	r	δ
1	0.47	0.57	0	5.438
2	0.18	0.83	0	4.46
3	0.69	0.29	0	5.13

DOI 10.17148/IJARCCE.2016.54282

Below shown are some test cases tested for application:

TABLE III: TEST CASES

Test ID	Test Case	Input	Output	Pass/ Fail
1.	To check whether application is up and accessible.	First of all start the application then provide the credentials.	The application window should show open after providing the correct credentials.	Pass
2.	To check whether the file is getting uploaded successfully.	Upload the file and check the result for the same	The application window should upload the file successfully.	Pass
3.	File pre-processing should happen.	After successfully uploading the file, system should calculate the threshold and n-grams for the file.	The application window should show the threshold, n -grams for the uploaded file.	Pass
4.	To check whether the author is verified or not for uploaded file.	System should compare the n–grams for uploaded file with the database value and provide the required information.	The application window should show author verification result as per the input provided	Pass

V. CONCLUSION

Thus we have created a system that learns the writing styles of various authors. In the process of Author Identification the important aspects of a document remains unknown and it is difficult to analyse the document completely. The learning system is based on statistical analysis of the text document. This has been one of the major issues faced in identifying the authors. The technique used is based on a combination of supervised learning and n-gram analysis hence identifying the anonymous document.

REFERENCES

[1] Vishal Chandani, Ninad Deshmane, Kshitij Buva, Suvrat Apte, Dr. R.S. Prasad, "Study of Different Methods for Author Identification", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4, Issue 01,January-2015.

[2] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", Dept. of Information and Communication Systems Eng.University of the Aegean Karlovassi, Samos – 83200, Greece.

[3] Abdur Rahman, Haroon A. Babri, Mehreen Saeed, "Feature Extraction Algorithms for Classification of Text Documents", ICCIT 2012, pp. 231-236.

[4] C. E. Chaski, "Who's at the keyboard: Authorship attribution in digital evidence investigations", International Journal of Digital Evidence, 4(1), Spring 2005.

[5] Akhil Gokhale, Kunal Brkar, Dr. Rajesh S. Prasad, "A Proposed System for Author Identification Using Statistical Method", International Journal of Engineering Research & Technology(IJERT), ISSN: 2278-0181, Vol.2, Issue 9, September-2013.

[6] Prof. Nihar Ranjan, Dr. R.S. Prasad, "Author Identification in Text Mining For used in Forensics", International Journal of Research in Advent Technology, E-ISSN: 2321-9637, Vol. 1, Issue 5, December 2013.

[7] Mubin Shaukat Tamboli, Dr.Rajesh S. Prasad, "Authorship Analysis and Identification Techniques: A Review", International Journal of Computer Applications(0975-8887), Vol.77-No.16, September 2013.

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

[9] A. Abbasi and H. Chen. Writeprints, "A stylometric approach to identity level identification and similarity detection in cyberspace", ACM Trans. Inf. Syst., 26:7:1–7:29, April 2008.

[10] Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, Isaac Woungang, "Authorship Verification for Short Messages using Stylometry".

[11] Vishal Chandani, Ninad Deshmane, Kshitij Buva, Suvrat Apte, Dr. R.S. Prasad, "Author Identification Method using Hybrid Technique", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4, Issue 04,April-2015.