ISSN: 0067-2904

An Improved Outlier Detection Model for Detecting Intrinsic Plagiarism

Nasreen J. Kadhim*, Maysaa I Abdulhussain Almulla khalaf
Department of Computer Science, College of Science, University of Baghdad, Baghdad Iraq

Received: 3/3/2022 Accepted: 15/6/2022 Published: 30/12/2022

Abstract

In the task of detecting intrinsic plagiarism, the cases where reference corpus is absent are to be dealt with. This task is entirely based on inconsistencies within a given document. Detection of internal plagiarism has been considered as a classification problem. It can be estimated through taking into consideration self-based information from a given document.

The core contribution of the work proposed in this paper is associated with the document representation. Wherein, the document, also, the disjoint segments generated from it, have been represented as weight vectors demonstrating their main content. Where, for each element in these vectors, its average weight has been considered instead of its frequency.

The proposed work has been evaluated in terms of Precision, Recall, F-measure, Granularity, and Plagdet. It is shown that the attained results are comparable to the ones attained by the best state-of-the-art methods. Where, through applying the proposed method to PAN-PC-09 and PAN-PC-11 for the detection of intrinsic plagiarism, a Recall scores of 0.4503 and 0.4303 have been recorded, even though further improvement for Precision (0.3308 and 0.2806) and Granularity (1.1765 and 1.1111) needs to be made. Concerning f-measure, the proposed approach has recorded 0.3814 and 0.3397. In terms of the total performance of a plagiarism detection approach, Plagdet, the proposed method has recorded 0.3399 and 0.3151.

Keywords: Intrinsic plagiarism detection, document representation, weight vectors, main content vectors.

نموذج كشف محسن لاكتشاف السرقة الأدبية الذاتية

نسرين جواد كاظم* , ميساء أبراهيم عبد الحسين الملا خلف
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

في مهمة الكشف عن السرقة الأدبية الداخلية ، يتوجب التعامل مع الحالات التي لا يتوفر فيها المراجع. هذه المهمة مبنية بالكامل على التناقضات داخل وثيقة معينة. تم اعتبار الكشف عن السرقة الأدبية كمشكلة تصنيف. يمكن تقديرها من خلال الأخذ بالاعتبار المعلومات ضمن وثيقة معينة.

الأسهام الرئيسي لهذا العمل يرتبط بطريقة التمثيل حيث تم تمثيل المستند بالأضافة للمقاطع المنفصلة الناتجة عنه كمتجهات للوزن لتوضح محتواها الرئيسي حيث أن كل عنصر في هذه المتجهات تم تمثيله من خلال معدل وزنه وليس من خلال تكراره.

*Email: nasreen.kadhim@sc.uobaghdad.edu.iq

تم تقييم العمل المقترح من خلال قياس Precision, Granularity, Recall, F-measure, Plagdet.
يتضح أن النتائج التي تم الحصول عليها قابلة للمقارنة مع تلك التي تم الحصول عليها بأفضل الطرق الحديثة.
حيث ، من خلال تطبيق الطريقة المقترحة على PAN-PC-09 و PAN-PC-11 للكشف عن السرقة الأدبية الجوهرية، تم تسجيل Recall 0.4503 و 0.4303 ، على الرغم من وجود الحاجة الى أجراء التحسين الإضافي لل Precision (0.3308 و 0.2806) وGranularity (1.1765 و 1.1111). فيما يتعلق بالقياس F-measure ، فقد سجل النهج المقترح 0.3814 و 0.3397. من حيث الأداء الكلي لنهج كشف الانتحال ، Plagdet ، سجلت الطريقة المقترحة 0.3399 و 0.3151.

1. Introduction

The usage of another's language, writing, or information, once performed with improper acknowledgment to the original source, causes plagiarism to arise [1]. In textual documents, plagiarism occurs in numerous forms: exact copying of the plagiarized text may be made, passages may be modified to a bigger or smaller extent, or translation may even be performed [2]. Billions of web pages became easily accessible to anyone as a result of the speedy development of World Wide Web (WWW), which in turn has provided plenty of possible sources for plagiarism. Accordingly, growing attention has been given to automated plagiarism analysis and detection in both academia and software industry [3]. Several authors have been motivated to work for describing this phenomenon [4, 5].

Researchers have considered two main strategies for detecting plagiarism [6]: Intrinsic plagiarism detection wherein no reference given and its aim is to discover plagiarism through the examination of the input document only, giving a decision whether portions of it are not written by the main author. The other is the strategy wherein a comparison is performed between suspicious documents and a collection of sources for detecting plagiarized sections and the documents that they came from and named as external plagiarism detection.
Formulating the problem of the traditional internal plagiarism detection is as in what follows [7]: The mission is determining whether a given document encompasses plagiarized segments or written by a single author. The detection process should be accomplished without comparing suspicious document against external sources. An essential condition exists in the traditional internal plagiarism setting: At least 70% of the considered text document written by one main author. The common scheme for detecting intrinsic plagiarism in what follows has been nominated by the «one-main-author» condition [8, 9, 1, 10 and 6]: 1) a text document is divided into a set of text segments, 2) a set of segment features is developed and combined to an author style function for the sake of measuring correspondence of an author-style for each text segment, and 3) for detecting plagiarized segments, critical values in the author style function is discovered. A *sliding window* approach was proposed by the authors in [8], wherein a text document is divided into a set of intersecting segments and as the main component of an author style function, character 3-gram frequencies was used. The additional well-thought-out examples of style function are the n-gram classes [4], pronouns, punctuation and part-of-speech tags count [9], and normalized word frequency class [1]. A style function counting an n-gram frequency relative deviation from its typical value was proposed to be constructed.

Outlier detection techniques on text-based data have been used to improve the detection strategies of internal plagiarism by means of deviation parameters concerning the writing style of a given document. For the work proposed in this paper, the text document has been separated into disjoint segments considering the original paragraphs exist. Also, a different representation has been put forward wherein the document has been represented as a weight

vector demonstrating its main content. Next, a relative deviation is computed for an n-gram average weight from its representative value through building a style function.
This paper is organized as follows: First, the works related to the proposed work have been presented in Section 2. In Section 3, a description for the intrinsic plagiarism detection approach proposed in this research paper has been stated. Section 4 presents the performance evaluation results for the proposed system in addition to the comparison results with state of the art methods. Finally, the conclusions have been presented in Section 5 together with future work directions.

2. Related work

Once performing a texts comparison against a reference set of probable sources, the complication of electing the true set of documents for comparing with will arise. Furthermore, this task becomes more complex to accomplish with the opportunities brought to plagiarists through the Internet. For this, analysis of the writing style can be accomplished within the document, and examining inconsistencies can be performed. The main idea is to define a criterion for determining if enough change has occurred to the style to give an indication of plagiarism. The analysis of text style and complexity can be accomplished based on certain parameters such as part-of speech features, structural features, syntactic features, text statistics, and closed-class word sets, as stated by [11].

A method in [8] for intrinsic plagiarism detection. A function of style variation constructed on a suitable dissimilarity measure firstly proposed for author identification and character n-gram profiles are attempted by his approach for quantifying the style change within a document. Initial construction of style profiles was performed by means of a sliding window. The use of character n-grams was proposed by the author for constructing those profiles. The aim for the use of n-grams was to get writer's style information. Then deviations on the profiles were analyzed for determining if a significant enough change occurred for indicating a style of another author [8].

In [12], Kolmogorov Complexity measures were introduced by Seaward & Matwin as a way to extract text's structural information for detecting Intrinsic Plagiarism. To detect style shifts within a single document, they experimented with complexity features based on the Lempel-Ziv compression algorithm, hence revealing probable plagiarized sections [12].

Text representation is considered one of the key building blocks of any application of natural language processing. Through using character n-grams of a text, its representation necessitates to decompose it into all the probable sequences of n consecutive characters. For instance, the word *probable* 3-grams are: *pro, rob, oba, bab, abl, ble*. The set of all the $n-grams$ of a predefined length, n, extracted along with their frequencies from a given text, is referred to as the text's $n-gram$ profile. Methods for intrinsic plagiarism detection; wherein character $n-grams$ are used, were summarized in [13].

For [8] (2009b) method in [15], the suspicious document and its segments were represented using 3-gram profiles. Obtaining the segments was performed by applying a sliding window, of about 1000 characters, in each step, the movement was done by 200 characters. Then, calculation of a style variation function was performed centring on the divergence between the n-gram profile of the entire document and the n-gram profile of each segment. Through the comparison of the standard deviation of the style change function values with a threshold parameter, the suspicious document was predicted whether containing

plagiarized portions or written by one main author. If plagiarism was detected, a portion was marked as plagiarized if its style variation value is higher than a defined threshold [15].

In [14], the authors held the understanding that when long texts have to be dealt with, the representation of documents by all their n-grams is computationally expensive. Thus, a predefined set of 3-grams with high-frequency was employed through their approach for representing the fragments of the suspicious document. The motivation of this idea was by authorship attribution research where the use of n-grams with high-frequency succeeded [15]. The dissimilarity measure of [8] was used by this method for detecting outliers, but computation was performed between each pair of fragments of the suspicious document.
The work of [6] was totally constructed on demonstrating the variation word frequency as a main indication of stylistic difference [6].

In [7] a set of features including the mean of n-grams relational frequency, the most frequent n-grams frequency, and the rarest n-grams frequency, were used for the representation of each sentence. The mean of n-grams relational frequency was a new feature and the computation was done for each n-gram in a sentence. N-gram's relational frequency became higher if it was more specific to a sentence. When doing experiments with different lengths, the authors stated that the optimal lengths of n-grams (1, 3, and 4) were determined. Then, for generating a model for combining features and predicting for each sentence, a score representing its mismatch degree with the main author style, the gradient boosting regression trees were used. Lastly, a plagiarized mark was given to the sentences having a score higher than a certain threshold [16].

3. Proposed intrinsic plagiarism detection approach

For detecting intrinsic plagiarism, some ideas that had been investigated by [6] have been considered in this work. His ideas had led to the intuition in what follows for his algorithm development: "*If some of the words used on the document are author-specific, one can think that those words could be concentrated on the paragraphs (or more general, on the segments) that the mentioned author wrote".* The work proposed in this paper considers a different document representation that focuses on the significance of words considering weight instead of frequency wherein the document, also the generated segments from decomposing it have been represented by their main contents instead of the frequency of words that they involve. Furthermore, the document has been segmented into disjoint segments concerning its paragraphs instead of applying a sliding window to form overlapping segments as in [6] work. The proposed work uses the same approach used in [6] work for comparing the segments generated against the entire document for detecting deviations from the style of the author who wrote the majority, if not all, the text.

In this paper, an approach to detect intrinsic plagiarism has been described. A suspicious document is separated into a sequence of disjoint segments considering its paragraphs. The entire document and the spawned segments are represented through the calculation of weight vectors containing the importance of all the unigram words that they involve after pre-processing the suspicious document. Importance of the words is represented through computing their average weights over the formed segments and the original document. Subsequently, a comparison is made between each segment and the entire document through their main content vectors. Finally, an algorithm for outlier detection is applied for detecting the plagiarized sections.

To detect deviations in the author's writing style given, a document D composed of n paragraphs such that $D = \{p_1, p_2, p_3, \dots, p_n\}$. Firstly, D is pre-processed wherein, numbers are

excluded, all non-alphabetic characters are removed, all characters are lowercased, and then the words unigram are considered without excluding stop words wherein, V =$\{u_1, u_2, u_3, \dots, u_m\}$ will be the resulted m uni-grams from D after applying pre-processing. Next, u_i is weighted using $tf - idf$ weighting scheme [17, 18]. After that, a main content vector $\underline{V} = \{\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_m\}$ is computed considering the main content of D represented by average $tf - idf$ weight for all words resulted from pre-processing step. The j^{th} coordinate, $\underline{v}_j$ of the main content vector $\underline{V}$is computed as in $Eq.1$ in what follows:

$$\underline{v}_j = \frac{\sum_{i=1}^{n} wt_{ij}}{n} \qquad j = 1, 2, 3, \dots, m \tag{1}$$

Where wt_{ij} is the $tf - idf$ weight of word unigram j at sentence$i$.

Then, the entire document is segmented considering paragraphs exist in it for the segmentation process wherein segments seg are created initially where $seg \in S$. Next, a new weight vector $\underline{v}_{seg}$ is computed as in Eq. 1 above for each segment $seg \in S$ that imitates average words $tf - idf$ weight. Wherein for each segmentseg, only the words unigram within it are considered. Afterwards, document self-similarity is tested using the proposed algorithm 1 in what follows:

Algorithm 1 Intrinsic plagiarism detector

Input:

Document $D = \{p_1, p_2, p_3, \dots, p_n\}$

Threshold : δ

Step 0: Pre-process $D = \{p_1, p_2, p_3, \dots, p_n\}$ $\Rightarrow$ V =$\{u_1, u_2, u_3, \dots, u_m\}$

Step 1: Weight each unigram in V using $tf - idf$ weighting scheme

Step 2: Build for $V$the main content vector $\underline{V} = \{\underline{v}_1, \underline{v}_2, \underline{v}_3, \dots, \underline{v}_m\}$ wherein each element in $\underline{v}_j$ is calculated as in Eq. 1

Step 3: Segment D into disjoint segments $seg \in S$ wherein each seg corresponds to a paragraph

Step 4: for each $seg \in S$ do

Step 5: $df_{seg} \leftarrow 0$

Step 6: Build main content vector $\underline{v}_{seg}$ for each segment $seg \in S$ wherein each element in $\underline{v}_{seg}$ is calculated as in Eq. 1 considering words exist in seg

Step 7: for each word unigram $j \in v_{seg}$ do

$$df_{seg} \leftarrow df_{seg} + \frac{|\underline{v}_j - \underline{v}_{segj}|}{|\underline{v}_j + \underline{v}_{segj}|}$$

end for

end for

Step 8: calculate $style \leftarrow \frac{1}{|S|} \sum_{seg \in S} df_{seg}$

Step 9: for each $seg \in S$ do

Step 10: if $df_{seg} < style - \delta$ then

Mark segment seg as an outlier.

end if

end for

Algorithm 1 presents the general document style which is characterized by the average of all differences calculated for each segment s and the complete document. This algorithm takes into account the intuition; a low value will result from the comparison of segment seg against the entire document if certain words are only used on a certain segment because the average

$tf - idf$ weight of those words would be the same in both the full document and in the segment. Finally, segment classification is performed according to its distance with respect to the document's style. Document's main style is represented by the average value resulted from comparing all segments with the entire document. This value is roughly calculated by the difference on the words' average $tf - idf$ weight between vectors v and v_{seg}, $\forall seg \in S$. If the difference is significant, in this case, the value of the style function will be lower than the threshold, and then the segment is classified as suspicious.

4. Performance evaluation and discussion

The proposed approach has been evaluated on two corpora (intrinsic part) that were used in the international competition of plagiarism detection in 2009 and 2011 (PANPC-09 and PAN-PC-11, respectively). The two document collections involve XML explanations that indicate the plagiarized segments positions. Macro-averaged F-measure, Recall, Precision, Granularity, and Plagdet have been used as evaluation measures as they were defined in [19]. The plagiarism case length does not affect the macro-averaged recall and precision. F-Measure is the weighted harmonic mean of recall and precision. Recall and precision are considered equally weighted since there is presently nothing that indicates either of the two is more significant. The power of a detection algorithm, that is, whether detecting a plagiarism case is achieved in one piece or in several pieces, is captured through granularity. To obtain an absolute order, Precision, Recall, and Granularity must be combined to an overall score named, Plagdet, because they allow for a partial ordering among plagiarism detection algorithms.

Table 1 and Table 2 clarify the comparison results of the method proposed in this paper implemented using $PAN - PC - 09$ and $PAN - PC - 11$ respectively to the one in [6] being the winner in PAN 2011 competition and considered one of the best intrinsic plagiarism detection methods.

Table 1: Performance comparison of the proposed model against [6] model in terms of $Precision, Recall, F - measure, Granularity,$ and $Plagdet$ evaluated using $PAN - PC - 09$ corpora.

Evaluation metric	Method in [6]	Proposed model
Precision	0.3897	0.3308
Recall	0.3109	0.4503
F-measure	0.3458	0.3814
Granularity	1.0006	1.1765
Plagdet	0.3457	0.3399

Table 2: Performance comparison of the proposed model against [6] model in terms of $Precision, Recall, F - measure, Granularity,$ and $Plagdet$ evaluated using $PAN - PC - 11$ corpora.

Evaluation metric	Method in [6]	Proposed model
Precision	0.3398	0.2806
Recall	0.3123	0.4303
F-measure	0.3255	0.3397
Granularity	1	1.1111
Plagdet	0.3255	0.3151

The results are based on detection quality, wherein only the information on each document itself is to be considered. Table 1 and 2 reveal that the proposed approach outperforms the work introduced in [6] in terms of Recall, even though further improvement

for Precision and Granularity needs to be made. However, precision satisfied is still closer to [6] results than the best results achieved in PAN-PC-11 competition on plagiarism detection. Concerning f-measure, the proposed approach has achieved results that are comparable to that of [6]. A good combination of recall and precision has been achieved for the proposed method. In terms of the total performance of a plagiarism detection approach, Plagdet, the proposed method has achieved comparable results to that of [6]. Moreover, it is realized that the performance stability of the proposed approach has been satisfied for both corpora.

5. Conclusions

This research paper has introduced an approach for detecting intrinsic plagiarism bases on making a comparison of the writing style on a specific document for determining if the text writing process has been performed through using the writing style of one or more authors. Experimental results achieved by the proposed detector have revealed that there is a positive impact for the used text representation scheme that focuses on significance of words through building main content vector, on improving the detection process.

As a future work, for improving granularity, other segmentation strategies are going to be tried. Besides, other text representation schemes focusing on words significance are to be tried for improving total number of detections. Moreover, as any of the language-dependent features have not been utilized in the proposed model, thus experiments may be performed on other languages. Results satisfied in the task of detecting intrinsic plagiarism reveal that much work is still needed, and new approaches need to be developed for modelling the writing style. Also, because the recall satisfied till now is still low, new approaches are to be investigated for improving it.

References

[1] Z. Eissen, and S. Benno, "Intrinsic plagiarism detection," In *European conference on information retrieval*, 2006, pp. 565-569. Springer, Berlin, Heidelberg.

[2] M. Mohammed, N. Kadhim, and A. Ibrahim, "Improved VSM Based Candidate Retrieval Model for Detecting External Textual Plagiarism," *Iraqi Journal of Science*, pp. 2257-2268, 2019.

[3] A. Maurer, K. Frank, and Z. Bilal, "Plagiarism-A survey," *Journal of Universal Computer Science*. vol. 12, no. 8, pp.1050-1084, 2006.

[4] R. Hunt, "Let's hear it for internet plagiarism." Teaching Learning Bridges vol. 2, no. 3, pp. 2-5, 2003.

[5] C. Park, "In Other (People's) Words: Plagiarism by university students--literature and lessons," *Assessment & Evaluation in Higher Education*, vol. 28, no. 5, pp. 471-488, 2003.

[6] G. Oberreuter, L. Gabriel, R. Gaston, A. Sebastián, and V. Juan, "Approaches for intrinsic and external plagiarism detection," *Proceedings of the PAN*, vol. 4, no. 5, 2011.

[7] M. Kuznetsov, M. Anastasia, K. Rita, and S. Vadim, "Methods for Intrinsic Plagiarism Detection and Author Diarization," In *CLEF*, pp. 912-919. 2016.

[8] E. Stamatatos, "Intrinsic plagiarism detection using character n-gram profiles," *threshold*, vol. 2, no. 1,pp.500, 2009.

[9] M. Zechner, M. Markus, K. Roman, and G. Michael, "External and intrinsic plagiarism detection using vector space models," In *Proc. SEPLN*, vol. 32, pp. 47-55. 2009.

[10] L. Bensalem, R. Paolo, and C. Salim, "Intrinsic plagiarism detection using n-gram classes," In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1459-1464. 2014.

[11] S. Eissen, S. Benno , and K. Marion, "Plagiarism detection without reference collections," In *Advances in data analysis*, pp. 359-366. Springer, Berlin, Heidelberg, 2007.

[12] L. Seaward, and M. Stan,. "Intrinsic plagiarism detection using complexity analysis," In *Proc. SEPLN*, pp. 56-61, 2009.

[13] L. Bensalem, Imene, R. Paolo, and C. Salim, "On the use of character n-grams as the only intrinsic evidence of plagiarism," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 363-396, 2019.

[14] CM. Kestemont, L. Kim, and D. Walter, "Intrinsic plagiarism detection using character trigram distance scores,*" Proceedings of the PAN*, vol. 63, 2011.

[15] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information*, vol. 60, no.3, ,pp. 538–556, 2009b. doi:10.1002/asi.

[16] M. Kuznetsov,M. Anastasia, K. Rita, and S Vadim, "Methods for Intrinsic Plagiarism Detection and Author Diarization," In *CLEF (Working Notes)*, pp. 912-919, 2016.

[17] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 25,no.5, pp, 513–523, 1988.

[18] N. Kadhim, H. Saleh, and B. Attea, "Improving Extractive Multi-Document TextSummarization Through Multi-Objective Optimization," *Iraqi Journal of Science* vol. *59,no.*4B, pp.2135–2149.

[19] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection,",In *Coling*, 2010.