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

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

KEYWORDS

Aggregation Rules, Cockroaches, Plagiarism of Idea, Plagiarism with Synonymy, Plagiarism with Translation, Security Function, Supervised Classification, Visualisation

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 Culwin in 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

DOI: 10.4018/IJIRR.2016100104

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
minister of defense URSULA VON DER LEYEN who was considered as a possible heiress of Angela Merkel, 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 (Bouarara1, 2015). An example of plagiarized text is presented in the next figure:

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

1.1. Aims and Objectives

- The main objective of our work is to improve the plagiarism detection systems existed in literature. The classical systems of plagiarism detection have proven their limits in terms of performance quality, execution time and their inability to detect all types of plagiarism.
- Modelling a new swarm intelligence algorithm called grouping cockroaches classifier (GCC) for the construction of an automatic plagiarism detection system using the n-gram and bag of word for the text representation.
- Verify the effectiveness of our approach through a comparative study with the works existed in literature.

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

Query Expansion based on Central Tendency and PRF for Monolingual Retrieval
[www.igi-global.com/article/query-expansion-based-on-central-tendency-and-prf-for-monolingual-retrieval/163130?camid=4v1a](www.igi-global.com/article/query-expansion-based-on-central-tendency-and-prf-for-monolingual-retrieval/163130?camid=4v1a)

Semantic Role Labeling Approach for Evaluation of Text Coherence
[www.igi-global.com/article/semantic-role-labeling-approach-for-evaluation-of-text-coherence/101360?camid=4v1a](www.igi-global.com/article/semantic-role-labeling-approach-for-evaluation-of-text-coherence/101360?camid=4v1a)

XML Documents Normalization Using GN-DTD
[www.igi-global.com/chapter/xml-documents-normalization-using-dtd/75901?camid=4v1a](www.igi-global.com/chapter/xml-documents-normalization-using-dtd/75901?camid=4v1a)
An Efficient and Flexible Approach of Keyword Search in XML
[www.igi-global.com/chapter/efficient-flexible-approach-keyword-search/69125?camid=4v1a](www.igi-global.com/chapter/efficient-flexible-approach-keyword-search/69125?camid=4v1a)