Detecting near-duplicates for web crawling

Ziawasch Abedjan / Tobias Flach

- Simhash: Hashwert wird mit Hilfe von Dokumenteigenschaften bestimmt (ähnliche Dokumenten erhalten ähnliche Hashwerte)
- Beispiel:

  (a) Sein oder nicht sein.

  (b) Wenn jeder an sich denkt, ist an alle gedacht.

  (c) Wenn jeder an sich denkt, ist an jeden gedacht.

(a)	1	1	1	1	0	0	0	1	0
(b)	1	0	0	1	1	1	1	1	0
(c)	1	0	0	1	1	1	0	1	0

- $k_{ab} = 5$

  $k_{ac} = 4$

  $k_{bc} = 1$ (wahrscheinlich ein Fast-Duplikat)

- Sortierte Liste von Hashwerten H
- Für eine Liste von Anfragen, finde Fast-Duplikate in H (Hamming-Distanz ≤ k)
- Erzeuge einen Index über den d signifikantesten Bits

0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1
0	1	1	0	0	1	1	1
0	1	1	0	1	1	0	1
1	0	0	1	1	0	0	1
1	0	1	1	1	1	1	0
1	1	0	1	0	0	1	0
1	1	1	0	0	0	1	0

d = 3

k = 2

- Zeilen in Blöcke mit (fast) gleicher Größe aufteilen
- Erstelle alle Kombinationen aus n Blöcken (Kombination enthält ca. d Bits und mindestens k Blöcke sind einer Kombination nicht enthalten, hier: d = 3, k = 2)

0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1
0	1	1	0	0	1	1	1
0	1	1	0	1	1	0	1
1	0	0	1	1	0	0	1
1	0	1	1	1	1	1	0
1	1	0	1	0	0	1	0
1	1	1	0	0	0	1	0

Blockgröße = 2
Kombinationen aus 2 Blöcken
6 Möglichkeiten

0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1
0	1	1	0	0	1	1	1
0	1	1	0	1	1	0	1
1	0	0	1	1	0	0	1
1	0	1	1	1	1	1	0
1	1	0	1	0	0	1	0
1	1	1	0	0	0	1	0

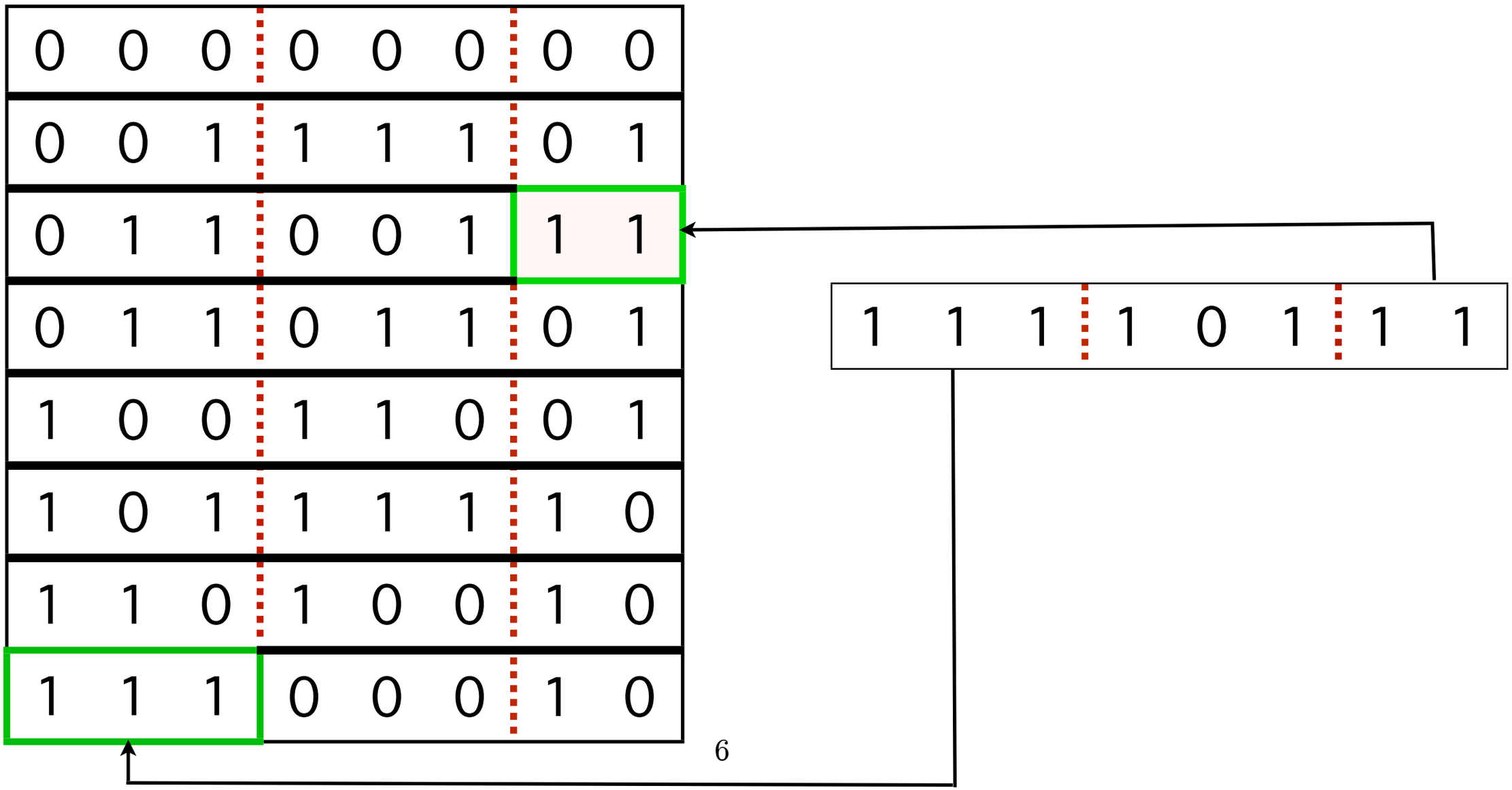
Blockgröße = 2 oder 3
Kombinationen aus 1 Block
3 Möglichkeiten

- Erstelle Indizes basierend auf den Bits der Kombinationen

0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	1
0	1	1	0	0	1	1	1
0	1	1	0	1	1	0	1
1	0	0	1	1	0	0	1
1	0	1	1	1	1	1	0
1	1	0	1	0	0	1	0
1	1	1	0	0	0	1	0

Blockgröße = 2 oder 3
Kombinationen aus 1 Block
3 Indizes

- Erstelle Blockkombinationen für die Anfragen
- Finde mögliche Fast-Duplikate (Kombinationsbits sind gleich)
- Vergleiche die verbleibenden Bits und prüfe ob Hamming-Distanz im Toleranzbereich ist



- MAP: Liste der Hashwerte in Blöcke aufteilen und an jeden Cluster einen Block + Anfrageliste übermitteln
- Alle Fast-Duplikate pro Block finden

- REDUCE: Fast-Duplikate der disjunkten Blöcke in einer Map mit der Anfrage als Schlüssel zusammenfügen