Journal of Organizational and End User Computing
Volume 32 • Issue 2 • April-June 2020

Unsupervised Model for Detecting Plagiarism in Internet-based Handwritten Arabic Documents

Mahmoud Zaher, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt
Abdulaziz Shehab, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt
Mohamed Elhoseny, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt
Farahat Farag Farahat, Sadat Academy, Cairo, Egypt

ABSTRACT

Due to the rapid increase of internet-based data, there is urgent need for a robust intelligent documents security mechanism. Although there are many attempts to build a plagiarism detection system in natural language documents, the unlimited variation and different writing styles of each character in Arabic documents make building such systems challenging. Based on its position in a word, the same Arabic letter can be written three different ways, which makes the handwritten character recognition a cumbersome process. This article proposes an intelligent unsupervised model to detect plagiarism in these documents called ASTAP. First, a handwritten Arabic character recognition system is proposed using the Grey Wolf Optimization (GWO) algorithm. Then, a modified Abstract Syntax Tree (AST) is used to match the contents of the Arabic documents to detect any similarity. Compared to the state-of-the-art methods, ASTAP improves the effectiveness of the plagiarism detection in terms of the matched similarity ratio, the precision ratio, and the processing time.

KEYWORDS

Abstract Syntax Tree, Gray Wolf Optimization, Handwritten Character Recognition, Hash value, Internet Data Security, Plagiarism Detection, Similarity Index, Unsupervised Documents Analysis

AN INTRODUCTION

The ever-increasing smart information processing services and applications offered by the Internet have explosively widened the span of the global inter-network. The recent advancements in designing low-cost small scaled devices have harbingered a great surge in the number of Internet-enabled devices which generate a big amount of data. Accordingly, internet data management for discovering plagiarized documents plays a vital role in many applications such as file management, copyright saving, and electronic theft prevention (Lam, et al., 2016; Abdi et al., 2015). Plagiarism not only depends on the content ratio that is copied but dramatically relates to using the work of others, i.e., ideas; without proper citation (Kahloula & Berri, 2016; Abdelrahman & Khalid, 2014).

In Internet-based document processing applications (Chen & Zhao,2017), the Arabic language is considered one of the most complicated languages, especially if the document contains handwritten words. The features of Arabic alphabets have various shapes of the written form based on their position and can be extended by making a dash between the **two** letters. For Arabic in electronic or printed media, no pronouncement makes misunderstanding for some words in an inevitable situation.

DOI: 10.4018/JOEUC.2020040103

These challenges make the plagiarism detection in Arabic documents an arduous task. Dependently, many machine learning and artificial intelligence based methods have been developed (Hussein, 2016; Wise, 2012). For example, an online Arabic plagiarism detection tool called APD (Alzahrani & Salim, 2015) is proposed to detect the plagiarism on the Arabic web pages. However, this tool does not handle the synonyms alternations or the rewording problem. To avoid that, another system called Plaggie (Ahtiainen et al., 2011) is proposed. Besides its disability to handle the handwritten documents, Plaggie needs a long processing time to manage a computerized Arabic document.

Due to the Hugging of information, and correlation networks, the discovery of electronic thefts is a difficult task, and the discovery of the thefts started in the Arabic language and the most difficult task no doubt. And in light of the growing e-learning systems in the Arab countries, this requires special techniques to detect thefts electronic written in Arabic. And although it could use some search engines like Google, it is very difficult to copy and paste the sentences into the search engines to find these thefts. For this reason, it must develop a good tool for the discovery of electronic thefts written the Arabic language to protect e-learning systems, and to facilitate and accelerate the learning process, where it can automatically detect electronic thefts automatically by this tool.

This paper shows, ASTAP, a system that works on the Internet to enable specialists to detect thefts of electronic texts in Arabic so it can be integrated with e-learning systems to ensure the safety of students and research papers and scientific theses of electronic thefts.

The paper also describes the major components of this system, including stage outfitted, and in the end, we will establish an experimental system on a set of documents and Arabic texts and compared the results obtained with some of the existing systems, particularly TurnItIn.

Accordingly, a new plagiarism_detection system has been proposed in this paper which can handle the internet based handwritten Arabic documents called ASTAP (Abstract Syntax Tree Arabic Plagiarism). ASTAP consists of two main phases. The first one aims to provide an Optical Character Recognition (OCR) tool for internet-based handwritten Arabic Documents. The proposed OCR has two primary functions, feature extraction and feature selection. The feature extraction process aims to remove redundancy from handwritten Arabic characters. While in the selected feature the most relevant are only reserved for improving the accuracy classification. The proposed OCR is implemented using a well-known optimizer called GWO (Seyedali, et al., 2014) that is used to optimize a character features selection. The second phase of ASTAP aims to detect the similarity of Arabic documents using a modified AST (Aiken, 2015; El Bachir & Bagais, 2014). For each node of the AST, the algorithm determines the hash value of AST and compares it with the other nodes in the form of node by node. Also, the algorithm compares sub_trees based on the tree_structure with some reduction in the execution. We have improved the way of syntax tree similarity and proposed a plagiarism detection algorithm that rearranges the nodes of AST to the longitudinal framework. The modified AST consists of five components: AST construction, hash value computation, node classification, hash comparison, and degree of similarity evaluation.

There are two main contributions of that paper. First, an intelligent unsupervised model for internet-based handwritten Arabic character recognition system is proposed using GWO algorithm. Second, a modified AST is proposed for matching the contents of the Arabic documents to detect any similarity. The proposed system improves the similarity accuracy for the plagiarized document by replacing the word synonyms and minimizing time consumption by enhancing the performance of AST algorithm.

The rest of this paper is organized as follow: Section 2 presents an overview of the different related works. Section 3 describes the working steps of the proposed system ASTAP. Section 4 explains the methodology and the internal ASTAP components and their roles in detecting a document similarity. Section 5 presents a discussion of the results. Finally, Section 6 concludes the paper.