西 南 交 通 大 学 学 报

第 54 卷 第 6 期
2019 年 12 月

JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY

Vol. 54 No. 6
Dec. 2019

ISSN: 0258-2724 DOI：10.35741/issn.0258-2724.54.6.30

Research Article

Computer Science

ONLINE MULTILINGUAL PLAGIARISM DETECTION SYSTEM USING MULTI SEARCH ENGINES

多搜索引擎的在线多语言抄袭检测系统

Maytham Alabbas [a,*], Raidah S. Khudeyer [b], Mustafa Radif [c], Hassan Khalid Hameed [a]
[a] Department of CS, College of CSIT, University of Basrah
Al-Ashar–Corniche St., Basrah, Iraq, ma@uobasrah.edu.iq, gruceing@gmail.com
[b] Department of CIS, College of CSIT, University of Basrah
Al-Ashar–Corniche St., Basrah, Iraq, raidah.khudayer@uobasrah.edu.iq
[c] Department of IS, College of CSIT, University of Al-Qadisiyah
P.O. Box 88, Al Diwaniyah, Al-Qadisiyah, Iraq, mustafa.radif@qu.edu.iq

Abstract

Using someone else's work or ideas without attribution is plagiarism, whether you meant to do it or not. Unintended plagiarism of snippet of text can have serious consequences and be a serious form of ethical misconduct. The current system is a web application that enables you to check a multilingual text, with special focus on Arabic, for duplicate contents on the World Wide Web. In this system, you can simply input or paste your text through the online system and for each sentence in the text it will go through three popular search engines: Google, Bing, and Yandex SERP and try to find the top three results on the first page for each search engine where duplicate contents already exist. This system is getting data from the three-search engines custom search APIs. Then, the system uses a text similarity technique between the suspicious sentence and the retrieved text snippet for all nine results. The result is the one that gives the highest similarity rate. The results were encouraging and will open doors for new and innovative techniques for researchers in this field.

Keywords: Plagiarism, Plagiarism Detection, Multilingual Plagiarism Detection, Arabic Plagiarism Detection, Search Engine API

摘要 不管您是否打算使用他人的作品或想法而没有归属，这都是窃。文本片段的意外抄袭可能会导致严重的后果，并且是严重的道德不端行为。当前系统是一个网页应用程序，使您可以检查多语言文本（尤其是阿拉伯语）在全球资讯网上的重复内容。在此系统中，您可以简单地通过在线系统输入或粘贴文本，对于文本中的每个句子，它将经过三个流行的搜索引擎：谷歌，必应和扬德克斯 SERP，并尝试在第一个搜索结果中找到前三个结果已存在重复内容的每个搜索引擎的页面。该系统正在从三个搜索引擎的自定义搜索应用程式介面获取数据。然后，系统针对所有九个结果在可疑句子和检索到的文本片段之间使用文本相似性技术。结果是给出最高相似率的结果。

结果令人鼓舞，并将为该领域的研究人员打开新的创新技术之门。

关键词: 窃，普拉窃检测，多语言普拉窃检测，阿拉伯语 gi 窃检测，搜索引擎应用程式介面

I. INTRODUCTION

Plagiarism is an act of fraud that involves both stealing someone else's work and lying about it afterward [1]. It is not only the copying of words but also includes the taking of other expressions, concepts, and other works. While the Internet's exponential growth has made it easier than ever to achieve plagiarism, it has also made it much easier to check or detect. There are many reasons why people plagiarize such as that they want to get things done conveniently and quickly, their lack of knowledge about the definition and forms of plagiarism, and the lack of an environment that has controls over plagiarism [2]. There are different types of plagiarism. We have defined the most common types according to intensity: (i) substantial plagiarism: this type is the most common in academia. The plagiarist here rephrases the original and replaces the words with their synonyms; (ii) minimal plagiarism: in this type, some information is added in the text, and the text patterns are altered; and (iii) complete plagiarism: in this type, everything is copied from other sources with no changes, and it is presented as the writer's own creation. Here are some examples of plagiarism such as failing to put quotation marks for the copied text, copying someone else's work or words without citing sources properly, and the incorrect citation [3]. However, most cases of plagiarism can be avoided by citing sources properly [1].

The rest of the paper is organized as follows: in Section 2 some related web-based plagiarism detection systems are presented. An explanation of the current system is given in Section 3. Section 4 explains the experiments performed using the current system. Finally, Section 5 is the conclusion of the paper.

II. RESEARCH AIM

This paper reports on the design of amulti-lingual web-based plagiarism detection system with special focus on Arabic. This system applies Google, Bing, and Yandex search engines APIs to retrieve candidate source texts from the Web in order to feed them to the matching stage to determine if the given text was plagiarized from the Web or not. The main contributions of the current work are applying a combination of various search engines, instead of one only like Google in the previous work, to exploit the unique advantage of each one and reduce some of the random mistakes. Moreover, even if the given text is written in several different languages, the current system is capable of detecting plagiarism for all languages that are supported by Google, Bing, and Yandex.

III. LITERATURE REVIEW

The growth and success of our academic communities are threatened by increasing plagiarism rate, especially with the advent of the Web, which is why steps to prevent it need to be taken. Plagiarism detection, therefore, has become an essential part of the academic community nowadays to prevent this notorious problem.

Plagiarism detection can be defined as the process of locating partial or full instances of plagiarism within original sources. It can be either manual or software-assisted. Manual detection requires huge effort and super memory and has become impractical and almost impossible in cases where a vast collection of documents are to be compared, or original documents are missing. On the other hand, software-assisted detection allows a substantial amount of documents to be compared to each other, making the results of detection much more reasonable.

Plagiarism detection can be divided into extrinsic (external) and intrinsic (internal) techniques [4]. The extrinsic plagiarism detection uses a number of measures to compute the similarity between a suspicious document and a reference collection [5]. Different similarity measures are used in this regard, such as Jaccard similarity, Euclidean distance, Cosine similarity, and others to compute the similarity between vectors that represent the documents [6]. On the other hand, in the intrinsic plagiarism detection, the suspicious document is analyzed single-handedly, by using different techniques, without being compared with any sources. As such, it does not take a reference collection into account.

Plagiarism detection has largely been applied to English. There is, so far, very little work on applying plagiarism detection techniques to Arabic, and little evidence that the existing approaches will work for it. The key problem for Arabic is that it is more ambiguous than English, where we are faced with an exceptional level of

structural and lexical ambiguity. Many of the existing techniques to plagiarism detection, therefore, are likely to be inapplicable.

Some related web-based plagiarism detection systems are reviewed here. The authors in [7] design a software tool called SNITCH (Spotting and Neutralizing Internet Theft by CHeaters), which is Google API-based plagiarism detection algorithm. SNITCH uses a sliding window to scan a document and determine candidate texts that might be plagiarized. Each text is searched for on the Web. A brief summary as an annotated HTML document is output containing the original document with hypertext links, statistics about the percentage of plagiarism, and the time taken to complete the checking.

In [8] the authors described a web-based detection tool for cross-language plagiarism, also known as translation plagiarism, that arises especially in academic works. The system considers documents written in Bahasa Melayuas. These are translated into English using Google Translate API. Google AJAX Search API is used to detect retrieved documents throughout the Web by considering the top ten sources as the candidate documents. This system also used the Stanford Parser and WordNet to determine the similarity level between the given documents with the candidate documents. Then, a similarity analysis is performed and a final report is generated.

In [5] the authors proposed an ongoing project for Arabic plagiarism detection framework with global and local components. In the global component, a suspicious document is used to construct different representative queries by using various high-performing heuristics. Then, the queries are used by Google's search API to retrieve source documents from the Web. Next, the local component combines different similarity techniques to detect if the suspicious document was plagiarized from the Web or not. The global part is completely evaluated, but the local part is only partially implemented so far and hence the overall quality of this system is not evaluated yet.

The authors in [9] described a web-based anti-plagiarism approach for the academic level, such as student's assignments, seminar reports, teacher's research papers, theses submitted in different research fields done at postgraduate level. The aim is to bring a halt to the copy and paste culture in academia. The proposed tool attempted to match parts of the given document to the parts of those in the large collection, i.e. the Local database (local drive), Distributed database (LAN), and Global database (WWW) sequentially. Google search API is used for certain keywords or key sentences from a given document on the Web. The tool displays the results in the form of the URL.

In [10] the authors evaluated the efficiency level of online academic plagiarism detection tools (PlagScan, iThenticate, and Check For Plagiarism) in detecting different plagiarism patterns' amounts in Arabic. Their experiments have shown that the most effective online plagiarism checker is the iThenticate system compared with the others.

IV. RESEARCH METHOD

The system described in this paper, namely Basrah Plagiarism (BasPlag), is a web-based plagiarism detection system. It is used to detect multilingual text plagiarism online using three popular search engines: Google, Bing, and Yandex APIs. It accepts as input a text with a plagiarism threshold and outputs each sentence with a score that tells you if the sentence is unique or plagiarized, along with additional information such as URL and the percentage of similarity in case of the online existence of the test sentence [26].

BasPlag proceeds in nine steps, as follows:

1. This step is responsible for inputting a text from the keyboard or pasting it from another source. In addition to a plagiarism threshold is specified at this point;
2. The given text is split into sentences using regular expressions;
3. Each sentence is checked to see whether it is Arabic or not. If it is not Arabic, goto step 5, otherwise the processing continues;
4. In order to normalize the Arabic sentence for the next step, the sentence undergoes the following preprocessing:
   a. removing all diacritics (Diacritization in the Arabic language is done by adding special symbols called vowel points to help in spoken language.) such as Damma (ضمة), Fatha (فتحة), Kasra (كسرة);
   b. removing all stop words such as هذا، في، عن;
   c. replacing يءwith ئ;
   d. replacing ةwith ه;
   e. normalizing Alif variants (أ,آ, and إ) toا;
   f. removing all punctuation marks (؟ ( )] [...etc) except (.,;).
5. This step uses three popular search engines: Google [11], Bing [12], and Yandex [13] APIs and tries to find the top three results for each search engine where duplicate contents already exist;
6. The text similarity rates between the sentence and the returned snippet of text for all

nine results are computed. The one that gives the highest similarity rate is selected as the final result. Here, if the detected language is not Arabic then the built-in PHP *similar_text* function [14] is used, otherwise a hybrid text similarity technique that combines cosine similarity [15] and built-in PHP *similar_text* function is used to find the final similarity score between the two strings (S_1, S_2) as in Eq. 1.

$$sim(S_1, S_2) = 0.5(cos(S_1, S_2) + similar\_text(S_1, S_2)). \quad (1)$$

Cosine similarity is calculated by measuring the cosine of the angle between two vectors (q,t) that represent a suspicious sentence and a source text respectively as in Eq. 2.

$$\cos(q,t) = \frac{qt}{\|q\|\|t\|} = \frac{\sum_{i=1}^{n} q_i t_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2}\sqrt{\sum_{i=1}^{n}(t_i)^2}}, \quad (2)$$

where q_i is the *tf-idf* weight of term *i* in S_1 and t_i is the *tf-idf* weight of term *i* in S_2.

One way to convert sentences into vectors involves using a bag of words with TF-IDF (term frequency - inverse document frequency) as in Eq. 3.

$$tf - idf_{i,j} = tf_{i,j} \times log\frac{N}{df_i}, \quad (3)$$

where $tf_{i,j}$ is the number of occurrences of *i* in *j*, idf_i is the number of documents containing *i*, *N* is the total number of documents;

7. The similarity rate is compared with the given plagiarism threshold. In case of the similarity rate is greater than or equal the threshold, the sentence, URL, the similarity percentage that tells you how originality or uniqueness of the sentence is displayed, followed by 'Plagiarized' are displayed in red color. Otherwise, the sentence and 'Unique' are displayed in green color;
8. If there are more sentences, go to step 3;
9. More results about the given text are shown on the final report, such as number of characters, number of words, the percentage of plagiarism, the percentage of originality, and the plagiarism report as PDF file contains all details above.

The general framework of BasPlag system is illustrated in Figure 1.

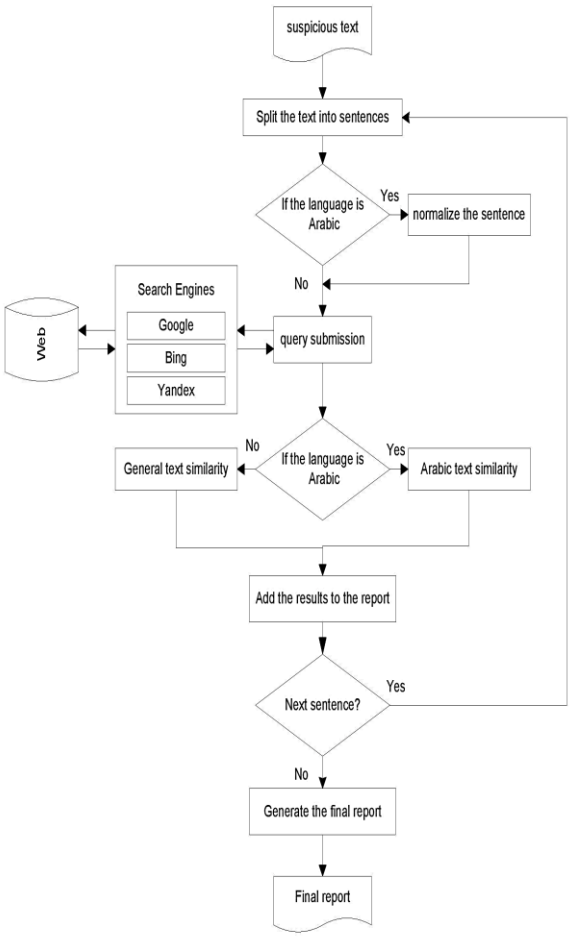

Figure 1. The general framework of BasPlag system

V. RESULTS AND DISCUSSION

Two types of experiment are presented here; Arabic text only and multilingual text. In both cases, the outcome or the expected final similarity report for the suspicious text will be generated, where the plagiarized sentences will be highlighted with red indicating the source, while the unique sentences will be highlighted with green as in [16] and [17].

A. Test 1: Arabic Text

In this experiment, an Arabic text, which contains four plagiarized sentences and one original sentence, is used. The given text is shown in Figure 2.

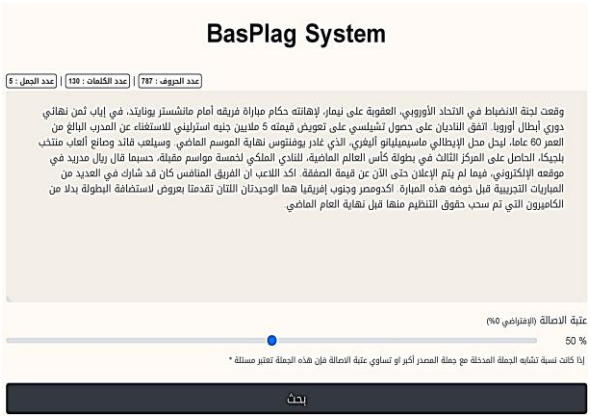

Figure 2. Input text user interface Test 1

The final similarity report of this test is shown in Figure 3.

Figure 4. Final plagiarism report Test 1

B. Test 2: Multilingual Text

In this experiment, a multilingual text, which contains four plagiarized sentences in Arabic, English, Russian and Franch, is used. The given text is shown in Figure 3.

Figure 4. Input text user interface Test 2

The final similarity report of this test is shown in Figure 5.

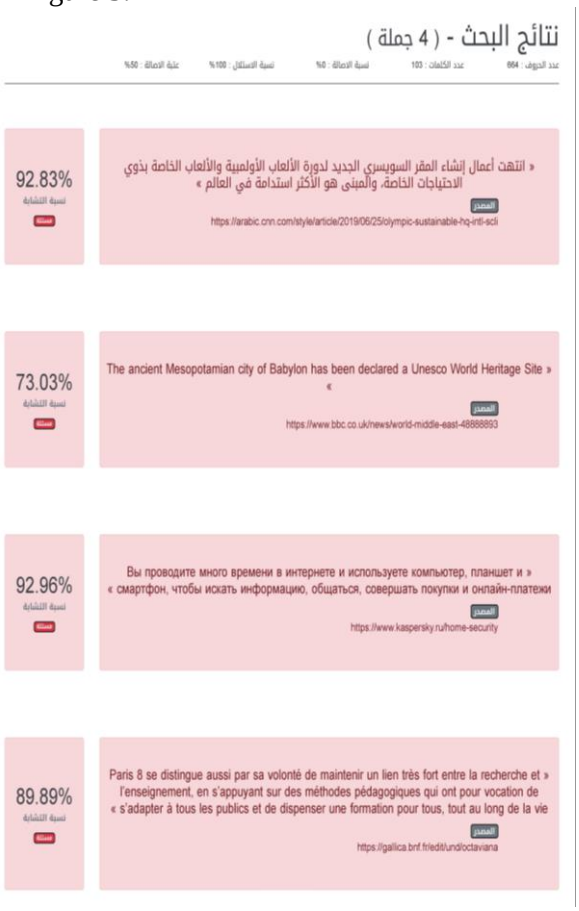

Figure 5. Final plagiarism report Test 2

As it can be seen from the above figures, BasPlag was successfully able to identify the original and stolen sentences even though the suspicious text was multilingual--written in four different languages.

Over the past two decades, there has been considerable interest in building automatic plagiarism detection systems. Many systems have been developed in this regard, however, the challenge still stands: how to identify the plagiarism more effectively. The challenge is even worse for Arabic where we are faced with an exceptional level of lexical and structural

ambiguity. In the current work, BasPlag plagiarism detection system for multilingual is applied.

The process of retrieving a correct related text to each query on the Web (text retrieved stage) using the search engine is considered an essential step in the current system. In general, however, search engines make mistakes and these mistakes will lead to problems in all subsequent stages of the system. It is thus important to obtain the highest possible accuracy at this stage of processing. One popular technique for improving search engine search accuracy involves system combination, which has been applied for different natural language processing (NLP) tasks such as tagging [18], parsing [19], and character recognition [20], [21] and it seemed *prima facie* plausible that it would also work for our task. This technique is concerned with combining different search engine results to exploit the unique properties of each search engine and reduce some of the random errors. We evaluate here the combination of three search engine results for several languages. Applying this technique gives encouraging results because each one uses a different search strategy. Google, for instance, is working on ways to understand the context behind the query. Bing uses targeted keywords as a ranking parameter and takes meta keywords into account when ranking websites. Yandex takes into consideration the distance between words and the relevance of documents to a query. Each search engine, therefore, will produce different results depending on the nature of the sentence and the language in which the sentence was written.

The findings are quite encouraging. But still there are some limitations must be overcome in order to make BasPlag system more accurate and applicable. The most important of these limitations are:

- The performance of BasPlag system mainly depends on the accuracy of the three search engines results (i.e. Google, Bing, and Yandex). This is because through stage 3 we retrieve potential source texts from the Web using search engine API. We believe that the improvement of search engines' strategies (using synonyms, automatic error correction, and context-sensitive help) can eliminate some of the errors. This will significantly improve the accuracy and speed of BasPlag system because the retrieved texts will be relevant thus the accuracy of subsequent stages would increase accordingly;
- The current search engines which are used here have both limitations on the maximum number of submissions per free subscription account and limitations on query length. For instance, Google has 100 free queries per day, Bing has 3000 free queries per month, and Yandex has 10000 free queries per day for account verified by telephone number.
- In Arabic, sentences tend to be rather long. The typical sentence length of Arabic is 30 to 40 words, and sentences whose length exceeds 200 words are not uncommon because Arabic writing rarely contains punctuation marks even though the language has them [22]. This makes Arabic NLP challenging in general and specifically the task of text similarity, which is the core of the Arabic plagiarism detection systems that are mainly based on readability score, significantly more difficult than it already is for the English language. The situation is much worse as we test longer examples. We have reasonably solved this problem by using regex-based rules to split the long sentences into shorten ones.

VI. CONCLUSION

We have presented here BasPlag system, which is a search engines-based online multilingual plagiarism detection tool that enables you to check a text for duplicate contents on the Web. BasPlag system accepts a text as input. The given text should be in a language supported by Google, Bing, and Yandex APIs. Then, the text is split into sentences to construct a set of representative queries that are submitted to three popular search engines (Google, Bing, and Yandex) via search engine API to retrieve candidate source texts from the Web. Next, the text similarity technique is applied to detect if the given text was plagiarized from the texts retrieved from the Web or not. Finally, a similarity report for the given text will be generated, where the plagiarized text will be highlighted with various colors indicating the original source.

The current findings are encouraging and show that combining different search engine gives better results than each search engine alone. BasPlag system proves itself as an efficient, quick and simple system, query a sentence on a search engine quickly brings back different sources related to the query.

Further experimental investigations are needed to extend BasPlag system by investigating the performance of different similarity measures [6]. We also speculate that further work by using improving text similarity using popular off-the-shelf word embedding

models such as Google Word2Vec [23], Stanford GloVe [24], and Facebook fastText [25].

ACKNOWLEDGMENT

The author's special thanks and appreciations are given to Professor Allan Ramsay (UK) for productive discussions.

REFERENCES

[1] PLAGIARISM (2017) *What is plagiarism?* [Online]. Available from: https://www.plagiarism.org/article/what-is-plagiarism [Accessed 17/06/19].

[2] MENAI, M.E.B. and BAGAIS, M. (2011) APlag: A plagiarism checker for Arabic texts. In: *Proceedings of the 6th International Conference on Computer Science & Education, Singapore, August 2011*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 1379-1383.

[3] KHORSI, A., CHERROUN, H., and SCHWAB, D. (2018) 2L-APD: A Two-Level Plagiarism Detection System for Arabic Documents. *Cybernetics and Information Technologies*, 18 (1), pp. 124-138.

[4] GUPTA, D. (2016) Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *Journal of Engineering Science & Technology Review*, 9 (5), pp. 150-164.

[5] KHAN, I.H., SIDDIQUI, M.A., and MANSOOR, K. (2015) A framework for plagiarism detection in Arabic documents. In: *Proceedings of the Conference on Computer Science & Information Technology*, pp. 1-9.

[6] GOMAA, W.H. and FAHMY, A.A. (2013) A survey of text similarity approaches. *International Journal of Computer Applications*, 68 (13), pp. 13-18.

[7] NIEZGODA, S. and WAY, T.P. (2006) SNITCH: a software tool for detecting cut and paste plagiarism. In: *Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, Houston, Texas, March 2006.* New York: Association for Computing Machinery, pp. 51-55.

[8] KENT, C.K. and SALIM, N. (2010) Web based cross language plagiarism detection. In: *Proceedings of the 2nd International Conference on Computational Intelligence, Modelling and Simulation, Bali, September 2010.* Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 199-204.

[9] KHATRI, J. and MOHAN, V. (2016) An Approach for Implementing Web-Based Tool for Plagiarism Detection. *International Journal of Engineering and Management Research*, 6 (3), pp. 57-60.

[10] ADEL, G.M.A. and WANG, Y. (2019) Effectiveness Level of Online Plagiarism Detection Tools in Arabic. *Internet of Things and Cloud Computing*, 7 (1), pp. 19-24.

[11] GOOGLE DEVELOPERS (2019) *Custom search JSON API: Introduction.* [Online]. Available from: https://developers.google.com/custom-search/v1/introduction [Accessed 15/03/19].

[12] MICROSOFT AZURE (2019) *Bing web search API.* [Online]. Available from: https://docs.microsoft.com/en-us/java/api/overview/azure/cognitiveservices/client/bingwebsearchapi?view=azure-java-stable [Accessed 01/04/19].

[13] YANDEX TECHNOLOGIES (2019) *Yandex.XML*. [Online]. Available from: https://tech.yandex.com/xml/ [Accessed 10/04/19].

[14] THE PHP GROUP (2019) *PHP: similar_text - Manual.* [Online]. Available from: https://www.php.net/manual/en/function.similar-text.php [Accessed 18/05/19].

[15] MANNING, C., RAGHAVAN, P., and SCHÜTZE, H. (2010) Introduction to information retrieval. *Natural Language Engineering*, 16 (1), pp. 100-103.

[16] iThenticate. [Online]. Available from: http://www.ithenticate.com/ [Accessed 18/06/19].

[17] Turnitin. [Online]. Available from: https://www.turnitin.com [Accessed 18/06/19].

[18] ALABBAS, M. and RAMSAY, A. (2012) Improved POS-Tagging for Arabic by Combining Diverse Taggers. In: ILIADIS, L., MAGLOGIANNIS, I., and PAPADOPOULOS, H. (eds.) *Artificial Intelligence Applications and Innovations.*

AIAI 2012. IFIP Advances in Information and Communication Technology, Vol. 381. Berlin, Heidelberg: Springer, pp. 107-116.
[19] ALABBAS, M. and RAMSAY, A. (2014) Improved Parsing for Arabic by Combining Diverse Dependency Parsers. In: VETULANI, Z., and MARIANI, J. (eds.) *Human Language Technology Challenges for Computer Science and Linguistics. LTC 2011. Lecture Notes in Computer Science*, Vol. 8387. Cham: Springer, pp. 43-54.
[20] ALABBAS, M., KHUDEYER, R.S., and JAF, S. (2016) Improved Arabic characters recognition by combining multiple machine learning classifiers. In: *Proceedings of the 2016 International Conference on Asian Language Processing, Tainan, November 2016.* Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 262-265.
[21] JAF, S., ALABBAS, M., and KHUDEYER, R.S. (2018) Combining Machine Learning Classifiers for the Task of Arabic Characters Recognition. *International Journal of Asian Language Processing*, 28 (1), pp. 1-12.
[22] ALABBAS, M. and RAMSAY, A. (2011) Evaluation of dependency parsers for long Arabic sentences. In: *Proceedings of the 2011 International Conference on Semantic Technology and Information Retrieval, Putrajaya.* Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 243-248.
[23] MIKOLOV, T., CHEN, K., CORRADO, G., and DEAN, J. (2013) Efficient estimation of word representations in vector space. [Online]. Available from: http://arxiv.org/pdf/1301.3781.pdf [Accessed 15/10/19].
[24] PENNINGTON, J., SOCHER, R., and MANNING, C. (2014) *GloVe: Global vectors for word representation.* In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, October 2014.* Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1532-1543.
[25] BOJANOWSKI, P., GRAVE, E., JOULIN, A., and MIKOLOV, T. (2017) Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135-146.
[26] HUSSEIN, S.A. and KAREEM, M.R. (2019) A Proposed Arabic Text and Text Image Classification Technique Using a URL Address. *Journal of Southwest Jiaotong University*, 54 (5). Available from http://jsju.org/index.php/journal/article/view/411.

参考文:

[1] GI 窃（2017）什么是窃？ [线上]。可从以下网站获得：https://www.plagiarism.org/article/what-is-plagiarism [访问时间：19/06/19]。
[2] MENAI，硕士和 BAGAIS，M.（2011）标记：阿拉伯文字抄袭检查工具。于：2011 年 8 月在新加坡举行的第六届国际计算机科学与教育国际会议论文集。新泽西州皮斯卡塔维：电气与电子工程师协会，第 1379-1383 页。
[3] KHORSI，A.，CHERROUN，H. 和 SCHWAB，D.（2018）2L-APD：阿拉伯文档的两级普拉窃检测系统。控制论与信息技术，18 (1)，第 124-138 页。
[4] GUPTA，D。（2016）外在文本抄袭检测技术和工具研究。工程科学与技术评论，9（5），第 150-164 页。
[5] KHAN，I.H.，SIDDIQUI，M.A. 和 MANSOOR，K.（2015）阿拉伯文件中抄袭检测的框架。在：计算机科学与信息技术会议论文集，第 1-9 页。
[6] GOMAA，W.H。和 FAHMY，A.A.（2013）文本相似性方法调查。国际计算机应用杂志，68（13），第 13-18 页。
[7] NIEZGODA，S. 和 WAY，T.P.。（2006）告密者：用于检测抄袭和粘贴抄袭的软件工具。在：第 37 届上证所计算机科学教育技术研讨会论文集，得克萨斯州休斯顿，2006 年 3 月。纽约：计算机协会，第 51-55 页。
[8] KENT，C.K. 和 SALIM，N.（2010）基于网页的跨语言窃检测。于：2010 年 9 月在巴厘岛举行的第二届国际计算智能，建模与仿真国际会议论文集。新泽西州皮

斯卡塔维：电气与电子工程师协会，第 199-204 页。
[9] KHATRI，J. 和 MOHAN，V.（2016）一种实现基于网页的普拉窃检测工具的方法。国际工程与管理研究杂志， 6（3），第 57-60 页。
[10] ADEL，G.M. 和 WANG，Y.（2019）阿拉伯语在线普拉窃检测工具的有效性水平。物联网与云计算，7（1），第 19-24 页。
[11] 谷歌开发人员（2019）自定义搜索格式应用程式介面：简介。 [线上]。可从以下网站获得：https://developers.google.com/custom-search/v1/introduction [访问时间 15/03/19]。
[12] 微软蔚蓝（2019）Bing 网页搜索应用程式介面。[线上]。可从以下网站获得：https://docs.microsoft.com/zh-cn/java/api/overview/azure/cognitiveservices/client/bingwebsearchapi?view=azure-java-stable [已访问 19/04/19]。
[13] 扬德克斯技术（2019）Yandex.XML。[线上]。可从以下网址获得：https://tech.yandex.com/xml/ [访问时间：19/04/19]。
[14] PHP 组（2019）PHP：相似文本-手册。[线上]。可从以下网站获得：https://www.php.net/manual/zh/function.similar-text.php [访问时间：19/05/19]。
[15] MANNING，C.，RAGHAVAN，P. 和 SCHÜTZE，H.（2010）信息检索简介。自然语言工程，16（1），第 100-103 页。
[16] iThenticate。[线上]。可从以下网站获得：http://www.ithenticate.com/ [访问时间：19/06/19]。
[17] Turnitin。[线上]。可从以下网站获得：https：//www.turnitin.com [访问时间：19/06/19]。
[18] ALABBAS，M. 和 RAMSAY，A.（2012）通过结合不同的匕首改进了阿拉伯语的销售点标签。在：ILIADIS，L.，MAGLOGIANNIS，I. 和 PAPADOPOULOS，H.（编）人工智能应用与创新。联合会 2012。《联合会信息和通信技术的进步》，第 1 卷。 381.柏林，海德堡：施普林格，第 107-116 页。
[19] ALABBAS，M. 和 RAMSAY，A.（2014）通过组合多样的依赖性解析器来改进阿拉伯语的解析。在：VETULANI，Z. 和 MARIANI，J.（编辑）《计算机科学和语言学的人类语言技术挑战》中。LTC2011。计算机科学讲义，第 1 卷。 8387. 湛：施普林格，第 43-54 页。
[20] M. ALABBAS，R.S.的 KHUDEYER 和 S. JAF（2016）通过组合多个机器学习分类器，改进了阿拉伯字符的识别。于：2016 年亚洲语言处理国际会议论文集，台南，2016 年 11 月。新泽西州皮斯卡塔维：电气与电子工程师学会，第 262-265 页。
[21] S. JAF，M。ALABBAS 和 R.S. KHUDEYER。（2018）结合用于阿拉伯字符识别任务的机器学习分类器。国际亚洲语言处理杂志，28（1），第 1-12 页。
[22] ALABBAS，M. 和 RAMSAY，A.（2011）对长阿拉伯文句子的依存解析器进行评估。于：2011 年国际语义技术与信息检索会议论文集，布城。新泽西州皮斯卡塔维：电气与电子工程师协会，第 243-248 页。
[23] MIKOLOV，T.，CHEN，K.，CORRADO，G. 和 DEAN，J.（2013）向量空间中单词表示的有效估计。 [线上]。可从以下网站获得：http://arxiv.org/pdf/1301.3781.pdf [19/10/15 访问]。
[24] PENNINGTON，J.，SOCHER，R. 和 MANNING，C.（2014）手套：用于词表示的全局向量。 于：2014 年 10 月，多哈，自然语言处理中的经验方法会议论文集，宾夕法尼亚州斯特劳兹堡：计算语言学协会，第 1532-1543 页。
[25] BOJANOWSKI，P.，GRAVE，E.，JOULIN，A. 和 MIKOLOV，T.（2017）用子词信息丰富词向量。 计算语言学协会学报，5，第 135-146 页。
[26] HUSSEIN，S.A. 和 KAREEM，M.R.（2019）一种使用网址地址的拟议阿拉伯文本和文本图像分类技术。 西南交通大学

学 报 ， 54 （ 5 ） 。 可 从 http://jsju.org/index.php/journal/article/view/411 获得。