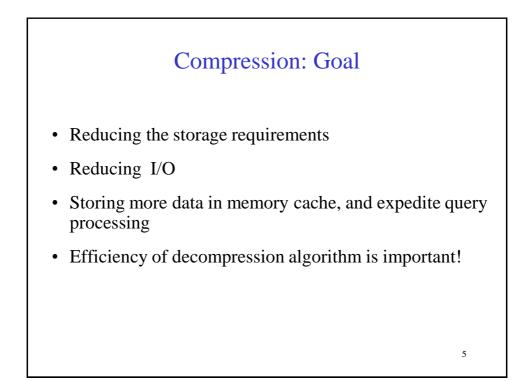
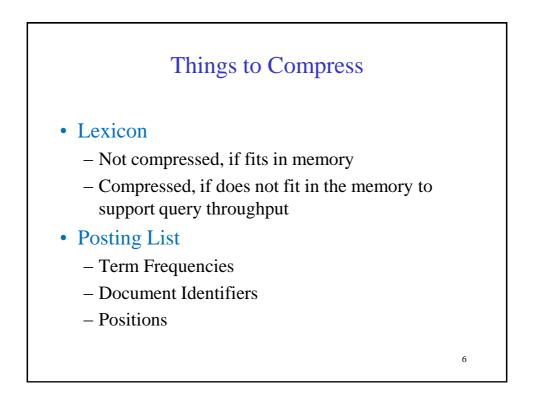


Sample Collection/Index Size before & after Compression (from: Information Retrieval, Buttcher, Clarke, Cormack)				
Collection	Collection Uncompressed	Collection Compressed (gzip)	Index Uncompressed	Index Compressed (vByte)
Shakespeare TREC4-5 Gov2	7.5 MB 1904.5 MB 425.8 GB	2.0 MB 582.9 MB 79.9 GB	10.5 MB 2331.1 MB 328.3 GB	2.7 MB 533.0 MB 62.1 GB
				4





Lexicon Compression

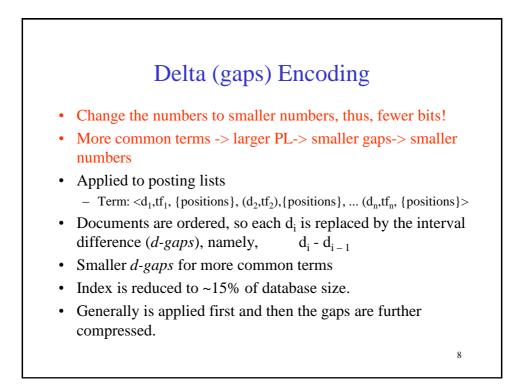
• Terms are in lexicographical order, sharing a common prefix. To prevent storing duplicates, use *Front coding (generally saves ~40%), as*

(preffix lenght, suffix lengh, suffix) <"book",(4,3,"ing"), (7,1,"s")><"book",(4,3,"let")...>

• Storing terms in lexicon as a string with pointers indicating end of a term and start of the next term.

7

• Hashing on terms



Compression Techniques of Inverted Indices

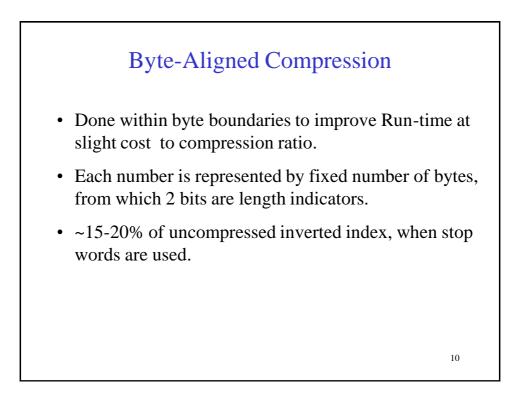
- Fixed Length
 - Byte Aligned

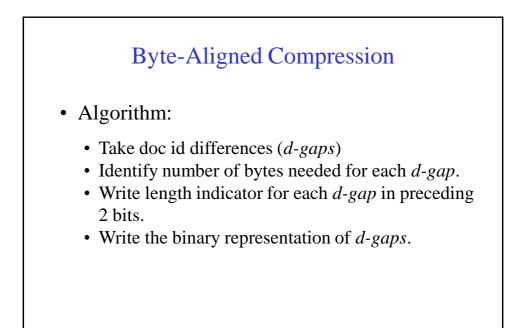
-

- Variable Length
 - Elias Encoding (γ), a family of universal codes

9

- ...





0 - 63	00xxxxxx
64 - (16K-1)	01xxxxxx xxxxxxxx
16K - (4M-1)	10xxxxxx xxxxxxxx xxxxxxxx
4M - (1G-1)	11xxxxxx xxxxxxxx xxxxxxx xxxxxxx
0	0000000
1	00000001
•••	
63	00111111
64	01000000 01000000
65	01000000 01000001
• The hope he	re is that the document distance between

