International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

A SURVEY OF PLAGIARISM DETECTION FOR ARABIC DOCUMENTS

Yahya A. Abdelrahman , PhD student, Department of Computer Science Sudan University of Science and Technology, Khartoum, Sudan
Ahmed Khalid, Assistant Professor, Department of Computer Najran University ,Najran KSA
Izzeldin M. Osman, Prof, Department of Computer Science, Sudan University of Science and Technology Khartoum, Sudan

Abstract

Plagiarism detection techniques and tools have developed mainly for English scripts. It has been found that different methods use different document descriptors ranging from characters to document structure. Arabic is a rich morphological language that poses special challenges to computational natural language processing systems. This paper presents a survey study on plagiarism detection methods, tools and algorithms for Arabic documents. At the end of this paper authors propose a system for Arabic document plagiarism detection in electronic resources.

I. Introduction

Plagiarism is defined as the unauthorized use or close imitation of the language and thought of authors and their representation as one's own original work [26]. It involves literary theft, stealing (by coping) the words or ideas of someone else and passing them off as one's own without recognizing the source. Many people think of plagiarism as copying another’s work, or borrowing someone else’s original ideas. However, terms like “copying” and “borrowing” can disguise the seriousness of the offense [27].

Plagiarism becomes one of the most important issues for universities, schools, and researchers [20]. It is so easy through the internet and due to using advanced search engine to find documents or journals by students. Some of the researchers are just copying and pasting others works without reference to the owner of the documents. Several types of plagiarism exist, including direct copying of phrases or passages from a published text without citing the sources, plagiarism of ideas, sources, and authorship. There are other types of plagiarism, such as translating content to another language, presenting the same content with other media like images, videos and texts, and using program code without permission. [2]

All these practices of plagiarism have a negative impact on the learning process. Thus, how can we ensure dealing with Plagiarism systems and how is plagiarism going to be detected. A critical issue needs solutions by computer scientists. [25]

We classified the survey into three categories:
1- Plagiarism in Arabic documents.
2- Arabic Plagiarism techniques.
3- Arabic Plagiarism algorithms.
These categories are explained as follows:

1- Plagiarism in Arabic Documents:

Plagiarized document detection plays important roles in many applications, such as file management, copyright protection, and plagiarism prevention. [27]. Plagiarism can take one of the popular types such as copying of the whole or some parts of the document, rewording the same content in different words, using others’ ideas or referencing the work to incorrect or non-existing sources [9]. Other ways of plagiarism include translated plagiarism wherein the content is translated and used without referencing the original work, artistic plagiarism in which different media such as images and videos are used to present other’s work without proper citation [15].

Most of the work in document plagiarism has been done for academic purpose. Detecting plagiarism is important to judge and mark students’ work, especially for postgraduates who are strictly prohibited from cheating, rewording, rephrasing, or restating without referencing. In this regard, numerous plagiarism detection systems have been developed for Arabic documents. Most of these systems use plagiarism techniques known as similarity detection techniques, which create special “fingerprints” for collecting files, including metrics, such as average line length, file size, average number of commas per line. The files with close fingerprints are treated as similar. Clearly, small fingerprint records can be compared rapidly, but this technique is now considered unreliable and rarely used nowadays [3]. Ameera Jadalla and Ashraf Elnagar proposed Iqtebas 1.0, which is a primary solid and complete piece of work for plagiarism detection in Arabic text files. It is similar to a search engine. The goal of the Iqtbas 1.0 is to compute the originality, value of the examined document, by computing the distance between each sentence in the text and the closest sentence in the suspected files [2]. Farahat F.

Farahat, etal [6] are tested experimentally ZPLAG, a prototype for detecting plagiarism in documents written in Arabic language, where some hidden plagiarism forms can be detected, such as change of sentence structure and replacement of synonym, The results show that ZPLAG system has excellent deal with Arabic scripts and allows students to submit assignments to their teachers in e-classrooms .The teacher, in turn, can retrieve the students' assignments in one of his/her classes and view a report that highlights the plagiarized parts in each submitted assignment. Several other tools and systems have been developed for Arabic plagiarism detection such as [27].

2- Plagiarism Techniques

Plagiarism techniques are known as similarity detection techniques [27]. Latent Semantic Analysis (LSA) [5] is a technique used to describe relationships between a set of documents and terms they contain. In this technique, words that are close in meaning are assumed to occur close together. A matrix is constructed in which rows represent words, and columns represent documents. Every document contains only a subset of all words. Singular Value Decomposition (SVD), a factorization method of real or complex matrix, is used to reduce the number of columns while preserving the similarity structure among rows. This decomposition is time consuming because of the sparseness of the matrix. Words are compared by taking the cosine of the angle between the two vectors formed by any two rows. Values close to 1 represent very similar words, while values close to 0 represent very dissimilar words this technique is suitable for Arabic plagiarism detection. Stanford Copy Analysis Mechanism (SCAM) [7] is based on a registration copy detection scheme. Documents are registered in a repository and then compared with the pre-registered documents. The architecture of the copy detection server consists of a repository and a chunker. The chunking of a document breaks up a document into sentences, words or overlapping sentences.

The most popular techniques include string tiling, finding the joint coverage for a pair of files [19, 20] and parse tree comparison [21,22]. Usually these techniques work in pairs of files, so the comparison routine should be called for each possible file pair found in the input collection.

Salha Mohammed Alzahrani and Naomie Salim present statement-based plagiarism detection technique in Arabic scripts using fuzzy-set IR model in which the degree of similarity is calculated and compared to a threshold value to judge whether two statements are the same or different. They construct and test documents with about 250 plagiarized statements, their results show that fuzzy set IR successfully detected not only exact but also similar statements that have different structure [23,24].

A fingerprint is a set of integers created by hashing subsets of a document to represent its key content. Techniques to generate fingerprints are mainly based on k-grams (a k-gram is a contiguous substring of length k) which serve as a basis for most fingerprint methods [17]. Fingerprinting technique is widely used for Arabic plagiarism detection. K-grams are central to fingerprinting techniques because fingerprinting divides the document into grams of certain length k [24]. This allows the fingerprints of two documents to be compared in order to detect plagiarism. The fingerprint matching approach differs based on the comparison unit (i.e., grams). This technique can be classified into two categories, namely full and selective fingerprinting [12].

3- Plagiarism Algorithms:

In this section a number of plagiarism detection algorithms will be discussed. The simple algorithm based on string comparisons by removing all comments, Ignore all blanks, and extra lines, except when needed as delimiters, perform a character string comparison between the two files, and maintain a count of percentages of character correlation. This algorithm is run for all possible program pairs. This simple algorithm will detect many cases of plagiarism.

APlag is built around a content-based method. It fulfills the three properties. The first property is handled by a preprocessing of any input text, including tokenization, stopword removal, Synonym replacement, Fingerprinting and similarity metrics [27].

Winnowing algorithm: The winnowing algorithm is an algorithm to select document fingerprints from hashes of k-grams [20]. To obtain the fingerprint of a document, the text is divided into k-grams, the hash value of each k-gram is calculated, and a subset of these values is selected to be the fingerprint of the document. The example below shows the steps to get the fingerprint for the text “Kuala Lumpur.”

Randa K. [21] has developed APD Tool stand-alone desktop tool base on Winnowing local document Fingerprinting Algorithm.it has been adaptive for Arabic and tested using three essays written by a class of Student. She has concluded that ADP is an efficient solution to minimizing student coping.

Mohamed .El Bachir in 2012 implemented a prototype of APlag in Java it is based on a new comparison algorithm that uses heuristics to compare suspect documents at different hierarchical levels to avoid unnecessary comparisons [23].

He evaluated its performance in terms of precision and recall on a large data set of Arabic documents, and show its capability in identifying direct and sophisticated copying, such as sentence reordering and synonym substitution [19]. He presented and discussed a series of experiments to demonstrate its effectiveness on a large set of Arabic documents. The results indicate that APlag has the capability to detect precisely exact copy, change in sentence structure, and synonym replacement [4].

Mohamed El Bachir Menai and Manar Bagais introduce APlag, a new plagiarism detection tool for Arabic texts, based on a logical representation of a document as paragraphs, sentences, and words, and new heuristics for text comparison. We describe its main attributes and present the results of some experiments conducted on a dummy test set. We demonstrate its effectiveness by comparing its performance to that of APD, a plagiarism detection tool for Arabic. Overall, preliminary results show that APlag significantly improves the results obtained by APD in terms of recall and precision metrics.

"Bing" a search engine, they developed a system to detect plagiarism in both Arabic and English languages using "Bing" search engine. The system which relies on plagiarism detection algorithm is effective and can support both Arabic and English languages. Through experiment and tests on our plagiarism detection algorithm, we found that this algorithm reduced the un-useful comparison between texts, since it compares only between cue-phrases surrounding words which forms the logical and natural boundaries of text sentences [13].

Alzahrani et al., 2009 have produced an Arabic plagiarized detection (APD) tool especially for working with Arabic language [11]. APD (Arabic Plagiarism Detection) tool use the Internet to help professors and teachers in e-learning systems identify stolen intellectual property by utilizing Google API to find similar documents on the web [10]. The typical workflow in APD paradigm has two major steps. The first step, students submit their assignments in Arabic to the system, which in turn will be stored into reports database. The second step, the teacher triggers APD tool via a user interface to check the assignments for plagiarism. Then, the tool will compare the documents against the intra corpus collection, which probably contains the previous assignments. Moreover, APD tool searches the web to give similar resources as well. An automatic report will be generated that contains highlighted plagiarized parts and a list of similar resources ranked from highest to lowest [1].

Modern plagiarism detection systems usually implemented using certain content-comparison techniques. The most popular techniques include string tiling, finding the joint coverage for a pair of files [13, 14] and parse trees comparison [15, 16 ,17].

Some of existing plagiarism detectors that employ structure-based methods such as plagues (one of the earliest structure-based detectors). [8]

II. PROPOSED SYSTEM

According to what has been discussed in the survey above, we propose a web based system for detecting plagiarism in electronic resources for Arabic documents. The proposed system depends on word stemming, fingerprint and heuristic method. The proposed system will be tested and evaluated using a set of Arabic documents.

III. CONCLUSIONS

Plagiarism is defined as the unauthorized use or closer imitation of the language and thought of another author and the representation of them as one's own original work. A survey on plagiarism detection systems for Arabic language has been introduced. Authors conclude that the need for plagiarism detection systems for Arabic language become very important issues. Authors propose a web based system that is able to detect many plagiarisms in Arabic text.

Acknowledgments

The authors are thankful to IJACT Journal for the support to develop this document.

References

[1] Alzahrani SM, Salim N. Statement-based fuzzy-set IR versus fingerprints matching for plagiarism detection in Arabic documents. In Proc. of the 5th Postgraduate Annual Research Seminar, Malaysia; 2009.

[2] Ameera Jadalla and Ashraf Elnagar "A Plagiarism Detection System for Arabic Text-Based Documents", Department of Computer Science, University of Sharjah, P.O. Box 27272, Sharjah, UAE, © Springer-Verlag Berlin Heidelberg 2012

[3] Alaa m. Riad , farahat f. Farahat , aziza s. Asem & mahmoud a. Zaher,"Studying Different Methods For Plagiarism Detection", International Journal of Computer Science and Engineering (IJCSE) ISSN(P): 2278-9960; ISSN(E): 2278-9979 Vol. 2, Issue 5, Nov 2013, 147-154 © IASET.

[4] B. Belkhouche, A. Nix, and J. Hassell, "Plagiarism detection in software designs," Proc. of the 42nd Annual Southeast Regional Conference, 2004.

[5] Dumais S.T. Latent Semantic Analysis [J]. Annual Review of Information Science and Technology, 2005: 38-188, doi:10.1002/aris. 1440380105.

[6] Farahat F. Farahat1, Aziza S. Asem2, Mahmoud A. Zaher3*and Ahmed M. Fahiem4, "Detecting Plagiarism in Arabic E-Learning Using Text Mining"British Journal of Mathematics & Computer Science 8(4): 298-308, 2015, Article no.BJMCS.2015.163 ISSN: 2231-0851

[7] F. Sanchez-Vega, E. Villatoro-Tello, M. Montes-y, L. Villase, P. Rosso, "Determining and characterizing the reused text for plagiarism detection" , Contents lists available at SciVerse ScienceDirect , Expert Systems with Applications 40 (2013) 1804–1813.

[8] G. Whale, "Plague : plagiarism detection using program structure,"Dept. of Computer Science Technical Report 8805, University of NSW,Kensington, Australia, 2008

[9] G. Oberreuter, and J. D. Velsquez, "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style", Contents lists available at SciVerse ScienceDirect, Expert Systems with Applications 40 (2013) 3756–3763

[10] Imtiaz Hussain Khan, Muazzam Ahmed Siddiqui, Kamal Mansoor Jambi and Abobakr Ahmed Bagais, " A FRAMEWORK FOR PLAGIARISM DETECTION IN ARABIC DOCUMENTS", Dhinaharan Nagamalai et al. (Eds) : CSEA, DKMP, AIFU, SEA – 2015 pp. 01–09, 2015. © CS & IT-CSCP 2015.

[11] Izzat Alsmadi1 , Ikdam AlHami2 and Saif Kazakzeh3," Issues Related to the Detection of Source Code Plagiarism in Students Assignments", International Journal of Software Engineering and Its Applications Vol.8, No.4 (2014), pp.23-34.

[12] J. A. Faidhi and S. K. Robison, "An empirical approach for detecting program similarity within a university programming environment,"Computers and Education, 2008.

[13] K. Omar, B. Alkhatib, M. Dashash ,"The Implementation of Plagiarism Detection System in Health Sciences Publications in Arabic and English Languages" International Review on Computers and Software (I.RE.CO.S.), Vol. 8, N. 4 ISSN 1828-6003 April 2013.

[14] L. Prechelt, G. Malpohl, and M. Philippsen, "Finding plagiarisms among a set of programs with JPlag," Journal of Universal ComputerScience, 2008.

[15] L. Romans, G. Vita, and G. Janis, "Computer-based plagiarism detection methods and tools: an overview", the 2007 international conference on Computer systems and technologies. 2007, ACM: Bulgaria.

[16] M. Mozgovoy, K. Fredriksson, and D. White, "Fast plagiarism Detection system," Lecture Notes in Computer Science, 2005.

[17] M. El Bachir Menai, " Detection of Plagiarism in Arabic Documents", I.J. Information Technology and Computer Science, 2012, 10, 80-89 Published Online September 2012 in MECS (http://www.mecs-press.org/)DOI:10.5815/ijitcs. 2012.10.10.

[18] Manuel Zini, Marco Fabbri, Massimo Moneglia, Alessandro Panunzi, " Plagiarism Detection Through Multilevel Text Comparison", Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06) 0-7695-2625-X/06 ,2006.

[19] M. Menai,and M. Bagais, "APlag: A Plagiarism Checker for Arabic Texts" The 6th International Conference on Computer Science & Education (ICCSE 2011), IEEE 2011.

[20] Plagiarism.org. "What is Plagiarism?" Web 4 Nov. 2015. <http://www.plagiarism.org/ plagiarism-101/what-is-plagiarism>.

[21] Randa. K., "A Plagiarism Detection Tool For Arabic Text Document", Thesis of Master Degree ,Sudan University of Science and Technology , 2010, Sudan.

[22] Shivakumar N., Garcia-Molina H. SCAM: a copy detection mechanism for digital documents [C]. In: Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, Austin,Texas, USA, June 1995

[23] Salha Mohammed Alzahrani, and Naomie Salim, "Plagiarism Detection In Arabic Scripts Using Fuzzy Information Retrieval", Proceedings of 2008 Student Conference on Research and Development (SCOReD 2008), 26-27 Nov. 2008, Johor, Malaysia.

[24] S. M. Alzahrani, N. Salim, "Statement-based fuzzy-set IR versus fingerprints matching for plagiarism detection in Arabic documents," In Proc. of the 5th

Postgraduate Annual Research Seminar (PARS09), Johor Bahru, Malaysia, 2009.

[25] Types of Plagiarism (n.d.) Retrieved Oct. 2, 2009, from http://www.plagiarism.org/plag_article_ types_ of _plagiarism.html Reprinted with permission.

[26] http://www.plagiarism.com/, visited 30 Apr 2014.

[27] UKessays.com, "A Survey Of Plagiarism Detection Methods Information Technology Essay.". 11 2013.

Biographies

Yahya A. Abdelrahman B.S from University of Science and Technology, Computer Science Master of Science (Computer Science) University Technology Malaysia (UTM) 2007 and PhD student research in Sudan University of Science and Technology now. Faculty members and lecturer at the Najran University College of computer Science and information system Saudi Arabia. Y.Ali Abdurrahman may be reached at yahyaali@gmail.com

AHMED KHALID B.Sc. (honor) in computer science and statistics from University of Khartoum 1983-1989, master of Computer Science (AI) Sudan University of science and technology 1998. PhD in computer science Sudan University of science and technology 2007 coordinate of information technology department, computer science and information technology college University of science and technology Sudan 2008-20012.professor assistant Najran University, computer science department now - 20012 .

Izzeldin M. Osman Emeritus professor of computer science at Sudan University of Science and Technology. Prof. Izzeldin Mohamed Osman.