**Dynamically Adjustable Approach through Obfuscation Type Recognition**

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

---

**TASK**

- **Source Retrieval**
  - Source documents
  - Collection of documents
- **Text Alignment**
  - Suspicious document
- **Seeding**
  - Seeds: pairs of similar sentences
- **Extension**
  - Step 1: Clustering
    - Cluster by distance between sentences ≤ maxGap
    - Cluster by left side
    - Cluster by right side
  - Step 2: Validation
    - if cos(left, right) ≤ th1 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

**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 ≥ th1
AND Dice similarity ≥ th2

**ADAPTIVE BEHAVIOR**

**RESULTS**

<table>
<thead>
<tr>
<th>Team</th>
<th>PlugDet</th>
<th>Recall</th>
<th>Precision</th>
<th>Granularity</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanchez-Perez15</td>
<td>0.9010</td>
<td>0.8957</td>
<td>0.9125</td>
<td>1.0046</td>
<td>00:25:35</td>
</tr>
<tr>
<td>Sanchez-Perez14</td>
<td>0.8781</td>
<td>0.8790</td>
<td>0.8816</td>
<td>1.0034</td>
<td>00:25:35</td>
</tr>
<tr>
<td>Oberreuter</td>
<td>0.8093</td>
<td>0.8577</td>
<td>0.8859</td>
<td>1.0036</td>
<td>00:05:31</td>
</tr>
<tr>
<td>Paklovskii</td>
<td>0.8680</td>
<td>0.8263</td>
<td>0.9222</td>
<td>1.0058</td>
<td>01:10:04</td>
</tr>
<tr>
<td>Glinos</td>
<td>0.8593</td>
<td>0.7933</td>
<td>0.9625</td>
<td>1.0169</td>
<td>00:23:13</td>
</tr>
<tr>
<td>Shrestha</td>
<td>0.8410</td>
<td>0.8378</td>
<td>0.8590</td>
<td>1.0070</td>
<td>09:51:15</td>
</tr>
<tr>
<td>R. Torrejon</td>
<td>0.8295</td>
<td>0.7690</td>
<td>0.9042</td>
<td>1.0027</td>
<td>00:00:42</td>
</tr>
<tr>
<td>Gross</td>
<td>0.8264</td>
<td>0.7662</td>
<td>0.9327</td>
<td>1.0251</td>
<td>00:03:00</td>
</tr>
<tr>
<td>Kong</td>
<td>0.8216</td>
<td>0.8074</td>
<td>0.8400</td>
<td>1.0030</td>
<td>00:05:26</td>
</tr>
<tr>
<td>Abnar</td>
<td>0.6722</td>
<td>0.6116</td>
<td>0.7783</td>
<td>1.0224</td>
<td>01:27:00</td>
</tr>
<tr>
<td>Alvi</td>
<td>0.6595</td>
<td>0.5506</td>
<td>0.9337</td>
<td>1.0711</td>
<td>00:04:57</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.4219</td>
<td>0.3422</td>
<td>0.9293</td>
<td>1.2747</td>
<td>00:30:30</td>
</tr>
<tr>
<td>Gillam</td>
<td>0.2830</td>
<td>0.1684</td>
<td>0.8863</td>
<td>1.0000</td>
<td>00:00:55</td>
</tr>
</tbody>
</table>

**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