



Author identification system

Prof. Manisha P. Mali, Kalyan Satpute, Aayush Sharma
Computer Engineering, VIIT, Pune, India.

Abstract: Every one of us has different approach to speak and write, and there exists a long history of linguistic and stylistic analysis into authorship attribution. In last year's, practical application for author identification have grown in area such as computer forensic(linking intercepted message to each other and to find rebel), criminal law(identifying author of payoff notes and harassing letter), civil law and computer security (tracking author of computer source code).

This paper proposes the implementation of author identification system. This proposed system is based upon the principles and concepts of text analysis. For ensuring maximum accuracy in identifying author of the document we will be using TF-IDF algorithm which consists of extraction of features from the text, scoring these features and comparing them with a set of scores stored in the corpus.

Keywords: -author identification, writer identification, Text mining, author stylometry, author database.

I. Introduction

Author identification is a process in which the author of a text is identified. Most known literary texts can easily be attributed to a certain author because they are, for example, signed. Yet sometimes we find unfinished pieces of work or a whole bunch of manuscripts with a wide variety of possible authors. Examples of this are some of Shakespeare's plays. In order to assess the importance of such a manuscript it can be vital to know who wrote it. This can be done through different statistical techniques that focus on analyzing an author's 'characteristics'. Further on I will explain what is meant by this. Author identification is made easier when you have a group of possible authors. Analyses can be made of their respective works and those analyses can then, through different techniques, be compared to the analysis of the manuscript in dubio. Sometimes, as is the case with J.R.R. Tolkien's works, the 'principal' author is known. Therefore the author is 'identified' but you want to verify it is actually him. Author verification isn't very different from 'author identification, because you still have to compare different authors, in order to verify that the manuscript is attributed to the right one.

II. Proposed system

We propose a system which will accept a input document from a particular author, then some preprocessing task will be performed in order to scan the input document. During preprocessing is being performed, we learn about the contents of the document and extracts some of the important pattern and features of the document. As the processing continues, most of the features are extracted and stored in the feature extraction phase. Now we score this features and use this score of the document to compare it with other documents in the database.

In our paper, TF-IDF algorithm is used to identify the author, where TF-IDF represents the term frequency of the word and the inverse document frequency of the term in the corpus. TF & IDF are multiplied to score the input document. After scoring the document, comparison is done with other documents in corpus and result is displayed. The result of the author identification consists of similarity percentage with author's details. Since implementing the author identification system leads to some limitations like proper formatted text document, use of US English only can be resolved but inappropriate use of language would lead to unexpected results.

The author identification system works as follow:

Preprocessing:

In preprocessing module, we are going to perform Stop-words Removal. Stop-words are the unnecessary words which do not add any meaning to the text. They simply connect important words to form meaningful sentence. Some of the stop-words are like 'and', 'a', 'the', 'which' and other words like 'for', 'of' etc. Hence these stop-words are removed before some feature extraction on the input document can be applied.

Feature Extraction:

An author's identification is done from his way of writing in the document. In particularly, an author has a unique and distinct way of writing which cannot be changed consciously. It remains some more or less same for an author. Hence extracting these features and analyzing them gives precise information of about his writing skills and style. Features to be extracted can be based on the number of words in a sentences, sentences in text or upon individual words. We will be scoring each of features by help of TF-IDF algorithm and based on scoring a report will be generated.

Expected Results:

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 final result can be represented in both textual and graphical manner.

Architecture:

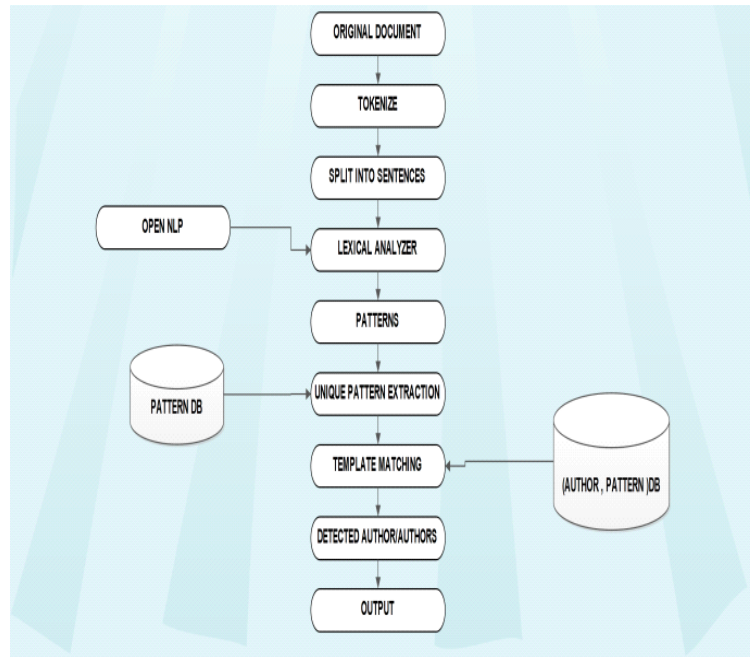


Fig. 1. Step by step flow for author identification

This diagram shows how document will be processed in our system. First, we will load original document as input to the system, then we will collect token, after splitting text document into sentences and pass document to lexical analyzer for pattern recognizing. We will have pattern database stored with us so we can match current template to the existing template and here you will find your author.

III. Conclusion

A lot of study has been done in the field of text mining and natural language processing till today. We thought of constructing the above proposed system for author identification and to gain maximum accuracy with efficiency we will be using TF-IDF algorithm. In our system, TF-IDF algorithm will be recognizing important pattern features of input text document then scores them, and relates scoring of input documents with database to identify author. Although this system has some practical limitations, all possible efforts will be taken for constructing efficient system and minimize the limitations. The system will give outputs in the format of a similarity percentage of the author. 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. Szymanski, "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] L. A. Fletcher and R. Kasturi. A robust algorithm for textstring separation from mixed text/graphics images. IEEE Trans. Pattern Anal. Mach. Intell., 10(6):910-918, 1988.
- [6] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger. Script-independent text line segmentation in freestyle handwriting documents. IEEE Trans. Pattern Anal. Mach. Intell., 30(8):1313-1329, August 2008.
- [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.