Signature Extraction for Overlap Detection in Documents

Raphael Finkel
University of Kentucky

Arkady Zaslavsky
Krisztián Monostori
Heinz Schmidt

School of Computer Science and Software Engineering,
Monash University
Overview

- Applications of overlap detection
- Overlap detection procedure
- Signature of documents
- Performance results
- Alternative approaches
- Conclusion and future work
Applications of Overlap Detection

- Plagiarism-detection
- Finding related documents in a document set
- Filtering search-engine results
- Finding illegal copies of documents
Overlap Detection Procedure

1. Partition
2. Retain representative chunks
3. Digest chunks to short byte strings
4. Store byte strings in hash table
5. Compare byte strings
Partition

Chunking strategies have a significant effect on the accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Partition

Chunking strategies have a significant effect on accuracy.

- Fix number of words
- Overlapping chunks
- Sentence chunking
- Hashed breakpoint chunking
Retain Representative Chunks

- Fingerprinting (culling)
- Other systems – random e.g. mod 25
- Sqrt
  - $\left\lceil \sqrt{n} \right\rceil$ number of chunks closest in length to the median chunk size

- Variance
  - $|L - M| \leq bs$, $s$ is the standard deviation of chunk sizes
Similarity Measures

- **Asymmetric**
  \[
  a(F,G) = \frac{|d(F) \cap d(G)|}{|d(F)|}
  \]

- **Symmetric**
  \[
  s(F,G) = \frac{|d(F) \cap d(G)|}{|d(F)| + |d(G)|}
  \]

- **Global**
  \[
  g(F) = \frac{|d(F) \cap (\bigcup_G d(G))|}{|d(F)|}
  \]
Performance Results

- 2591 RFC documents (112MB)
- Index 5.3MB
- >90%  5
- 80-90%  24
- 70-80%  30
- 60-70%  60
## Performance Results

<table>
<thead>
<tr>
<th>RFC 1</th>
<th>RFC 2</th>
<th>MDR 1</th>
<th>MDR 2</th>
<th>SE 1</th>
<th>SE 2</th>
<th>OV 1</th>
<th>OV 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1596</td>
<td>1604</td>
<td>99</td>
<td>99</td>
<td>91</td>
<td>92</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>2264</td>
<td>2274</td>
<td>99</td>
<td>99</td>
<td>96</td>
<td>95</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>1138</td>
<td>1148</td>
<td>96</td>
<td>95</td>
<td>93</td>
<td>92</td>
<td>91</td>
<td>89</td>
</tr>
<tr>
<td>1065</td>
<td>1155</td>
<td>96</td>
<td>91</td>
<td>71</td>
<td>68</td>
<td>84</td>
<td>79</td>
</tr>
<tr>
<td>1084</td>
<td>1395</td>
<td>86</td>
<td>84</td>
<td>58</td>
<td>64</td>
<td>79</td>
<td>75</td>
</tr>
<tr>
<td>1600</td>
<td>1410</td>
<td>72</td>
<td>77</td>
<td>52</td>
<td>48</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>2497</td>
<td>2394</td>
<td>19</td>
<td>17</td>
<td>33</td>
<td>27</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>2422</td>
<td>2276</td>
<td>18</td>
<td>3</td>
<td>23</td>
<td>6</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>2392</td>
<td>2541</td>
<td>16</td>
<td>12</td>
<td>27</td>
<td>17</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>
Alternative Approaches

- Statistical data: number of syllables in words, frequency of passive constructions, number of dependent clauses.
  - k-d tree nearest neighbour search
- Compression

\[
s(F, G) = 2 - \frac{2|\text{compress}(F + G)|}{|\text{compress}(F)| + |\text{compress}(G)|}
\]
Conclusion and Future Work

- Overlap detection procedure
- Culling methods
- Performance results
- Future work
  - more datasets
  - prototype system