Modern Information Retrieval

Index compression

Hamid Beigy

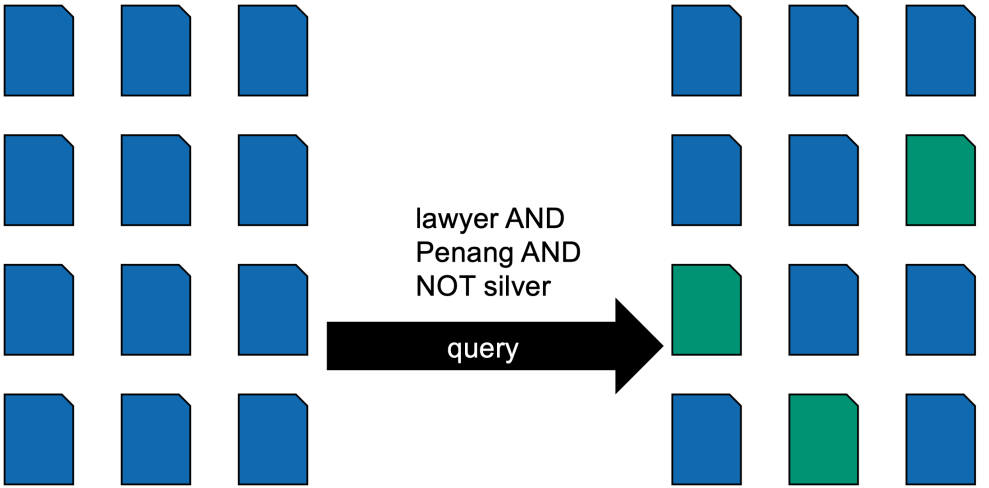
Sharif university of technology

February 23, 2025

Table of contents

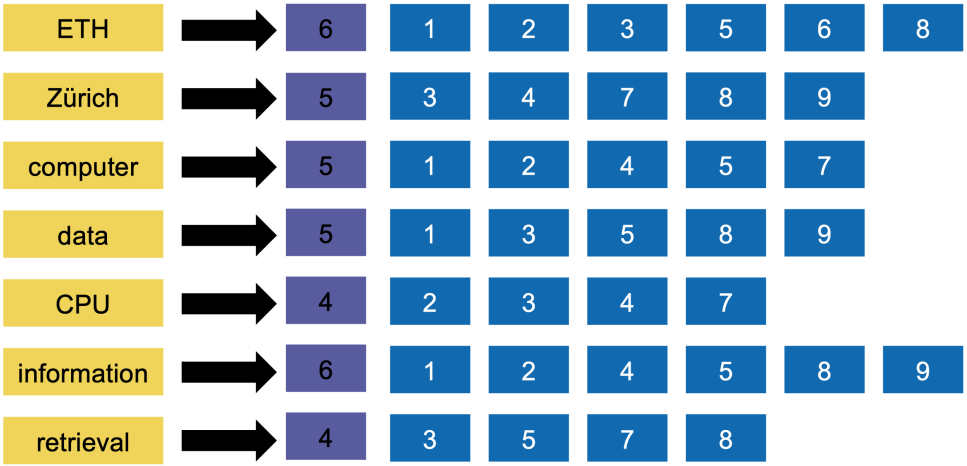
Introduction

Boolean model[1]



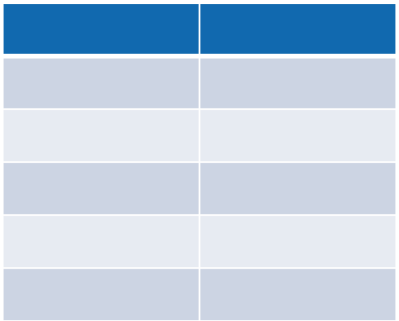
[1] Credit: Ghislain Fourny

Inverted index[2]

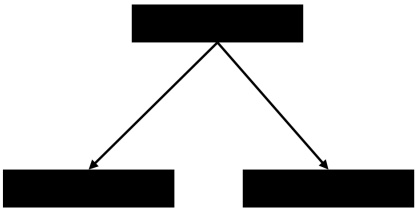


[2] Credit: Ghislain Fourny

Dictionary structures[3]



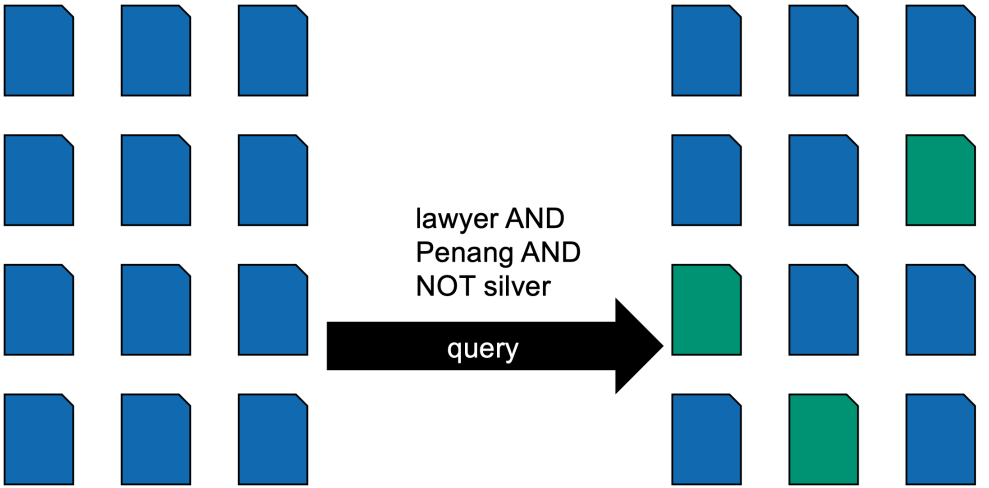
Hash tables



Trees (B, B+)

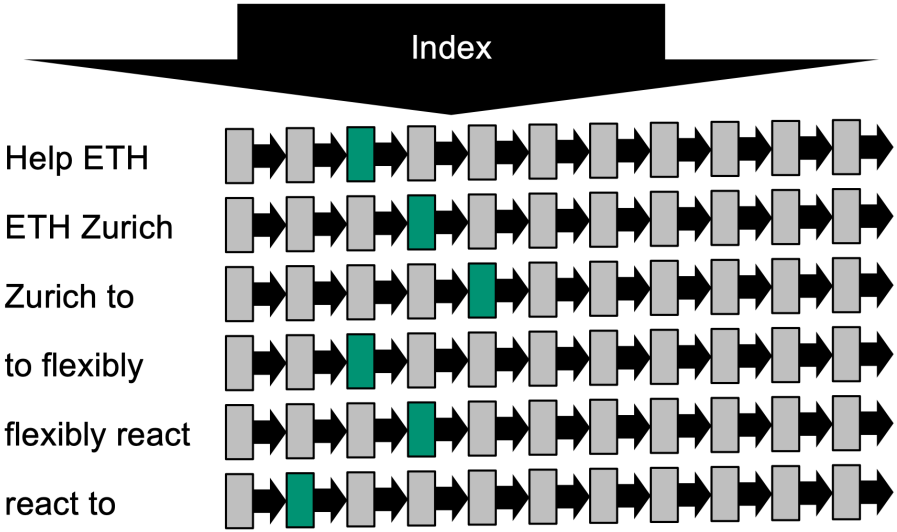
[3] Credit: Ghislain Fourny

Term mapping[4]



[4] Credit: Ghislain Fourny

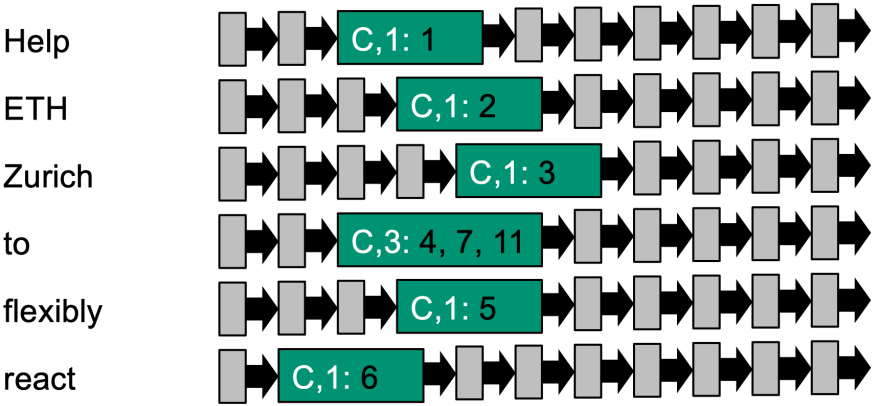
Help ETH Zurich to flexibly react to new challenges and to set new accents in the future.



[5] Credit: Ghislain Fourny

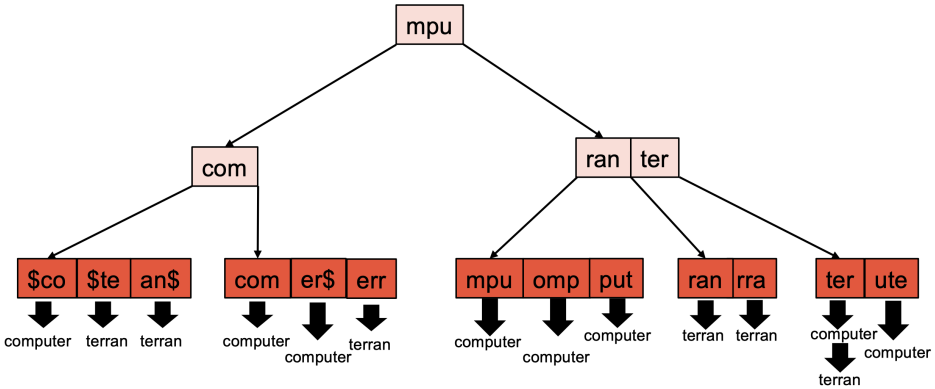
Positional index[6]

"ETH Zurich"|



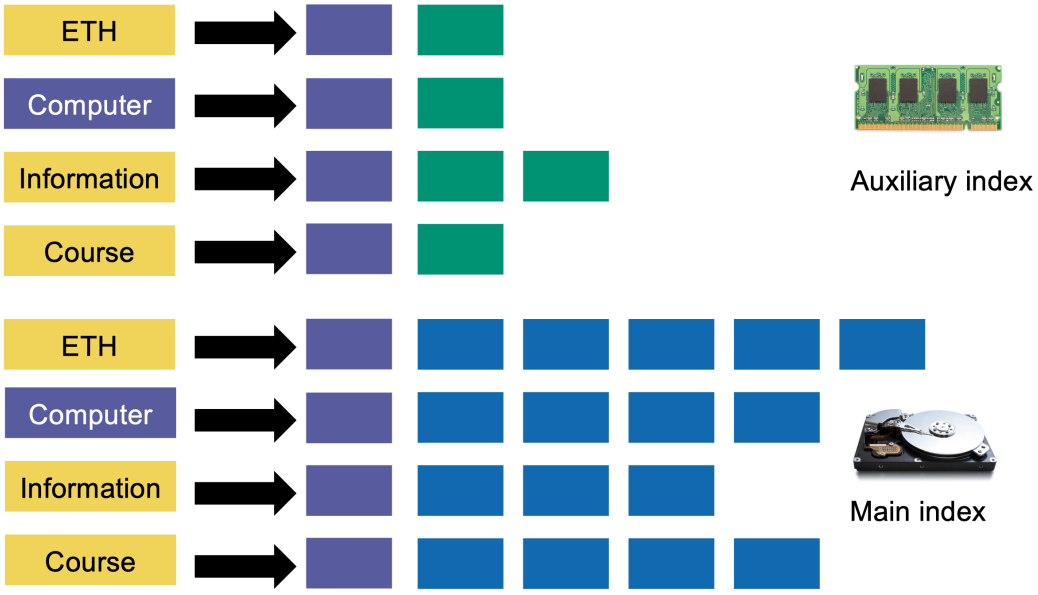
[6] Credit: Ghislain Fourny

Trigram index[7]



[7] Credit: Ghislain Fourny

Auxiliary index[8]



[8] Credit: Ghislain Fourny

Characterization of an index

Characterization of an index

Considering the Reuters-RCV1 collection

size of	dictionary			non-positional index			positional index		
	size	Δ	cum.	size	Δ	cum.	size	Δ	cum.
unfiltered	484,494			109,971,179			197,879,290		
no numbers	473,723	-2%	-2%	100,680,242	-8%	-8%	179,158,204	-9%	-9%
case folding	391,523	-17%	-19%	96,969,056	-3%	-12%	179,158,204	-0%	-9%
30 stop words	391,493	-0%	-19%	83,390,443	-14%	-24%	121,857,825	-31%	-38%
150 stop words	391,373	-0%	-19%	67,001,847	-30%	-39%	94,516,599	-47%	-52%
stemming	322,383	-17%	-33%	63,812,300	-4%	-42%	94,516,599	-0%	-52%

1. The vocabulary grows with the corpus size
2. Empirical law determining the number of term types in a collection of size M (Heap's law)

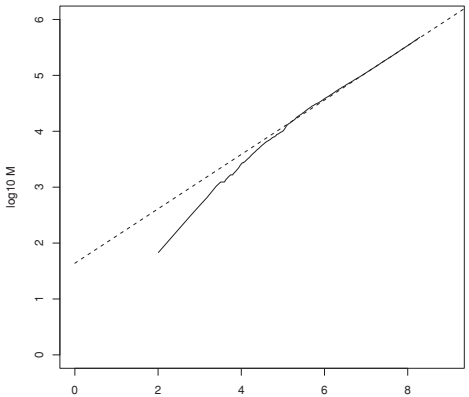
$$M = kT^b$$

T is the number of tokens, and k (growth-rate) and b are two parameters defined as:

$$b \approx 0.5 \qquad 30 \leq k \leq 100$$

3. On the REUTERS corpus for the first $1,000,020$ tokens (taking $k = 44$ and $b = 0.49$):

$$M = 44 \times 1,000,020^{0.5} = 38,323$$ The actual number: 38,365



Modeling the distribution of terms (Zipf's law)

1. We want understand how terms are distributed across documents.
2. We want know how many frequent vs. infrequent terms.
3. In natural language, there are a few very frequent terms and very many very rare terms.
4. **Zipf's law:** The i^{th} most frequent term has frequency cf_i as

$$\text{cf}_i \propto \frac{1}{i}$$

   cf_i is collection frequency: the number of occurrences of the term t_i in the collection.

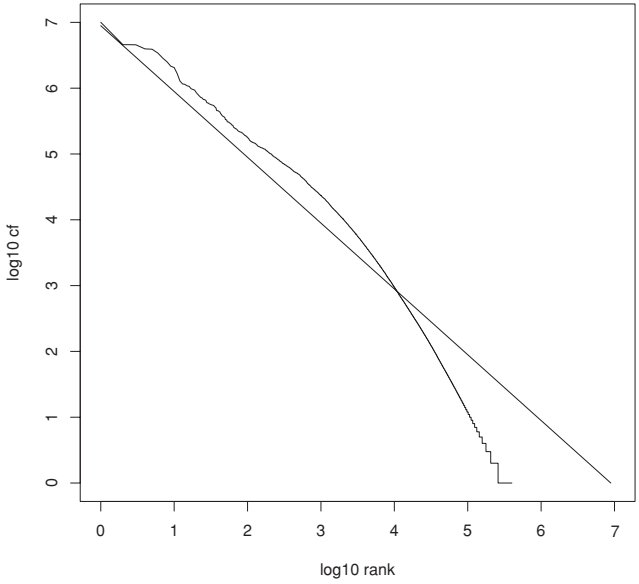
5. It means: rank of a word (cf_i) times its frequency (i) is approximately a constant (k).
6. So if the most frequent term (*the*) occurs cf_1 times, then

$$\text{cf}_2 = \frac{1}{2}\text{cf}_1$$
$$\text{cf}_3 = \frac{1}{3}\text{cf}_1$$
$$\vdots$$
$$\text{cf}_k = \frac{1}{k}\text{cf}_1$$

1. Equivalently, we can write Zipf's law as

$$\mathrm{cf}_i = ci^k$$

$$\log \mathrm{cf}_i = \log c + k \log i \quad \text{for} \quad k = -1$$

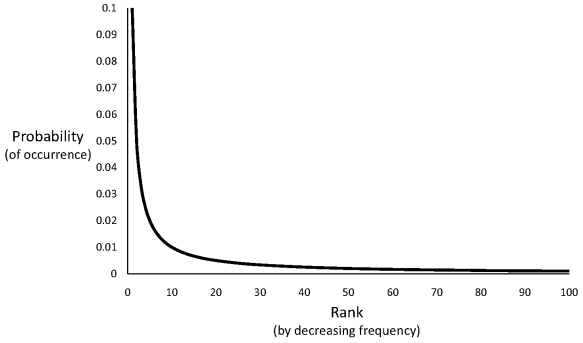


Modeling the distribution of terms (Zipf's law)

1. How about the probability of occurrence of a word?
2. The probability is:

   the frequency of the word divided by the total number of word occurrences in the text.
3. For Zipf's law, we have

$$p_i = \frac{c}{i} \qquad \text{for English, } c \approx 0.1.$$



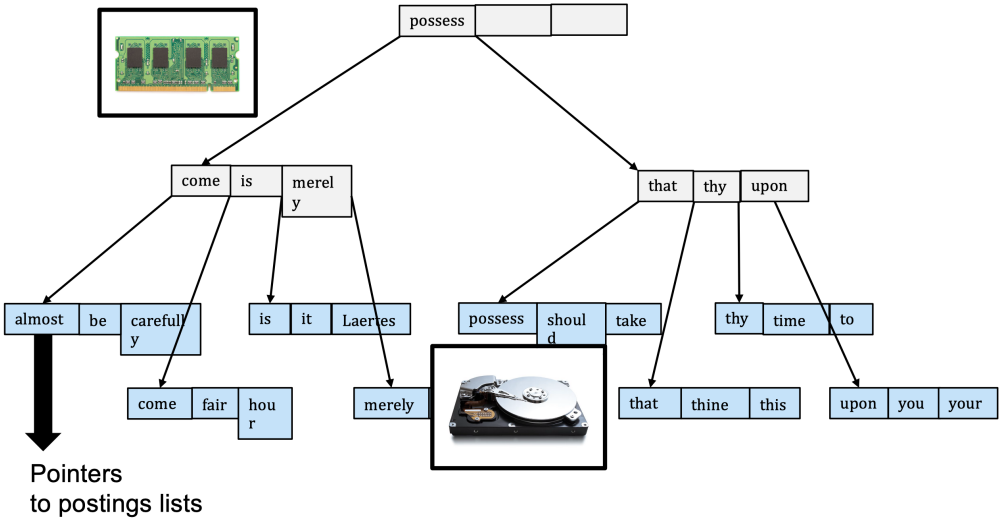
4. This figure shows how the frequency of word occurrence falls rapidly after the first few most common words.

Most frequent 50 words from AP89

Word	*Freq.*	*r*	$P_r(\%)$	$r.P_r$	*Word*	*Freq*	*r*	$P_r(\%)$	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Dictionary compression

B-Tree for Dictionary[9]



[9] Credit: Ghislain Fourny

Dictionary compression

1. The dictionary is small compared to the postings file.
2. But we want to keep it in memory.
3. We compress the dictionary because of
   - Reduce the response time of an IR system
   - We want design the search system for systems with limited hardware such as cell phones, onboard computers.
   - Fast startup time
   - Sharing resurces with other applications.
4. So compressing the dictionary is important.

Index format with fixed-width entries

term	document frequency	pointer to postings list	postings list
a	656,265	$\longrightarrow$	...
aachen	65	$\longrightarrow$	...
...	...	...	...
zulu	221	$\longrightarrow$	...
40	4	4	space needed

1. Total space for using Unicode and fixed-width entries (term-length=20):

$$M \times (2 \times 20 + 4 + 4) = 400{,}000 \times 48 = 19.2 \text{ MB}$$

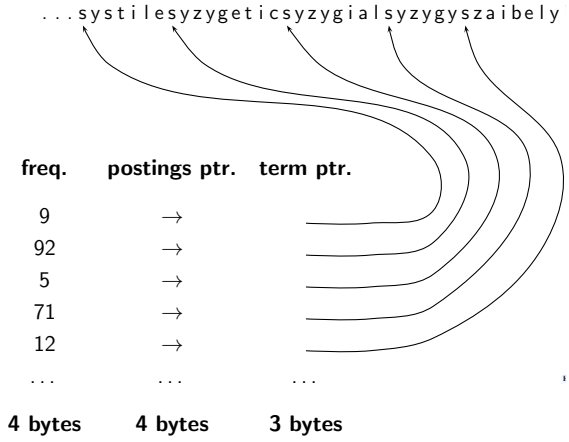
2. Without using Unicode:

$$M \times (20 + 4 + 4) = 400{,}000 \times 28 = 11.2 \text{ MB}$$

3. Remarks
   - The average length of a word type for REUTERS is 7.5 bytes
   - With fixed-length entries, a one-letter term is stored using 20 bytes!
   - Some very long words (such as hydrochlorofluorocarbons) cannot be handled.
   - How can we extend the dictionary representation to save bytes and allow for long words?

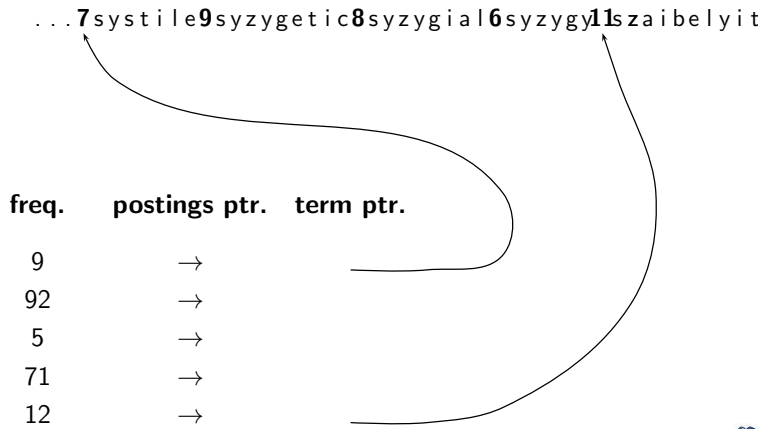
Compressing the dictionary

Dictionary as a string



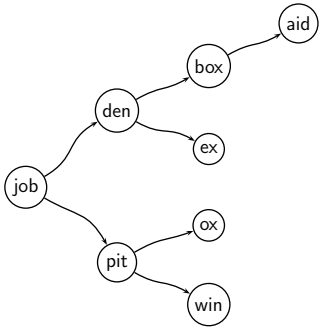
1. 3 bytes per pointer into string (need $\log_2(400000 \times 8) \approx 22$ bits to resolve 400,000 positions).
2. 8 chars (on average) for term in string
3. Using Unicode: $400{,}000 \times (4 + 4 + 3 + 2 \times 8) = 10.8$MB (compared to 19.2 MB for fixed-width)
4. Without using Unicode: $400{,}000 \times (4 + 4 + 3 + 8) = 7.6$ MB (compared to 11.2 MB for fixed-width)

Block storage



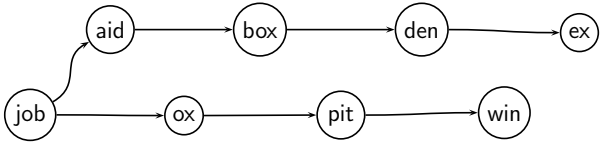
1. Let us consider blocks of size k
2. We remove $k-1$ pointers, but add k bytes for term length
3. Example: $k=4$, $(k-1)\times 3$ bytes saved (pointers), and 4 bytes added (term length) $\rightarrow$ 5 bytes saved
4. Space saved: $400{,}000 \times (\frac{1}{4}) \times 5 = 0.5$ MB (dictionary reduced to 10.3 MB and for non-Unicode 7.1MB)
5. Why not taking $k > 4$?

1. Uncompressed dictionary



Average search cost: $(0+1+2+3+2+1+2+2)/8 \approx 1.6$ steps

2. Compressed dictionary with blocking



Average search cost: $(0+1+2+3+4+1+2+3)/8 \approx 2$ steps

Front coding

1. Many words have the same prefix. We can write common prefix once.
2. One block in blocked compression ($k = 4$)

   8automata8automate9automatic10automation
3. Compressed with front coding.

   8automat∗a1⋄e2⋄ic3⋄ion
4. End of prefix marked by $*$
5. Deletion of prefix marked by $\diamond$

Dictionary compression for Reuters

representation	size (unicode)	size (non-unicode)
dictionary, fixed-width	19.2MB	11.2MB
dictionary as a string	10.8MB	7.6MB
$\sim$, with blocking, $k = 4$	10.3MB	7.1MB
$\sim$, with blocking & front coding	7.9MB	5.9MB

Compressing the posting lists

Compressing the posting lists

1. The REUTERS collection has
   - about 800 000 documents,
   - each having 200 tokens
2. Since tokens are encoded using 6 bytes, the collection's size is 960 MB
3. A docId must cover all the collection, *i.e.* must be $log_2 800,000 \approx 20$ bits
4. If the collection includes about $100,000,000$ postings, the size of the posting lists is $100,000,000 \times 20/8 = 250MB$
5. How to compress these postings ?
6. Idea: most frequent terms occur close to each other.
7. **We encode the gaps between occurrences of a given term**

Gap encoding

	encoding	postings list									
the	docIDs	...		283042		283043		283044		283045	.
	gaps				1		1		1		.
computer	docIDs	...		283047		283154		283159		283202	.
	gaps				107		5		43		.
arachnocentric	docIDs	252000		500100							
	gaps	252000	248100								

Furthermore, small gaps are represented with shorter codes than big gaps.

Two techniques

- Variable-length byte-codes (Byte-level)
- γ-codes (Bit-level)

Compressing the posting lists

Using variable-length byte-codes

Using variable-length byte-codes

1. Variable-length byte encoding uses an integral number of bytes to encode a gap
   - First bit := *continuation byte*
   - Last 7 bits := *part of the gap*
2. The first bit is set to 1 for the last byte of the encoded gap, 0 otherwise
3. Example: a gap of size 5 is encoded as 10000101

Example

docIDs	824	829	215406
gaps		5	214577
VB code	00000110 10111000	10000101	00001101 00001100 10110001

What is the code for a gap of size 1283?

4. The posting lists for the REUTERS collection are compressed to 116 MB with this technique (original size: 250 MB)
5. The idea of representing gaps with variable integral number of bytes can be applied with units that differ from 8 bits
6. Larger units can be processed (decompression) quicker than small ones, but are less effective in terms of compression rate

Compressing the posting lists

Using γ-codes

Using γ-codes

1. Idea: representing numbers with a *variable bit* code
2. **Unary code:** the number n is encoded as: $\overbrace{11\ldots}^{\text{n times}}\ 0$
   (not efficient)
3. γ**-code:** variable encoding done by splitting the representation of a gap as follows:

length	offset

   - *offset* is the binary encoding of the gap (without the leading 1)
   - *length* is the unary code of the offset size

Unary and γ-codes

number	unary code	length	offset	γ code
0	0			
1	10	0		0
2	110	10	0	10,0
3	1110	10	1	10,1
4	11110	110	00	110,00
9	1111111110	1110	001	1110,001
13		1110	101	1110,101
24		11110	1000	11110,1000
511		111111110	11111111	111111110,11111111
1025		11111111110	0000000001	11111111110,0000000001

Example

1. Given the following γ-coded gaps:

   1110001110101011111101101111011

2. Decode these, extract the gaps, and recompute the posting list
3. γ-decoding :
   - first reads the length (terminated by 0),
   - then uses this length to extract the offset,
   - and eventually prepends the missing 1

   1110001 - 11010 - 101 - 11111011011 - 11011

Compression of Reuters: Summary

representation	size in MB Unicode	size in MB non-unicode
dictionary, fixed-width	19.2	11.2
dictionary, term pointers into string	10.8	7.6
$\sim$, with blocking, $k = 4$	10.3	7.1
$\sim$, with blocking & front coding	7.9	5.3
collection (text, xml markup etc)	3600.0	3600.0
collection (text)	960.0	960.0
term incidence matrix	40,000.0	40,000.0
postings, uncompressed (32-bit words)	400.0	400.0
postings, uncompressed (20 bits)	250.0	250.0
postings, variable byte encoded	116.0	116.0
postings, γ encoded	101.0	101.0

Conclusion

Conclusion

1. γ-codes achieve better compression ratios (about 15 % better than variable bytes encoding), **but** are more complex (expensive) to decode
2. This cost applies on query processing $\rightarrow$ trade-off to find
3. The objectives announced are met by both techniques, recall:
   - reducing the disk space needed
   - reducing the time processing, by using a *cache*
4. The techniques we have seen are *lossless compression* (no information is lost)
5. *Lossy compression* can be useful, *e.g.* storing only the most relevant postings (more on this in the ranking lecture)

References

Reading

1. Chapters 5 of Information Retrieval Book[10]
2. Sections 4.2 and 5.4 of Search Engines - Information Retrieval in Practice Book[11]

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

[11] W. Bruce Croft, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.

References


Croft, W. Bruce, Donald Metzler, and Trevor Strohman (2009). *Search Engines - Information Retrieval in Practice*. Pearson Education.


Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Questions?