Chapter 19
A New Algorithm of Grouping Cockroaches Classifier (GCC) for Textual Plagiarism Detection

Hadj Ahmed Bouarara
Tahar Moulay University of Saida, Algeria

Reda Mohamed Hamou
Tahar Moulay University of Saida, Algeria

ABSTRACT

In the last decade with the new technology, it is important to allow users to access information freely, while at the same time, restrict them from illegal copying and distribution of information. In the age of information technologies plagiarism has become a topical subject in the digital world and turned into a serious problem. The author's work deals with the development of a new system for combating this phenomenon using a new insect behaviour algorithm called Groping cockroaches classifier GCC. Each suspicious text (cockroach) will be classified (hidden) in a class (shelter) that can be plagiarism or no-plagiarism, using a security function that is based on the attractiveness of each class (calculated using the aggregation operators (shelter darkness, congeners attraction and security quality)) and the displacement probability (calculated using the naive Bayes algorithm). The experimental results performed on the Pan 09 dataset and using the validation measures (recall, precision, f-measure, and entropy), have demonstrated that GCC has clear advantages over others plagiarism detection techniques existed in literature. Finally, a set of service was added in order to detect the different cases of plagiarism such as plagiarism with translation, plagiarism of idea, plagiarism with synonymy, and plagiarism paraphrase.

1. INTRODUCTION AND PROBLEMATIC

We are living in the age of information technologies that rendered digital libraries a concrete possibility. The easy access to the information via electronic resources, such as the Web has made billions of web pages easily accessible to anyone providing plenty of potential sources for the plagiarists and rendere the digital documents more vulnerable to be copied by turning cheating extremely easy. Lancaster and

DOI: 10.4018/978-1-5225-8057-7.ch019

Culwin in [13] state that "plagiarism is theft of intellectual property, it is not only dishonest, but also an offense that may result in sanctions". It is considered as one of the biggest problems in publishing, science, and education.

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 Minister of defense, 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 defense URSULA VON DER LEYEN who was considered as a possible heiress of Angela Mmarkel, was suspected by a site specialized in the analysis of theses written by politicians to have plagiarized 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 (Bouarara[1], 2015).An example of plagiarised text is presented in Figure 1.

In order to give you a global view about our work the plagiarism is mentioned as a lack of moral, civil or commercial, which can be subject to criminal penalty. It can be defined in three points:

- Appropriating the creative work of someone else and present it as his own;
- To grab snippets of text, images, data, etc. from external sources and integrate them into his own work without citing the source;
- Summarize even the original idea of an author by expressing it in his own words, but omitting to mention the source.

Depending on the behaviour of plagiarist, we can distinguish several plagiarism types such as the plagiarism verbatim, The paraphrase and the cases of plagiarism the most difficult to detect are plagiarism with translation and plagiarism of ideas.

Plagiarism detection, the automatic identification of plagiarism and the retrieval of the original sources, is developed and investigated as a possible countermeasure. Although humans can identify cases of plagiarism in their areas of expertise quite easily, it requires much effort to be aware of all potential sources on a given topic and to provide strong evidence against an offender. The manual analysis of text with respect to plagiarism becomes infeasible on a large scale, so that automatic plagiarism detection attracts considerable attention

Figure 1. (a) Plagiarised Text VS (b) Source Text (Potthast, 2010)

Plagiarized Passage s_{plg}

The people who left their countries and sailed with Gilbert were more suited for fighting the crusades than for leading a settled life in the colonies. They were bitterly disappointed as it was not the America that they had expected. Since they did not immediately find gold and silver mines, many deserted. At one stage, there were not even enough man to help sail the four ships. So the Swallow was sent back to England carrying the sick. The other fleet was supplied with fish and the other necessities from St. John. On August 20, Gilbert had sailed as far as forty-four degrees to the north latitude. His ship known as the Delight, which bore all the required supplies, was attacked by a violent storm near Sable Island. The storm had driven it into a rock shattering it into pieces.

Source Passage s_{src}

The emigrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted; and soon there were not enough sailors to man all the four ships. Accordingly, the Swallow was sent back to England with the sick; and with the remainder of the fleet, well supplied at St. John's with fish and other necessaries, Gilbert (August 20) sailed south as far as forty-four degrees north latitude. Off Sable Island a storm assailed them, and the largest of the vessels, called the Delight, carrying most of the provisions, was driven on a rock and went to pieces.

[Excerpt from "*Abraham Lincoln: A History*" by John Nicolay and John Hay.]