Determining Window Size from Plagiarism Corpus for Stylometric Features

Šimon Suchomel(✉) and Michal Brandejs

Faculty of Informatics, Masaryk University, Brno, Czech Republic
{suchomel,brandejs}@fi.muni.cz

Abstract. The sliding window concept is a common method for computing a profile of a document with unknown structure. This paper outlines an experiment with stylometric word-based feature in order to determine an optimal size of the sliding window. It was conducted for a vocabulary richness method called 'average word frequency class' using the PAN 2015 source retrieval training corpus for plagiarism detection. The paper shows the pros and cons of the stop words removal for the sliding window document profiling and discusses the utilization of the selected feature for intrinsic plagiarism detection. The experiment resulted in the recommendation of setting the sliding windows to around 100 words in length for computing the text profile using the average word frequency class stylometric feature.

1 Problem Statement

In automated plagiarism detection, the task for the computer system is to highlight potentially plagiarized passages from input suspicious documents and ideally, to match the highlighted passage with the original document from a set of all documents. This style of detection is referred as external plagiarism and one needs a reference corpus of source documents in order to match the suspicious document with the original document [1].

If no reference corpus is available, the task shifts into the detection of anomalies inside the text itself. This is called intrinsic plagiarism detection [9], which in this case can be viewed as a one-class classification problem [2]. The text portion is either classified as written by the same author or classified as not written by the same author and therefore, suspicious. In this concept the task is closely related to the author identification problem [8].

It is generally believed that each writer has a specific writing style and if a text contains copied passages, they would probably deviate from the writing style of the putative author. The various methods used for this task try to detect changes in the writing style of the text being analyzed and are called stylometric features [6]. Such features are based on statistical likelihood estimation, therefore, the more statistical data they compute with, the more precision they can achieve. This means that generally the longer the analyzed text is the better the feature distinguishes between text characteristics. However, in plagiarism detection there is often a need for detection of relatively short passages, which is a hard problem to achieve without a reference corpus for text comparison.

J. Mothe et al. (Eds.): CLEF 2015, LNCS 9283, pp. 293–299, 2015.
DOI: 10.1007/978-3-319-24027-5_31

In general document analysis there is usually no prior information about the position and the length of which different passage should be detected, which comprises the most challenging part of the plagiarized passage detection. This problem is usually addressed by the moving or shifting window computation concept.

The most widespread method is to compute the feature for the whole document as a reference value. Thereafter, the feature is computed for the portion of the document defined by the window size and compared with the reference value. If the current window-size feature differs significantly from the reference value, that part is said to be suspicious according to the feature description. However, this method has several difficulties. The moving windows should ideally, precisely overlap with the plagiarized passage in order to produce unbiased characteristics of that passage. Any misalignment in this manner produces more biased results towards the surrounding text.

The right setting of the moving window size and position is important for the stylometric feature to produce accurate results. While moving the window through the document, the adjacent windows can be overlapping in order to minimize the probability of a misplaced window. Small shifting intervals ensure that the beginning of some windows will be close enough to the beginning of the plagiarized passage. The maximum deviation from the optimal placement is half of the window shifting interval.

Moreover, the size of the window is more important and more difficult to set. In order to compute some text features, a sufficient amount of statistical data is needed, therefore a bigger window size might seem advantageous. On the other hand, if the plagiarized text is shorter than the window size, the calculated feature from that part would be distorted by the redundant text contained in that window .Various window sizes can be used, it may depend on many variables such as stylometric features used, input data type, or purpose of analysis. Examples can be less than 200 words [3,12], 250 words [10], 500 words [2], 1000 characters [7].

Window feature comparison against the reference value from the whole document assumes that the reference value describes the whole document correctly, and relatively small textual anomalies inside the document could be detected. However, if the document contains lot of plagiarism the reference value is too affected of it and the feature would then describe a mashup created from plagiarism and from the original text of the alleged writer. In such cases the values obtained from moving window should be compared only to each other, while the character of the document is determined by changes among those values.

For our experiment we have chosen a lexical word-based feature called 'Average Word Frequency Class' (AWFC), which is a statistical vocabulary richness method [3]. This method is supposed to be accurate for short text passages and is also said to be consistent with the length of passages, which makes it suitable for plagiarism detection. We wanted to extrapolate an optimal window size for the AWFC. The experiment was conducted on training corpus for PAN competition on a plagiarism detection [4] for source retrieval subtask. Another contribution

is to analyse whether this feature is suitable for intrinsic plagiarism detection within the PAN source retrieval corpus.

The PAN 2015 source retrieval corpus is a set of intentionally plagiarized documents by semiprofessional writers [5]. The task of source retrieval is part of the automated plagiarism detection process [11], which is conducted before actual textual similarities computation within the reference corpus of all known documents. If a plagiarized passage can be detected in this stage, it could be used as a template for search engine queries for original document retrieval.

2 Methodology

The PAN source retrieval corpus contained 98 plagiarized documents written manually on a random topic and each document followed only one theme. The corpus were based on texts retrieved from ClueWeb[1] corpus by querying a search engine for topic related documents. The size of the plaintexts were 30 KB on average and each document contained around five thousand words on average. The plagiarism in the corpus was wide spread, which results in understanding this corpus as a simulation of highly plagiarized seminar papers or similar types of documents.

The plagiarism cases were annotated with the assigned *id* of the case and also with some metadata, such as the URL of the original document. Each case, according to its *id*, referred to one source, therefore, the assumption is that texts from one source should hold a common textual feature.

From each document in the corpus all passages under a given *id* were extracted and concatenated. Resulting texts from all plagiarism cases in each document formed a base for calculations of a feature result.

The AWFC feature is defined as follows [3]: Let C be a reference text corpus, and let $f(w)$ be the frequency of a word w in that corpus. The class of each word w in the suspicious document is defined as:

$$c(w) = \lfloor log_2 \frac{f(w^*)}{f(w)} \rfloor, \quad w \in C \ , \tag{1}$$

where w^* denotes the most frequent word in C. Finally, the averaged word frequency class for a text passage (chunk) u is calculated as an averaged value of classes $c(w)$ of all words $w \in u$:

$$\mathrm{AWFC}(u) = \frac{\sum_{i=1}^{|u|} c(w_i)}{|u|} \ . \tag{2}$$

For all based texts the referencing AWFC were also calculated. Each text was subsequently divided into smaller chunks, simulating the length of the resulting text window. All smaller chunks were of the same size and not overlapping over

[1] http://lemurproject.org/clueweb09.php

the based text in order not to average the feature among the chunks. A resulting length for each plagiarism case was calculated as follows: Divide based text into the chunks $u \in U$ of length n in words. Find the minimal windows size n for which chunks u_i and u_j have the same AWFC value for all i and j:

$$\forall i, j : \mathrm{AWFC}(u_i) = \mathrm{AWFC}(u_j), \quad |U| > 1 \ . \tag{3}$$

The division process was considered successful, if all the chunks AWFC values were equal to the referencing AWFC, so the feature held for the whole passage and for all the windows of size n within the passage. The experiment was carried out for both the texts with removed stop words and for the unchanged texts.

3 Analysis and Results

Table 1 shows plagiarism cases and their portion of success and failure. Only cases for which the extracted text was at least 20 words long were considered. Unsuccessful cases were those for which no n complying with (3) was found. Cases labelled inconsistent didn't comply with the reference AWFC value, despite their successful division and meeting the (3) requirements.

Figure 1 depicts resulting chunk sizes of successful cases. The x axis shows $|U|$, which is the number of chunks into which the text was divided. The resulting chunk lengths for stop words clean texts depicted in the left plot of Fig. 1 were lower, which results in the fact that AWFC converges faster for vocabulary richer texts. However, the stop words removal significantly reduces the size of an original text passage. In terms of word count, for the PAN corpus, it was a reduction of 69% from the original size.

The final recommended window size was calculated as a weighted arithmetic mean from chunk sizes of successful cases. The higher weight was assigned to

Table 1. Plagiarism cases.

plagiarism cases	unchanged text	without stop words
in total	1263	1101
successful	75.3%	77.3%
unsuccessful	13.9%	13.5%
inconsistent	10.8%	9.2%

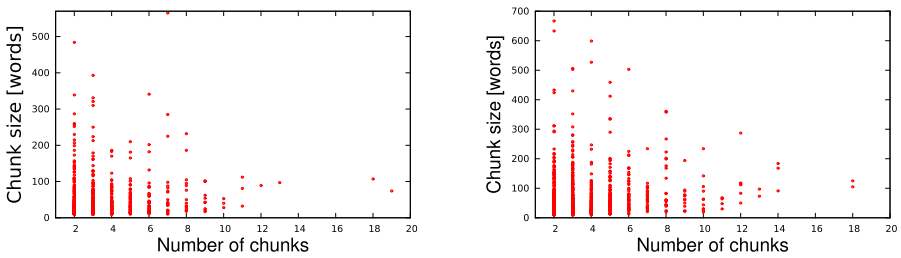

Fig. 1. Sizes of chunks, left with removed stop words.

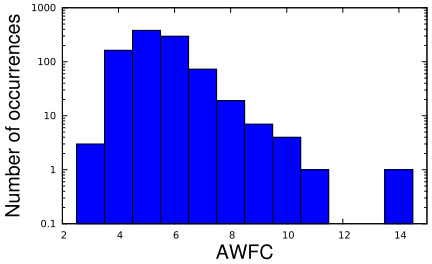

Fig. 2. Occurrences of classes.

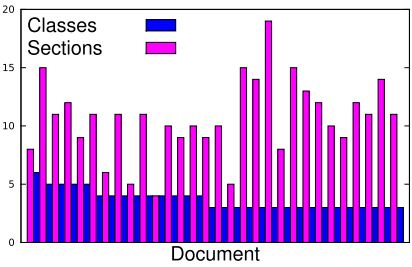

Fig. 3. Number of different plagiarism sections vs. number of different classes.

sizes which stem from higher chunk count of divided text. For example, one of the most statistical data, which Fig. 1 shows in the right plot, contains the two cases, which based texts were successfully divided into 18 chunks of length 105 and 125 words, while complying with (3). Let X be the sorted sequence in descending order of defined chunk sizes ($|U_j|$). The average size $\bar{n}$ of all chunks of all successful cases were calculated as:

$$\bar{n} = \frac{\sum_{i=1}^{n} n_i w_i^2}{\sum_{i=1}^{n} w_i^2}, \quad w \in (0, 1\rangle, \quad w = (1 - \frac{X.\text{index}(|U_j|)}{|X|}) \tag{4}$$

For the original text, the average size was **101.67**, for the stop words clean text it was **62.28** words, which makes a window size decrease of 39%.

In terms of average word frequency class, the most frequent of successful classifications for unchanged text was in class 5, with 40% of all occurrences, and for stop words clean text in class 7, occupying 30% of all classifications. Figure 2 shows class distribution of the unchanged texts, please note that the scale of the y axis is logarithmic, thus showing a single occurrence of classes 11 and 14. Figure 3 shows only 30 selected documents from the input corpus with the highest diversity of occurred classes. The number of different classes is compared with the number of different plagiarism cases in each document. Due to the fact that AWFC has a relatively sparse classification domain, it hardly distinguishes among all plagiarism cases in largely plagiarized documents.

4 Conclusion

This paper presented an experiment with a stylometric statistical vocabulary richness method called 'Average Word Frequency Class' (AWFC) conducted on PAN source retrieval training corpus for plagiarism detection, with both the stop words removed and not removed texts. The benefit of the corpus is that the documents were written manually and not automatically generated, thus creating quality testing environment.

The purpose of the experiment was to determine the size of a text passage, a window, a chunk into which it is profitable to divide the input text for computing the characteristic profile of the text in order to detect style anomalies, which may indicate plagiarism. The resulting recommendation is to apply the sliding windows of length around 100 words, on unchanged text. If stop words are removed, one needs chunks nearly twice as long[2] than the original document for the method to produce comparable results.

However, the AWFC seems not to be suitable for detecting intrinsic plagiarism in the PAN source retrieval corpus. In the corpus, the plagiarism cases are usually distributed across the whole document and sometimes form passages shorter than 100 words. The number of plagiarism cases outnumbers the number of different classes into which a text is classified. On the other hand, if a class change between two neighbouring plagiarized passages is detected the intrinsic plagiarism detection is successful, and so there is no need for the classification method to have a different class for each plagiarism case inside one document. The main purpose of the AWFC is to detect a change of writing style in an otherwise consistent text, for example, to distinguish a brilliant passage that has been copied, in otherwise average seminar work. The performance of the method on the PAN corpus is a matter of future work.

References

1. Kasprzak, J., Brandejs, M., Křipač, M.: Finding Plagiarism by Evaluating Document Similarities. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 24–28. CEUR Workshop Proceedings, August 2009
2. Koppel, M., Schler, J.: Authorship Verification as a One-class Classification Problem. In: Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, July 4–8 (2004)
3. Meyer zu Eissen, S., Stein, B., Kulig, M.: Plagiarism Detection Without Reference Collections. In: Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, pp. 359–366 (2006)
4. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, pp. 845–876 (2014)
5. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: ACL (1), pp. 1212–1221. The Association for Computer Linguistics (2013)
6. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology **60**(3), 538–556 (2009)

[2] By stop words removal, the original text's word count is reduced to 31%, but the resulting windows size of stop words clean experiment were reduced to 61% of the unchanged text window size.

7. Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, pp. 38–46 (2009)
8. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the Author Identification Task at PAN 2014. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, pp. 877–897 (2014)
9. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. Language Resources and Evaluation **45**(1), 63–82 (2011)
10. Stein, B., Meyer zu Eissen, S.: Intrinsic Plagiarism Analysis with Meta Learning. In: Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, Amsterdam, Netherlands, July 27 (2007)
11. Suchomel, Š., Brandejs, M.: Approaches for Candidate Document Retrieval. In: 2014 5th International Conference on Information and Communication Systems (ICICS), pp. 1–6. IEEE, Irbid (2014)
12. Suchomel, Š., Kasprzak, J., Brandejs, M.: Three Way Search Engine Queries with Multi-feature Document Comparison for Plagiarism Detection. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy (2012)