A Framework for Plagiarism Detection Based on Author Profiling

Notebook for PAN at CLEF 2013

Seifeddine Mechti,Maher Jaoua,Lamia Hadrich Belguith

ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia
mechtiseif@gmail.com,maher.jaoua@fsegs.rnu.tn,l.belghith@fsegs.rnu.tn

Abstract. In this paper, we describe a method for the detection of plagiarism based on author profiling [1]. After having segmented a document into a set of texts, we apply the technique of predicting the age and gender of the author on these texts. In case the predictions are heterogeneous, the probability of the existence of plagiarism becomes really great. Predicting the gender and age of the author was done by machine learning using decision trees.

Keywords: plagiarism, author profiling, machine learning, decision trees.

1 Introduction

The word plagiarism, from the Latin "plagiarus", means stealing someone's slave [2]. According to [3], many people consider plagiarism as copying another's work, or borrowing someone's original ideas. However, these approaches can be classified into two major groups: the approach based on the contents of the document and the approach based on the style of the author. Several techniques have been used in the comparison of the contents of the documents: if a suspect document A is part of a similar document B, then A is plagiarized from B. Among the methods used, we can mention the ANLP (automatic natural language processing) method, n-grams, hash functions, the Levenshtein algorithm, the Jaroll-Winkler algorithm, etc.

The question here is: if no plagiarized documents have been found, can we assert that there was no plagiarism? The answer to this question comes as a follow up of the definition of the concept of "invisible web". In fact, the hidden webs include the web sites which are accessible but not indexed by the known search engines; also, the documents which are protected by passwords are inaccessible. Therefore, the IRS (Information Retrieval System) for plagiarism will fail in the face of invisible documents, paying documents and documents that are not digitized (books).

To overcome this problem, a local analysis of the suspect document seems inevitable. The most recent method is one based on the detection of the author's style. This means that each author has his own style, and you can retrieve the characteristics of the style from his documents. As a result, to determine whether a document matches a certain writer, his style can be compared to the characteristics of the style of the document; if they are different, we conclude that there is plagiarism [4, 5, 6, 7]. A new stylistic trend educes the user's profile ((gender) male or female, approximate age, language, etc.)[8] from a document; thus, two documents that have the same user profile will have a greater chance of being similar.

2 Our Approach

Our approach is based on educing the author's profile by focusing on the age and gender dimensions. Our system takes as input a document which is written in English or in Spanish and generates the age and the gender of its author [1]. First, we computed a ranked list of words that occur in the corpus and we grouped them into classes according to their similarities. Then, we calculated the TF * IDF score of each class for each document in order to find the stylistic differences between men and women, on the one hand, and those between different age intervals on the other hand. After that, we applied the learning process on 66% of the English and of the Spanish corpuses using decision trees via the J48 algorithm. In fact, we got the second place in the competition for the English corpus; our system has shown a high level of accuracy and effectiveness in treating the gender dimension. Our method provides one of the following six classes: 10s Male, 20s Male, 30s Male, 10s Female, 20s Female and 30s Female.

Below is the plagiarism detection system based on machine learning of the author:

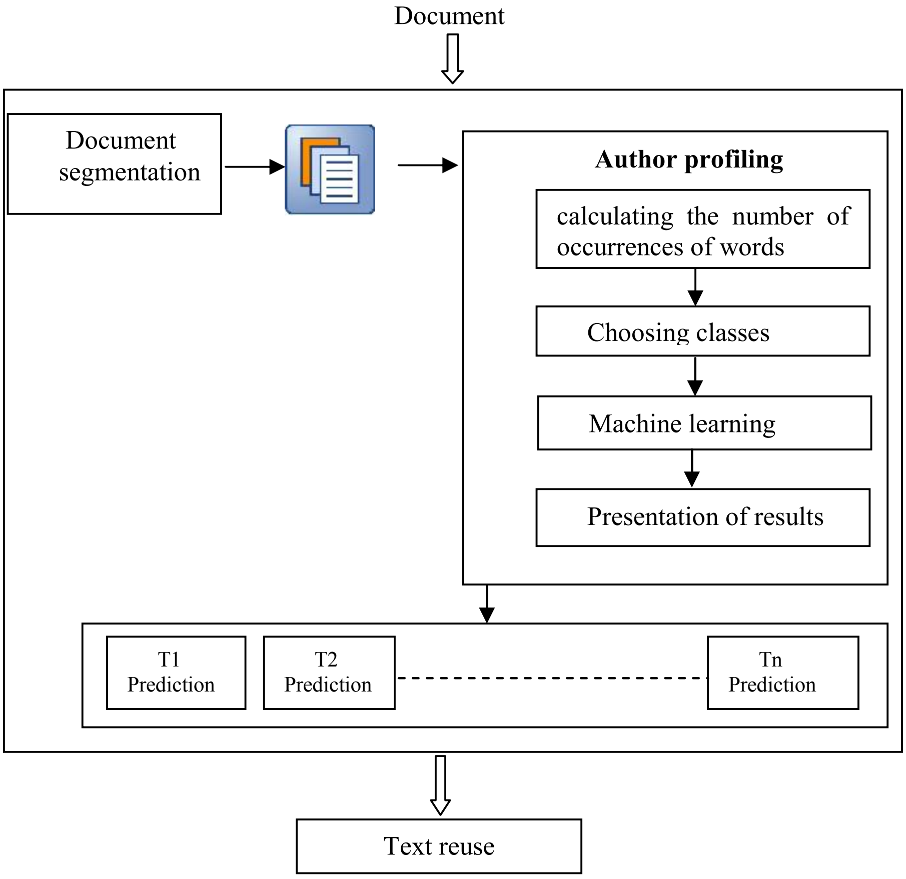

Figure1. Plagiarism detection based on author profiling

3 Plagiarism Detection based on author profiling

As shown in Figure 1, author profiling can detect the classification of all parts of any document. In addition, each fragment of the document will have a prediction regarding the author's age and gender. But in a normal case, we have the same prediction for all the parts of a document written by the same person.

Below is an example of author profiling (age) of "document X".

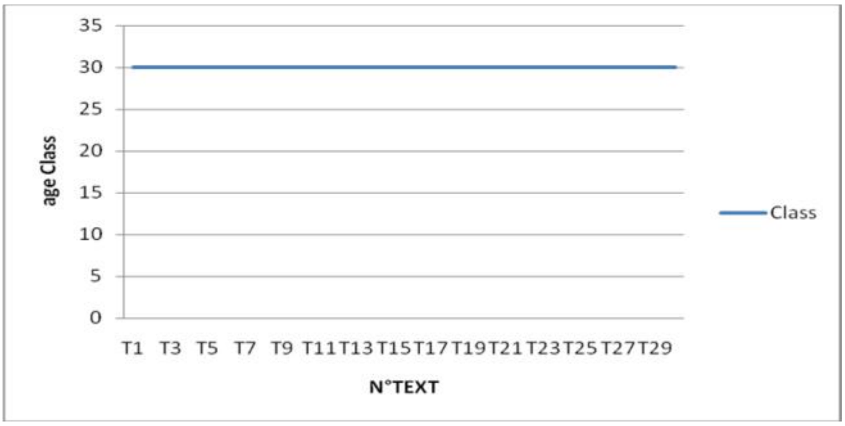

Figure 2. Age prediction for document 'X'

As shown in Figure 1, the predictions of this document are homogeneous. Through author profiling, this document had the prediction 30s (30 years) for all its parts. We can assert that there is no plagiarism in this document.

As shown in Figure 2, "document Y" is different from "document X":

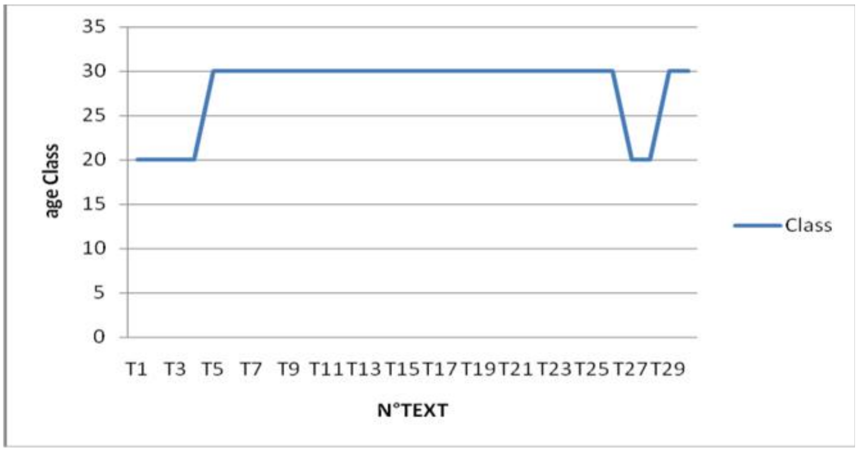

Figure 3. Age prediction for document 'Y'

We notice that the prediction of this document is 30s. Nevertheless, what is striking is that parts T1, T2, T3 and T4 as well as parts T27, T28 and T29 had the result 20s (20 years). These parts correspond to the introduction and conclusion of the document. This difference in prediction reflects a difference at the stylistic level as author profiling can detect the stylistic differences between the different age classes and also between a man and a woman. Therefore, the introduction and the conclusion of this document have a high probability of containing plagiarized passages.

4 Conclusion

The author profiling technique can locally detect plagiarism. Compared with the text alignment method, our method does not use external data such as databases, but neither is there a need to use the Internet to check if the document is plagiarized from the web.

We hope to apply this method to detect plagiarism in Arabic texts. We also aim to compare this method with conventional methods for comparing texts like text alignment.

References

[1] Mechti S., Jaoua M. and Hadrich Belghith L, Author profiling using style based features. PAN@CLEF (Online Working Notes/Labs/Workshop) Valencia, Spain, 2013. (to appear).

[2] Pericault F, L'appropriation du processus créatif. Université Lumière Lyon, 2011.

[3] http://www.turnitin.com/research_site/e_what_is_plagiarism.html.

[4] Bonsall B, The Automatic Detection of Plagiarism, rapport de projet de recherche, 2003-2004.

[5] Kimler M, Using style markers for detecting plagiarism in natural language documents, Rapport de these, Université de Skovde, Suède, 2003.

[6] Parvati I and Singh A, Document similarity analysis for a Plagiarism Detection System*,* *2*nd Indian international conference on artificial intelligence, IICAI05, India, 2005.

[7] Gruner S and Naven S, Tool support for plagiarism detection in text documents*,* ACM symposium on applied computing, New Mexico, USA, 2005.

[8] Argamon S., Koppel M., Pennebaker J. and Schler J., Automatically profiling the author of an anonymous text, Communications of the ACM 119–123. New York, USA, 2009.