

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
 - $\lceil \sqrt{n} \rceil$ 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

$$g(F) = \frac{|d(F) \cap (\cup_G d(G))|}{|d(F)|}$$

Performance Results

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

Performance Results

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

Alternative Approaches

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

$$s(F, G) = 2 - \frac{2|\mathit{compress}(F + G)|}{|\mathit{compress}(F)| + |\mathit{compress}(G)|}$$

Conclusion and Future Work

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