Chapter 7
Multi-Agents Machine Learning (MML) System for Plagiarism Detection

Hadj Ahmed Bouarara
Dr. Tahar Moulay University of Saida, Algeria

ABSTRACT

*Day after day the cases of plagiarism increase and become a crucial problem in the modern world caused by the quantity of textual information available in the web. As data mining becomes the foundation for many different domains, one of its chores is a text categorization that can be used in order to resolve the impediment of automatic plagiarism detection. This chapter is devoted to a new approach for combating plagiarism named MML (Multi-agents Machine Learning system) composed of three modules: data preparation and digitalization, using n-gram character or bag of words as methods for the text representation, TF*IDF as weighting to calculate the importance of each term in the corpus in order to transform each document to a vector, and learning and vote phase using three supervised learning algorithms (decision tree c4.5, naïve Bayes, and support vector machine).*

INTRODUCTION AND BACKGROUND

The information revolution jostled by the large-scale development of Internet / Intranet of networks access has detonated the amount of textual data available online or offline and the popularization of computer science in the world of business, government and individuals, has created large volumes of electronic documents written in natural language. It is very difficult to estimate the quantities of textual data created each month in government, corporations, institutions, or the amount of scientific publications in various search fields. This resolution gave the birth to a big problem called plagiarism which has received much attention from both the academic and commercial communities, in recent years. Instead of producing original work, some students or researcher prefer to directly take ideas things or content found in books, encyclopedias, newspaper articles or previous work submitted by others or on one of the many cheat sites available recently on the Internet, sites with existing equipment perfectly organized by

DOI: 10.4018/978-1-5225-3004-6.ch007

subject and level for easy access, either intentionally or un-intentionally without putting it in quotation marks and / or without citing source.

Many examples of text reuse surround us today, including the creation of literary and historical texts, summarization, translation or revision of existing texts. Many factors influence plagiarism including translating an original text into a different language, restyling an original to fit different authorial or consumer needs (e.g. rewriting a scientific text to be readable by the layman), reducing or expanding the size of the original text and the competency and production requirements of the writer. Recent advances in technology are making plagiarism much easier. For example, the Google web search engine claims to index over 3 billion web pages1 providing a large variety of source texts on a diverse range of topics in many different languages. Word processors have also become more sophisticated, enabling users to easily cut and paste, merge and format pre-existing texts from a variety of sources.

Depending on the behavior of plagiarist, we can distinguish several types of plagiarism as a 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 or buy a work online, the paraphrase when the words or the syntax of sentence copied are changing and finally the cases of plagiarism the most difficult to detect are plagiarism with translation and plagiarism of ideas when summarizing the original idea of the author expressed in his own words partially or completely (Stein, 2007).

The Plagiarism problem can be regarded as a categorization problem for this we have studied the different learning algorithms for the supervised classification of texts (two classes plagiarism or no-plagiarism), which produce a prediction model. These techniques based on a single agent are face to limitations when we sought to develop more complex models for the classification of a gigantic textual database. These limitations can be easily felt by a considerable performance degradation of the best classifiers in response time that increases proportionally with the size of the volumes treated and even the quality of results. In our work we had the idea of decentralizing the classification process using multiple agents that will communicate with each other to share knowledge to improve the performance and efficiency of our classification system. We have developed a model based on an architecture composed of several parts; each part will deal with the problem of categorization of a textual database in a specific way, but can communicate and share their knowledge.

Aims of our work:

- Construction of a multi-agents supervised classification system for the detection of the plagiarism cases.
- Find the best representation of the data set 09 Pan used in our experiment
- Find the number of ideal agent for each learning algorithm used
- Vary the algorithm used by each agent in order to find the best combination.
- Compare the obtained results with the results of the classical technique
- Make decisions that can be used by another researcher in the fields of plagiarism detection, classification and multi-agents systems.
- The Construction of a visualization method to help experts in the field of analyzing the results obtained.