Research Journal of Applied Sciences, Engineering and Technology 13(4): 277-282, 2016
DOI:10.19026/rjaset.13.2943
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 8, 2015 Accepted: April 22, 2016 Published: August 15, 2016

Research Article

Role of Text Mining in Detection of Plagiarism in Arabic Texts: An Architectural Perspective

Abdullah Al Hussein
College of Computer and Information Sciences, Majmaah University, P.O. Box 66, Al Majmaah 11952, Kingdom of Saudi Arabia

Abstract: The aim of the study of to design and text mining tool for plagiarism detection in Arabic document. Plagiarism is an act or instance of using or closely imitating the language and thoughts of another author without authorization and the representation of that author's work as one's own, as by not crediting the original author (El-Matarawy *et al*., 2013). Plagiarism detecting in Arabic language documents are difficult because of the complex linguistic structure of this language. Text mining in data mining has a very good role in the natural language processing. In this study, we present plagiarism detection architecture for comparison of Arabic texts to identify similarities using text-mining methods. A new text-mining algorithm namely Text Mining Algorithm for Plagiarism Detection (TMA-PD) is proposed to generate tokens from the given Arabic document. A new framework, which combines the familiar text mining methods *Topic Tracking*, *Clustering* and *Concept Linkage*, is also proposed in this research study. This tool thus reduces the time in preliminary part in the detection of plagiarism. As the present text mining tools do not have the feature to process Arabic documents, a new add-on will be developed and integrated in it. Software agents are also used for better comparisons and to find out more texts that are similar. The performance of this tool will be evaluated on a large data set of Arabic texts.

Keywords: Plagiarism, text mining

INTRODUCTION

Several language-centered tools for the detection of plagiarism is in place, particularly for English. Plagiarism detecting in Arabic language documents are particularly a challenging task because of the complex linguistic structure of this language. Text mining in data mining has a very good role in the natural language processing. In this study, we present plagiarism detection architecture for comparison of Arabic texts to identify similarities using text-mining methods. A new text-mining algorithm namely Text Mining Algorithm for Plagiarism Detection (TMA-PD) is proposed to generate tokens from the given Arabic document. A new framework, which combines the familiar text mining methods *Topic Tracking*, *Clustering* and *Concept Linkage*, is also proposed in this research study. This tool thus reduces the time in preliminary part in the detection of plagiarism. As the present text mining tools do not have the feature to process Arabic documents, a new add-on will be developed and integrated in it. Software agents are also used for better comparisons and to find out more texts that are similar. The performance of this tool will be evaluated on a large data set of Arabic texts.

LITERATURE REVIEW

In order to meet the objective specified in Objective and methodology, the review of literature has been carried out in the following contexts:

- Plagiarism detection system
- Plagiarism detection techniques
- Role of text mining in plagiarism detection techniques

Osman *et al*. (2012) introduces a plagiarism detection technique which is based on the semantic role labelling. This technique analyses and compares text based on the semantic allocation for each term inside the sentence.

Barrón-Cedeño *et al*. (2013) propose a freely available architecture for plagiarism detection across languages covering processes like heuristic retrieval, detailed analysis and post-processing.

El-Matarawy *et al*. (2013) explores the potential of data mining techniques in plagiarism detection. In particular, the research proposed a plagiarism technique based on Sequential Pattern Mining (SPM), words/ statements are treated as a sequence of transactions

Table 1: Abstract of reviewed literature

2013	Freely available architecture for plagiarism across language covering heuristic retrieval, detail analysis and post-processing (*Alberto* Barrón-Cedeño *et al*., 2013)
	Plagiarism technique based on sequential pattern mining (El-Matarawy *et al*., 2013)
	Discover deviation in the style, looking for segments of the document that could have been written by another person (Oberreuter and Velásquez, 2013)
	Discusses different textual features that characterize different plagiarism types (Ramya and Venkatalakshmi, 2013)
	Approach based on reducing the author's profile by focusing on the age and gender dimensions (Mechti *et al*., 2013)
2012	Novel plagiarism detected system for Arabic text-based documents, Iqtebas 1.0 (Jadalla and Elnagar, 2012)
	Plagiarism detection technique based on the semantic Role Labelling (Osman *et al*., 2012)
	Survey on plagiarism detection system, a summary of several plagiarism types, techniques and algorithms is provided (Bin-Habtoor and Zaher, 2012)
	Preliminary study on intrinsic plagiarism detection in Arabic textual documents (Bensalem *et al*., 2012)
	Plagiarism detection tool for comparison of Arabic documents to identify potential similarities (Menai and Bagais, 2011)
	New taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism (Alzahrani *et al*., 2012)
2011	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 (Menai and Bagais, 2011)
2010	New approach to detect plagiarism which integrates the use of fingerprints matching technique with four key features to assist in the detection process (Kent and Salim, 2010)
2008	Statement-based plagiarism detected approach in Arabic script using fuzzy-set IR model (Alzahrani *et al*., 2012)

processed by the SPM algorithm to find frequent item sets. The research submits an experiment to discover copy/paste in the text source and produces good results in a reasonable and acceptable time.

Jadalla and Elnagar (2012) present a plagiarism detection system for Arabic text-based documents, Iqtebas 1.0. For given input of suspected files, the goal is to compute the originality value of the examined document (s). Winnowing n-gram fingerprinting algorithm is used for indexing to reduce the index size.

Bin-Habtoor and Zaher (2012) present a survey on plagiarism detection systems, a summary of several plagiarism types, techniques and algorithms is provided. Authors propose a web enabled system to detect plagiarism in documents, code and images, also this system could be used in E-Learning, E-Journal and E-Business.

Oberreuter and Velásquez (2013) try to discover deviations in the style, looking forsegments of the document that could have been written by another person/researcher. This research demonstrates that this feature shows promise in this area, achieving reasonable results compared to the available benchmark models.

Bensalem *et al*. (2012) presents a preliminary study on intrinsic plagiarism detection in Arabic textual documents. A set of experiments were conducted to gain an insight into the effect of some well-known language-independent stylistic features on Arabic text discrimination. Used Stylysis tool to measure these features on our small-sized corpus.

Menai and Bagais (2011) introduces *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. This study presents results of some experiments conducted on a dummy test set.

Kent and Salim (2010) propose a new approach to detect plagiarism which integrates the use of fingerprint matching technique with four key features to assist in the detection process. In this study, a method in text similarity detectionis presented.

Ramya and Venkatalakshmi (2013) discusses different textual features that characterize different plagiarism types. Systematic frameworks and methods of monolingual, extrinsic, intrinsic and cross-lingual plagiarism detection are surveyed and correlated with plagiarism types, which are listed in the taxonomy.

Alzahrani *et al*. (2012) presents a new taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism, from the plagiarist's behavioural point of view. Different textual features that characterize different plagiarismtypes are discussed.

Mechti *et al*. (2013) and Rajan and Saravanan (2008) proposes an approach based on reducing the author's profile of a given document by focusing on the age and gender attributes. System takes as input a document, which is written in English or in Spanish and generates the age and the gender of its author.

Table 1 summarizes the above reviewed literatures.

It is very crystal clear from the above reviewed literatures that, no authors were using the text-mining in the plagiarism detection techniques while processing Arab text documents. The proposed model uses text-mining methods for token formation.

OBJECTIVE AND METHODOLOGY

The objective is to propose atext-mining algorithm and architecture for plagiarism detection for processing Arabic text document(s). The Arabic text documents are processed by text miner algorithm and methods initially. To implement the proposed architecture, an open source tool *Rapidminer* is chosen initially which is used for text mining. As this tool, does not have a

feature to process Arabic documents, a new *add-on* will be developed and attached with *Rapidminer*. The tool then processes the Arabic documents and stores the resultant tokens in the archive. If the results found to be of good quality, the developed add-on will be integrated with many open source data mining tools.

ROLE OF TEXT MINING, PROPOSED FRAMEWORK AND ALGORITHM

Many text-mining tools do not have the ability to process Arabic documents. For example, text mining in *Rapidminer* extracts the tokens from text-based content such as word documents and other related documents for English. The proposed system eliminated this limitation by introducing a new text-mining algorithm and architecture. This new algorithm processes the Arabic documents and generates token based on a *word*, *sentence or paragraph*. The text mining thus reduces the time in the initial processing. In addition, the token generated by the text-mining tools will be clearer and valid. The role of text-mining in the proposed system is outlined in the Fig. 1a.

The following text-mining technologies as given in Gupta and Lehal (2009), Al Hussein *et al*. (2014) and Jayabrabu *et al*. (2012a, 2012b) are reviewed that to be considered in the proposed framework:

- Information Extraction
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Concept Linkage
- Information Visualization
- Question Answering
- Association Rule Mining

The technologies topic tracking, clustering and concept linkageare considered in this research study by considering the objective as provided in session 3. These technologies are combined in order to enhance the quality of token produced by the proposed model. This is described in detail in the Topic tracking, clustering and concept linkage.

Topic tracking, clustering and concept linkage: A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles, Gupta and Lehal (2009). The proposed system uses topic tracking to read the archive for a given set of topics. The topic tracking output is used by the clustering technique. The clustering technique is used to remove stop words, stem and for filtering. Concept linkage is used connect the related documents by identifying the commonly shared concepts. This is used in plagiarism detection to retrieve all the similar related documents. The above concept is illustrated in Fig. 1b.

Proposed framework for text mining methods: The topic tracking method extract the keywords from the given inputted Arabic document. The extracted keywords are then ranked against its number appearances and stored in the dictionary (Fig. 1c). The clustering method retrieves the keywords and clusters the words, sentences and paragraphs. These clusters are

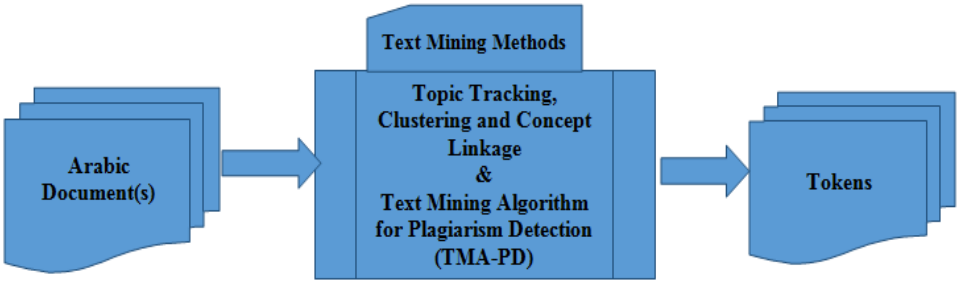

Fig. 1a: Role of text mining

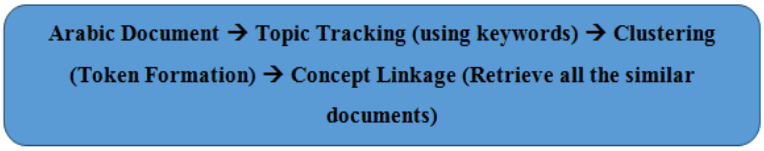

Fig. 1b: Combined text mining technologies

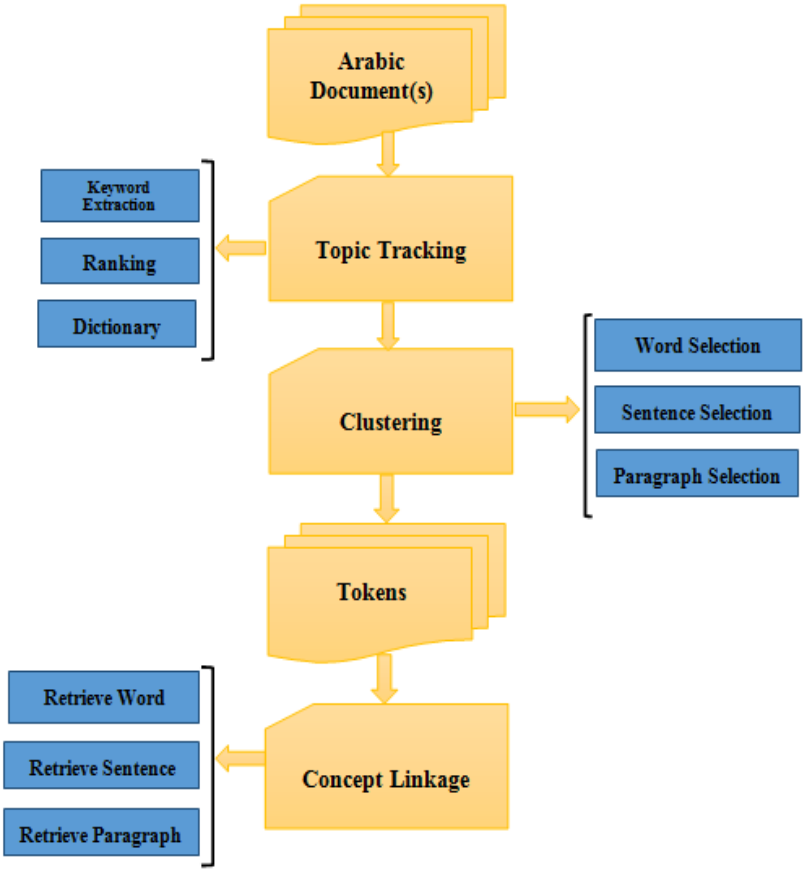

Fig. 1c: Framework of text mining methods

also referred as tokens in data preprocessing. These tokens are then used by the concept linkage methods to retrieve similar words, sentences or/and paragraphs from the archive to find out the similarity index. The existing winnowing finger-printing algorithm is used for calculating the similarity index.

Text Mining Algorithm for Plagiarism Detection (TMA-PD):

Step 1: Start to upload document file in text format
Step 2: Read the input
Step 3: Use the combined text mining methods (topic tracking, clustering and concept linkage)
Step 4: Select the appropriate parameter to construct a token (*word*, *sentence* and *paragraph*)
Step 5: Form tokens for the given input based upon the parameter considered in step 3
Step 6: Check the quality of the token using metrics
Step 7: If (quality = "good") then call similarity_index ();
Step 8: Else go to step 2
Step 9: Store the results (as a Report) in the archive system
Step 10: Stop

The developed algorithm will mine the similarity of the given text from the user inputted Arabic document. The parameter to be used for token formation is *word*, *sentence*, *or paragraph*. The algorithm constructs the token based on an appropriate parameter. The core part of this proposed system is to generate the tokens as provided in the step 5. The text similarity using similarity index function with respect to quality of the text is calculated in step 7. If the generated token quality is not good, the process is repeated to get quality tokens. During text-mining the given text is compared with other text using tokenization.

PROPOSED ARCHITECTURE OF THE SYSTEM

In Fig. 2, the user selects the source file to be checked for plagiarism and submits it to the text-mining model. The text-mining tool processes the Arabic texts and generated the tokens; the generated tokens are then compared with the archive database to find out the similarity index. Software agents are used in this architecture for better interaction between the various components. The detailed discussion on text mining methods and the algorithm is provided in session 4.

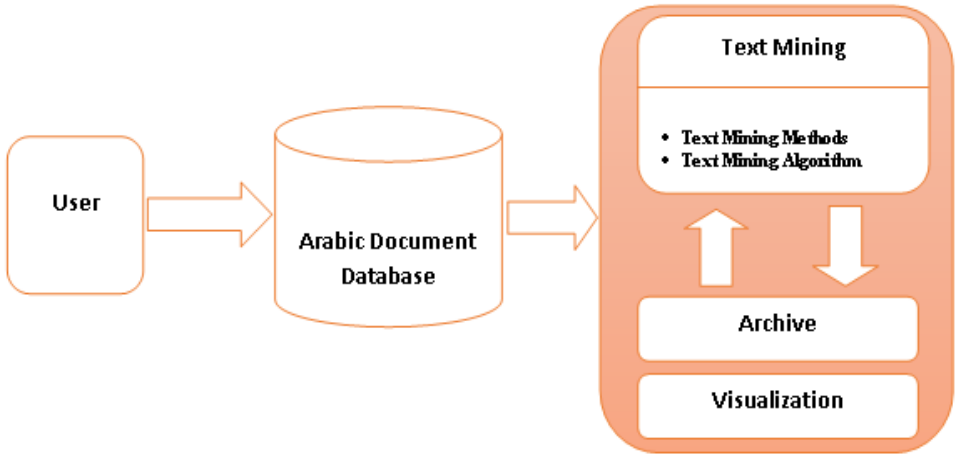

Fig. 2: Architecture of the proposed model

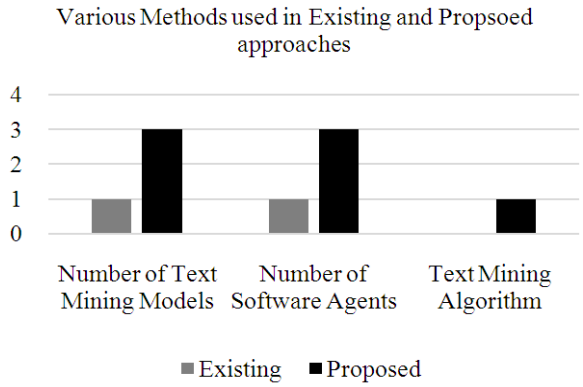

Fig. 3a: Comparision of various methods

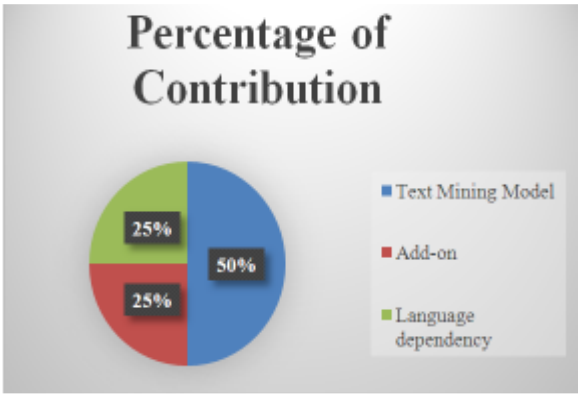

Fig. 3b: Percentage of contribution

OUTCOME

After processing the tool provides the percentage plagiarism involved quoting the respected documents from which the plagiarism is initiated. It will be a combination of many text-mining methods, which will for sure will be a better model for plagiarism check compared to the traditional classical models. This tool can be made with flexible with various algorithmic combinations, which can be added into the tool according to the dataset which is to be processed by supervised learning methods. Figure 3a shows the various models used in the proposed model and the Fig. 3b shows the percentage of contribution of various models in the proposed approach.

CONCLUSION

This is a white study approach such that the architecture proposed is about to be implemented in *Rapidminer* environment with the software agents. Through this, plagiarism in Arabic texts can be accurately detected and the percentage of plagiarism can be analyzed. Three dimensional approaches would be the added advantage for this tool. Hence forth, irrespective to the type of Arabic dataset, this tool can analyze them using different text mining methods which provide much more better reliable results for the initialized dataset which can be analyzed by statistical metrics and measures for further processing.

REFERENCES

Al Hussein, A., S. Venkataraman and R. Jayabrabu, 2014. Detection of plagiarism in Arabic texts using text mining: A software agent based approach. Proceeding of the 6th International Integrity and Plagiarism Conference. Newcastle Gateshead, UK.

Alzahrani, S.M., N. Salim and A. Abraham, 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE T. Syst. Man Cy. C, 42(2): 133-149.

Barrón-Cedeño, A., P. Gupta and P. Rosso, 2013. Methods for cross-language plagiarism detection. Knowl-Based Syst., 50: 211-217.

Bensalem, I., P. Rosso and S. Chikhi, 2012. Intrinsic plagiarism detection in Arabic text: Preliminary experiments. Proceeding of the 2nd Spanish Conference on Information Retrieval (CERI-2012). Valencia, Spain.

Bin-Habtoor, A.S. and M.A. Zaher, 2012. A survey on plagiarism detection systems. Int. J. Comput. Theor. Eng., 4(2): 185-188.

El-Matarawy, A., M. El-Ramly and R. Bahgat, 2013. Plagiarism detection using sequential pattern mining. Int. J. Appl. Inform. Syst., 5(2): 24-29.

Gupta, V. and G.S. Lehal, 2009. A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell., 1(1): 60-76.

Jadalla, A. and A. Elnagar, 2012. A Plagiarism Detection System for Arabic Text-based Documents. In: Chau, M. *et al*. (Eds.), Intelligence and Security Informatics. Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 7299: 145-153.

Jayabrabu, R., V. Saravanan and K. Vivekanandan, 2012a. Software agents paradigm in automated data mining for better visualization using intelligent agents. J. Theor. Appl. Inform. Technol., 39(2): 167-177.

Jayabrabu, R., V. Saravanan and K. Vivekanandan, 2012b. A framework: Cluster detection and multidimensional visualization of automated data mining using intelligent agents. Int. J. Artif. Intell. Appl., 3(1): 125-138.

Kent, C.K. and N. Salim, 2010. Features based text similarity detection. J. Comput., 2(1): 53-57.

Mechti, S., M. Jaoua and L.H. Belguith, 2013. A Framework for Plagiarism Detection Based on Author Profiling. Notebook for PAN at CLEF 2013. ANLP Research Group-MIRACL Laboratory, University of Sfax, Tunisia.

Menai, M.E.B. and M. Bagais, 2011. APlag: A plagiarism checker for Arabic texts. Proceeding of the 6th International Conference on Computer Science and Education (ICCSE, 2011), pp: 1379-1383.

Oberreuter, G. and J.D. Velásquez, 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. Expert Syst. Appl., 40(9): 3756-3763.

Osman, A.H., N. Salim, M.S. Binwahlan, R. Alteeb and A. Abuobieda, 2012. An improved plagiarism detection scheme based on semantic role labeling. Appl. Soft Comput., 12(5): 1493-1502.

Rajan, J. and V. Saravanan, 2008. A framework of an automated data mining system using autonomous intelligent agents. Proceeding of the International Conference on Computer Science and Information Technology, pp: 700-704.

Ramya, L. and R. Venkatalakshmi, 2013. Intelligent plagiarism detection. Int. J. Res. Eng. Adv. Technol., 1(1): 1-4.