Overview of the 4th International Competition on Plagiarism Detection

Martin Potthast
Tim Gollub
Matthias Hagen
Jan Graßegger
Johannes Kiesel
Maximilian Michel
Arnd Oberländer
Martin Tippmann
Benno Stein

Parth Gupta
Paolo Rosso

NLEL Group
Universitat Politècnica de València
www.dsic.upv.es/grupos/nle

Alberto Barrón-Cedeño

Webis Group
Bauhaus-Universität Weimar
www.webis.de

LSI Group
Universitat Politècnica de Catalunya
www.lsi.upc.edu
Introduction
Introduction
Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.

2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.

3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.

4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.
Introduction

Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.

2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.

3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.

4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.
Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.

2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.

3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.

4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.
Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.

2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.

3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.

4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.
Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.

2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.

3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.

4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.
Candidate Retrieval

Considerations:

1. PAN’12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB

2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.

3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).
Considerations:

1. PAN’12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB

2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.

3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).
Candidate Retrieval

Considerations:

1. PAN'12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB

2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.

3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).
Candidate retrieval task:

- Humans write essays on given topics, plagiarizing from the ClueWeb, using the ChatNoir search engine for research.
- Detectors use ChatNoir to retrieve candidate documents from the ClueWeb.
- Detectors are expected to maximize recall, but use ChatNoir in a cost-effective way.
Candidate Retrieval

About ChatNoir  [chatnoir.webis.de]

Filters
- Reading level basic (45%)
- expert (55%)
- More text
- Reset filters

271593 other results (15.026 seconds)

The Family of Barack Obama in Kenya
http://www.aztlan.net/barack_obama_family.htm
Los Angeles, Alta California July 25, 2008 LA VOZ DE AZTLAN The Family of Barack Obama in Kenya The father of Barack Obama was married to Kezia Obama in Kenya prior to marrying the senator's mother Ann

Barack Obama and Family on Election Night
http://www.gazotube.com/mnMIO7/RxACh.html
- Mickey N - TubeFire.com - 3GP.FM Barack Obama Sheds Tears for His Beloved Grandmother
0:37 Behind The Scene Pic! Barack Obama Family Election Night! 0:41 Raw Images: Obama on Election Night 1:30 Obama and family snorkel at Hanauma Bay 3:03 Barack Oba

Bargain Prices On First Families
http://www.retrostats.com/stats/First-Families.aspx
1st Steps for Family Fren $ 10.95 The First Family Obama pin! 3" NEW LOWER PRICES!!!! $ 0.99 *OFFICIAL* THE FIRST FAMILY BARACK OBAMA 3" BUTTON PIN $ 4.99 Second City: First Family of Comedy - New DVD LOW PRICE $ 15.55 Pokemon the First Movie-Mewtwo V

Bargain Prices On First Families
- ER PRICES!!!!!! $ 0.99 *OFFICIAL*
- THE FIRST FAMILY BARACK
- OBAMA 3" BUTTON PIN $ 4.99
- Second City: First Fa

>> more results from www.aztlan.net

Yes We Can - Blog
http://www.yeswecanviralbands.us/blog/tag/obama-family
- the current White House pooh, Barney Bush! Woof. Woof Jason | Post a Comment | Share
- Article tagged obama dog, obama family barack obama white house, dog in U.S. POLITICS This
Candidate Retrieval

About ChatNoir [chatnoir.webis.de]

- employs BM25F retrieval model
  (CMU’s Indri search engine is language-model-based)
- provides search facets capturing readability issues
- own index development based on externalized minimal perfect hash functions
- index built on a 40 nodes Hadoop cluster
- search engine currently running on 11 machines
Barack Obama's Family

The Family of Barack Obama is an extended clan of African American, English, Indonesian, and Kenyan heritage. They are best known through the writings and political career of Barack Obama, the current President of the United States of America. His immediate family is the First Family of the United States. The Obamas are the first First Family of African American descent in the United States and the youngest to enter the White House since the Kennedys. Obama's young, energetic family harks back today of Camelot.

http://web15.medien.uni-weimar.de/dament/chkweb/NJ=10001719993&cokewor=0911001.qnt

In what follows, we give a detailed overview of Barack Obama's family. We shed light on himself, his immediate and extended family, including maternal and paternal relations. Moreover, we give insights into the relations of Michelle Obama, Barack Obama’s wife, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. Obama was the junior United States Senator from Illinois from 2005 until he resigned following his election to the presidency. Obama is a graduate of Columbia University and Harvard Law School. He worked as a community organizer in Chicago prior to earning his law degree, and practiced as a civil rights attorney in Chicago before serving three terms in the Illinois Senate from 1997 to 2004. He also taught Constitutional Law at the University of Chicago Law School from 1992 to 2004. Following an unsuccessful bid for a seat in the U.S. House of Representatives in 2000, Obama was elected to the Senate in November 2004.

Barack Obama was born on 4 August 1961 at the Kapiolani Medical Center for Women & Children in Honolulu, Hawaii, to Ann Dunham of Wichita, Kansas and Barack Hussein Obama Sr. of Nyangoma Kogelo, Nyanza Province, Kenya. His parents met while both were attending the East-West Center of the University of Hawaii at Manoa, where his father was enrolled as a foreign student. The couple married only 6 months earlier on February 2, 1961.

Childhood and Youth

When Barack Obama was two years old, his parents divorced and his father moved to Connecticut to continue his education before returning to Kenya. He saw his son only once more before dying in an automobile accident in 1982.

When Obama was six, his mother married Lolo Soetoro, an Indonesian oil manager. In 1967, when Soetoro's student visa was revoked because of political unrest in Indonesia, Dunham and Barack then in first grade, accompanied him to Jakarta.
Candidate Retrieval

About Corpus Construction

- an essay has approx. 5000 words which means 8-10 pages
- own web editor was developed for essay writing
- the writing is crowdsourced via oDesk

→ full control over:
  - plagiarized document
  - set of used source documents
  - annotations of paraphrased passages
  - query log of the writer while researching the topic
  - search results for each query
  - click-through data for each query
  - browsing data of links clicked within ClueWeb
  - edit history of the document covering all keystrokes
  - work diary and screenshots as recorded by oDesk

→ insights on how humans work when reusing text
Candidate Retrieval
Survey of Approaches

An analysis of the participants’ notebooks reveals a candidate retrieval process:

1. Chunking
   Given a suspicious document, it is divided into (possibly overlapping) passages of text. Each chunk of text is then processed individually.

2. Keyphrase Extraction
   Given a chunk (or the entire suspicious document), keyphrases are extracted from it in order to formulate queries with them.

3. Query Formulation
   Given sets of keywords extracted from chunks, queries are formulated which are tailored to the API of the search engine used.

4. Search Control
   Given a set of queries, the search controller schedules their submission to the search engine and directs the download of search results.

5. Download Filtering
   Given a set of downloaded documents, all documents are removed that are not worthwhile for detailed comparison to the suspicious document.
<table>
<thead>
<tr>
<th>Team</th>
<th>Total Workload</th>
<th>Time to 1st Detection</th>
<th>Reported Sources</th>
<th>Downloaded Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Queries</td>
<td>Dwnlds</td>
<td>Queries</td>
<td>Dwnlds</td>
</tr>
<tr>
<td>Gillam</td>
<td>63</td>
<td>527</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>Jayapal</td>
<td>67</td>
<td>174</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Kong</td>
<td>551</td>
<td>327</td>
<td>81</td>
<td>28</td>
</tr>
<tr>
<td>Palkovskii</td>
<td>63</td>
<td>1027</td>
<td>27</td>
<td>319</td>
</tr>
<tr>
<td>Suchomel</td>
<td>13</td>
<td>95</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

- Suchomel et al. implement the best tradeoff between cost and quality.
- Jayapal implements the best approach in terms of precision and recall.
Detailed Comparison

1. Document collection
2. Candidate retrieval
3. Candidate documents
4. Suspicious document

Detailed comparison
Suspicious passages
Knowledge-based post-processing
Suspicious document
Thesis
Candidate retrieval
Candidate documents
Detailed comparison

PAN
Detailed comparison task:

- Detectors are presented with a suspicious and a candidate document, and are asked to extract the plagiarized passages.
- Developers submit their detection softwares instead of detection results.
- This allows for re-evaluating detectors, as well as to measure runtime and to use private corpora.
Eleven participants, about the average number from last years.

Software submissions do not distract people from participating.

<table>
<thead>
<tr>
<th>Team</th>
<th>Submission Size [MB]</th>
<th>Operating System</th>
<th>Programming Language</th>
<th>Average Runtime [sec/comparison]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rodríguez Torrejón</td>
<td>1.80</td>
<td>Linux</td>
<td>sh, C/C++</td>
<td>0.19</td>
</tr>
<tr>
<td>Sánchez-Vega</td>
<td>0.04</td>
<td>Linux</td>
<td>C++</td>
<td>2.48</td>
</tr>
<tr>
<td>Oberreuter</td>
<td>0.19</td>
<td>Linux</td>
<td>Java</td>
<td>2.58</td>
</tr>
<tr>
<td>Palkovskii</td>
<td>68.20</td>
<td>Windows</td>
<td>C#</td>
<td>4.51</td>
</tr>
<tr>
<td>Grozea</td>
<td>1.90</td>
<td>Linux</td>
<td>Perl, Octave</td>
<td>4.82</td>
</tr>
<tr>
<td>Suchomel</td>
<td>0.02</td>
<td>Linux</td>
<td>Perl</td>
<td>5.36</td>
</tr>
<tr>
<td>Kong</td>
<td>2.60</td>
<td>Linux</td>
<td>Java</td>
<td>5.91</td>
</tr>
<tr>
<td>Jayapal</td>
<td>37.20</td>
<td>Linux</td>
<td>Java</td>
<td>8.43</td>
</tr>
<tr>
<td>Gillam</td>
<td>0.48</td>
<td>Linux</td>
<td>Python 2.7</td>
<td>9.40</td>
</tr>
<tr>
<td>Küppers</td>
<td>42.90</td>
<td>Linux</td>
<td>Java</td>
<td>27.64</td>
</tr>
<tr>
<td>Ghosh</td>
<td>554.50</td>
<td>Linux</td>
<td>sh, Java</td>
<td>–</td>
</tr>
</tbody>
</table>

Congratulations to Rodríguez Torrejón et al. for submitting the most efficient detailed comparison program.
Detailed Comparison

Survey of Approaches

An analysis of the participants’ notebooks reveals a detailed comparison process:

1. Seeding
   Given a suspicious document and a source document, matches (also called „seeds”) between the two documents are identified using some seed heuristic. Seed heuristics either identify exact matches or create matches by changing the underlying texts in a domain-specific or linguistically motivated way.

2. Match Merging
   Given seed matches identified between a suspicious document and a source document, they are merged into aligned text passages of maximal length between the two documents which are then reported as plagiarism detections.

3. Passage Filtering
   Given a set of aligned passages, a passage filter removes all aligned passages that do not meet certain criteria.
TIRA takes locally executable programs and turns them into web services.

TIRA assumes responsibility for storing and indexing of execution results.

For the PAN evaluation, TIRA servers are provided for two operating systems, Windows and Ubuntu.

Participants submit their plagiarism detection software for deployment on the appropriate TIRA server.

A third TIRA server controls the overall evaluation of all deployed submissions on the private test set and provides the overall results.
Detailed Comparison

Evaluation Corpus Construction

- Like in last years based on books from Project Gutenberg.
- Divided into seven sub-corpora:

<table>
<thead>
<tr>
<th>Evaluation Corpus Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sub-Corpus</strong></td>
</tr>
<tr>
<td>Real Cases</td>
</tr>
<tr>
<td>Simulated</td>
</tr>
<tr>
<td>Translation ({de, es} → en)</td>
</tr>
<tr>
<td>Artificial (High)</td>
</tr>
<tr>
<td>Artificial (Low)</td>
</tr>
<tr>
<td>No Obfuscation</td>
</tr>
<tr>
<td>No Plagiarism</td>
</tr>
<tr>
<td>Overall</td>
</tr>
</tbody>
</table>

- Similarity of document pairs was taken into account this year.
- Real Cases were taken from the Web. Cross-Language cases were constructed using the multi-lingual Europarl corpus.
## Detailed Comparison

### Evaluation Results: Overall Performance

<table>
<thead>
<tr>
<th>Rank / Team</th>
<th>PlagDet</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Kong</td>
<td>0.738</td>
<td>0.824</td>
<td>0.678</td>
<td>1.01</td>
</tr>
<tr>
<td>2  Suchomel</td>
<td>0.682</td>
<td>0.893</td>
<td>0.552</td>
<td>1.00</td>
</tr>
<tr>
<td>3  Grozea</td>
<td>0.678</td>
<td>0.774</td>
<td>0.635</td>
<td>1.03</td>
</tr>
<tr>
<td>4  Oberreuter</td>
<td>0.673</td>
<td>0.867</td>
<td>0.555</td>
<td>1.00</td>
</tr>
<tr>
<td>5  Rodríguez Torrejón</td>
<td>0.625</td>
<td>0.834</td>
<td>0.500</td>
<td>1.00</td>
</tr>
<tr>
<td>6  Palkovskii</td>
<td>0.538</td>
<td>0.574</td>
<td>0.523</td>
<td>1.02</td>
</tr>
<tr>
<td>7  Küppers</td>
<td>0.349</td>
<td>0.776</td>
<td>0.282</td>
<td>1.26</td>
</tr>
<tr>
<td>8  Sánchez-Vega</td>
<td>0.309</td>
<td>0.537</td>
<td>0.349</td>
<td>1.57</td>
</tr>
<tr>
<td>9  Gillam</td>
<td>0.308</td>
<td>0.898</td>
<td>0.190</td>
<td>1.02</td>
</tr>
<tr>
<td>10 Jayapal</td>
<td>0.045</td>
<td>0.622</td>
<td>0.075</td>
<td>6.93</td>
</tr>
</tbody>
</table>

→ Congratulations to Kong et al. for submitting the most effective detailed comparison program.
Summary and Outlook

PAN 2012:

- Task-wise evaluation of plagiarism detectors.
- Candidate document retrieval at Web scale using ChatNoir.
- Software submissions for sustainable / repeatable evaluation using TIRA.
- More realistic plagiarism corpus.
- New performance measures in addition to the traditional ones.

→ A lot of fun!
   Thanks to everyone who volunteered to test our new setup!

PAN 2013 and beyond:

- Improvement and consolidation of the new tools.
- Use of the plagiarism corpus for detailed comparison as well.
- Community process to collect more plagiarism (real and manual).

→ Fully automatic plagiarism detection evaluations.