Strategies for Retrieving Plagiarized Documents

Contribution:
A tailored index for plagiarism analysis

For the identification of plagiarized passages in large document collections we present retrieval strategies which rely on stochastic sampling and chunk indexes.

Using the entire Wikipedia corpus we compile n-gram indexes and compare them to a new kind of fingerprint index in a plagiarism analysis use case. Our index provides an analysis speed-up by factor 1.5 and is an order of magnitude smaller, while being equivalent in terms of precision and recall.

Plagiarism Challenge:
"Did the author of a document dq commit a plagiarism offense?"

Approaches for computer-based plagiarism analysis break down this question into manageable parts:
"Given a collection D of documents, does dq contain a section sq for which one can find a document dq’ ∈ D that contains a section sx such that under some retrieval model R the similarity jR between sq and sx is close to 1?"

Process Overview

1. Heuristic retrieval (based on chunk index)
2. Detailed analysis (based on VSM)
3. Knowledge-based post-processing

Index Construction:

- A prior probabilities of prefix classes in BNC
- Distribution of prefix classes in sample
- Normalization and difference computation
- Fuzzification
- Fingerprint

Results

Characteristics of the Wikipedia chunk index:

<table>
<thead>
<tr>
<th>Chunk index type</th>
<th>avg. postlist length</th>
<th>Size</th>
<th>Size ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_{FF}, l=40, g=0.8</td>
<td>1.17 entries</td>
<td>0.35 GB</td>
<td>0.07</td>
</tr>
<tr>
<td>D_{FF}, l=40, g=0.5</td>
<td>2.43 entries</td>
<td>3.42 GB</td>
<td>0.65</td>
</tr>
<tr>
<td>D_{FF}, l=40, g=0.3</td>
<td>1.44 entries</td>
<td>5.25 GB</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Related Technology [Koppel/Schler 2004]

Unmasking applied to very short documents (4-8 pages):

Characteristics of the Wikipedia chunk index: