On Implementation of Word-based Compression Methods

Jakub Jaroš Petr Procházka Jan Holub

Department of Computer Science and Engineering, FEE CTU Prague,
Karlovo náměstí 13, 121 35 Praha 2, Czech Republic

Memics 2008 + PSC

Table of Contents

1 Introduction
- Basic Notions
- Related Work

2 Open Dense Code
- Basic Ideas
- Experiments

3 Summary

The Data Compression

Process to reduce time and/or space.

- Lossless data compression—fully reversible process.
- Dictionary methods:
  - use (build up) the dictionary,
  - LZ77, LZ78, LZW, WBLZW, WLZW.
- Statistical methods:
  - use statistical information (probability of occurrences),
  - Huffman code, Arithmetic code, WAC, Dense codes.
- Character-based X Word-based approach.

The Word-based approach

- Efficient for textual data (natural languages, formal languages...)
- Use words instead of characters as the symbols of the alphabet.
- Strictly alternating sequence of *words* and *non-words*.
- Faster adaptation to the encoded data.

Table of Contents

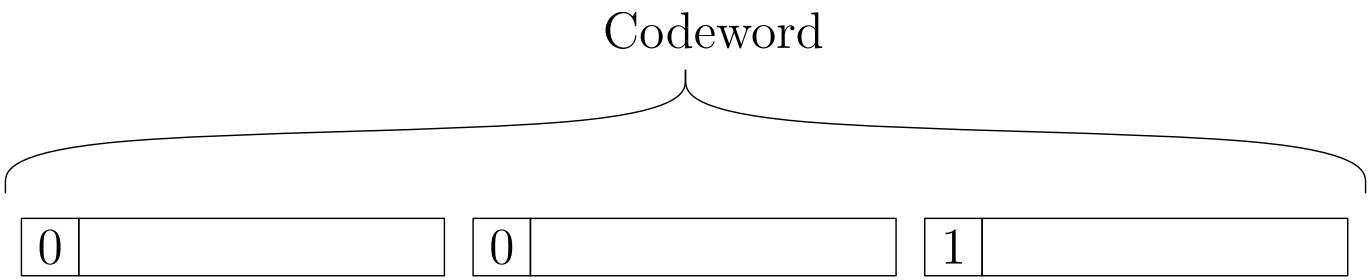
1 Introduction
- Basic Notions
- Related Work

2 Open Dense Code
- Basic Ideas
- Experiments

3 Summary

End-Tagged Dense Code (ETDC)

- Word-based compression method proposed by Brisaboa et al. (2003).
- Byte-oriented—improves the speed of compression and decompression.
- Uses the probability to define the rank of each word.
- Shorter codeword is assigned to the word with higher rank.
- Structure of codewords:
  - sequence of blocks,
  - the most important bit—to define the last block.



(s, c)-Dense Code (SCDC)

- Word-based compression method proposed by Brisaboa et al. (2007).
- Similar to ETDC.
- Differs between *Continuers* and *Stoppers*.
- Codeword is the sequence *Continuers* closed by one *Stopper*.

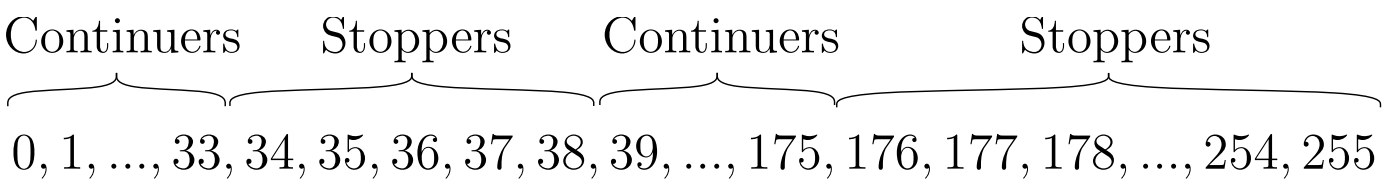


Table of Contents

1 Introduction
- Basic Notions
- Related Work

2 Open Dense Code
- Basic Ideas
- Experiments

3 Summary

Open Dense Code (ODC)

- Generalized concept of dense coding.
- Covers ETDC and SCDC.
- Provides a frame for definition of many other dense code schemas.

Definition

The b-ary Open Dense Code (ODC) is a couple $\langle b, G \rangle$ where b is a size of block and $G = (N, T, P, S)$ is a grammar defining syntax of the code. ODC assigns to the r-th most frequent symbol (starting with $r = 0$) a codeword c_r of k blocks, which satisfies following conditions:

1. $c_r \in L(G)$,
2. c_r is not a prefix of any other codeword $c_i \in L(G)$,
3. $\sum_{i=1}^{k-1} \prod_{j=1}^{i} v_j \leq r < \sum_{i=1}^{k} \prod_{j=1}^{i} v_j$, where v_j is number of codewords covered by block j.

Dynamic ODC (*dodc3*)

- Schema strongly adjusted to natural language compression.

$b = 8;\ G(N, T, P, S):$
$N = \{Codeword\}$
$T = \{a, b, c, d, e\}$
$S = Codeword$

$P:$ $Codeword \rightarrow a$
$Codeword \rightarrow b\ c$
$Codeword \rightarrow d$
$Codeword \rightarrow e$

a is 1B word, $a \in \{1, ..., 159\}$
b is 1st byte of 2B word, $b \in \{160, ..., 223\}$
c is 2nd byte of 2B word, $c \in \{0, ..., 255\}$
d is 1B non-word, $d \in \{224, ..., 255\}$
e is 1B ESC symbol, $e = 0$

Byte	Meaning	# codewords
$\langle 00000000 \rangle$	ESC	1
$\langle 00000001 \rangle - \langle 10011111 \rangle$	1B alphanumeric word	159
$\langle 10100000 \rangle - \langle 11011111 \rangle$	2B alphanumeric word	16384
$\langle 11100000 \rangle - \langle 11111111 \rangle$	1B non-alphanumeric word	32

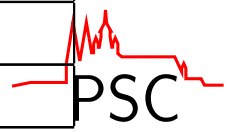


Table of Contents

1 Introduction
- Basic Notions
- Related Work

2 Open Dense Code
- Basic Ideas
- Experiments

3 Summary

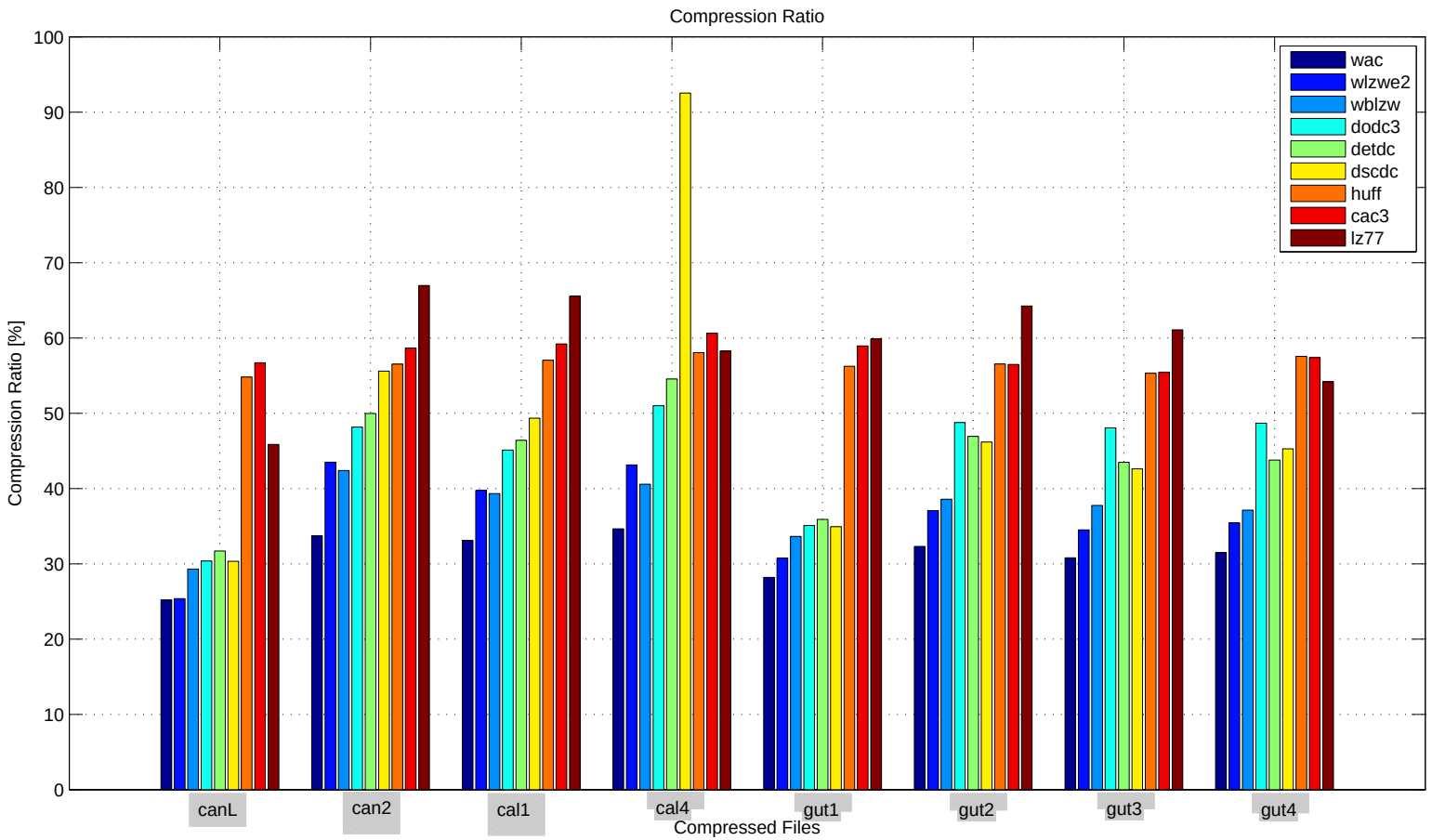
Tested Corpora

- Textual files in natural languages.
  - Canterbury and Large Canterbury Corpus.
  - Calgary Corpus.
  - Gutenberg Corpus.

File	Notation	Language	Source	Size [B]
bible.txt	canL	English	Large Cantebury	4,047,392
alice29.txt	can1	English	Cantebury	152,089
plrabn12.txt	can2	English	Cantebury	481,861
book1	cal1	English	Calgary	768,711
book2	cal2	English	Calgary	610,856
paper1	cal3	English	Calgary	53,161
paper2	cal4	English	Calgary	82,199
wrnpc11.txt	gut1	English	Gutenberg	3,217,389
17073-8.txt	gut2	Spanish	Gutenberg	1,805,493
2donq10.txt	gut3	Spanish	Gutenberg	2,106,147
8va5810.txt	gut4	Spanish	Gutenberg	979,749

Compression ratio

- Substantially better than character-based methods.
- Comparable to Word-based dictionary methods.
- Better than ETDC (SCDC) at large files.



Compression speed

- The best of Word-based methods.
- Comparable to Huffman encoding.

Alg./File	canL	can1	can2	cal1	cal2	cal3	cal4	gut1	gut2	gut3	gut4
wac	2.40	2.14	0.60	0.64	1.26	2.13	1.79	1.04	0.31	0.39	0.88
wlzwe2	0.91	0.94	0.75	0.70	0.78	0.80	0.79	0.89	0.42	0.42	0.45
wblzw	1.14	1.13	0.96	0.92	0.98	1.07	1.06	0.85	0.51	0.50	0.54
dodc3	**24.12**	**16.86**	**17.97**	**18.33**	**21.58**	**11.52**	**14.25**	**22.23**	**16.95**	**18.74**	**19.92**
detdc	22.71	2.68	9.35	12.02	7.87	1.01	1.54	20.05	13.27	13.48	10.27
dscdc	21.44	3.46	10.70	13.09	9.25	1.41	2.06	18.82	12.30	13.39	11.13
huff	**32.99**	**30.22**	**35.35**	**33.32**	**28.84**	**28.17**	**29.03**	**33.72**	**28.89**	**29.45**	**28.49**
cac3	8.04	7.65	7.88	7.72	7.42	7.04	7.47	7.77	7.83	8.03	7.79
lz77	0.03	0.03	0.02	0.02	0.03	0.08	0.05	0.02	0.02	0.02	0.03

Table: Compression speed in MB/s, general comparison

Summary

- Conclusion
  - Defined a generalized concept of dense coding called Open Dense Code (ODC).
  - Designed *dodc3* schema focused on natural language compression:
    - compression and decompression speed is comparable to character-based algorithms,
    - substantially better compression ratio.
  - Generally better compression ratio of word-based methods was proven.
- Future work
  - Tests on large and multilingual files.
  - Apply ODC concept to formal languages etc.

PSC

Summary

- Conclusion
  - Defined a generalized concept of dense coding called Open Dense Code (ODC).
  - Designed *dodc3* schema focused on natural language compression:
    - compression and decompression speed is comparable to character-based algorithms,
    - substantially better compression ratio.
  - Generally better compression ratio of word-based methods was proven.
- Future work
  - Tests on large and multilingual files.
  - Apply ODC concept to formal languages etc.

PSC

End of the presentation

Thank you for your attention.