Chapter 14
Agents Oriented Genetic-K-Means (AOGK) System for Plagiarism Detection

Hadj Ahmed Bouarara
Tahar Moulay University of Saida, Algeria

Yasmin Bouarara
Tahar Moulay University of Saida, Algeria

ABSTRACT

In the last decade, the plagiarism phenomenon has widely spread and become a topical problem in the modern scientific world, caused by the wide availability of electronic documents online and offline. This work will be devoted to describe a new plagiarism detection system named AOGK « Agents Oriented Genetic-K-means » based on a multi-agents architecture composed of three modules: text parsing to transform documents into vectors; Learning module using genetic algorithms to build a prediction model; Test module using k-means for the final classification of suspicious document; To evaluate their system the authors have used a range of reference metrics (precision, recall, f-measure and entropy) and the benchmark PAN 09. They have compared the results obtained with the performance of other systems found in literature; the authors' aim is the preservation of copyright.

1. INTRODUCTION AND BACKGROUND

Recently, the internet was continually enriched by new contents, where Google web search engine contains more than 3 billion of web pages providing a wide variety of free source texts in many different languages. Unfortunately, the electronic documents are vulnerable to being copied and the cases of plagiarism have been increased tremendously in the last few years. It is a big problem in the scientific community, which represents the reuse of ideas, words, images or expressions of others persons without making citations (Basile, 2009). We can found different forms of plagiarism such as:

DOI: 10.4018/978-1-5225-8057-7.ch014

- **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.

Existing approaches to safeguarding intellectual properties can be roughly categorized into two-families:

- **The Supervised Plagiarism Detection (SPD):** The SPD is based on the external information. It allows comparing the content of each suspicious document (document to be analysed) against a repository of reference documents (external information) in order to detect the plagiarised parts (Stein, 2007).
- **The Unsupervised Plagiarism Detection (UPD):** The UPD does not require the use of reference documents. It is based on the stylometry technique to analyse the content of the suspicious document to detect the change in the styles between the paragraphs. It is very difficult to achieve because an author can have different styles.

Nowadays, several classical plagiarism detection systems have seen the light, but they are face to many drawbacks in terms of (Quality of detection, the parameters selected (similarity measure and text representation method), response time and presentation of results).

We are thus naturally led to seek better performance by using a decentralized architecture. Our work consists to deal with all the problems previously cited through:

- The development of a new system for plagiarism detection based on the hybridization between meta-heuristic method (genetic algorithm) and machine learning method (k-means).
- Using a set of agents to orient the final decision by a vote.
- Verify the effectiveness of our system through a comparative study with the works existed in literature.
- Construct a visualization method, which allows us to interact with the system to retrieve decisions and have a global view / detail of the detection results.
- Help the scientific world to limit the phenomenon of plagiarism.

Our paper takes place in the intersection of different domain like shown in Figure 1.

The paper is organized as follows: Section 2 describes some automatic plagiarism systems. Section 3 gives a detailed view around the HGK-PD system. Section 4 exposes the obtained results after the different tests realized on the PAN 09 dataset. Section 5 presents a comparative study between our system and others plagiarism detection systems existed in literature. Finally in section 6, we conclude the paper and give some future perspectives.