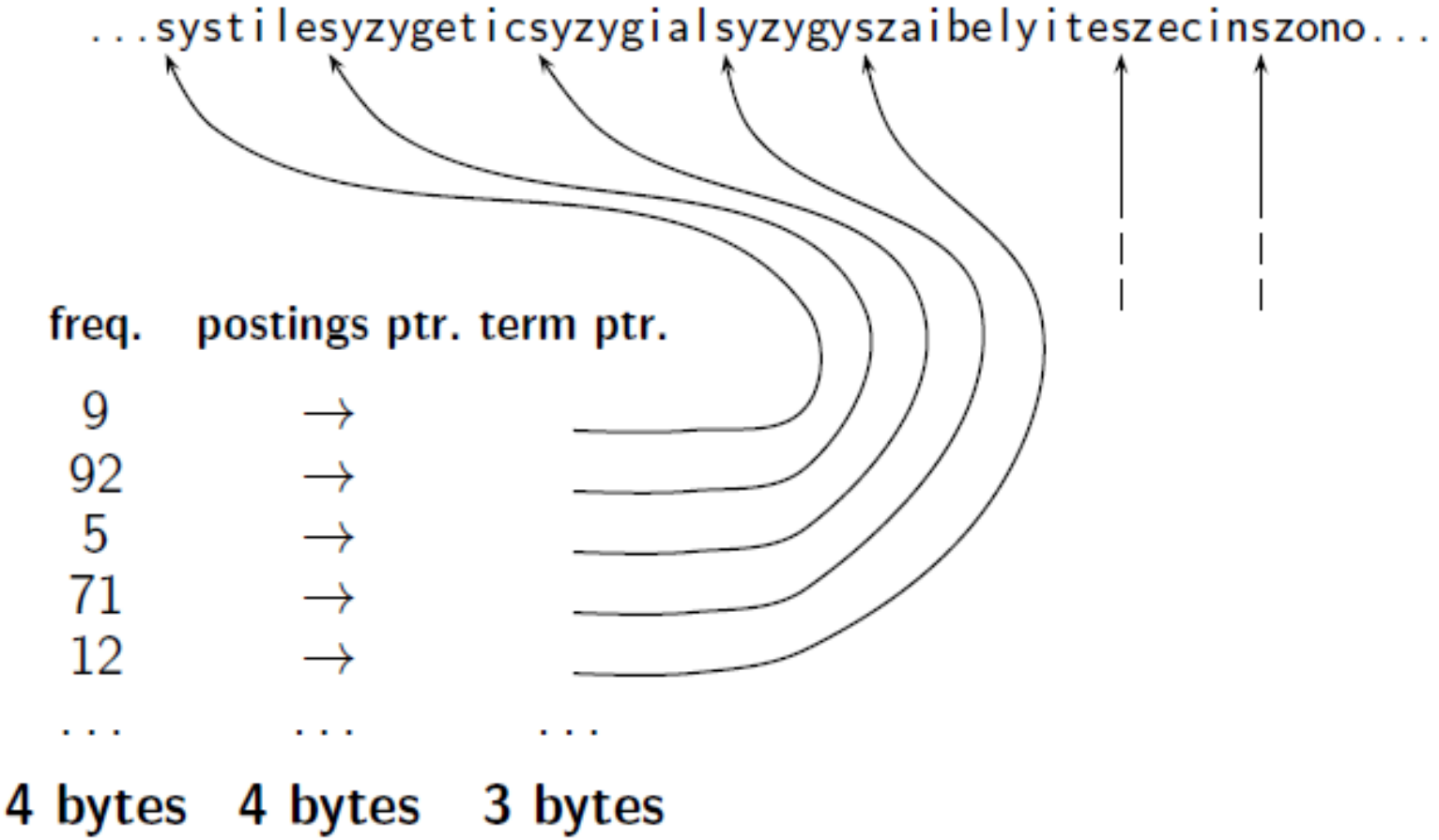
Information Retrieval
Topic: Index Compression (Part-2)
Lecture-22

Prepared By

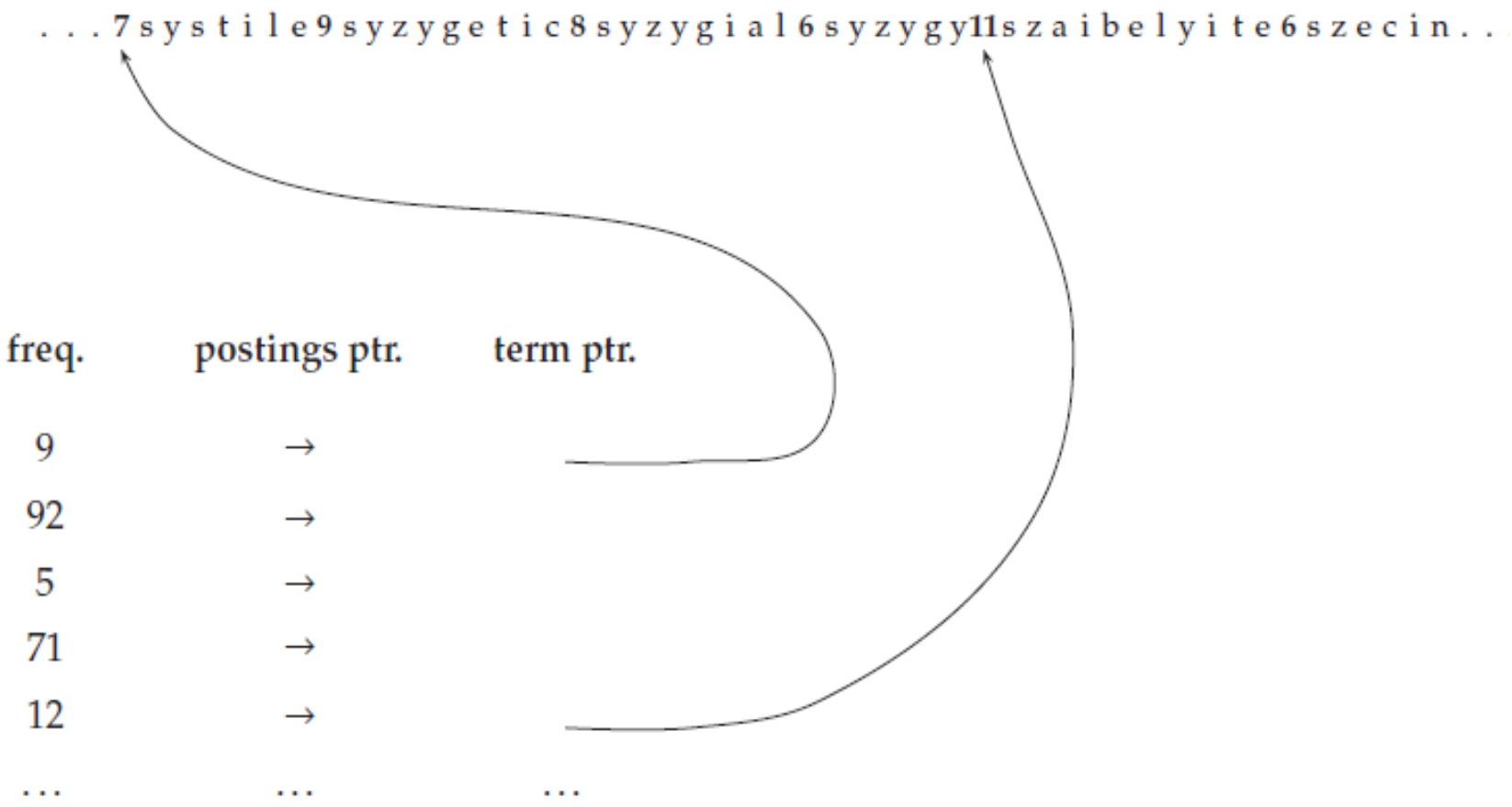
Dr. Rasmita Rautray & Dr. Rasmita Dash
Associate Professor
Dept. of CSE

Dictionary as a string



we need 400,000 × (4 + 4 + 3 + 8) = 7.6 MB for the Reuters-RCV1 dictionary: 4 bytes each for frequency and postings pointer, 3 bytes for the term pointer, and 8 bytes on average for the term. So we have reduced the space requirements by one third from 11.2 to 7.6 MB

Dictionary compression as Blocked storage



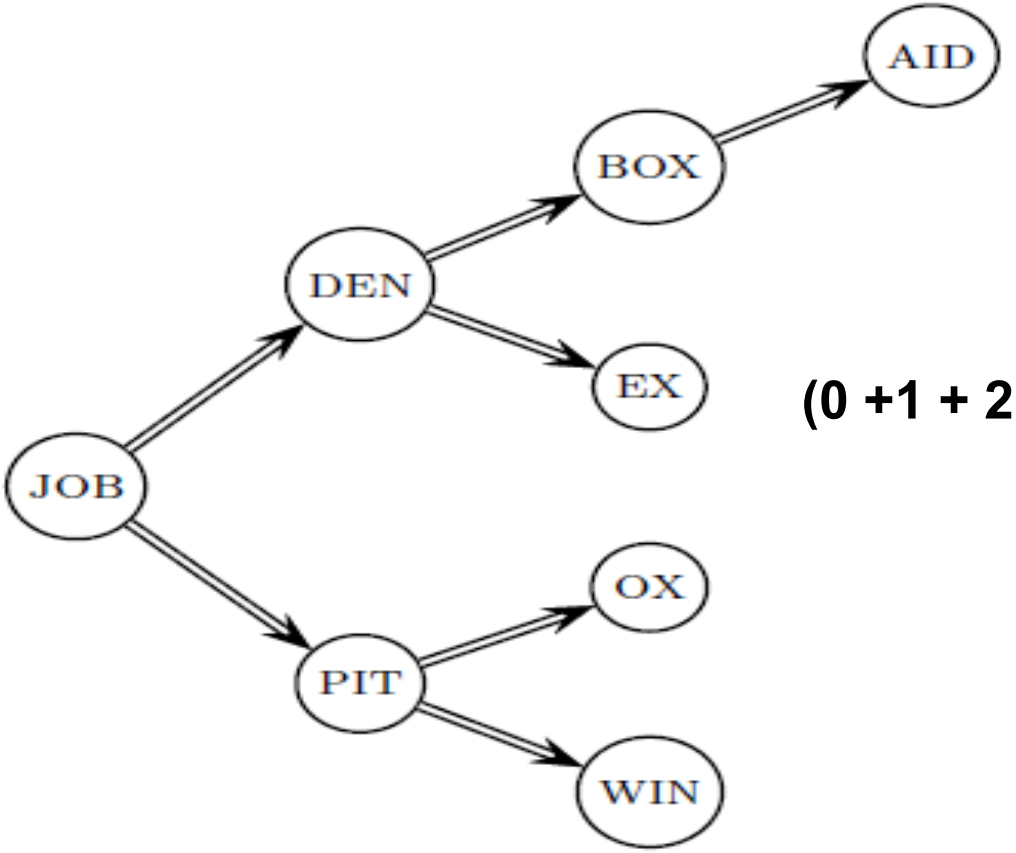
Space for dictionary as a string with blocking

Example: block size k = 4

- Where we used 4 × 3 bytes for term pointers without blocking . . .
- . . .we now use 3 bytes for one pointer plus 4 bytes for indicating the length of each term.
- We save 12 − (3 + 4) = 5 bytes per block.
- Total savings: 400,000/4 ∗ 5 = 0.5 MB
- This reduces the size of the dictionary from 7.6 MB to 7.1 MB.

- By increasing the block size *k,* we get better compression.
- There is a tradeoff between compression and the speed of term lookup.

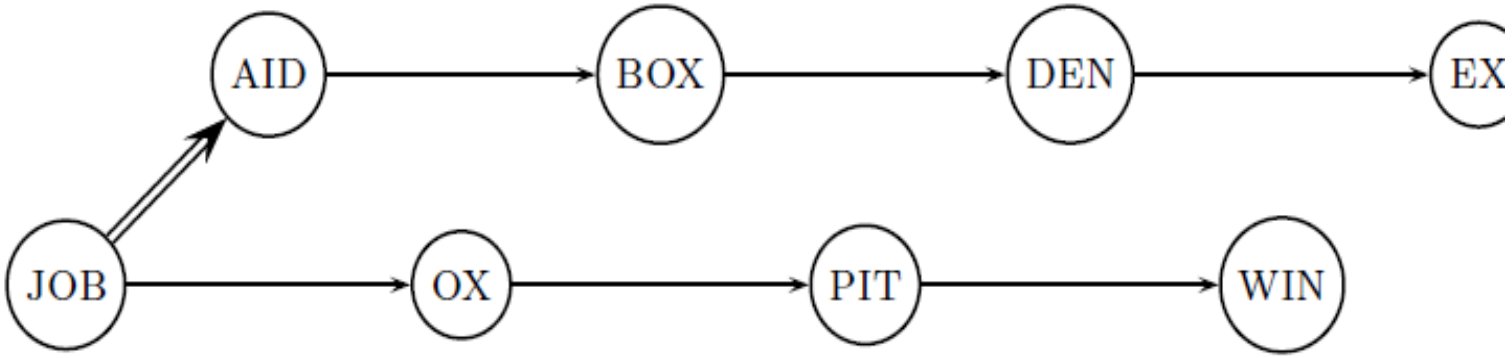
Lookup of a term without blocking



(0 +1 + 2+ 3+ 2 +1 +2 + 2)/8 ≈ 1.6 steps

Search of the uncompressed dictionary

Lookup of a term with blocking: (slightly) slower



(0+1+2+3+4+1+2+3)/8 = 2 steps on average, ≈ 25% more

Dictionary compressed by blocking with $k = 4$

Front coding

- One block in blocked compression (k = 4) . . .

8 a u t o m a t a 8 a u t o m a t e 9 a u t o m a t i c 10 a u t o m a t i o n

⇓

- . . . further compressed with front coding.

8 a u t o m a t ∗ a 1 ⋄ e 2 ⋄ i c 3 ⋄ i o n

Dictionary compression for Reuters: Summary

data structure	size in MB
dictionary, fixed-width	11.2
dictionary, term pointers into string	7.6
$\sim$, with blocking, $k = 4$	7.1
$\sim$, with blocking & front coding	5.9

Thank You