Approaches for Intrinsic and External Plagiarism Detection

Notebook for PAN at CLEF 2011

Gabriel Oberreuter Gallardo
goberreu@ing.uchile.cl
University of Chile - September 2011

Group members : Gastón L'Huillier, Sebastián Ríos and Juan D. Velásquez

Who are we?

- Group of students, professionals and professors from University of Chile
- We work on studying plagiarism in academia
- www.docode.cl

Who we are

Docode Engine

DOCODE Engine
Integration of document copy detection algorithms,
collecting and indexing documents,
queuing large scale copy detection petitions.

Docode ASP

Application Service DOCODE
Web application for DOCODE,
design of document copy detection reporting tools,
software engineering, features and requirement analysis.

Docode Impact

DOCODE Impact Analysis
Surveys and interviews in schools,
social analysis of copy and paste phenomenon,
evaluation of the impact of such tools in education.

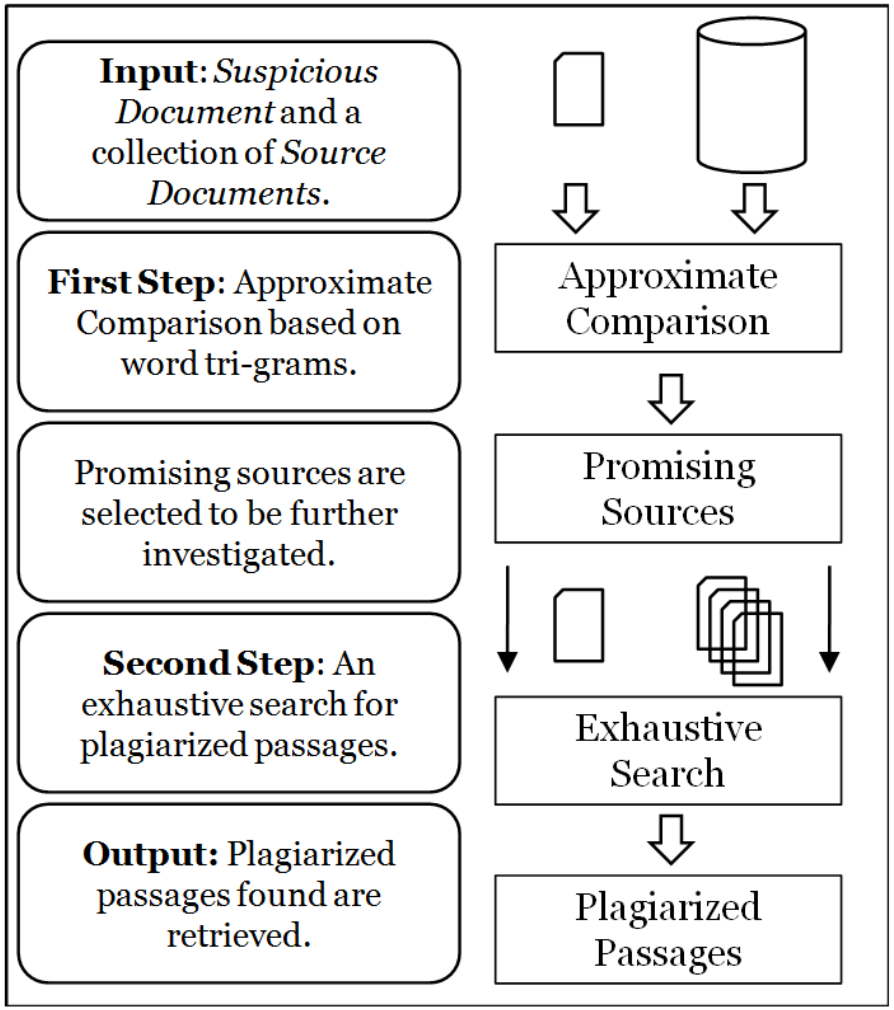
Index

1. Where we were @2010
2. External Plagiarism Detection
3. Intrinsic Plagiarism Detection
4. Results @PAN2011
5. Conclusions

Where we were @2010

Where we were @2010

- Focused on external plagiarism detection.
- No cross-lingual consideration.
- Based on word bi-grams and word tri-grams.
- Selecting *samples* for each document in order to reduce compute time.

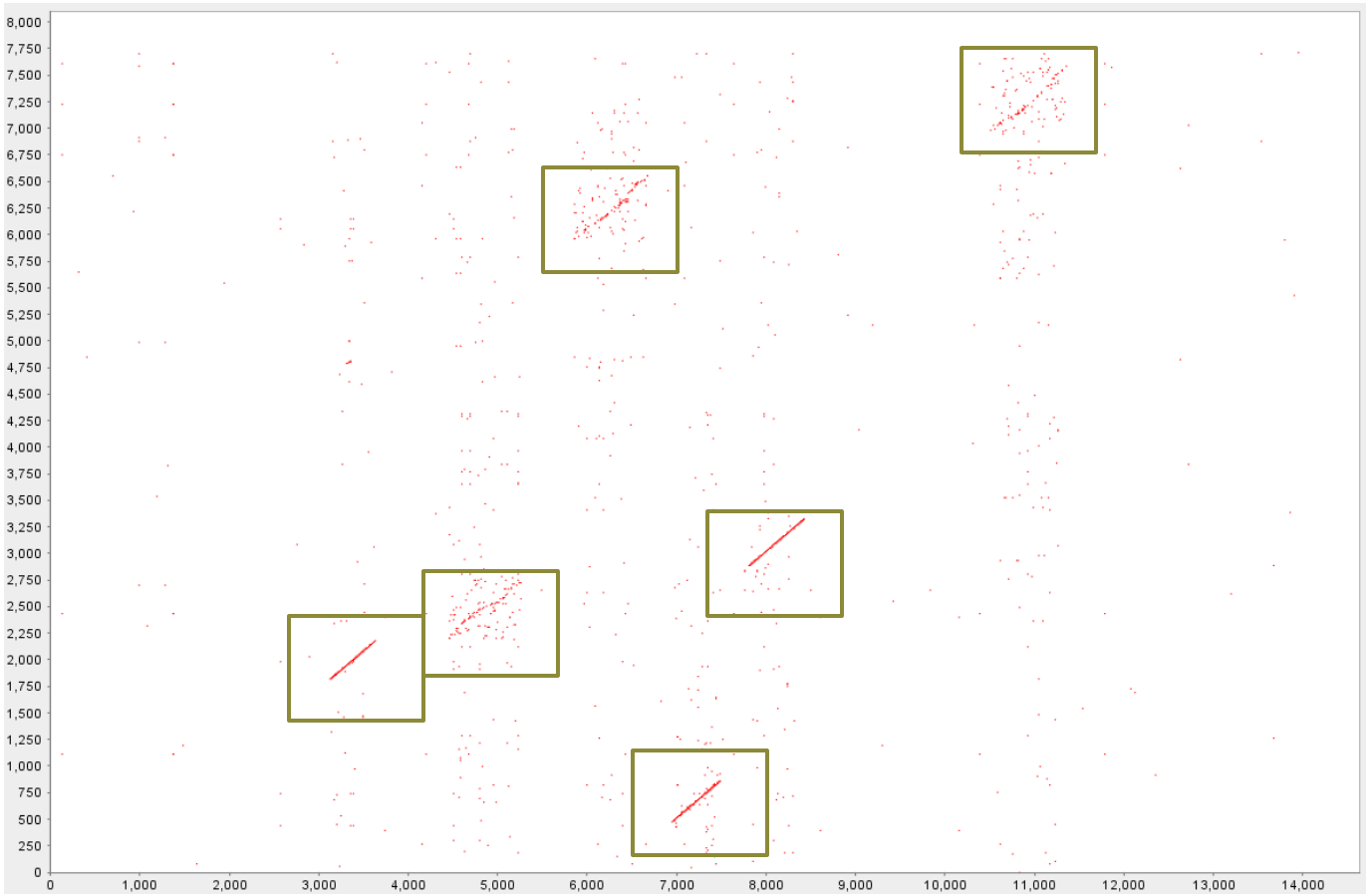


Where we were @2010

External Comparison between 2 documents

Where we were @2010

External Comparison between 2 documents



Where we were @2010

- Monolingual
- Based on word bi-grams and word tri-grams.
- No intrinsic detection

Acceptable precision, detecting half of cases and good granularity.

Plagiarism Detection Performance

Rank	Overall	Recall	Precision	Granularity	Participant
1	0.7971	0.6917	0.9414	1.0006	J. Kasprzak and M. Brandejs Masaryk University, Czech Republic
2	0.7090	0.6299	0.9055	1.0675	D. Zou, W. Long, and Z. Ling South China University of Technology, China
3	0.6948	0.7057	0.8417	1.1508	M. Muhr, R. Kern, M. Zechner, and M. Granitzer Know-Center Graz, Austria
4	0.6209	0.4808	0.9085	1.0177	C. Grozea* and M. Popescu° *Fraunhofer FIRST, Germany °University of Bucharest, Romania
5	0.6066	0.4768	0.8479	1.0086	G. Oberreuter, G. L'Huillier, S.A. Ríos, and J.D. Velásquez University of Chile, Chile

Comparison computed on two eight-core Servers, each with 6 GB of RAM.
Java Implementation.
Reducing Search Space: ~20 Hours.
Finding Plagiarized Passages: ~12 Hours.

External Plagiarism Detection

External Plagiarism Detection

From 2010 experience, we decided to focus on:

- Better precision
- Better recall
- Reduce processing time

External detector @2011

- Uses word 4-grams, removing SW for search space reduction
- Uses word 3-grams for exhaustive search

External Plagiarism Detection

Some results on 2010 PAN Corpus (includes intrinsic and external plagiarism):

Algorithm Version	Overall	Recall	Precision	Granularity
2010	0.61	0.48	0.85	1.001
2011	0.73	0.6	0.94	1.001

- Dual core notebook with 4GB RAM.
- Java Implementation.
- Reducing Search Space :~2 Hours. (0.3% promising doc pairs)
- Exhaustive Search :~1 Hour.

Intrinsic Plagiarism Detection

Intrinsic Plagiarism Detection

Given a document, determine whether one of its paragraphs belong to the average writing style

Intrinsic Plagiarism Detection

Following Stamatatos' (2009) approach:

- Divide the document in partitions
- Compare each partition's *writing style characterization* against the whole document's style
- If a partition deviates from the mean value past some threshold, flag it

Intrinsic Plagiarism Analysis Task						
Rank	Overall score	F-measure	Precision	Recall	Granularity	Participant
1	0.2462	0.3086	0.2321	0.4607	1.3839	E. Stamatatos University of the Aegean, Greece
2	0.1955	0.1956	0.1091	0.9437	1.0007	B. Hagbi and M. Koppel Bar Ilan University, Israel
3	0.1766	0.2286	0.1968	0.2724	1.4524	M. Granitzer, M. Muhr, M. Zechner, and R. Kern Know-Center Graz, Austria
4	0.1219	0.1750	0.1036	0.5630	1.7049	L. M. Seaward and S. Matwin University of Ottawa, Canada

Intrinsic Plagiarism Detection

On the characterization of writing style…

If some of the words used on the document are author-specific, one can think that those words could be concentrated on the paragraphs (or more general, on the segments) that the mentioned author wrote

Intrinsic Plagiarism Detection

Fundamentals:

- Divide document in partitions of equal length
- Word Frequencies
- No stopword removal
- Only chars from a-z

Intrinsic Plagiarism Detection

Comparing the partitions against the whole document…

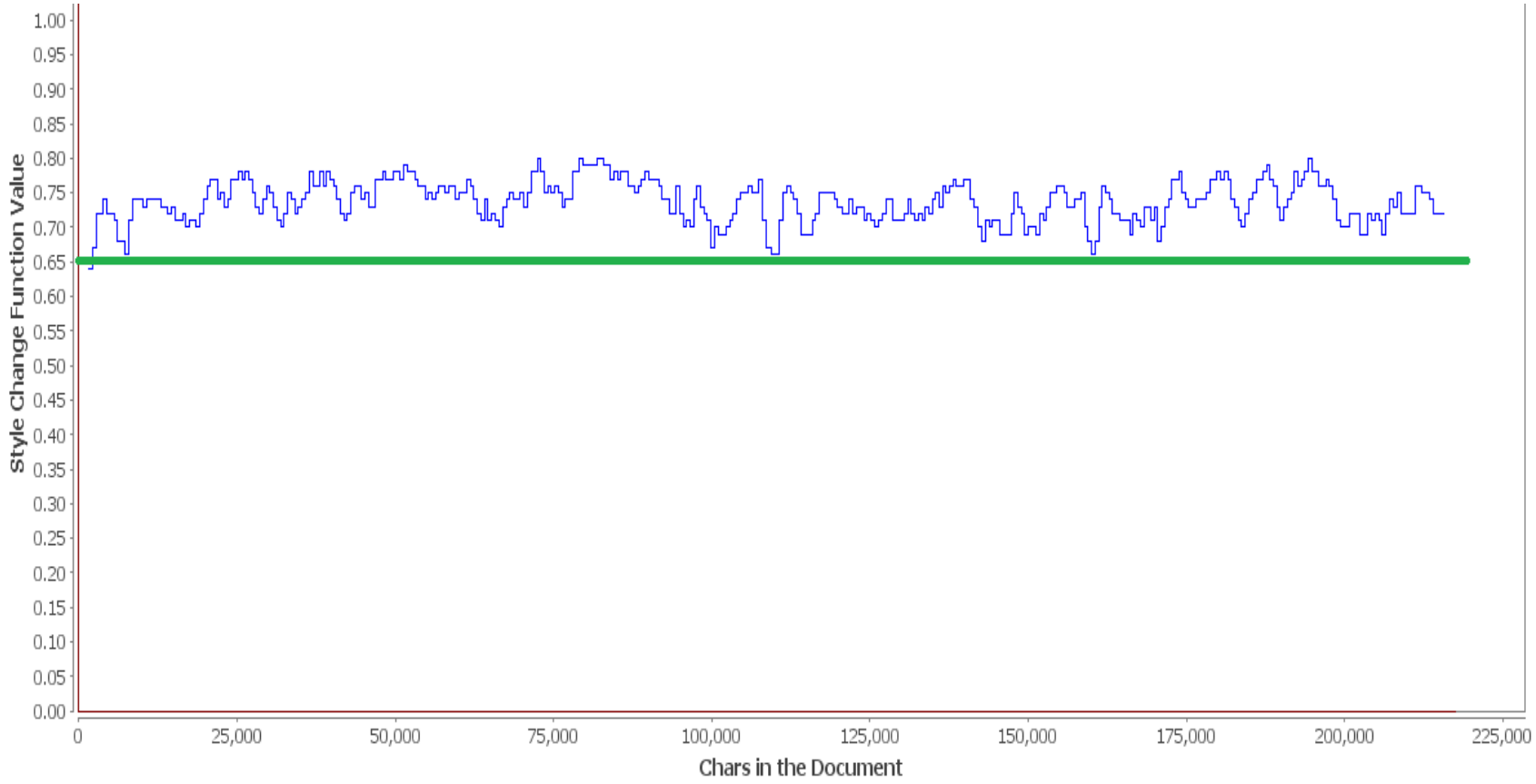
Algorithm 1 Intrinsic plagiarism evaluation

Require: $\mathcal{C}, \mathbf{v}, m, \delta$

for $c \in \mathcal{C}$ **do**
  $d_c \leftarrow 0$
  build v_c using term frequencies on segment c
  for word $w \in v_c$ **do**
    $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v}) - freq(w,v_c)|}{|freq(w,\mathbf{v}) + freq(w,v_c)|}$
  end for
end for
style $\leftarrow \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} d_c$
for $c \in \mathcal{C}$ **do**
  if $d_c < style - \delta$ **then**
    Mark segment c as outlier and potential plagiarized passage.
  end if
end for

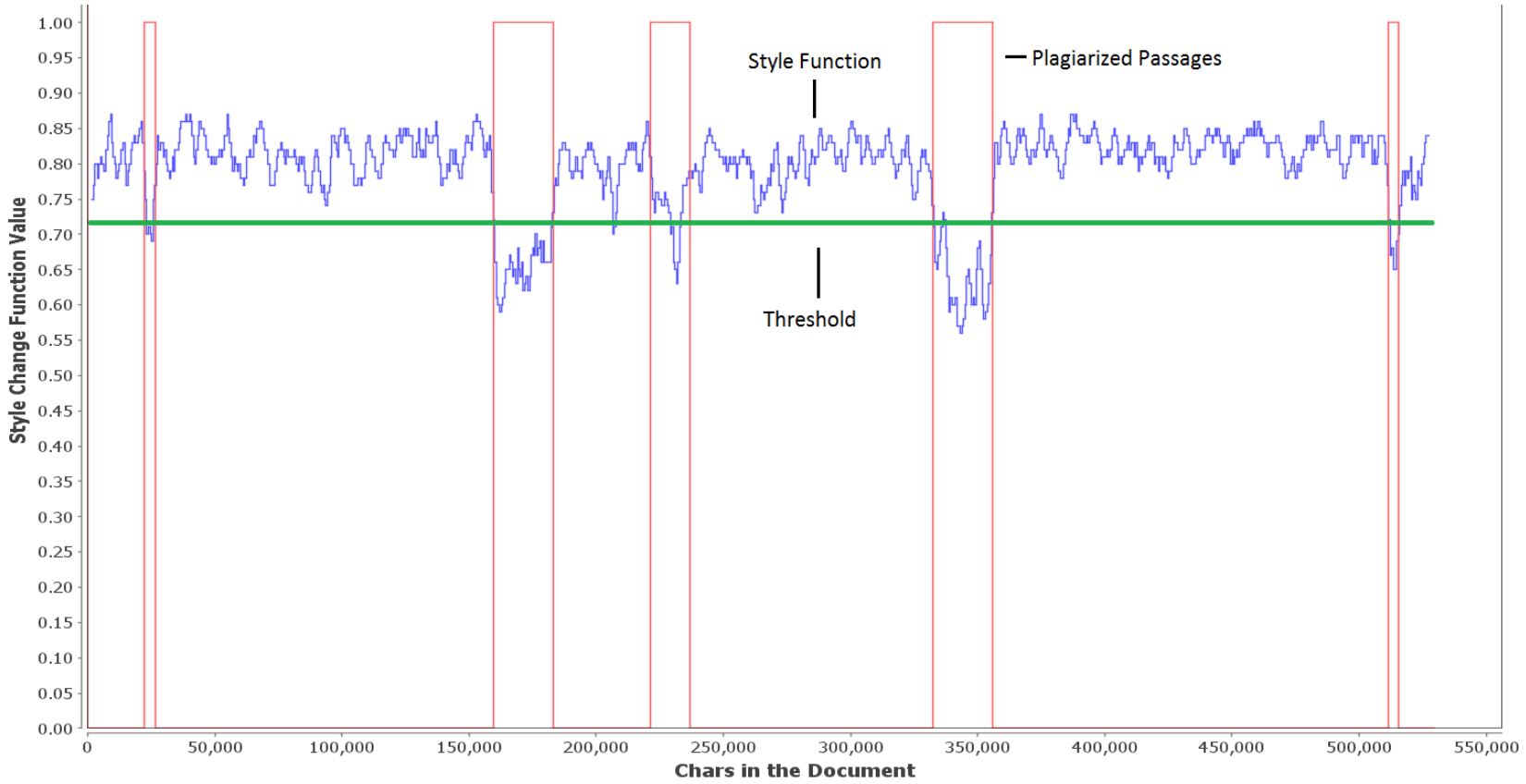
Intrinsic Plagiarism Detection

Example 1: Document written by single author



Intrinsic Plagiarism Detection

Example 2: Document written by multiple authors



Intrinsic Plagiarism Detection

Some results on 2009 PAN Intrinsic Corpus:

	Overall	Recall	Precision	Granularity
Stamatatos	0.25	0.46	0.23	1.38
Oberreuter	0.34	0.31	0.39	1.01

- Dual core notebook with 4GB RAM.
- Java Implementation.
- Run under 10 minutes (~6.000 documents).

Results @PAN2011

Results @PAN2011

External Plagiarism Detection Performance

- Ranked third after Grman&Ravas (0.56) and Grozea&Popescu (0.42)
- Overall good precision, but low recall for obfuscated plagiarism and simulated plagiarism

	plagDet	Recall	Precision	Granularity
overall	0.3468605	0.2257937	0.9116530	1.0611984
translated-obfuscation	0.0012375	0.0006658	0.2176871	1.1034483
translated-manual-obfuscation	0.0000000	0.0000000	0.0000000	1.0000000
simulated-obfuscation	0.4710712	0.3132938	0.9826403	1.0119190
obfuscation-no	0.9079297	0.8840460	0.9411057	1.0057405
obfuscation-low	0.5533509	0.4242483	0.9327743	1.0762728
obfuscation-high	0.0497227	0.0262977	0.6698578	1.0248521

Results @PAN2011

Intrinsic Plagiarism Detection Performance

- Ranked first with a good recall-precision balance
- Overall score of 0.32, with better results with medium- and long-length documents

	plagDet	Recall	Precision	Granularity
overall	0.3254817	0.3397965	0.3123243	1.0000000
doc-length-long	0.3787308	0.3828166	0.3747313	1.0000000
doc-length-medium	0.4001631	0.3660643	0.4412672	1.0000000
doc-length-short	0.2811900	0.2479395	0.3247399	1.0000000
translated-obfuscation	0.3131128	0.2789482	0.3568141	1.0000000
translated-manual-obfuscation	0.1095166	0.1276579	0.0958898	1.0000000

Conclusions

- Word tri-grams and word 4-grams can be used effectively as tokens for external plagiarism detection
- The effectiveness of the approach is strongly correlated to the ability to detect those *dense coincidence* zones
- When no sources are available, the use of words appear to be a good starting point to model the writing style present in documents
- Best result in self-information task, but the scores are overall still too low

Approaches for Intrinsic and External Plagiarism Detection

Notebook for PAN at CLEF 2011

Gabriel Oberreuter Gallardo
goberreu@ing.uchile.cl
University of Chile - September 2011

Group members : Gastón L'Huillier, Sebastián Ríos and Juan D. Velásquez