The ENCOPLOT Similarity Measure for Automatic Detection of Plagiarism

Cristian Grozea[1] Marius Nicolae Popescu[2]

cristian.grozea@brainsignals.de

Fraunhofer Institute FIRST – Berlin

University of Bucharest Romania

September 23, 2011

I'll be short...

Thank you!

.

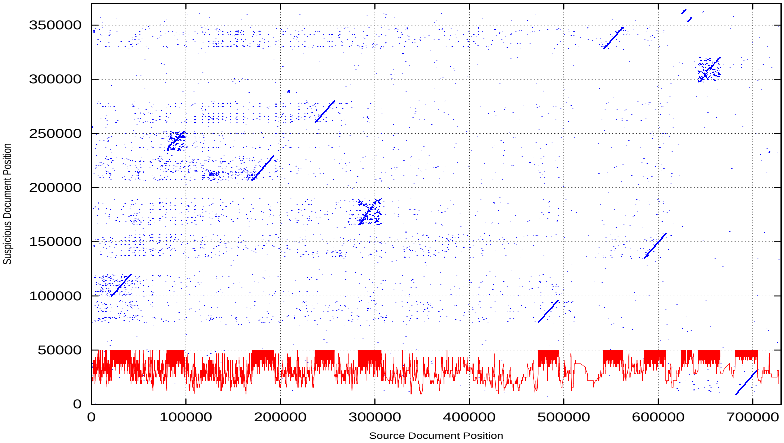
Our extended paper
http://brainsignals.de/encsimTR.pdf

Results

External plagiarism, same language.

- 2009: 1st
- 2010: 4th (2nd w. vers.2011)
- 2011: 2nd (1st?)
  - best score on the **manual paraphrasing**
  - best recall on the non-translated corpus

Encoplot and the Similarity Measure



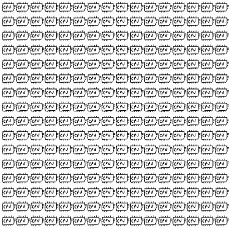
Encoplot Features

- Guaranteed linear time – Dotplot is quadratic.
- Extremely fast highly optimized open-source implementation, for N-grams up to N=16, on 64 bit CPUs.

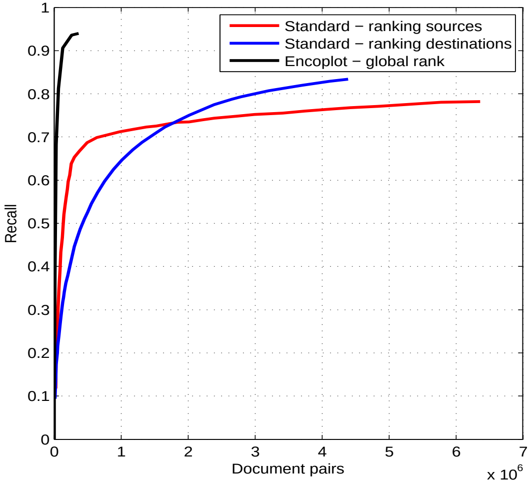
Grozea et. al. (PAN 2009)

The Parallel Encoplot

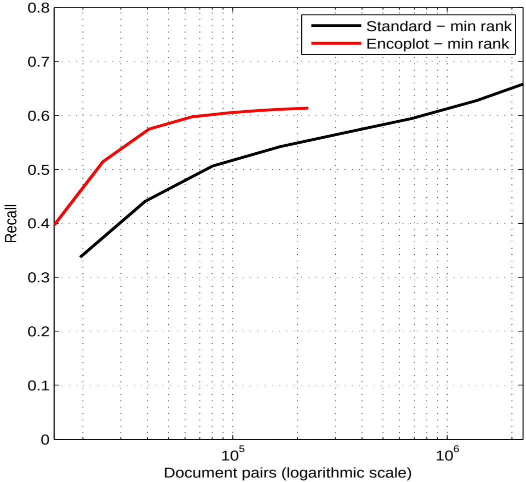
- Open source, licensed under Apache APL http://code.google.com/p/parallel-encoplot/
- Includes the parallelization with BSC SMPSs
- Scalable, tested on a machine with 256 cores
  HPC Europa2 - You can have that too!



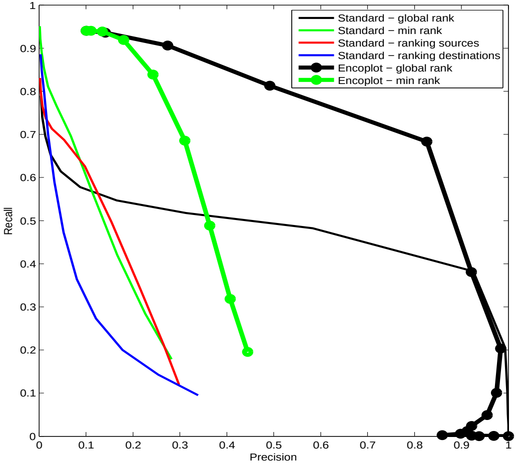
Ranking - 2010



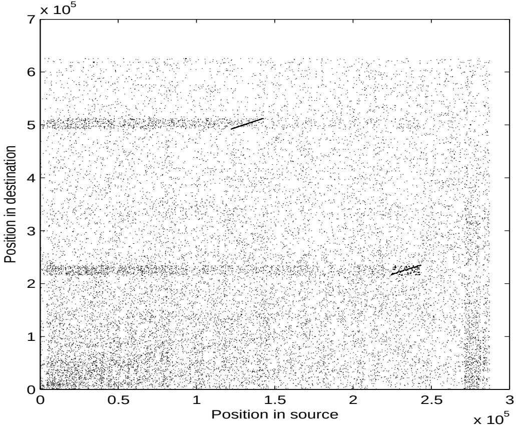
Ranking - 2011



Ranking - 2010 P-R



Who's the Thief?



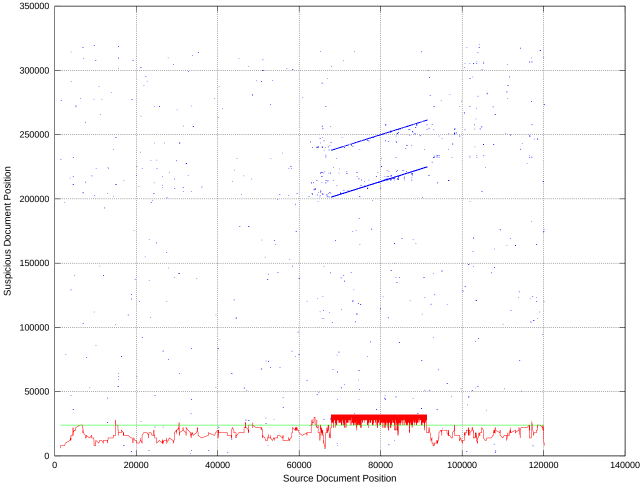
Grozea and Popescu (CICLING 2010) – 75%

Found anything useful to you?

Thank you again!

Reserve slides

2010 duplicates



2011 Corpus

Table: Results on 2011 Competition Data

Subset	Size	Recall	Precision	F-score	Granularity	Plagdet score
Entire corpus	49,621	0.34	0.81	0.48	1.22	0.42
No paraphrasing	976	0.90	0.84	0.86	1.02	0.85
Manual paraphrasing	4,609	**0.36**	**0.96**	**0.53**	1.06	**0.50**
Automatic low	19,779	**0.58**	0.90	0.71	1.27	0.60
Automatic high	19,115	**0.08**	0.64	0.14	1.19	0.13
Manual translation	433	0.08	0.25	0.12	1.01	0.12
Automatic translation	4,709	0.23	0.40	0.29	1.07	0.28

Other Bits

2011 no obfuscation: 976 = 1.97% of 49 621 total (vs. 40%).
The 18% includes about 10 000 from the intrinsic corpus.
2010 multiplicity problem:
Maximum multiplicity =17 (source 8584, suspicious 3283).
55 723 external plagiarism instances
10 694 of which with multiplicity ≥ 2 (20% of total).
3 483 with multiplicity at least 3.
Being able to handle multiplicity up to 4 would leave out only 506 instances.
2010 performance: plagdet score 0.72 (first team - 0.78), with recall 0.66 and precision 0.86, without handling the translated cases (14%).

N-Gram Coincidence Plot

Algorithm

Input: Sequences A and B to compare

Output: list (x,y) of positions in A, respectively B, where there is exactly the same N-gram

Steps

1. Extract the N-grams from A and B
2. Sort these two lists of N-grams
3. Compare these lists in a modified mergesort algorithm. Whenever the two smallest N-grams are the equal, output the position in A and the one in B.

Small example

A=abcabd
B=xabdy

N=2

Encoplot pairs	Dotplot pairs
1 2 ab	1 2 ab
	4 2 ab
5 4 bd	5 4 bd

N=3

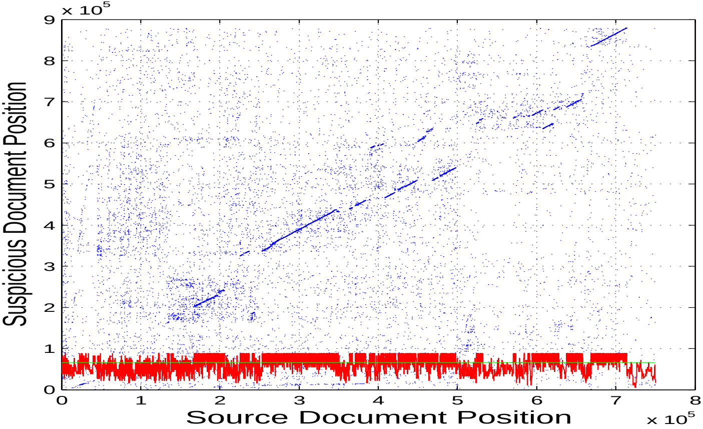
Encoplot pairs	Dotplot pairs
4 2 abd	4 2 abd

Fast Radix Sort for N-Grams

```
for(i,NN)ix[i]=i;
//radix sort, the input is x,
// the output rank is ix
for(k,RANGE)counters[k]=0;
for(i,NN)counters[*(x+i)]++;
for(j,DEPTH){
   int ofs=j;//low endian
   t_int sp=0;
   for(k,RANGE){
      startpos[k]=sp;
      sp+=counters[k];
   }
   for(i,NN){
      unsigned char c=x[ofs+ix[i]];
      ox[startpos[c]++]=ix[i];
   }
   memcpy(ix,ox,NN*sizeof(ix[0]));
   //update counters
   if(j<DEPTH-1){
      counters[*pout++]--;
      counters[*pin++]++;
   }
}
```

- Who's the Thief? Automatic Detection of the Direction of Plagiarism, C.Grozea and M.Popescu, CICLING 2010 , LNCS 6008, DOI 10.1007/978-3-642-12116-6, 2010
- ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection, C.Grozea, C.Gehl, and M.Popescu – In Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, San Sebastian, Spain, 2009. Universidad Politecnica de Valencia 2009
- Encoplot – Performance in the Second International Plagiarism Detection Challenge, C. Grozea and M. Popescu, Lab Report for PAN at CLEF 2010
- Plagiarism Detection with State of the Art Compression Programs, C.Grozea Report CDMTCS-247, Centre for Discrete Mathematics and Theoretical Computer Science, University of Auckland, Auckland, New Zealand, 2004.

Self-plagiarism


x 10^5
Suspicious Document Position
Source Document Position
x 10^5