Outlier-Based Approaches for Intrinsic and External Plagiarism Detection

Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez

Web Intelligence Consortium Chile Research Centre
Department of Industrial Engineering
University of Chile
goberreu@ing.uchile.cl, {glhuilli,srios,jvelasqu}@dii.uchile.cl

Abstract. Plagiarism detection, one of the main problems that educational institutions have been dealing with since the massification of Internet, can be considered as a classification problem using both self-based information and text processing algorithms whose computational complexity is intractable without using space search reduction algorithms. First, self-based information algorithms treat plagiarism detection as an outlier detection problem for which the classifier must decide plagiarism using only the text in a given document. Then, external plagiarism detection uses text matching algorithms where it is fundamental to reduce the matching space with text search space reduction techniques, which can be represented as another outlier detection problem. The main contribution of this work is the inclusion of text outlier detection methodologies to enhance both intrinsic and external plagiarism detection. Results shows that our approach is highly competitive with respect to the leading research teams in plagiarism detection.

Keywords: Text Classification, Outlier Detection, Search Space Reduction, External Plagiarism Detection, Intrinsic Plagiarism Detection.

1 Introduction

Plagiarism in academia is rising and multiple authors have worked to describe this phenomena [6,9]. As commented by Hunt in [6], "Internet Plagiarism" is referred sometimes as a cataclysmic consequence of the "Information Technology revolution", as it proves to be a big problem in academia. In [9], plagiarism is analyzed from various perspectives and considered as a problem that is growing bigger over time. To tackle this problem, the most common approach so far is to detect plagiarism using automated algorithms based on rules and string matching algorithms.

Two main strategies for plagiarism detection have been considered by researchers [10]: Intrinsic and external plagiarism detection. Intrinsic plagiarism detection aims at discovering plagiarism by examining only the input document, deciding whether parts of the input document are not from the same author.

A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 11–20, 2011.

External plagiarism detection is the approach where suspicious documents are compared against a set of possible references. From exact document copy, to paraphrasing, different levels of plagiarism techniques can been used in several contexts [16].

The main contribution of this work is the usage of outlier detection techniques on text-based data to enhance two plagiarism detection strategies, one for intrinsic plagiarism detection using deviation parameters with respect of the writing style of a given document, and another one to reduce the search space for external plagiarism detection based on the generation of segments of n-gram for approximated plagiarism decision where unrelated documents are discarded efficiently.

This paper is structured as follows: In Section 2 an overview of intrinsic and external plagiarism detection algorithms is presented. Then, in Section 3 the proposed plagiarism detection methods are introduced. Afterwards, in Section 4, the experimental setup and evaluation performance criteria are described. In Section 5 results are discussed. Finally, in Section 6 the main conclusions are presented.

2 Related Work

According to Schleimer et al. [12], copy prevention and detection methods can be combined to reduce plagiarism. While copy detection methods can only minimize it, prevention methods can fully eliminate it and decrease it. Notwithstanding this fact, prevention methods need the whole society to take part, thus its solution is non trivial. Copy or plagiarism detection methods tackle different levels, from simple manual comparison to complex automatic algorithms [11,10].

2.1 Intrinsic Plagiarism Detection

When comparing texts against a reference set of possible sources, comes the complication of choosing the right set of documents to compare. And now more than ever, with the possibilities that Internet bring to plagiarists, this task becomes more complicated to achieve. For this, the writing style can be analyzed within the document and an examination for incongruities can be done. The complexity and style of each text can be analyzed based on certain parameters such as text statistics, syntactic features, part-of-speech features, closed-class word sets, and structural features [16]. Whose main idea is to define a criterium to determine if the style has changed enough to indicate plagiarism.

Stamatatos [14] presented a new method for intrinsic plagiarism detection. As described by it's author, this approach attempts to quantify the style variation within a document using character n-gram profiles and a style change function based on an appropriate dissimilarity measure originally proposed for author identification. Style profiles are first constructed using a sliding window. For the construction of those profiles the author proposed the use of character n-grams. These n-grams are used for getting information on the writer's style.

The method then analyzes changes on the profiles to determine if a change is significative enough to indicate another's author style.

Other approaches have been proposed, such as presented by Seaward & Matwin [13] introduce Kolmogorov Complexity measures as a way of extracting structural information from texts for Intrinsic Plagiarism Detection. They experiment with complexity features based on the Lempel-Ziv compression algorithm for detecting style shifts within a single document, thus revealing possible plagiarized passages.

2.2 External Plagiarism Detection

In terms of external plagiarism detection algorithms, the use of n-grams have shown to give some flexibility to the detection task, as reworded text fragments could still be detected [8]. Other approaches focus on solving the plagiarism detection problem as a traditional classification problem from the machine learning community [1,4]. Bao et al. in [1], proposed to use a Semantic Sequence Kernel (SSK), and then using it into a traditional Support Vector Machines (SVMs) formulation based on the Structural Risk Minimization (SRM) principle from statistical learning theory [15], where the general objective is finding out the optimal classification hyperplane for the binary classification problem (plagiarized, not plagiarized).

In [7] the authors introduced their model for automatic external plagiarism detection. It consist of two main phases; the first is to build the index of the documents, while in the second the similarities are computed. This approach uses word n-grams, with n ranging from 4 to 6, and takes into account the number of matches of those n-grams between the suspicious documents and the source documents for computing the detections. The algorithm have the authors won the first place at the PAN@2010 competition [10].

2.3 Outlier Identification Approaches and Plagiarism Detection

As described in [5], an outlier is an observation which deviates from other observations as to become suspicious that was generated by a different statistical process. In general terms, outlier detection can be classified in proximity approaches, such as distance-based or density-based, and model-based approaches, such as statistical tests, depth-based, and deviation-based methodologies [5]. In plagiarism detection, deviation-based methodologies have been previously used for intrinsic plagiarism detection [14], and distance-based for probability distribution approaches for external plagiarism [2].

3 Proposed Methods

In this section, the main contribution of our work is described. In the first place, a search space reduction algorithm using outlier detection techniques for external plagiarism detection is presented. Then, an intrinsic plagiarism detection algorithm is proposed based on variance of n-gram content on sliding windows over the whole document.

Let us introduce some concepts. In the following, let $\mathcal{V}$ a vector of words that defines the vocabulary to be used. We will refer to a word w, as a basic unit of discrete data, indexed by $\{1, ..., |\mathcal{V}|\}$. A document d is a sequence of S words ($|d| = S$) defined by $\mathbf{w} = (w^1, ..., w^S)$, where w^s represents the s^{th} word in the message. Finally, a corpus is defined by a collection of $\mathcal{D}$ documents denoted by $\mathcal{C} = (\mathbf{w}_1, ..., \mathbf{w}_{|\mathcal{D}|})$.

3.1 Distance-Based Outlier Detection for External Plagiarism Detection

For a corpus $\mathcal{C} = \{\mathcal{D}_{\text{source}}, \mathcal{D}_{\text{suspicious}}\}$, the idea is to find all plagiarized documents in the suspicious partition, using as search space the source partition. In general terms, the algorithm first reduces the search space by using an approximated search of segments of n-grams, and then within selected pairs of documents, using an exhaustive search algorithm, finds the offset and its length.

First, for each document d_i, after stop-words removal, a set t_i of n-grams with the structure $(w_i, w_{i+1} \ldots, w_{i+n}), \forall i \geq 1, n \leq S$ must be created. Then, to compute the difference between these document vector's, groups κ_i of k n-grams are created. The basic idea is to test the closeness between a pair of documents using a distance-based outlier detection approach, which firstly indicates whether chunks of words between n-gram representations of the document have at least θ_κ exact matches, and then, checks if all n-grams of both documents have at least θ_t matches, as shown in Algorithm 1.

Algorithm 1. Approximate comparison between two documents

Require: $\kappa_i, \kappa_j, \mathbf{t}_i, \mathbf{t}_j, \theta_\kappa, \theta_t$
1: **if** sMatch$(\kappa_i, \kappa_j, s \geq \theta_\kappa)$ **then**
2: **if** sMatch$(\mathbf{t}_i, \mathbf{t}_j, s \geq \theta_t)$ **then**
3: return true
4: **end if**
5: **else**
6: return false
7: **end if**

As presented in Algorithm 1, once documents d_i and d_j are processed in n-grams and segments of k n-grams, t_i, t_j and κ_i, κ_j respectively, a set of conditions are evaluated in order to set the relation that document d_i has with document d_j, that is, if they are somehow related (algorithm 1 returns true), or if it is not worthy to keep finding further relationships (Algorithm 1 returns false). In this sense, this is an approximated finding procedure that considers both n-grams and their k segments to decide if there is enough information to classify as plagiarism or not, and using the distance function sMatch, which checks for thresholds θ_κ and θ_t.

Condition $\text{sMatch}(\kappa_i, \kappa_j, s \geq \theta_\kappa)$ states that at least θ_κ n-grams must match in between segments κ_i and κ_j. If this is hold, the next condition $\text{sMatch}(\mathbf{t}_i, \mathbf{t}_j, s \geq \theta_2)$ is associated to find whether at least θ_t n-grams matches between $\mathbf{t}_i$ and $\mathbf{t}_j$.

After reducing search space, it is possible now to go into a further algorithm for finding the needed offset and its length. More details on offset and length finding algorithm, please refer to Oberreuter et al. in [8], which is intentionally omitted by authors due to lack of space.

3.2 Intrinsic Plagiarism Detection

Using uni-grams and without removing stop-words, a frequency-based algorithm to test self-similarity of document is proposed. First, a frequency vector $\mathbf{v}$ is built for all words on a given document. Then, the complete document is clusterized creating groups $\mathcal{C}$. As a first approach, these groups or segments $c \in \mathcal{C}$ are created using a sliding window of length m over the complete document. Afterwards, for each segment $c \in \mathcal{C}$, a new frequency vector v_c is computed, which is used in further steps to compare whether a segment is deviated with respect to the footprint of the complete document. This is performed by using the Algorithm 2.

Algorithm 2. Intrinsic plagiarism evaluation

Require: $\mathcal{C}, \mathbf{v}, m, \delta$
for $c \in \mathcal{C}$ **do**
  $d_c \leftarrow 0$
  build v_c using term frequencies on segment c
  for word $w \in v_c$ **do**
    $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v}) - freq(w,v_c)|}{|freq(w,\mathbf{v}) + freq(w,v_c)|}$
  end for
end for
style $\leftarrow \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} d_c$
for $c \in \mathcal{C}$ **do**
  if $d_c < style - \delta$ **then**
    Mark segment c as outlier and potential plagiarized passage.
  end if
end for

As presented in Algorithm 2, the general footprint or style of the document is represented by the average of all differences computed for each segment and the complete document. Finally, all segments are classified according to its distance with respect to the document's style. As an example, in Figure 1, a graphical representation of this evaluation is presented.

In this case, the average value (represented by yellow line), is compared against the style function, which is roughly computed by the difference on the frequency of words between vectors $\mathbf{v}$ and $v_c, \forall c \in \mathcal{C}$. In any case that the style function is lower than the average value minus δ, the segment is classified as suspicious.

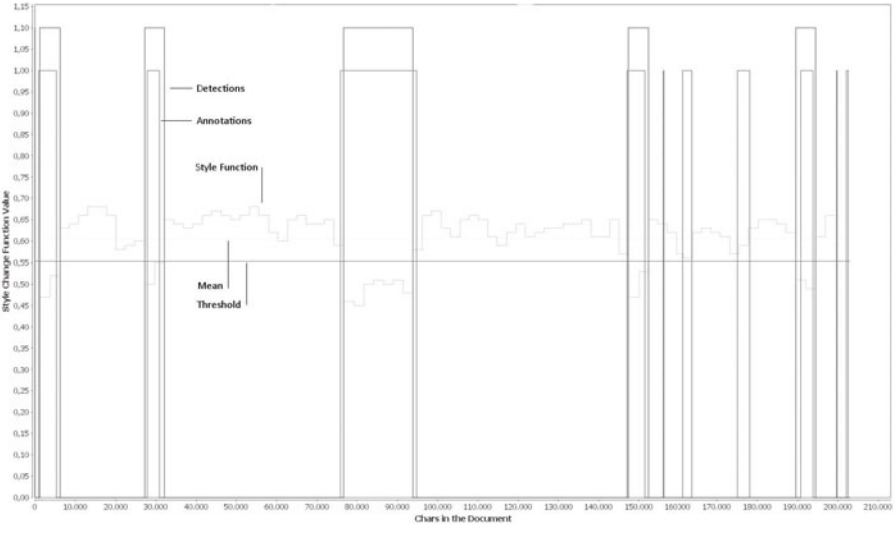

Fig. 1. Intrinsic plagiarism detection example

In this example, real plagiarized annotations are presented by red areas, and all classified passages are presented by blue areas, indicating that for all 10 cases of real plagiarized annotations, the proposed method achieved to classify 5 of them, without mistakes.

4 Experiments

In this section, the experimental setup and the evaluation criteria is presented. For external plagiarism detection evaluation, the PAN@2010 plagiarism detection corpus [10] was used, which considers a set of 11,148 source documents and another set of 15,925 suspicious documents, with 68,558 plagiarism cases. For the evaluation of intrinsic plagiarism detection, the PAN@2009 plagiarism detection corpus [11] was used, which considers a set of 6,183 suspicious documents.

4.1 Evaluation Criteria

As described in [10,11], let S be the set of all plagiarized passages, R the set of all detections made by a given plagiarism detection algorithm, and S_R a subset of S whose detections are presented in R. Let the function $|\cdot| : \mathcal{P}(\mathcal{V}) \rightarrow \mathbb{R}^1$ which states the number of chars in a given string generated from vocabulary $\mathcal{V}$. The evaluation metrics proposed for plagiarism detection are described as follows:

$$\text{Recall} = \frac{1}{|S|} \sum_{i=1}^{|S|} \left(\frac{\#\text{ detected chars of } s_i}{|s_i|} \right) \tag{1}$$

$$\text{Precision} = \frac{1}{|R|} \sum_{i=1}^{|R|} \left(\frac{\#\text{ plagiarized chars of } r_i}{|r_i|} \right) \tag{2}$$

$$\text{Granularity} = \frac{1}{|S_R|} \sum_{i=1}^{|S_R|} (\#\text{ of detections of } s_i \in R) \tag{3}$$

$$\text{Overall score} = \frac{\text{F-measure}}{log_2(1 + \text{granularity})} \tag{4}$$

where F-measure is the harmonic mean of precision and recall. The Granularity was introduced in order to quantify the number of detections of a given plagiarized passage. If the model detects the passage more than once, the score gets penalized.

4.2 Intrinsic Plagiarism Detection Experimental Setup

For intrinsic plagiarism detection, evaluation was conducted by using the PAN@2009 intrinsic competition, for a detailed description of benchmark algorithms, please refer to Potthast et al. [11].

4.3 External Plagiarism Detection Experimental Setup

In this case, the evaluation setup is characterized by all performance metrics described in Section 4.1, and considering as benchmark all contestants of the PAN@2010 plagiarism detection competition[1]. For more information on external plagiarism detection benchmark algorithms, please refer to Potthast et al. [10].

5 Results and Discussions

5.1 External Plagiarism Detection

As shown in Table 1, the best results for the retrieval task were achieved by Kasprzak and Brandejs approach [7]. The overall score was 0.80, and their method achieved good results at the three metrics: precision, recall and granularity. The next top results show similar characteristics, being well balanced in the three metrics. Our proposed model took fifth place, with an overall score of 0.61, precision of 0.85 and recall of 0.48. The granularity of the top performers were all close to 1.

5.2 Intrinsic Plagiarism Detection

The results for the intrinsic task are shown in Table 2. The results are based on the quality of the detection, which only considers the information on each document itself. The second best can be considered as the baseline, as it classified almost every segment as plagiarized [11]. This lead to a reduced precision, thus

[1] `http://pan.webis.de/` [last accessed 01-03-2010].

Table 1. Results for ranking, overall score, F-measure, precision, recall, granularity, and name of lead developer [10]

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead developer
1	0.80	0.80	0.94	0.69	1.00	Kasprzak et al.
2	0.71	0.74	0.91	0.63	1.07	Zou et al.
3	0.69	0.77	0.84	0.71	1.15	Muhr et al.
4	0.62	0.63	0.91	0.48	1.02	Grozea et al.
5	0.61	0.61	0.85	0.48	1.01	**Proposed Model**
6	0.59	0.59	0.85	0.45	1.00	Torrejón et al.
7	0.52	0.53	0.73	0.41	1.00	Pereira et al.
8	0.51	0.52	0.78	0.39	1.02	Palkovskii et al.
9	0.44	0.45	0.96	0.29	1.01	Sobha et al.
10	0.26	0.39	0.51	0.32	1.87	Gottron et al.
11	0.22	0.38	0.93	0.24	2.23	Micol et al.
12	0.21	0.23	0.18	0.30	1.07	Costa-jussá et al.
13	0.21	0.24	0.40	0.17	1.21	Nawab et al.
14	0.20	0.22	0.50	0.14	1.15	Gupta et al.
15	0.14	0.40	0.91	0.26	6.78	Vania et al.
16	0.06	0.09	0.13	0.07	2.24	Suárez et al.
17	0.02	0.09	0.35	0.05	17.31	Alzahrani et al.
18	0.00	0.00	0.60	0.00	8.68	Iftene et al.

Table 2. Results for rank, overall score, F-measure, precision, recall, and granularity for each algorithm presented in section 4

Rank	Overall Score	F-Measure	Precision	Recall	Granularity	Lead Developer
1	0.2462	0.3086	0.2321	0.4607	1.3839	Stamatatos (2009)
2	0.1955	0.1956	0.1091	**0.9437**	1.0007	Hagbi and Koppel (2009)
3	0.1766	0.2286	0.1968	0.2724	1.4524	Muhr et al. (2009)
4	0.1219	0.1750	0.1036	0.5630	1.7049	Seaward and Matwin (2009)
	0.3457	**0.3458**	**0.3897**	0.3109	**1.0006**	**Proposed Model**

obtaining an overall score of 0.1955. The winner was Stamatatos approach ([14]), with a recall of 0.4607, precision of 0.2321 and granularity of 1.3839. His method achieved a good combination of precision and recall, and a not top performer granularity. Our proposed method gets an overall score of 0.3457, greater that any other approach, with a positive difference of 0.0995 with the winner's approach. Our model gets the best result at F-measure, precision and granularity.

6 Conclusions and Future Work

In this work we have introduced two new models for outliers classification applied to plagiarism in digital documents. In the intrinsic plagiarism detection task, our model uses information only from the given document, and select those segments from the text that deviate significantly from the general style. The algorithm

achieves remarkable results, being the best at precision and overall score, using as a benchmark other approaches from PAN@2009 intrinsic plagiarism competition. Also, the algorithm does not utilize language-dependent features such as stopwords, and is simple and straightforward. The model's effectiveness is to be studied in languages other than English, for which as future work, it would be interesting to study other applications, e.g. extracting different topics in a given document, or for author identification.

For the external plagiarism detection task, the proposed model remains competitive against other approaches, obtaining the fifth place in the PAN@2010 external plagiarism detection competition. The task of classifying whether a document is plagiarized or not, and comparing it against a set of possible sources, remains to be a compelling task, as it would be interesting to include Web search for the retrieval of additional source candidates.

Acknowledgment. Authors would like to thank continuous support of "Instituto Sistemas Complejos de Ingeniería" (ICM: P-05-004- F, CONICYT: FBO16; `www.isci.cl`); and FONDEF project (DO8I-1015) entitled, DOCODE: Document Copy Detection (`www.docode.cl`).

References

1. Bao, J.-P., Shen, J.-Y., Liu, X.-D., Liu, H.-Y., Zhang, X.-D.: Semantic sequence kin: A method of document copy detection. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 529–538. Springer, Heidelberg (2004)
2. Barrón-Cedeño, A., Rosso, P., Benedí, J.-M.: Reducing the plagiarism detection search space on the basis of the kullback-leibler distance. In: Gelbukh, A. (ed.) CICLing 2009. LNCS, vol. 5449, pp. 523–534. Springer, Heidelberg (2009)
3. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, Padua, Italy (September 22-23, 2010)
4. Chow, T.W.S., Rahman, M.K.M.: Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. Trans. Neur. Netw. 20(9), 1385–1402 (2009)
5. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
6. Hunt, R.: Let's hear it for internet plagiarism. Teaching Learning Bridges 2(3), 2–5 (2003)
7. Kasprzak, J., Brandejs, M.: Improving the reliability of the plagiarism detection system - lab report for pan at clef 2010. In: Braschler, et al. (eds.) [3] (2010)
8. Oberreuter, G., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Fastdocode: Finding approximated segments of n-grams for document copy detection - lab report for pan at clef 2010. In: Braschler, et al. (eds.) [3] (2010)
9. Park, C.: In other (people's) words: plagiarism by university students – literature and lessons. Assessment and Evaluation in Higher Education (5), 471–488 (2003)
10. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd international competition on plagiarism detection. In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy (September 22-23, 2010)

11. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st international competition on plagiarism detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 1–9. CEUR-WS.org (September 2009)
12. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: local algorithms for document fingerprinting. In: SIGMOD 2003: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 76–85. ACM, New York (2003)
13. Seaward, L., Matwin, S.: Intrinsic plagiarism detection using complexity analysis. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 56–61. CEUR-WS.org (September 2009)
14. Stamatatos, E.: Intrinsic plagiarism detection using character n-gram profiles. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009), pp. 38–46. CEUR-WS.org (September 2009)
15. Vapnik, V.N.: The Nature of Statistical Learning Theory (Information Science and Statistics). Springer, Heidelberg (1999)
16. Eissen, S.M.z., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Decker, R., Lenz, H.-J. (eds.) GfKl. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 359–366. Springer, Heidelberg (2006)