Boosting Algorithm and Meta-Heuristic Based on Genetic Algorithms for Textual Plagiarism Detection

Hadj Ahmed Bouarara, GeCode Laboratory, Department of Computer Science, Tahar Moulay University of Saida Algeria, Saïda, Algeria

Reda Mohamed Hamou, GeCode Laboratory, Department of Computer Science,T ahar Moulay University of Saida Algeria, Saïda, Algeria

Amine Rahmani, GeCode Laboratory, Department of Computer Science, Tahar Moulay University of Saida Algeria, Saïda, Algeria

Abdelmalek Amine, GeCode Laboratory, Department of Computer Science, Tahar Moulay University of Saida Algeria, Saïda, Algeria

ABSTRACT

Day after day, the plagiarism cases increase and become a crucial problem in the modern world, caused by the quantity of textual information available in the web and the development of communication means such as email service. This paper deals on the unveiling of two plagiarism detection systems: Firstly boosting system based on machine learning algorithm (decision tree C4.5 and K nearest neighbour) composed on three steps (text pre-processing, first detection, and second detection). Secondly using genetic algorithm based on an initial population generated from the dataset used a fitness function fixed and the reproduction rules (selection, crossover, and mutation). For their experimentation, the authors have used the benchmark pan 09 and a set of validation measures (precision, recall, f-measure, FNR, FPR, and entropy) with a variation in configuration of each system; They have compared their results with the performance of other approaches found in literature; Finally, the visualisation service was developed that provides a graphical vision of the results using two methods (3D cub and a cobweb) with the possibility to have a detailed and global view using the functionality of zooming and rotation. The authors' aims are to improve the quality of plagiarism detection systems and preservation of copyright.

Keywords: *Bag of Word, Boosting, Decision Tree, Entropy, F-Measure, Genetic Algorithms, K Nearest Neighbour, N-Gram, Plagiarism Detection, Visualisation*

DOI: 10.4018/IJCINI.2015100105

1. INTRODUCTION AND PROBLEMATIC

Nowadays, with the increasing numbers of documents available on the web and the development of communication means, find the possessor of the information has become a crucial subject. In the recent few years, we have observed clearly that the cases of plagiarism in the works of scholars and researchers have been increased. The basics of this problem are numerous and crossed because there are many websites which articles and ready documents are available, these sites are ideal for the plagiarists. For this reasons developing an automatic plagiarism detector tool has become a necessity.

The most relevant case of plagiarism was designed by the Germany minister of education and research SCHAVAN ANNETTE who put his resignation because the Dusseldorf University revoked her doctorate that contains too many passages "borrowed from others". In a country where the title of Doctor is a valuable, we do not mess with plagiarism.

In order to give you a global view about our work, the plagiarism is defined as the wrongful misuse of stealing thoughts, ideas or words from the original work of someone, in the same language or in a different language (Basile, 2009). Depending on the behaviour of plagiarist, we can distinguish several plagiarism types such as:

- **The Plagiarism Verbatim:** When the plagiarist copied the words or sentence from a book, magazine or web page as like it, without putting it in quotation marks and / or without citing source.
- **The Paraphraser:** When the words or the syntax of sentence copied are changing.
- The cases of plagiarism the most difficult to detect are plagiarism with translation and plagiarism of ideas.

In the former years, the classical method to detect plagiarism is to examine manually each document that represents a slow process. Recently, two automatic plagiarism detection families have emerged:

- The external plagiarism detection, which allows comparing the suspicious document with the reference documents, based on external information (Stein, 2007).
- The internal plagiarism detection based on stylometry method. Each document has a specific style will be compared to a base of style. The case of plagiarism will be detected depending on how the document is writing and if there is a change in style between the paragraphs (Meyer, 2007).

The classical plagiarism detection systems are face to many limits:

- **Detection Errors:** the detection errors (classification of text plagiarised as no-plagirised and classification of no-plagiarised text as plagiarised) can cause many problems for researchers and students. For e.g. a researcher send a paper (really this paper is no-plagiarised) to a journal. The plagiarism detector system used by this journal detects that that this paper is plagiarised then this researcher will be blacklist automatically. It is a big problem in our scientific life.
- The selection of parameters (similarity measure and text representation method)

Emotional Axes: Psychology, Psychophysiology and Neuroanatomical Correlates

Didem Gökçay (2011). *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives (pp. 56-73).*
www.irma-international.org/chapter/emotional-axes-psychology-psychophysiology-neuroanatomical/49529