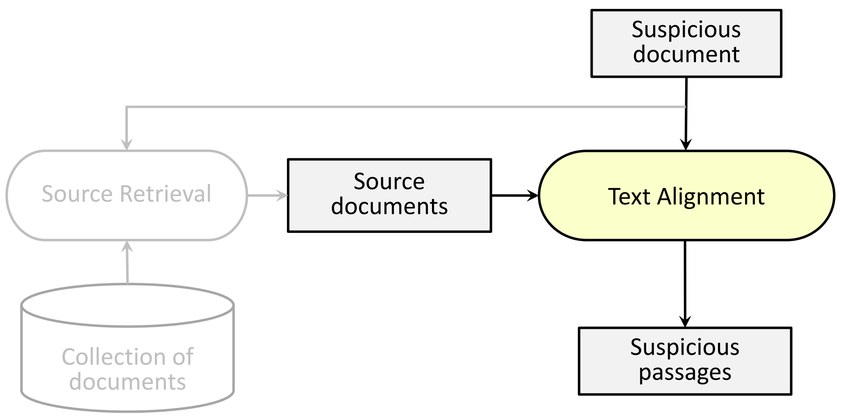
Dynamically Adjustable Approach through Obfuscation Type Recognition

Miguel A. Sanchez-Perez, Alexander Gelbukh and Grigori Sidorov

Centro de Investigación en Computación
Instituto Politécnico Nacional

TASK



METHODOLOGY

Preprocessing

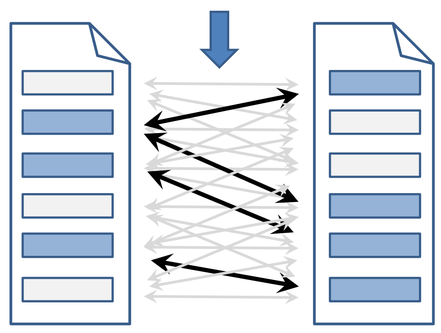
Sentence splitting, tokenizing, removal of tokens that do not start from a letter or digit, reducing to lowercase, stemming, joining small sentences (1-2 words) with the next one.

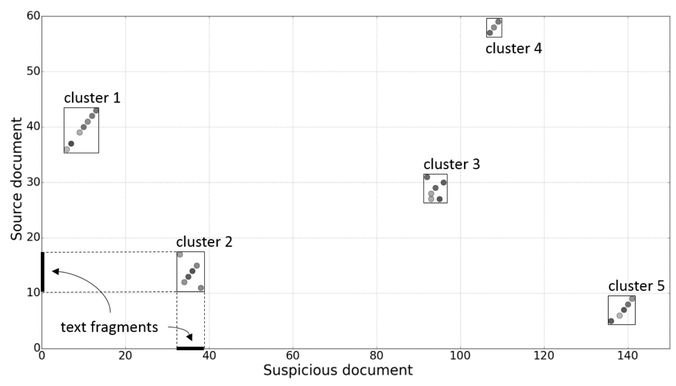
Seeding

Vector representation of sentences:
TF-IDF, where sentences are "documents", thus called TF-ISF: inverse sentence freq.
"Documents": union of sentences of both docs

Vector similarity:
Cosine similarity $\geq th1$
AND Dice similarity $\geq th2$

Seeds: pairs of **similar** sentences

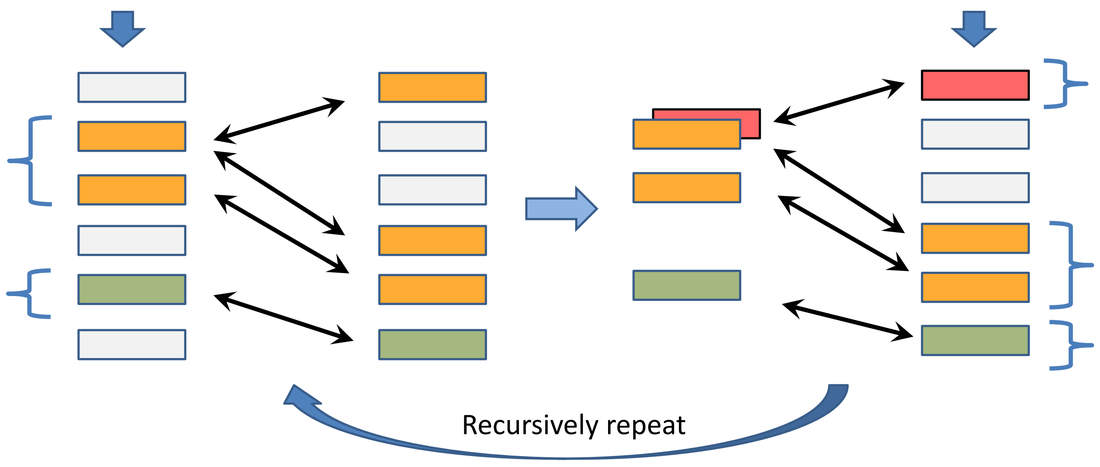




Extension

Step 1: Clustering
Cluster by distance between sentences $\leq$ ***maxGap***

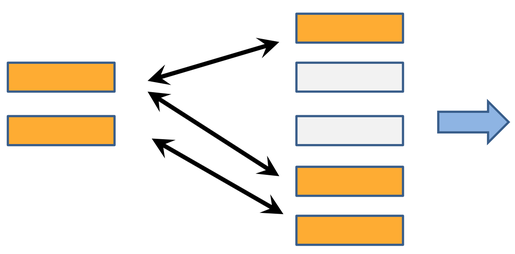
Cluster by **left side**



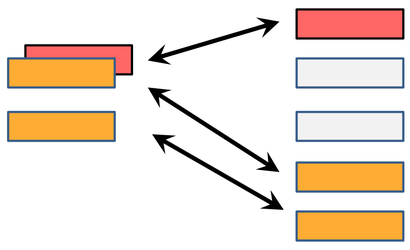
Cluster by **right side**

Step 2: Validation
for each cluster
if cos(left, right) $\leq$ *th3* **then**
cluster **again** with ***maxGap - 1***

Example:
After group left and right with ***maxGap = 2***



Grouping with ***maxGap = 1***



Resulting **clusters** are considered **plagiarism cases**

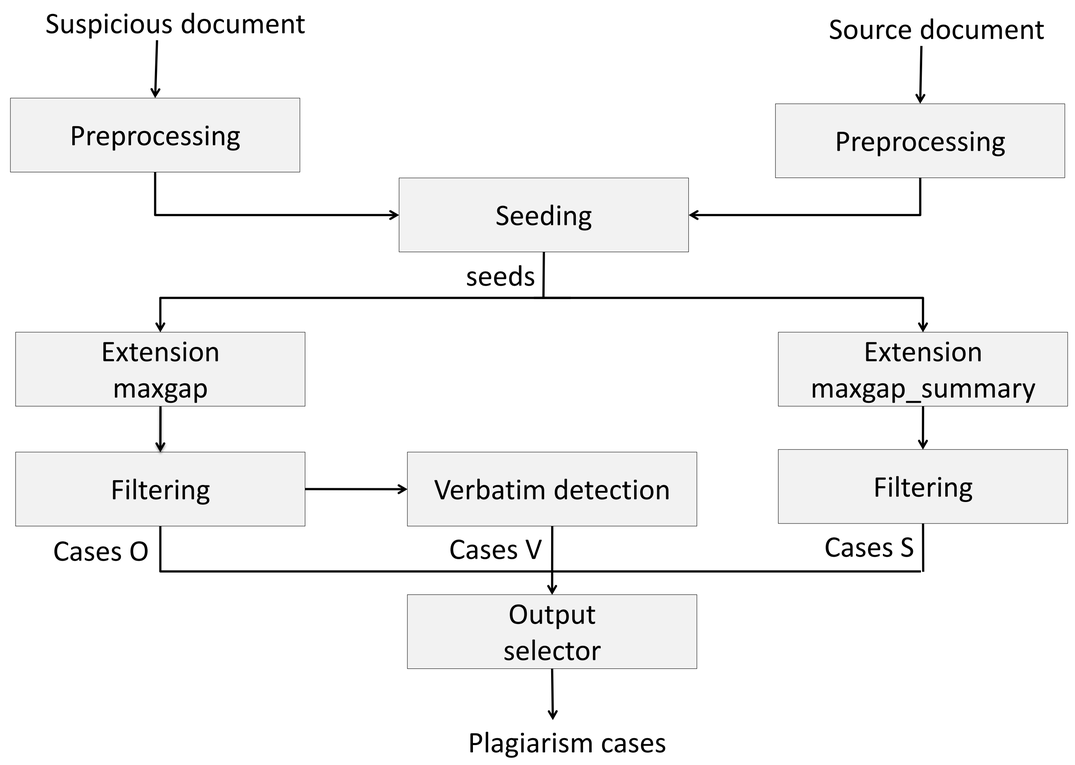
Filtering

1. While exists a case P ("pivot") that overlaps with some other case
   (a) Denote $\Psi(P)$ be the set of cases $Q \neq P$ overlapping with P
   (b) For each $Q \in \Psi(P)$, compute the quality $q_Q(P)$ and $q_P(Q)$
   (c) Find the maximum value among all obtained $q_y(x)$
   (d) Discard all cases in $\Psi(P) \cup \{P\}$ except the found x

where

$$q_y(x) = sim_{F^x_{src}}(O) + \left(1 - sim_{F^x_{src}}(O)\right) \times sim_{F^x_{src}}(N)$$

ADAPTIVE BEHAVIOR



RESULTS

Our approach compared to the PAN 2014 Official results

Team	PlagDet	Recall	Precision	Granularity	Runtime
Sanchez-Perez15	**0.9010**	**0.8957**	**0.9125**	**1.0046**	–
Sanchez-Perez14	**0.8781**	**0.8790**	**0.8816**	**1.0034**	**00:25:35**
Oberreuter	0.8693	0.8577	0.8859	1.0036	00:05:31
Palkovskii	0.8680	0.8263	0.9222	1.0058	01:10:04
Glinos	0.8593	0.7933	0.9625	1.0169	00:23:13
Shrestha	0.8440	0.8378	0.8590	1.0070	69:51:15
R. Torrejón	0.8295	0.7690	0.9042	1.0027	00:00:42
Gross	0.8264	0.7662	0.9327	1.0251	00:03:00
Kong	0.8216	0.8074	0.8400	1.0030	00:05:26
Abnar	0.6722	0.6116	0.7733	1.0224	01:27:00
Alvi	0.6595	0.5506	0.9337	1.0711	00:04:57
Baseline	0.4219	0.3422	0.9293	1.2747	00:30:30
Gillam	0.2830	0.1684	0.8863	1.0000	00:00:55

Sanchez-Perez15 approach evaluated in the PAN 2015 test corpus

Corpus	PlagDet	Recall	Precision	Granularity	Runtime
pan15_alvi	0.6554	0.8641	0.5279	1.0000	00:00:36
pan15_cheema	0.4780	0.8938	0.3262	1.0000	00:00:48
pan15_khoshnava	0.9355	0.9078	0.9649	1.0000	00:03:11
pan15_mohtaj	0.7537	0.7985	0.7136	1.0000	00:32:45

CONCLUSIONS

We improved the system proposed in PAN 2014 thanks to the following additions:
1. Verbatim detector module based on the longest common substrings algorithm.
2. Recursive clustering.
3. Parameters optimization

CONTACT

Miguel A. Sanchez-Perez
masp1988@hotmail.com
Centro de Investigación en Computación,
Instituto Politécnico Nacional, http://www.cic.ipn.mx