Applying Specific Clusterization and Fingerprint Density Distribution with Genetic Algorithm Overall Tuning in External Plagiarism Detection

Notebook for PAN at CLEF 2012

Yurii Palkovskii, Alexei Belov

Zhytomyr State University, MARS p.e., Plagiarism Detector Accumulator Project
palkovskiy@yandex.ru

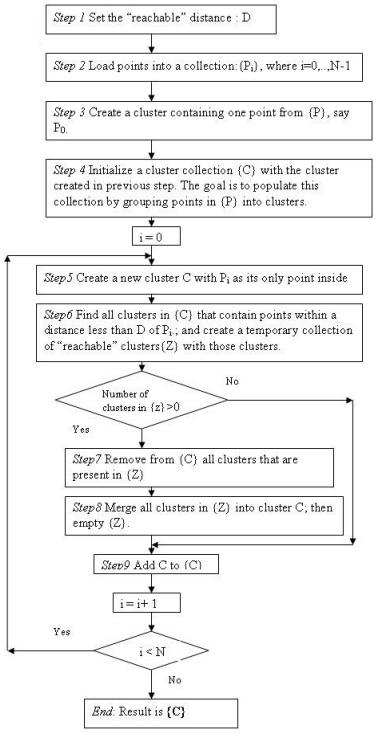
Abstract. One of the biggest challenges encountered at PAN'11 External Plagiarism Detection was the need for different clusterization methods for different types of plagiarism within the corpus. The existence of sparse sections of highly obfuscated, low obfuscated and translated plagiarism sections alongside with verbatim plagiarism parts, made single pass clusterization inefficient as it produced negative effects in one of the above cases. At PAN'11 we used a single pass fixed length clusterization algorithm with a fixed value defining the maximum distance for cluster formation. The main issue with the fixed cauterization value is that large numbers (1600-1800) perform best for high obfuscation, medium (900) for translated and low (40) for verbatim sections. We decided to develop the system that will be able to either dynamically adjust the clusterization distance depending on the type of detected sections or try out multi-pass clusterization with different distance value with the exclusion of already detected clusters and heuristic post processing. For each detected cluster in several clusterization runs we measured Diagonal Density Distribution (DDD) and Mean Average Diagonal Fingerprint Distance (MADFD). These two values reflect the relative distribution of detected equal fingerprints within the cluster diagonal and allows to effectively tell which type of plagiarism is actually there. One more important role that these values play is the negation of cluster merging if the resulting DDD is less than any of the two clusters merged. This was particularly effective preventing accidental fingerprints merging the resulting clusters. Additionally we discovered that the total number of parameters that affect the system performance is already large and decided to apply the genetic algorithm in order to tackle the best possible meta values instead of picking them by hand. In PAN'12 prototype application we employed a dot plot visualization with both detected clusters and master clusters overlay that allowed us to efficiently control the training process and to measure the overall progress for each separate document pair.

1 Introduction

In response to the challenges that have been brought forward by the PAN 2012 competition conditions we decided to focus our research on the detailed document comparison task as our previous experience at CLEF 2011 proved the existence of more effective plagiarism detection methods. One of the main issues we faced during the construction of the PAN 2011 prototype application for detailed document compare stage was the difficulty selecting the most efficient clusterization algorithm. After carefully investigating the possible options used by our competitive colleagues in PAN 2010 and PAN 2012, we come to the conclusion that using custom multi-layered clustering algorithm at clusterization stage may bring better efficiency than using generic clusterization engines such as WEKA, due to the exact nature of plagiarism distribution within the plagiarized passages.

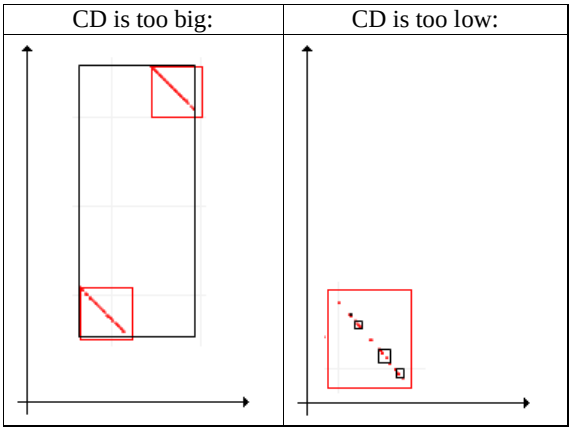
2 Methods

As a basis for the algorithm we took generic Euclidian distance clusterization implementation by Emilio Arp:



We included a number of "filtering layers" to affect the 8-th step that is responsible for cluster merging. During the analysis of the visualized results we discovered that the exact distribution of the shared fingerprints within any detected cluster is located within the cluster diagonal. Most problematic cases that has already been mentioned are the "Over clustering" and "Under clustering" effects produced by different Clusterization Distance for different types of Plagiarism.

"Over clustering" and "Under clustering:"

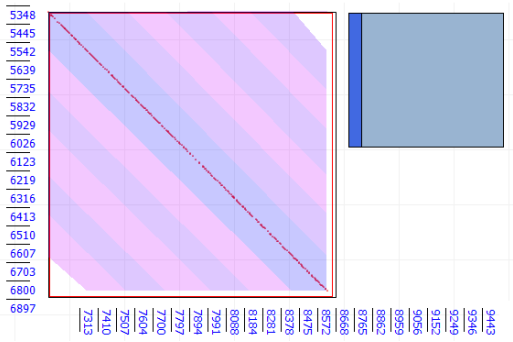


Thus one of the hypothesis to use was the ability to build density diagonal distribution histogram and use it is as an indicator of plagiarism type encountered and the cluster merging efficiency validator. So we ended up with two values that best describe these relations Diagonal Density Distribution and Diagonal Minimal Density Distribution Percent.

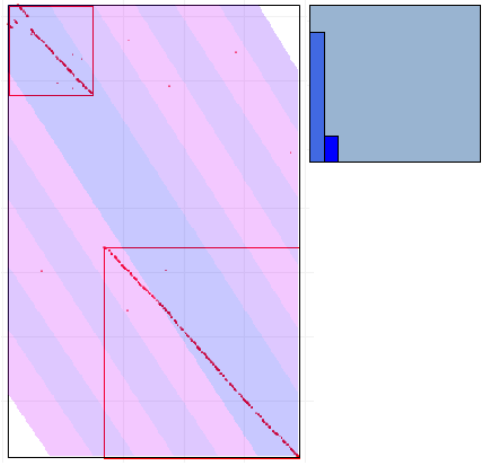
During each cluster formation its Diagonal Minimal Density Distribution Percent value is built automatically. During the cluster merging stage the above mentioned measure of the clusters that are going to be merged is compared to the resulting cluster Diagonal Minimal Density Distribution Percent. Is the resulting value is lower than any of the two merged clusters then this merge is not performed. This particular mechanism was used to fight "over clustering" and it proved to be most efficient. One more benefit of such approach is that it allowed 2 stages of clusterization process - first with a relatively small Clustering Distance of 40 targeting the cases of verbatim plagiarism and another clusterization pass with the Clustering Distance of 1500 that was targeting the detection of low and high obfuscated cases of plagiarism.

One more specific layer of cluster merging used is the relation of the newly detected cluster width to height named Cluster Dimensions Maximum Allowed Skew. This filtering was aimed at removing accidental clusters that affected the total clusterization in a negative way via "over clustering".

Diagonal Density Distribution at 99% in **suspicious-document01712-source-document03867.xml PAN 2012** document pair:



Diagonal Density Distribution - Cluster negatively affecting the cluster formation in **suspicious-document01801-source-document04208.xml** document pair:



3 Evaluation

Used as a basis of the compare mechanism we decided to more efficiently tackle the meta parameters for the developed system and instead manual adjustment we tried to run a genetic algorithm over these parameters in order to adjust the best possible values. We ran the multi-staged evaluation process over a limited pre-selected group of files visualizing the "dot-plot" like graphs, then trying to figure out why the investigated case does not produce the desired result and namely, why the clusterization algorithm failed to achieved any better performance.

Genetic Algorithm Genome Structure and Fittest Values:

Gene Name:	Data type:	Eff. Range:	Final Value:
DoStemming	boolean	0-1	1
RemovePuvctuatuion	boolean	0-1	0
SortFingerprintBeforeHashing	boolean	0-1	0
1StageClusteringDistance	integer	40-3000	49
2StageClusteringDistance	integer	40-3000	1851
2StageClusterMinLength	integer	100-max	6
1StageClusterMinimalFpsCount	integer	1-max	6
2StageClusterMinimalFpsCount	integer	1-max	26
2StageClusterMinimalLengthChars	integer	1-max	190
ClusterDimentionsMaximumSkew	percent	0-100	40
ExcludeMinimalAverageDensity	percent	0-100	3
MinimalDensityDistributionPercent	percent	0-100	8
FingerprintLength	integer	1-7	3
FingerprintStep	integer	1-10	1

We were not able to exhaustively run the complete corpus genetic search for the most optimal values due to the extreme runtime overhead. Instead we used "effective range" by educated guess and a separate sub-corpus for each individual type of plagiarism. The idea behind was twofold - to get the generation run data, visualize it thus tuning the algorithms appropriately and to get most effective p-det over the mixed corpus that represented the low-scaled tuning corpus of PAN 2012.

When evaluating our final performance at PAN 2012 in comparison to the previously achieved results it must be noted, that this year competition has lots of new conditions and completely new environment, thus the comparison is not straightforward - CDC and CR stages do not influence each other and the resulting number of false positives thus is much lower. This particular detail makes such comparison not feasible. Still we consider the achieved result reflects our efforts directed onto the project development.

PAN 2012 Performance:

PlagDet	Precision	Recall	Granularity	Runtime
0.5382163	0.5748453	0.5230450	1.0246376	4.5162973

Things to be noted - our best subcorpus result is p-det 0,74. Our current research focus is trying to determine why the achieved p-det is much lower than the one achieved during our tests. Secondly - as our later tests showed, a bug in the software implementation PAN 2012 prototype application failed to map the exact offsets of the translated plagiarism thus accumulating delta that negatively affects the results of translated plagiarism sections. Thirdly - the training corpus for PAN2012 is different from the test corpus in its structure, plagiarism distribution and some other characteristics we are not aware at the moment.

4 Conclusions

In this paper we introduced a new approach to the clusterization algorithm guided specifically to tackle the "over clustering" and "under clustering" negative effects that are usually produced by generic type clusterization. We investigated the benefits of different filtering layers within the cluster merging stage of the main clusterization algorithm and the usage of two staged clusterization in order to effectively handle the different types of plagiarism - namely highly obfuscated and translated ones. We applied genetic algorithm to effectively tune in the meta parameters that affect the total efficiency of plagiarism detection system.

References

1. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010)
2. Clough, P.: Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service (2003)
3. Grozea, C., Gehl, C., Popescu, M.: ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In: 3rd PAN WORKSHOP. UNCOVERING PLAGIARISM, AUTHORSHIP AND SOCIAL SOFTWARE MISUSE. p. 10 (2009)
4. Grozea, C., Popescu, M.: The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Extended Technical Report.
http://brainsignals.de/encsimTR.pdf (Aug 2011)
5. Grozea, C., Popescu, M.: Encoplot - Performance in the Second International Plagiarism Detection Challenge - Lab Report for PAN at CLEF 2010 . In: Braschler et al. [1]
6. Grozea, C., Popescu, M.: Who's the Thief? Automatic Detection of the Direction of Plagiarism. In: Gelbukh, A.F. (ed.) CICLing. Lecture Notes in Computer Science, vol. 6008, pp. 700–710. Springer (2010)
7. Planas, J., Badia, R.M., Ayguadé, E., Labarta, J.: Hierarchical task based programming with StarSs. International Journal of High Performance Computing 23(3), 284–299 (August 2009)
8. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 997–1005. Association for Computational Linguistics (2010)
9. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler et al. [1]