International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

Author Identification System using Hybrid Technique

Vishal Chandani, Ninad Deshmane, Kshitij Buva, Suvrat Apte, Dr. Rajesh Prasad
Department of Computer Engineering,
NBN Sinhgad School of Engineering,
Savitribai Phule Pune University, Pune, India

***Abstract*—Author identification of a document can be implemented using statistical or computational method. In author identification, an author can be distinguished by his unique writing style. The basic idea behind author identification using statistical or computational method authorship is to measure different textual features for determining the author. The statistical method allows us to analyse and explore aspects of a text that wouldn't be easy for us to identify. In this paper we have focused on CUSUM technique which is a statistical method. CUSUM technique is based on calculation of average sentence length and set of words used by the author frequently.**

Keywords— Part of speech (POS) tagging, CUSUM, Natural Language Processing (NLP) .

I. INTRODUCTION

In authorship identification problem, an undisputed author is identified from a given set of authors for whom samples of text are available. In the early 1990s, author identification research was dominated by various attempts to define features for evaluating writing style, a line of research known as 'Stylometry'. Hence, various measures such as sentence length, word length, word frequencies, character frequencies and vocabulary richness functions had been proposed. Since the late 1990s, there has been a significant change in author identification studies.[3] There is a necessity to handle large amount of electronic texts available through Internet media such as emails, blogs, online forums, etc. efficiently. This fact had a notable impact on areas such as information retrieval, machine learning, and natural language processing (NLP). The development in Author identification techniques can be stated as follows:

- Representation and classification of large amount of text developed efficiently with information retrieval research.
- Multidimensional and sparse data became easy to handle using powerful machine learning algorithms.
- Standard evaluation methodologies have been established to compare different approaches on the same benchmark data.

It is easy to analyse text efficiently using NLP research developed tools for representing the style such as syntax-based features).

In the last decade in numerous efforts have been taken in the field of author identification to develop practical application that deal with real world text.[7]

In this paper, we have focused on various features associated with author identification such as

1. Author verification-to decide whether a particular document was written by a particular author or not.
2. Plagiarism detection-finding similarities between two texts.
3. Author characterization-extracting details about the writing style of the author.[1]

Earlier studies in author identification intended various features to measure the writing style, known as style markers, under different circumstances. The text representation features for stylistic purposes is mainly based on the computational requirements for quantifying them.

Application specific features can only be defined in certain text domain or languages while syntactic and semantic features require deeper linguistic analysis.

II. RELATED WORK

The first significant attempt to measure the writing style of an author dates back to 19th century, with the innovative study of Mendenhall (1887) which was based on the plays of Shakespeare. In the beginning of 20th century, Yule and Zipf continued the further research. Mosteller and Wallace opposed the traditional methods based on human expert analysis and engendered initiated non-traditional author identification studies, as opposed to traditional human expert-based methods. Non-traditional approach to author identification includes various measures like word length, frequency of characters, sentence length and vocabulary richness functions. [3][8] The methods suggested during that era were not computer-based but were computer aided i.e. it did not aim at developing a fully automated system. There were some methods which resulted in solutions which people thought were very close. The most significant method for achieving such result is the CUSUM technique which was proposed by Morton and Michaelson in 1990.This method became renowned and was accepted in courts but the research community considered it as unreliable. The lack of objective evaluation proved to be the major concern during that period. The basis of testing was documents of unknown or disputed authors which led to inaccurate results to identify the author. The main limitations for identifying an author during that period were as follows:

4. The evaluation of the suggested techniques was mainly instinctual i.e. inspection of visual scatterplots.

5. The lack of benchmark data made it difficult to compare different techniques.
6. The database of the candidate author was too small.[6]

III. PARTS-OF-SPEECH TAGGER

A Parts-Of-Speech Tagger (POS Tagger) is software that reads text in some language and used for assigning corresponding parts of speech to each word such as noun, verb, adjective, etc. The computational applications such as author identification use more refined POS tags like 'noun-plural'. POS tagging is more difficult than just listing the words and their parts of speech, because some words can be represented with more than one part of speech at different times, and because some parts of speech are much more complex. Such instances can be observed in many situations like, "dogs", which is usually thought of as just a plural noun, can also be a verb.

In part-of-speech tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English. POS tagging work can be used in a variety of languages, and the set of POS tags used varies according to the language. [5][9]The tags used may lead to inconsistencies such as case-marking for pronouns but not nouns in English. The set of POS tags change the form of a word to express a particular grammatical function in languages such as Greek and Latin can be very large.

In the early days the tagging was done by humans which was very inefficient but now it is performed using computers. POS tagging is done with reference to computational languages, using algorithms which helps in determining discrete terms and hidden parts of speech in accordance with a set of descriptive tags. POS-tagging algorithms are classified into two different groups: rule-based and stochastic. The first and most widely used English tagger is the E. Brill's tagger that employs rule-based algorithms.[2]

In POS tagging method, the entire document is given as input to the system. The sentence is divided into different tags categorized according to its respective part of speech. The primary parts of speech in this tagger are co-ordinating conjunction, number, determiner, adjective, noun, pronoun, preposition, verb, wh-pronoun and adverb.

Percentage of each category is stored in the database using the following formula:

Let ‘P’ represent the percentage of POS category in a document.

P = (Total number of words belonging to category P / Total number of words in the document) * 100

IV. CUSUM METHOD

This method is based on the unique writing style of every author. Each author has a unique set of words which he may tend to use frequently. An author’s document may be quantified using various features out of which ‘average sentence length’ is the most important in CUSUM technique. [4][6]

In this method, the entire document is given as input to the system. Then the length of every sentence is calculated and average sentence length is calculated. A parameter ‘dx’ is set an arbitrary value which is usually small ($1 < dx < 5$). After the calculation of the parameter ‘dx’, a range value ‘R’ is calculated with the formula given below:

Range R = dx + average sentence length + dx

Length of every sentence is checked, and if it is in the range ‘R’, it is marked as a sentence with average length. If it is greater than ‘R’, it is marked with a ‘+’ symbol which indicates that it is greater than the average sentence length. And if it is less than ‘R’, it is marked with a ‘-’ symbol to denote that it is less than the average sentence length. After analysing the entire document, the total number of average length sentences, more than average length sentences and less than average length sentences is calculated and stored in the form of percentages.

Let the literal ‘A’ denote the percentage of number of sentences with length within range ‘R’.

A = (Total number of sentences with average length / Total number of sentences) * 100

Let the literal ‘G’ denote the percentage of number of sentences with length greater than the range ‘R’.

G = (Total number of sentences greater than the average length / Total number of sentences) * 100

Let the literal ‘L’ denote the percentage of number of sentences with length less than the range ‘R’.

L = (Total number of sentences less than the average length / Total number of sentences) * 100

The literals ‘A’, ‘G’ and ‘L’ are stored in a database.

V. EXPERIMENT & RESULTS

The document to be identified is given as input to the system. The result is calculated after applying POS tagging and the CUSUM method on the document. The evaluated result is compared with the average statistics of documents of stored authors in the database. The author of the document is the one whose statistics are close enough with the statistics of input document.

Example:

We have considered a set of documents written by the author Dale Carnegie. Five documents are given as input to the system and the average statistics is evaluated. These statistics are compared with the statistics of the input document. Based on this comparison, if the results of the statistics are close then the document is written by the author.

Let A1, A2, A3, A4, A5 be average statistics of five different authors.
Let T1, T2, T3 be the three test cases to identify the author of the particular document.
To determine the author based on the statistics we have used the following formula:
Let,
S1- Statistics of Training Document
S2- Statistics of Testing Document

$C_n = (1 - (|S1-S2| / S1)) * 100$

After applying the above formula for each and every textual feature, we calculate the average for all the values of 'C_n'.

Table 1: Statistics of Textual Features

Textual Features	Statistics of Training Document					Statistics of Testing Document		
	A1	A2	A3	A4	A5	T1	T2	T3
No. of words	416	365	429	359	144	469	267	387
No. of sentences	26	25	34	25	17	28	19	22
Avg. sentence length	15	15	13	13	8	16	14	14
No. of sentences less than avg. length	54	45	45	51	50	50	42	36
No. of sentences more than avg. length	34	38	45	34	30	42	36	28
No. of sentences with length more than 13	14	14	11	13	12	11	32	17
No. of sentences with length more than 18	34	35	30	30	9	43	22	24

Table 2: Parts-Of-Speech Statistics

POS	Statistics of Input Document					Average Statistics		
	A1	A2	A3	A4	A5	T1	T2	T3
Conjunctions	3	2	3	4	3	3	5	2
Numbers	0	0	1	0	0	1	1	2
Determiners	9	9	6	7	9	11	11	8
Prepositions	10	10	10	9	7	12	11	5
Adjectives	4	5	5	5	5	3	7	3
Nouns	14	16	14	13	17	17	20	17
Proper Nouns	3	4	6	5	6	7	11	3
Pronouns	14	8	9	11	10	14	5	6
wh-Pronouns	0	0	0	0	0	0	0	0
Adverbs	4	5	7	6	4	3	7	3
wh-Adverbs	0	0	0	0	0	1	0	1
Verbs	20	21	23	22	22	18	16	16
Foreign words	0	0	0	0	0	0	0	0
Phrases	0	0	0	0	0	0	1	1

By referring the above table, we get the following result:

Case 1: The test case T1 matches A1 because the matching percentage is 87. Hence, A1 is the author of test case T1. We have tested 10 documents and 7 were successfully identified. Hence 70 percent documents were identified.

Case 2: The test case T2 does not match with any of the authors. Hence, we do not have any documents of this author.

Case 3: The test case T3 does not match with any of the author even though the author's document is present in the database because the match percentage is 42. We have tested 10 documents and 3 were not identified. Hence 30 percent documents did not match with any author.

VI. CONCLUSION

Thus we have created a system that learns the writing style of various authors. The learning system is based on statistical analysis of the text document. This system can also identify the author of a certain document. Identification is done by comparing statistics of input document with the statistics present in the database. The statistical analysis is a combination of CUSUM technique and Parts-Of-Speech analysis.

REFERENCES

1. Joachim Diedrich, -Computational methods to detect plagiarism in assessment Paper No. 145: DiederichJ :Computational methods to detect plagiarism in assessment 2006 ITHET.
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 Royal Statistical Society, A, 1992.
4. Akhil Sanjeev Gokhale, Rajendra Krishnat Dalbhanjan, 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. Theorodis and K. Koutrombas –Pattern Recognition, New York: Academic Press, 1999.
6. Efstathios Stamatatos, --A Survey of Modern Authorship Attribution Methods, Dept. of Information and Communication Systems Engineering, 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, Classification 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 Volume1, Issue 5, Dec 2013.

AUTHORS


Kshitij Sandeep Buva
Pursuing B.E.(Computer Science & Engineering), Savitribai Phule Pune, NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune 411041, India.


Ninad Ravindra Deshmane
Pursuing B.E.(Computer Science & Engineering), Savitribai Phule Pune, NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune 411041, India.


Suvrat Rajesh Apte
Pursuing B.E.(Computer Science & Engineering), Savitribai Phule Pune, NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune 411041, India.


Vishal Sunil Chandani
Pursuing B.E.(Computer Science & Engineering), Savitribai Phule Pune, NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune 411041, India.


Rajesh Sharadanand Prasad
Has received masters (M.E. Computer) degree from College of Engineering, Pune in 2004 and his doctorate degree from SGGS, Nanded in 2012. He is working as professor and head in Computer Engineering department of NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune. He is having 18 years of experience. His area of interest is soft computing, text analytics and information management. He has published over 40 papers in national and international journals. He is a lifetime member of CSI and ISTE.