Improved implementation for finding text similarities in large collections of data

Notebook for PAN at CLEF 2011

Ján Grman and Rudolf Ravas

SVOP Ltd., Bratislava, Slovak Republic
{grman,ravas}@svop.sk

Abstract. In this article we describe a new algorithm method for the detection of plagiarism. The method removes numerous limitations of our older method, which has been used as part of a complex information system for the detection of plagiarism. The method has been tested using multiple corpora mainly in Slovak language. With the PAN-09 and PAN-10 corpora it was of great advantage that we could compare our results with the results of other methods. The very good initial results gave us motivation to implement multiple algorithm and parameter improvements.

1 Introduction

The issue of plagiarism is very serious in all areas, however, it is the most vivid in the field of education. In 2008 the Ministry of Education initiated the establishment of a Central Register of Thesis and Dissertations of the Slovak Republic, which serves as a central repository for all academic institutions. In 2010 a subsystem for comparison of documents and detection of plagiarism was added. As a producer and supplier of a library information system, we had long experience in the field of document and metadata collection.

Our first experiments with plagiarism detection solutions began in the spring of 2009. By the end of the year, our results were sufficient to allow the creation of a commercial system. We did not make it to the PAN-10 competition due to our busy work schedule related to the introduction of a system for the complex evaluation of thesis and dissertations of all 33 universities in the Slovak Republic.

The core of both methods is basically language independent. Detection quality, however, depends to a large extent on text pre-processing which is language dependent. We have been working with documents in the following languages: Slovak, Czech, Ukrainian, Hungarian and English.

For the PAN competition we also had to take into consideration Spanish, German and Greek. Because of this motivation, we succeeded in implementing a significant portion of the English WordNet. To translate texts from other languages into English we have used Google API. We've been dividing texts into smaller units and interpolating the translated word offsets with the original text offsets.

2 Overview of related work

Computer (machine) implementation of an anti-plagiarism system is very specific, because many excellent ideas and methods need to be implemented using the means offered by the computers (existing hardware and software equipment). The functioning of an actual anti-plagiarism system, as well as the competition in plagiarism detection, requires that a huge amount of data be processed within a "reasonable" time. The huge amount of data implies high demands on computer and memory subsystem performance. Our reflections about existing solutions will be based on methods and concepts which lead to algorithms (programs) which demonstrably and within a relatively short time produce very good results in the detection of plagiarised text sections. Contributions of workshop participants describe procedures and ideas, the efficiency of which is comparable with the score achieved within the competitions at the **PAN WorkShop** [6]**.**

Having analysed the methods successful in individual **PAN** competitions, it is possible to divide the basic problems into several groups, namely:

a. Editing of suspicious and source texts, which currently requires the comparison of multi-lingual texts, which when machine processed require language detection, machine translation, as precise a determination of word position in the original and translated text as possible [4], use of lexical semantics (synonyms, antonyms and similar), use of stop-words and unification of word representation

b.Transformation of text for representation, which allows the detection of matches within a certain part of text. Detection of plagiarised text passages based on short sub-sequences is only possible on the basis of the detection of matching parts of suitable representations of suspicious and source text. A method, which is currently very popular and backed by analyses, is the N-gram method [5], which can be defined by the number of subtext characters [2], or the number of subtext words [3], [7]. Normally, overlapping N-grams are generated. Methods usually differ in the value of N. The representation using N-grams can also be applied to transformed texts [1].

c.Detection of matching and similar passages in suspicious and source files. The passage (similarity) detection is usually based on heuristics which define the minimum number of consecutive positive representation matches or the minimum frequency of matches within an passage (window) (according to b.) [2] through [4]. Suitable (optimum) values are usually obtained and verified by experiments.

d.Extraction of relevant passages in suspicious and source files. This stage is usually the most complicated because it not only requires division and exclusion, but also the merging of passages from a phrase (c.). This is either done by heuristic methods [3], or methods of segmentation (extraction) which can be visualized in a 2D plane. The visualizations of non-plagiarized, non- obfuscated or obfuscated suspicious passages and source passages show certain typical patterns [2], [1], [10].

e.Post-processing. Some methods in stage (d.) produce overlapping significant passages in the suspicious text for a certain document pair [2], or "false" passages which can be caused, for example, by attempts to hide similarities in suspicious documents with the aim to increase precision or decrease the granularity of detection results.

f.Reducing analysis time. Comparing all suspicious and all source documents between each other is a task of quadratic computational complexity. That is why, before comparing documents, many of the more demanding methods focus on how to detect plagiarised extracts in PAN within a "reasonable time" by reducing the number of source texts to be compared with a particular suspicious document.

There are basically two possible approaches. The first is to parallelise the process of finding suspicious passage**s** using a suitable representation [3] in the form of hash tables. The second is to reduce the number of files to be compared by quantifying the degree of file similarity (usually by means of statistical methods), whereby a set of N most "similar" documents is selected (10 to 50 source files for each suspicious file).

3 Method principle

In principle, the anti-plagiarism system we developed can be divided into three main parts, namely: pre-processing of input data (in this case plain-text pre-processing), detection of passage pairs (plagiarism candidates) and post-processing (removal of overlapping passages, merging of passages, and exclusion of uncertain passage pairs)

Pre-processing

In real life the first stage of document pre-processing is its conversion to plain text. In the case of the PAN corpus this phase can be ignored. We continue working with the text on the level of individual words.

The unit we are interested in is the word and its variations (synonyms, basic forms, abbreviations and similar). Our objective during this stage is to transform the text to a form with the following properties: reduces the amount of data which needs to be processed and allows for efficient comparison of words while taking into account all their versions.

The result is a mapping function which allows streaming text processing, saving text into a more efficient, reduced, and morphology invariant data structure. The steps are: text translation into English (if needed), word extraction (chars, offset and length), and word normalization (stemming, synonym normalization). Original text consisting of words is transformed into binary file of word invariants (codes).

Suspicious passage detection

Our objective was to create a method for detecting similar or matching passages in a suspicious reference text so that the detection is invariant against a change of word order, against the occurrence of changed words, against omissions or additions of words in the passage in a suspicious document, whereby no passage length limits will be set (neither minimum nor maximum length). We assume that passage lengths don't have to be the same.

The method is based on quantification of the degree of concordance between tested passages. The quantification is based on a quick calculation (measurement) of the number of matching words in a pair of passages. The degree of concordance or similarity is defined as the number of elements in an intersection of sets of words from passages in a suspicious and reference text.

$$N_{MW}(I_S, I_R) = | I_S \cap I_R | \quad (1)$$

where N_{MW} is the number of matching words, I_S and I_R are the passages of the suspicious and reference text. The detector selects the area in which the value of N_{MW} exceeds the threshold N_{MWT} .Choice of threshold value N_{MWT} depends mainly on the nature of the compared files which determine the required minimum length of the detected match as expressed by the number of matching words. A pair of passages may be represented by area. For all pairs of representations of suspicious and references documents, which were divided into non-overlapping passages (subintervals) with constant number of words were calculated number of matching words and were thresholded so that it can detect at least 15 words consistently. In the first stage, if the detected areas are adjacent, then they are merged into a single area.

After that, the areas are divided into disjunct areas (pair of passages) so that the resulting passages have the following property. Let's mark the sub-passages I_{Si} and I_{Rj} of passages I_S and I_R , which either start or end in a word belonging to the set (1). If the ratios

$$q_{Si} = \frac{N_{MW}(I_{Si}, I_R)}{N_{MW}(I_{Si}, I_{Si})} \geq q_{\min} \wedge q_{Rj} = \frac{N_{MW}(I_S, I_{Rj})}{N_{MW}(I_{Rj}, I_{Rj})} \geq q_{\min} \quad (2)$$

exceed the selected threshold q_{min}, then the pair I_{Si} , I_{Rj} becomes plagiarism candidate passages for the validity of the assumption (3)

$$N_{MW}(I_{Si}, I_{Rj}) \geq N_{MWT1} \quad (3)$$

where N_{MWT1} is the minimum matching words of the detected passage. We used $q_{min}=0.5$ and $N_{MWT1}=15$.

Post-processing

In our case, two tasks were solved: removal of overlapping passages in suspicious document, if source text was the same and increasing of global score by reducing granularity and by increasing precision using methods described in next section.

Analysis of anti-plagiarism system properties

The system was tested on the PAN-10 corpus on files created from plain texts using synonyms acquired from WordNet. Testing was done using transformed data with and without stop-words. Table 1 shows the results of suspicious passage detection for data without stop-words. Row one shows the score for results without post-processing. The rows below show the score after post-processing, whereby the final results were thresholded to three monitored quantities - the T_1 threshold to t_1 share of word number (1) in all words of passages in the suspicious and reference text, the T_2 threshold to t_2 share of word number (1) in all words of passages in the suspicious and reference text, whereby the words were expressed by the number of characters, and T_3 was the threshold to t_3 minimum length of passages expressed by the number of characters. Conditions for the refusal of detected passages are specified in (4).

$$t_1 = \frac{2 * N_{MW}(I_S, I_R)}{| I_S | + | I_R |}.100 < T_1 \quad t_2 = \frac{2 * \sum_{w_i \in (I_S \cap I_R)} (| w_i | + 1)}{| I_S | + | I_R |}.100 < T_2 \quad | I_S | \leq T_3 \vee | I_R | \leq T_3 \quad (4)$$

where |x| means the length of word or passage expressed in the number of characters. The best score was achieved with the following threshold settings: T_1=70, T_2=60 a T_3=200. Row one shows the score for results without post-processing (marked **).

Table 1. Plagiarism detection score in PAN-10 (with synonyms and without stop-words) for different threshold settings for parameters T_1, T_2 and T_3.

PlagDet	Recall	Precision	Granularity	T_1	T_2	T_3
0.433957	0.737183	0.312248	1.015155	**		
0.811796	0.733454	0.910356	1.001009	50	50	150
0.812908	0.733206	0.913456	1.000951	60	50	150
0.82334	0.730341	0.944667	1.000761	70	50	150
0.823852	0.729678	0.947132	1.000762	70	60	150
0.824488	0.726819	0.953666	1.000746	70	60	200

The plagiarism detection results in the PAN-11 corpus can be described using two statements: satisfaction with the achieved rank and dissatisfaction with the achieved score. The options of our system are described by the results in table 2. With known correct results it is easy to set suitable system parameters. In our case we have used the post-processing settings which produced the best results for PAN-10 (Table 1).

It is apparent that the PAN-11 corpus was prepared with more precision. The number of incorrect detections with a relatively high percentage of similarity was significantly lower, which we could have expected considering the trend between PAN-09 and PAN-10, but we did not dare to apply this assumption to PAN-11 by decreasing the parameter values for post-processing (Table 2).

Table 2. Plagiarism detection score in PAN-11 (using synonyms without stop-words) for different threshold settings for parameters T_1, T_2 and T_3.

PlagDet	Recall	Precision	Granularity	T_1	T_2	T_3	Cases
0.5569	0.396916	0.938023	1.002249	70	60	200	22108
0.615389	0.473128	0.892744	1.006975	50	50	150	28781

4 Conclusions

Each plagiarism detection method faces one basic problem - a huge amount of data. That means only methods that are capable of processing a certain amount of data within a reasonable time limit are usable. Even though the performance of computers increases day by day, the biggest impact is the efficiency of the method used.

In our case we only needed one mainstream server to run the complex plagiarism detection system and collect data for the whole of Slovakia (two years of data collection). The PAN-11 was equally processed using a single server and even several times within the given short period of time. The main advantages of the new method

are better opportunity of detecting paraphrased text, extended support for different word forms significantly improved detection reliability for texts translated into foreign languages (translation through individual paragraphs, offset alignment of paragraphs – original and translated).

Of course, there are some issues that all creators of complex systems face. The basic one is the definition of plagiarism. How much identical and/or similar text can already be considered plagiarism. Should the computer decide, or should it just be a tool that helps decide?

There are also issues with the automatic recognition of citation marks or citation links in general. A specific question is the use of laws and standards in texts. These questions, however, become relevant only in the last stage of processing, which goes beyond the scope of this competition.

Our systems are currently used to compare thesis and dissertations with Internet, yet this is just one of the possible applications. There are plans to extend the system to process documents published by scientists, documents related to projects funded from the state budget or European projects, and documents from published monographs and university textbooks. In some cases it's not so much about looking for matches as indexing a document and thereby protecting it and the copyright of the author.

We would like to thank the competition organizers and the authors of the test corpus for the excellent opportunity to obtain a relatively objective and independent view of the detection capabilities of our solution. A great deal of development and work can be seen in the corpora PAN-09 through PAN-11. Some examples are not very realistic, we could even say marginal, yet this allows the testing of how robust the algorithms really are.

Bibliography

[1] Basile C., Benedetto D., Caglioti E., Cristadoro G. and Esposti M. D.: "A plagiarism detection procedure in three steps: Selection, Matches and "Squares"," in Proc. SEPLN'09, Donostia, Spain, 2009, pp. 19-23.

[2] Grozea C., Gehl C. and Popescu M.: "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," in Proc. SEPLN'09, Donostia, Spain, 2009, pp. 10-18.

[3] Kasprzak J., Brandejs M. and Kripac M., "Finding Plagiarism by Evaluating Document Similarities," in Proc. SEPLN'09, Donostia, Spain, 2009, pp. 24-28.

[4] Kasprzak J. and Brandejs M.: Improving the Reliability of the Plagiarism Detection System - Lab Report for PAN at CLEF 2010. Proceedings of the 2nd Competition on Plagiarism Detection PAN-2010.

[5] Lyon C., Barrett R. and Malcolm J.: Plagiarism Is Easy, But Also Easy To Detect. Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 2006, pp. 57-65.

[6] Potthast, M., Stein, B., Eiselt, A., Cedeño, A.B., Rosso, P.: Overview of the 2nd international competition on plagiarism detection. In: Proceedings of the CLEF'10 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. (2010)

[7] Zou Du, Long Wei-jiang and Ling Zhang: A Cluster-Based Plagiarism Detection Method - Lab Report for PAN at CLEF 2010. Proceedings of the 2nd Competition on Plagiarism Detection PAN-2010.