BHA2: Bio-inspired Algorithm and Automatic Summarisation for Detecting Different Types of Plagiarism

Hadj Ahmed Bouarara, GeCode Laboratory, Department of Computer Science, Tahar Moulay University of Saida Algeria, Saida, Algeria
Reda Mohamed Hamou, GeCode Laboratory, Department of Computer Science, Tahar Moulay University of Saida Algeria, Saida, Algeria
Amine Rahmani, GeCode Laboratory, Tahar Molay University of Saida, Algeria

ABSTRACT
In the last decade, the plagiarism cases were increased and become a topical problem in the modern scientific world, caused by the quantity of textual information available online/offline. The authors’ work deals on the development of a new plagiarism detector system called BHA2 which has as input the suspicious text (to be analysed) and the original texts (learning basis). It can detect the different forms of plagiarism based on: Google API to detect the cases of plagiarism with translation; text summarization to detect the plagiarism of idea; conceptual transformation to detect the plagiarism with synonymy; bag of phrases to detect the paraphraser plagiarism; the social worker bees algorithm that was inspired from the lifestyle of social worker bees (forager, guardian, and cleaner) to select the documents source of plagiarism; the output of the authors’ system are the plagiarised passages (the copied parts from the original texts) and the plagiarism percentage for each suspicious text. Their experiments were performed on the Pan 09 dataset and using the validation measures (recall, precision, accuracy, error, f-measure, and entropy, FPR, FNR, W-accuracy, ROC and TCR) in order to show the benefit derived from using such idea compared to the result of classical systems existed in literature. A comparative study in term of services was realised between their system and others commercial systems such as (check, Turnitin, and machine learning system) with their system. Finally, a visualization step was achieved for the purpose to see the outcome in graphical form (3d cub and cobweb) with more realism using the functionalities of zooming and rotation.

KEYWORDS
Automatic Summarization, Plagiarism Detection, Plagiarism of Idea, Plagiarism with Synonymy, Plagiarism with Translation, Social Worker Bees, Visualisation

1. INTRODUCTION AND BACKGROUND
In today’s world of globalization and borderless technology, the appearance of the Internet and the rapid development of telecommunication have made the world a global village. Nowadays, with the increasing numbers of documents available on the web and the copy/paste option, finding 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 (thesis, research papers … etc.) have been increased. The basics of this problem are numerous and crossed because there are many
websites where articles and ready documents are available, these sites are ideal for the plagiarists. A large-scale study on 18,000 students realised by McCabe shows that about 70% of the students admit to plagiarize from extraneous documents (Meuschke, 2013). According to a new recent American study, published in the journal for academic ethics it affirms that plagiarism case had been poured before the arrival of the digital era today. In this study 184 doctoral works published before 1994 and after 2012 were selected randomly from online universities. The result proves that over 50% of manuscripts contain plagiarized passages. For this reasons developing an automatic plagiarism detector tool has become a necessity.

Recently the plagiarism phenomenon has spread, where it has even touched the most popular politicians in the world designed by the Germans ministers like the defence Minister, the atypical KARL-THEODOR ZU GUTTENBERG that was resigned after accusations of plagiarism concerning the writing of his doctoral thesis in law at the University of Bayreuth (Bavière). In 2014 also the minister of defence URSULA VON DER LEYEN who was considered as a possible heiress of Angela Markel, was suspected by a site specialized in the analysis of theses written by politicians to have plagiarised a number of passages in his medical thesis. Without forgetting the minister of education and research SCHAVAN ANNETTE resigned in 2014 even due to plagiarism (Bouarara1, 2015).

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 forms such as:

- **Verbatim Plagiarism:** Copying directly sentences or passages from the work of other person.
- **Paraphraser:** Using the same sentences of another person, by changing the order of the words.
- **Blunt plagiarism (Copyright Plagiarism):** Stealing the work of another and put another name to it.
- **Plagiarism of Ideas:** The reuse of an original thought or idea (independent of the form) from a source text.
- **Plagiarism with Synonym:** Copying the same words of someone and replacing them by their synonyms.
- **Plagiarism with Translation:** It is to reuse the work of another person in another language by using the automatic translation technology.
- **Text Recycling (Self-Plagiarism):** When an author uses parts of his article already published in another article.

In the former years, the classical method to detect plagiarism is to examine manually each document which 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).

Unfortunately, the classical plagiarism detection systems are face to many limits in term of: