

Comparison of Overlap Detection Techniques

Krisztián Monostori, Arkady Zaslavsky,

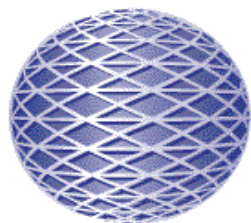
*School of Computer Science and Software Engineering
Monash University, Australia*

Raphael Finkel

Computer Science, University of Kentucky, USA

Gábor Hodász, Máté Pataki

*Department of Automation and Applied Informatics,
Budapest University of Technology and Economic
Sciences, Hungary*



**DISTRIBUTED
SYSTEMS
TECHNOLOGY
CENTRE**

ICCS 2002, Amsterdam, 21-24 April, 2002



Overview

- Overlap Detection
- Copy-Detection Methods
- Chunking Methods
 - Test Results
- Fingerprint Selection Strategies
 - Test Results
- Conclusion



Overlap Detection

- Applications
 - Plagiarism Detection
 - Copyright Law
 - Search-engines
- Methods
 - Prevention
 - Detection



Overlap Detection Steps

1. Chunking
2. Fingerprinting
3. Digesting
4. Storage
5. Comparison
(MatchDetectReveal)



Comparison

- *Asymmetric similarity*
- *Symmetric similarity*
- *Global similarity*



Chunking Methods

Copy detection methods use some kind of a hash-function to reduce space requirements.

■ Non-overlapping

Copy detection methods use some kind of a hash-function to reduce space requirements.

■ Overlapping

- Copy detection methods, detection methods use, ...

■ Word

Copy detection methods use some kind of a hash-function to reduce space requirements.

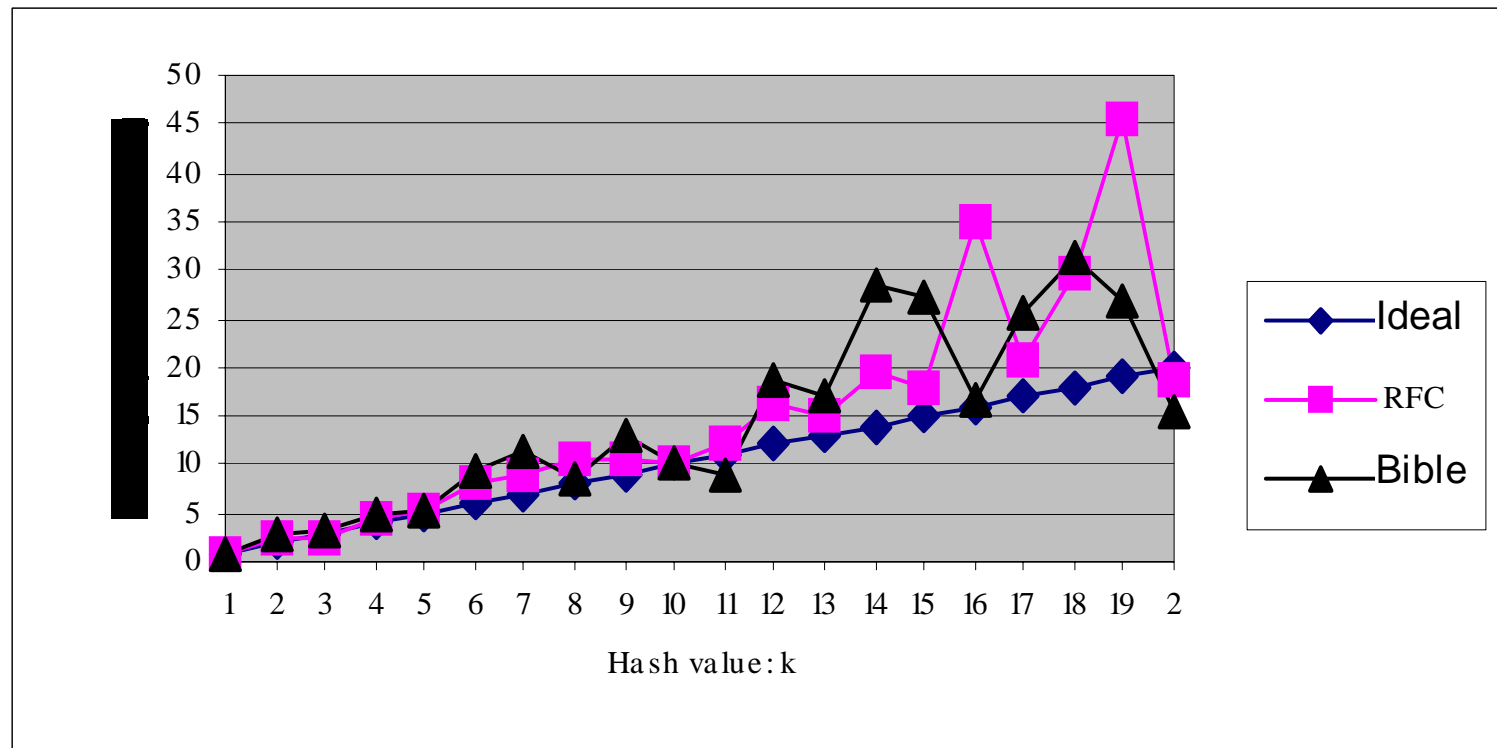
■ Hashed Breakpoint

Copy detection methods use some kind of a hash-function to reduce space requirements.

Test Results - Chunking Strategies

Hashed Breakpoint Chunking

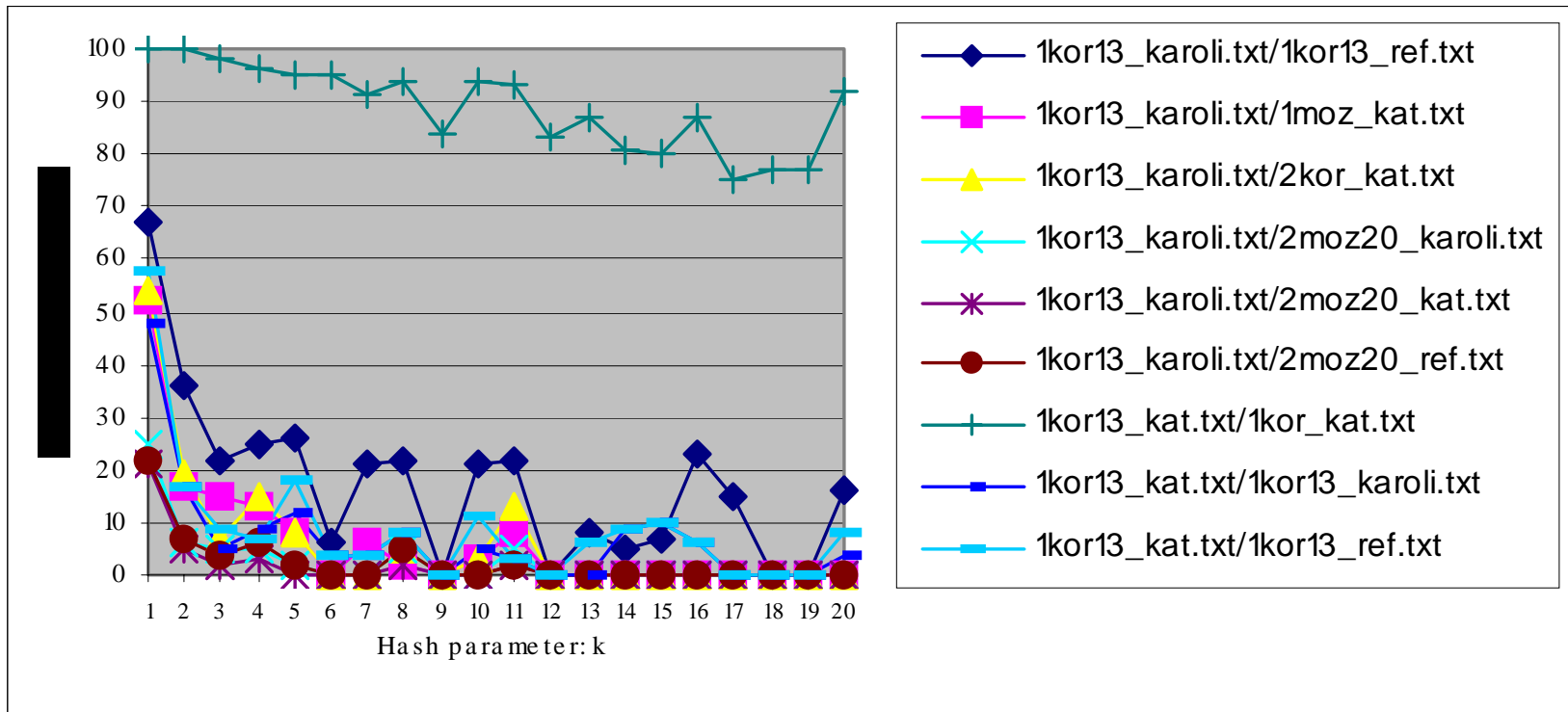
■ Average Chunk Length



Test Results - Chunking Strategies

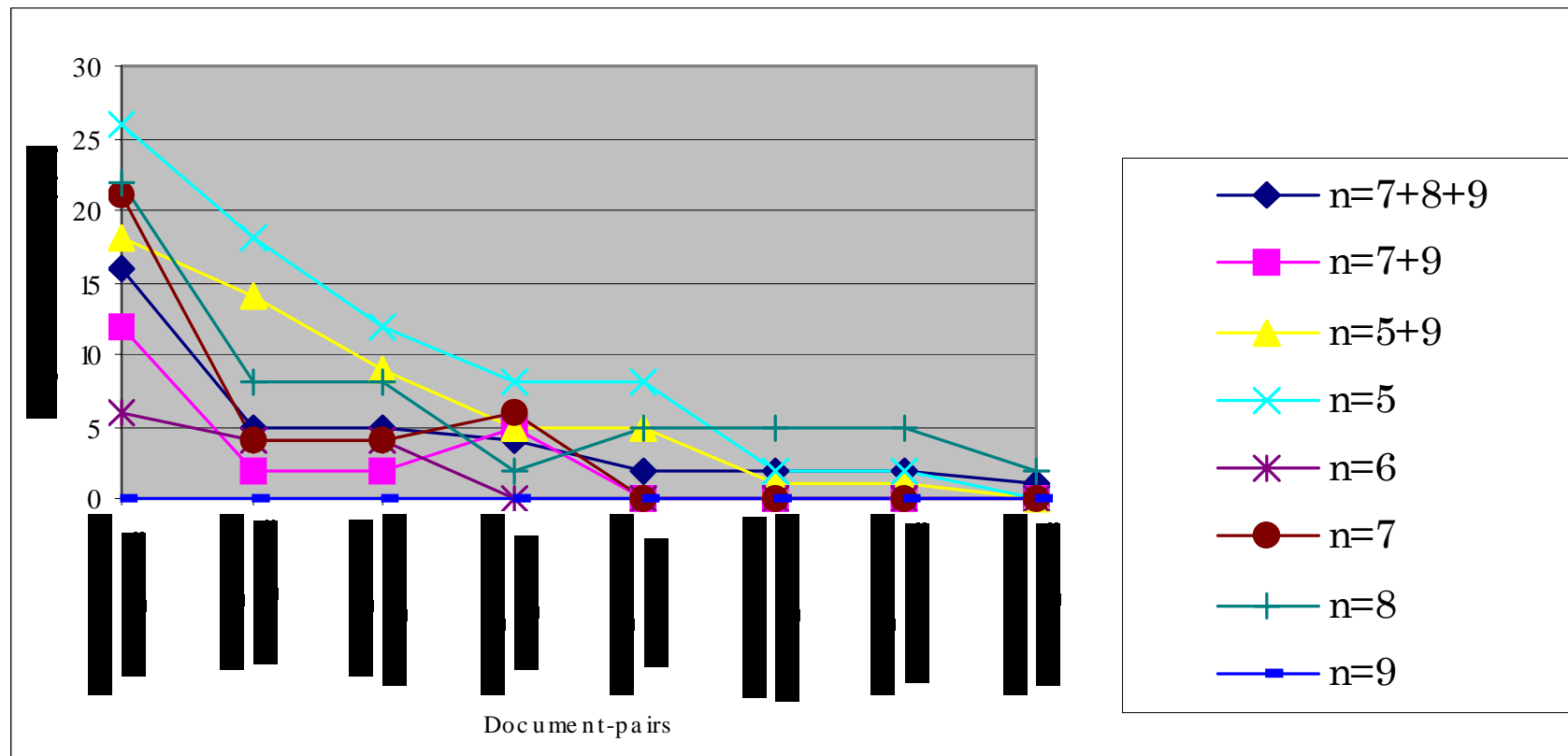
Hashed Breakpoint Chunking

■ Overlap



Multiple k-values

■ Overlap





False Positives

Method	bit-depth	false positives	false positive (%)
hashed breakpoint (k=6)	24	8434	1.6868
hashed breakpoint (k=9)	24	6790	1.3580
overlapping (k=6)	24	7115	1.4230
sentence	24	13954	2.7908
hashed breakpoint (k=6)	32	23	0.0046
hashed breakpoint (k=9)	32	21	0.0042
overlapping (k=6)	32	26	0.0052
sentence	32	15	0.0030



Fingerprint Selection

■ Sqrt

- Retain \sqrt{n} chunks whose lengths L are closest to m

■ Variance

- those chunks such that $|L-m| \leq bs$. Increase b , if necessary, until at least \sqrt{n} chunks are selected. Start with $b=0.1$



Test Results - Selection Strategy

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



Conclusion and Future Work

- Overlap Detection Techniques
- Hash-size
- Selection Strategy
- Future Work
 - Test Sets
 - System