[HCM-UIT] CS232.L21.KHCL – Tính Toán Đa phương tiện

NÉN DỮ LIỆU VỚI THUẬT TOÁN LZW

Lempel – **Z**iv – **W**elch Compression

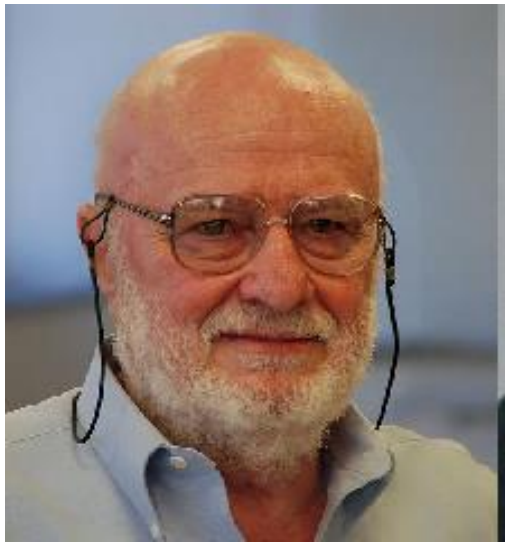
Sinh viên thực hiện: Hoàng Ngọc Bá Thi – MSSV:19522255 – KHCL2019.3

MỤC LỤC

I. THUẬT TOÁN LZW

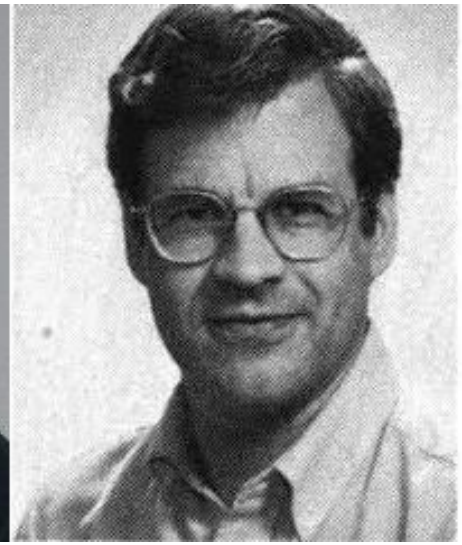
I.1 Giới thiệu về LZW

- Tên đầy đủ: **L**empel-**Z**iv-**W**elch.
- Năm 1977, Abraham Lempel và Jacob Ziv lần đầu tiên đưa ra khái niệm “nén từ điển”. Sau đó phát triển thành 1 họ giải thuật nén từ điển LZ (LZ77, LZ78).
- Năm 1984, Terry Welch cải tiến LZ thành một giải thuật mới hiệu quả hơn, đặt tên là LZW.



Abraham Lempel

Jacob Ziv



Terry Welch

I.1. Giới thiệu về LZW

- Phương pháp LZW dựa trên việc xây dựng từ điển cho các “chuỗi ký tự” đã từng xuất hiện trong văn bản, những “chuỗi ký tự” xuất hiện sau đó sẽ được thay thế bằng mã của nó trong bảng từ điển.

- Giải thuật LZW được sử dụng cho tất cả các loại file nhị phân. Nó thường được dùng để nén các loại văn bản, ảnh đen trắng, ảnh màu ... và là chuẩn nén cho các dạng ảnh GIF, TIFF...

I.2. Thuật toán nén LZW

- Hoạt động theo nguyên tắc là tạo ra một *từ điển động*.

- Từ điển là tập hợp những cặp *Khóa* và *Nghĩa* của nó. Trong đó *Khóa* được sắp xếp theo thứ tự nhất định, *Nghĩa* là một chuỗi con trong dữ liệu.

- Từ điển được xây dựng đồng thời với quá trình đọc dữ liệu. Sự có mặt của một chuỗi con trong từ điển khẳng định rằng chuỗi đó đã từng xuất hiện trong phần dữ liệu đã đọc. Thuật toán liên tục tra cứu và cập nhật từ điển sau mỗi lần đọc một ký tự ở dữ liệu đầu vào.

- Người ta thường dùng từ điển với kích thước 4096 (2^{12}) phần tử, độ dài chuỗi bit là 12 bits.

I.2. Thuật toán nén LZW

Cấu trúc từ điển có dạng như sau:

Khóa	Nghĩa	Ghi chú
0	<NULL>	Mã ASCII
1	<SOH>	Mã ASCII
…	…	…
255	<nbsp>	Mã ASCII
256	Chuỗi	Chuỗi trong dữ liệu
257	Chuỗi	Chuỗi trong dữ liệu
258	Chuỗi	Chuỗi trong dữ liệu
259	Chuỗi	Chuỗi trong dữ liệu
…	…	Chuỗi trong dữ liệu
4095	Chuỗi	Chuỗi trong dữ liệu

I.2. Thuật toán nén LZW

INPUT: ABCBCABCABCD => 12 * 8bits = 96 bits

*MÃ GIẢ:

Cài đặt từ điển với các chuỗi ký tự đơn (Bảng mã ASCII 0-255)

P = Ký tự đầu tiên của chuỗi INPUT

WHILE chưa đọc hết chuỗi INPUT

    C =ký tự tiếp theo trong chuỗi INPUT

    IF chuỗi P + C có trong từ điển

        P = P + C

    ELSE

        Xuất mã khóa cho P

        thêm P + C vào từ điển

        P = C

    END WHILE

Xuất mã khóa của P

INPUT	KHÓA	NGHĨA	OUTPUT
A (65)	65	A	
B (66)	256	AB	65
C (67)	257	BC	66
B	258	CB	67
C	-	-	-
A	259	BCA	257
B	-	-	-
C	260	ABC	256
A	261	CA	67
B	-	-	-
C	-	-	-
D (68)	262	ABCD	260
End of file	-	-	68

OUTPUT CODE: 65 66 67 257 256 67 260 68 => 8 * 9 bits = 72 bits

Tỉ số nén: 72 / 96 = 0,75

I.3. Phương pháp giải nén LZW

INPUT CODE: 65 66 67 257 256 67 260 68

* MÃ GIẢ:

Cài đặt từ điển với các chuỗi ký tự đơn
OLD = Mã khóa đầu tiên trong dãy INPUT
Xuất ra bản dịch của OLD
WHILE chưa đọc hết dãy INPUT
    NEW = Mã khóa tiếp theo trong dãy INPUT
    IF NEW không có trong từ điển
        S = bản dịch của OLD
        S = S + C
    ELSE
        S = bản dịch của NEW
    Xuất S
    C = ký tự đầu tiên của chuỗi S
    Thêm OLD + C vào từ điển
    OLD = NEW
END WHILE

INPUT	KHÓA	NGHĨA	OUTPUT
65 (A)	65	A	A
66 (B)	256	AB	B
67 (C)	257	BC	C
257 (BC)	258	CB	BC
256 (AB)	259	BCA	AB
67 (C)	260	ABC	C
260 (ABC)	261	CA	ABC
68 (D)	262	ABCD	D
End of file	-	-	-

OUTPUT STRING: ABCBCABCABCD

II. THỰC NGHIỆM THUẬT TOÁN LZW

II.1. Thực nghiệm trên văn bản

II.1. Thực nghiệm trên văn bản

Thực nghiệm 1: input nhỏ, có sự lặp lại các ký tự.

```
Input string: ABCBCABCABCD

[Encoding]
String  Output_Code     Addition
A       65              AB      256
B       66              BC      257
C       67              CB      258
BC      257             BCA     259
AB      256             ABC     260
C       67              CA      261
ABC     260             ABCD    262
D       68

Output Codes: 65 66 67 257 256 67 260 68
Compress Ratio = 0.75

[Decoding]
ABCBCABCABCD

Process returned 0 (0x0)   execution time : 0.047 s
Press any key to continue.
```

II.1. Thực nghiệm trên văn bản

Thực nghiệm 2: input lớn, có sự lặp lại các ký tự.

```
Input string: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis
nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nul
la pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
```

```
Output Codes: 76 111 114 101 109 32 105 112 115 117 260 100 111 108 257 32 115 105 116 32 97 109 101 116 44 32 99 111 110 115 101 99 116 278 117 114 275 100 262 105 115
 99 105 110 103 32 101 108 273 280 285 100 32 267 301 105 117 115 109 111 307 288 109 112 270 298 297 293 100 117 110 274 117 274 108 97 98 257 101 301 274 267 269 258
32 109 97 103 110 97 275 303 113 117 97 46 32 85 274 101 110 105 260 97 307 109 298 357 32 118 355 105 276 280 348 295 32 110 111 115 116 114 117 307 101 120 101 114 29
7 116 97 116 105 283 32 117 108 330 109 282 32 330 332 114 371 356 272 390 274 97 347 117 262 301 120 301 345 282 109 314 309 282 284 101 348 386 351 68 407 115 275 328
 334 105 377 339 337 320 110 32 258 112 258 104 355 100 382 273 261 434 118 268 117 112 385 288 364 302 443 101 115 285 281 105 392 265 308 268 333 301 117 32 102 117 1
03 367 274 110 391 330 32 112 97 399 386 290 351 69 120 99 101 449 101 290 271 298 274 111 99 99 97 286 386 281 448 105 100 386 498 373 434 437 111 501 355 279 271 325
274 298 499 108 477 32 370 32 111 102 102 105 297 345 441 285 377 326 340 268 303 404 356 260 501 301 375 396 331 257 265 46
Compress Ratio = 0.733146

[Decoding]
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exerci
tation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. E
xcepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
```

II.1. Thực nghiệm trên văn bản

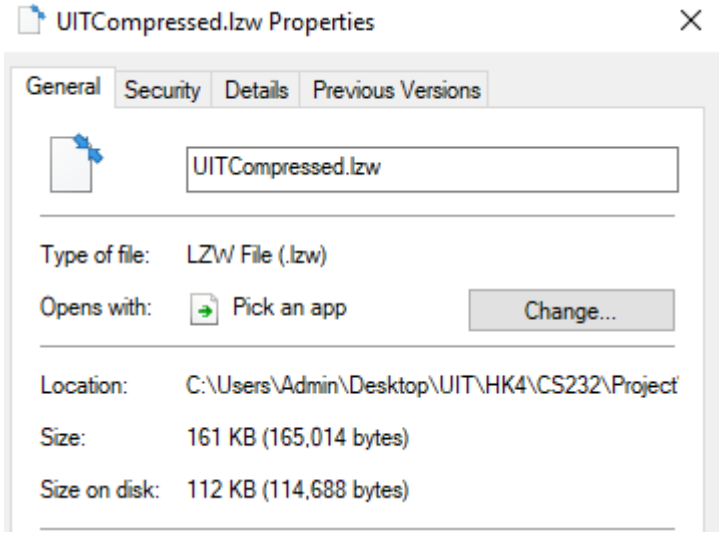
Thực nghiệm 3: input không có sự lặp lại các ký tự.

```
Input string: ABCDEFGabcdefg

[Encoding]
String  Output_Code     Addition
A       65              AB      256
B       66              BC      257
C       67              CD      258
D       68              DE      259
E       69              EF      260
F       70              FG      261
G       71              Ga      262
a       97              ab      263
b       98              bc      264
c       99              cd      265
d       100             de      266
e       101             ef      267
f       102             fg      268
g       103

Output Codes: 65 66 67 68 69 70 71 97 98 99 100 101 102 103
Compress Ratio = 1.125

[Decoding]
ABCDEFGabcdefg

Process returned 0 (0x0)   execution time : 0.030 s
Press any key to continue.
```

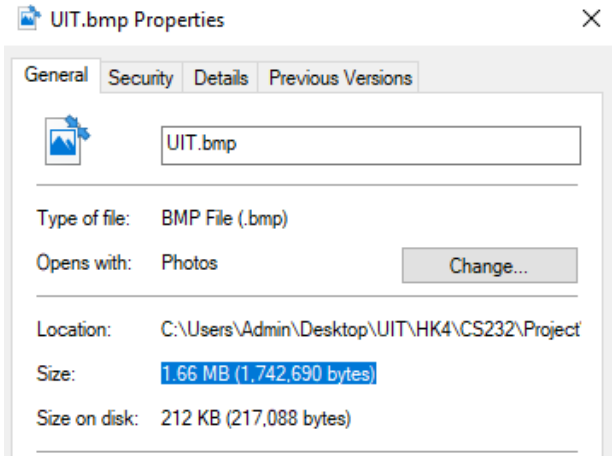
II.1. Thực nghiệm trên hình ảnh

Thực nghiệm 1: Ảnh trắng đen (Binary Image)

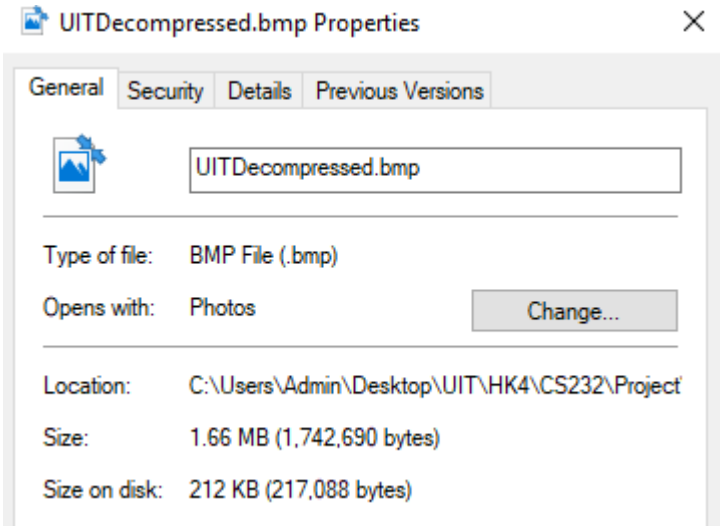
INPUT



Ảnh đã giải nén



- Kích thước hình ảnh ban đầu: **1.66 MB**
- Kích thước file nén: **161 KB**
- Kích thước ảnh giải nén: **1.66MB**
- Tỉ số nén: **0.09**



II.2. Thực nghiệm trên hình ảnh

Thực nghiệm 2: Ảnh có màu đơn giản vẽ bằng pixel.

Kích thước ảnh: 1.37 MB
Kích thước file nén: 188 KB
Kích thước ảnh giải nén: 1.37MB
Tỉ số nén: 0.13

Kích thước ảnh: 732 KB
Kích thước file nén: 167 KB
Kích thước ảnh giải nén: 732KB
Tỉ số nén: 0.22

Kích thước ảnh: 46.8 MB
Kích thước file nén: 5.62 MB
Kích thước ảnh giải nén: 46.8MB
Tỉ số nén: 0.12

II.2. Thực nghiệm trên hình ảnh

Thực nghiệm 3: Ảnh chụp thực tế, có độ phức tạp về màu sắc

Kích thước ảnh: 10.5 MB
Kích thước file nén: 22.1 MB
Kích thước ảnh giải nén: 10.5 MB
Tỉ số nén: 2.1
$\Rightarrow$ File nén có kích thước lớn hơn file gốc.

III. KẾT LUẬN

III. KẾT LUẬN

Ưu điểm:

- Đơn giản, dễ cài đặt, chạy nhanh.
- Là kỹ thuật nén không mất mát.
- Đạt hiệu quả cao với những văn bản tiếng Anh, vì sự lặp lại các ký tự, các chuỗi là thường xuyên.
- Không chỉ văn bản mà LZW còn đạt hiệu quả cao với hình ảnh trắng đen, ảnh pixel có ít độ phức tạp màu.
- Không cần kèm thêm từ điển vào dữ liệu đã nén.

Khuyết điểm:

- Với những thông tin không có sự lặp lại thì không thể nén được, thậm chí file nén còn có kích thước lớn hơn file gốc.
- Chỉ đạt hiệu quả cao với dữ liệu dạng văn bản, hình ảnh trắng đen, ảnh pixel đơn giản, còn đối với ảnh chụp thực tế với độ phức tạp cao thì không được như thế, thậm chí không nén được.
- Từ điển có thể chiếm nhiều tài nguyên vì sau rất nhiều lần xét nó có thể tăng kích thước rất nhanh.

IV. TÀI LIỆU THAM KHẢO

1. *LZW (Lempel–Ziv–Welch) Compression technique - https://www.geeksforgeeks.org/lzw-lempel-ziv-welch-compression-technique/*

2. *Lempel–Ziv–Welch - Lempel–Ziv–Welch – Wikipedia*

3. *LZW COMPRESSION AND DECOMPRESSION - Abstract.pdf (indstate.edu)*

V. TÀI LIỆU ĐÍNH KÈM

1. **LZW-Text-Compression |** Công cụ nén và giải nén văn bản bằng thuật toán LZW.
2. **LZW-Image-Compression |** Công cụ nén và giải nén hình ảnh bằng thuật toán LZW.
3. **File PDF:** “LZW Compression.pdf”
4. **File PowerPoint:** “LZW Compression.pptx”