On Web-based Plagiarism Analysis

A. Kleppe, D. Braunsdorf, Chr. Lössnitz, S. Meyer zu Eissen
Bauhaus University Weimar
Web-based Information Systems
What is Plagiarism?

“Plagiarism refers to the use of another’s ideas, information, language, or writing, when done without proper acknowledgment of the original source”

[Wikipedia]

11.9. 2005 Kleppe/Braunsdorf/Lössnitz/Meyer zu Eissen
What is Plagiarism?

“Plagiarism refers to the use of another’s ideas, information, language, or writing, when done without proper acknowledgment of the original source”

[Wikipedia]

Plagiarism analysis:

*Given.* A suspicious document.

*Task.* Find potentially copied parts, and provide references to original sources.
What is Plagiarism?

"Plagiarism refers to the use of another’s ideas, information, language, or writing, when done without proper acknowledgment of the original source”

[Wikipedia]

Plagiarism analysis:

*Given.* A suspicious document.

*Task.* Find potentially copied parts, and provide references to original sources.

Fact: About 50% of the students admit to plagiarize from Internet documents (study on 18,000 students).

[McCabe 2005]
Current Research on Plagiarism Analysis

Current research is corpus-oriented.

Given. A suspicious document and a corpus of original documents.

Task. Find potentially copied parts in the corpus, and provide references to original sources.
Current Research on Plagiarism Analysis

Current research is corpus-oriented.

Given. A suspicious document and a corpus of original documents.

Task. Find potentially copied parts in the corpus, and provide references to original sources.

Research questions:

- How can a corpus of potentially original documents be constructed from the Web?
- Can plagiarized parts be detected without a corpus?
Plagiarism may happen in manyfold variants:

- plagiarism delicts
  - accurate copy
    - part of document
      - outline
    - paragraph
    - sentence/subset
  - complete document
  - translation
    - outline
    - sentence/subset/word
  - transformation
    - synonym
Intuition:

- Partition each document in meaningful units (“chunks”).
- Compare them with a similarity function $\varphi$ (pairwise).
Plagiarism Analysis against a Corpus (1)

Intuition:

- Partition each document in meaningful units (“chunks”).
- Compare them with a similarity function $\varphi$ (pairwise).

Complexity:

$n$ documents in corpus, $c$ chunks per document on average

$\Rightarrow O(n \cdot c^2)$ comparisons
Plagiarism Analysis against a Corpus (2)

Text comparison with digital fingerprints:

- Partition each document in meaningful units (“chunks”).
- Compute fingerprints of the chunks using a hash function \( h \).
- Put all hashes into a hashtable. A collision indicates matching chunks.

---

Figure 1: Taxonomy of plagiarism forms. Orthogonal to plagiarism forms that copy the structure of other documents, new forms may also be seen as plagiarized:

1. **Copied text.** If text stems from a source that is known, it is not cited properly.
2. **Bibliography.** If the references in documents overlap significantly, the bibliography may be copied.
3. **Change in writing style.** A suspect change in the author’s style may appear.
4. **Change in formatting.** In copy-and-paste plagiarism cases, the formatting of the source may be altered.
5. **Change in setting style.** If the setting style of the source document changes, the target may be plagiarized.
6. **Change in method.** If the method changes, the plagiarism may be detected.
7. **Change in implementation.** If the implementation of the source changes, the target may be plagiarized.
8. **Change in presentation.** If the presentation of the source changes, the target may be plagiarized.
9. **Change in data.** If the data of the source changes, the target may be plagiarized.
10. **Change in explanation.** If the explanation of the source changes, the target may be plagiarized.
11. **Change in structuring.** If the structuring of the source changes, the target may be plagiarized.
12. **Change in overall structure.** If the overall structure of the source changes, the target may be plagiarized.

---

Abstract
The paper presents an application for the analysis of text documents against a corpus of documents. For this purpose, a retrieval method is introduced that computes digital fingerprints of the documents. These fingerprints are stored in a hash table, and a collision indicates matching chunks.

---

Conclusion
The presented method can be used to focus a search for potentially plagiarized sections within a single document. Apart from contributing to solve the outlined problem, the presented method can also be used to focus a search for potentially plagiarized sections within a single document.
Plagiarism Analysis against a Corpus (2)

Text comparison with digital fingerprints:

- Partition each document in meaningful units (“chunks”).
- Compute fingerprints of the chunks using a hash function \( h \).
- Put all hashes into a hashtable. A collision indicates matching chunks.

Complexity:

\( n \) documents in corpus, \( c \) chunks per document on average

\( \Rightarrow O(n \cdot c) \) operations (fingerprint generation, hashtable operations).
Plagiarism Analysis against a Corpus (3)

Discussion:

- Hashing is fast, but sensitive to (even small) changes:
  \[ h(c_1) = h(c_2) \Rightarrow c_1 = c_2 \] 
  (with very high probability).
- Pairwise comparisons based on similarity-function \( \varphi \) are too expensive.
  \[ \Rightarrow \] Past research focussed on chunking strategies.
Introduction

Technical Background

Style Analysis

Plagiarism Detection on the Web

Prototype

Plagiarism Analysis against a Corpus (3)

Discussion:

- Hashing is fast, but sensitive to (even small) changes:
  \[ h(c_1) = h(c_2) \implies c_1 = c_2 \] (with very high probability).
- Pairwise comparisons based on similarity-function \( \varphi \) are too expensive.
  \[ \rightarrow \text{Past research focused on chunking strategies.} \]

Current research:

- Focus on fuzzy hash functions \( h_F \):
  \[ h_F(c_1) = h_F(c_2) \implies \varphi(c_1, c_2) \geq 1 - \varepsilon \] \[\text{[Stein 2005]}\]
- Fuzzy hash functions allow for big chunk sizes (speed-up) and are not sensitive to changes.
Indications for Plagiarism

Text similarity is not the only indication for plagiarism. Indications include:

- Changes in formatting.
- *Changes in writing style.*
- Broken argumentation.
- Inconsistent spelling.
- Outmoded diction.

These indications can be detected (by humans) without corpora.

⇒ How can we operationalize the detection of these indications?
Q: How can writing style changes be measured?
A: Not directly, but divergences of word class distributions give hints.

Word class use in a document:
**Style Analysis**

Idea: Compare the word class distribution of each paragraph to the distribution of the entire document.

- **word-class frequencies for the document (global distribution)**
- **word-class frequencies for a single paragraph (local distribution)**
Style Analysis

The divergence can be measured by means of the Kullback-Leibler divergence.

Let $W$ denote the set of word classes, let $w \in W$, let $p(w)$ denote local word class frequencies, let $q(w)$ denote the global word class frequencies.
Style Analysis

The divergence can be measured by means of the Kullback-Leibler divergence.

Let $W$ denote the set of word classes, let $w \in W$, let $p(w)$ denote local word class frequencies, let $q(w)$ denote the global word class frequencies.

The Kullback-Leibler divergence measure is defined as

$$KL_W(p, q) = \sum_{w \in W} p(w) \log \frac{p(w)}{q(w)} = H(p, q) - H(p) \in \mathbb{R}_0^+$$

If $KL_W(p, q)$ is significant then the paragraph that is associated with $p$ may be copied.

We found $KL_W$ to work very well when single paragraphs are copied from one document to another.
Web-based Plagiarism Analysis

**Given.** A suspicious document, 
*and the Web as corpus of original documents.*

**Task.** Generate a *candidate document base* from the Web, 
find potentially copied parts in the base documents, 
and provide references to original sources.
Web-based Plagiarism Analysis

**Given.** A suspicious document, and the Web as corpus of original documents.

**Task.** Generate a candidate document base from the Web, find potentially copied parts in the base documents, and provide references to original sources.
Prototypic Implementation

This page allows you to upload a file on the Plagiarism Finder Server. Note that the size of the file is limited to 5MB.

File upload

Durchsuchen...

Submit Press the Submit Button to upload the file

Plagiarism Search Bauhaus-University Weimar

11.9. 2005 Kleppe/Braunsdorf/Lössnitz/Meyer zu Eissen
Introduction

Technical Background

Style Analysis

Plagiarism Detection on the Web

Prototype

Plagiarism Search Bauhaus-University Weimar

<table>
<thead>
<tr>
<th>Input Document</th>
<th>Source Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>title: inp2.txt</td>
<td>title: candidate31.txt</td>
</tr>
<tr>
<td>words: 3726</td>
<td>words: 3245</td>
</tr>
<tr>
<td>stopwords: 1806</td>
<td>stopwords: 1550</td>
</tr>
<tr>
<td>similarity: 0.104806451</td>
<td></td>
</tr>
</tbody>
</table>

11.9. 2005 Kleppe/Braunsdorf/Lössnitz/Meyer zu Eissen
Thank You!

Questions?