Information Compression

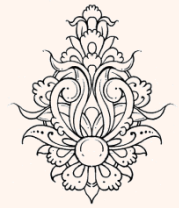
روش‌های مبتنی واژه‌نامه

فشرده‌سازی اطلاعات
۴۰-۱۰-۷۰۲-۰۱
بخش دوم

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
زمستان ۱۴۰۰
احمد محمودی ازناوه

فهرست مطالب

- معرفی
- روش‌های ایستا
  - کدهای دوتایی
- روش‌های وفقی(پویا)
  - الگوریتم LZ77
  - الگوریتم LZ78
  - الگوریتم LZW
- چند نمونه از کاربردهای کدگذاری واژه‌نامه‌ای



معرفی

- در روش‌هایی که تاکنون بررسی شد، فرض بر این بود که نمادها به صورت مستقل از یک توزیع تصادفی پیروی می‌کنند(.i.i.d).
  - اغلب بین نمادهای مختلف نوعی همبستگی وجود دارد.
  - معمولا پیش از کدگذاری یک مرحله ناهمبسته‌سازی (decorrelation) انجام می‌شود.
- در این بخش به مرور روشی می‌پردازیم که ساختار داده‌ها را نیز در نظر می‌گیرد.
- برای دنباله‌ای از نمادها که زیاد تکرار می‌شوند، یک کد در نظر گرفته می‌شوند.
  - چنین دنباله‌هایی در متون و دستورهای کامپیوتری زیاد یافت می‌شود.
  - در زندگی روزمره هم با چنین رویکردی مواجه می‌شویم!

روش‌های مبتنی بر واژه‌نامه

- در این شیوه واژه‌نامه‌ای از دنباله‌های پرتکرار تهیه شده و به جای کد کردن تک‌تک نمادهای چنین دنباله‌هایی، شاخص آن در واژه‌نامه در نظر گرفته می‌شود.
  - سایر نمادها به صورت تکی کدگذاری خواهند شد.
- برای این که این شیوه کارایی لازم را داشته باشد، تعداد مجموعه نمادهای پرتکرار و در نتیجه اندازه‌ی واژه‌نامه باید به صورت نسبی کوچک باشد.
- متنی شامل کلمه‌های چهارحرفی از متشکل از ۲۶ حرف کوچک انگلیسی و ۶ نماد برای علامت‌گذاری
  - ۲۵۶ کلمه پرتکرار در واژه‌نامه‌ی گردآوری می‌شوند.
  - در صورت استفاده از واژه‌نامه کلماتی که از شاخص واژه‌نامه استفاده می‌کنند با پیشوند « **1**» و سایر کلمه‌کدها با « **0**» آغاز می‌شوند.

مثال

the average number of bits per pattern

$$l_{av} = 9p + (1-p)21 = 21 - 12p \leq 20$$

روش‌های ایستا

- در این دسته از روش‌های یک واژه‌نامه‌ی ثابت در نظر گرفته می‌شود.
- استفاده از چنین روش‌های زمانی مفید است که از خصوصیات نمادها اطلاعات کاملی داشته باشیم.
  - این ویژگی‌ها به مرور زمان تغییر نکند.
- چنین روش‌هایی معمولا در کاربردهای خاصی می‌تواند مفید باشد.
  - به عنوان مثال در فشرده‌سازی مستندات مربوط به اطلاعات دانشجویان واژه‌هایی مانند شماره‌دانشجویی، معدل، واحدهای پاس‌شده، مشروطی و ... زیاد تکرار می‌شوند.

کدهای دوتایی

- در این شیوه‌ی دنباله‌های دوتایی پرتکرار کد می‌شوند.
  - به عنوان نمونه برای کدگذاری کاراکترهای قابل چاپ اسکی، می‌توان یک جدول ۲۵۶تایی در نظر گرفت، ۹۵ مدخل آن به کاراکترهای قابل چاپ اختصاص یابد و مابقی آن به دنباله‌های دوتایی پرتکرار

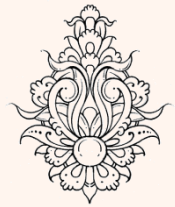
مثال

$$A = \{a, b, c, d, r\}$$

Code	Entry	Code	Entry
000	*a*	100	*r*
001	*b*	101	*ab*
010	*c*	110	*ac*
011	*d*	111	*ad*

abracadabra

$$101-100-110-111-101-100-000$$



مثال

LaTeX document

Pair	Count	Pair	Count
e␣	1128	ar	314
␣t	838	at	313
␣␣	823	␣w	309
th	817	te	296
he	712	␣s	295
in	512	d␣	272
s␣	494	␣o	266
er	433	io	257
␣a	425	co	256
t␣	401	re	247
en	392	␣$	246
on	385	r␣	239
n␣	353	di	230
ti	322	ic	229
␣i	317	ct	226

collection of C programs

Pair	Count	Pair	Count
␣␣	5728	st	442
nl␣	1471	le	440
; nl	1133	ut	440
in	985	f(	416
nt	739	ar	381
= ␣	687	or	374
␣i	662	r␣	373
t␣	615	en	371
␣ =	612	er	358
);	558	ri	357
, ␣	554	at	352
nlnl	506	pr	351
␣f	505	te	349
e␣	500	an	348
␣*	444	lo	347

مطلوب است روشی اتخاذ شود که واژه‌نامه بر اساس محتوای نمادها به صورت وفقی تشکیل شود.


Abraham Lempel


Jacob Ziv

روش‌های وفقی مبتنی بر واژه‌نامه

- واژه‌نامه، بخشی از داده‌های کد شده است.

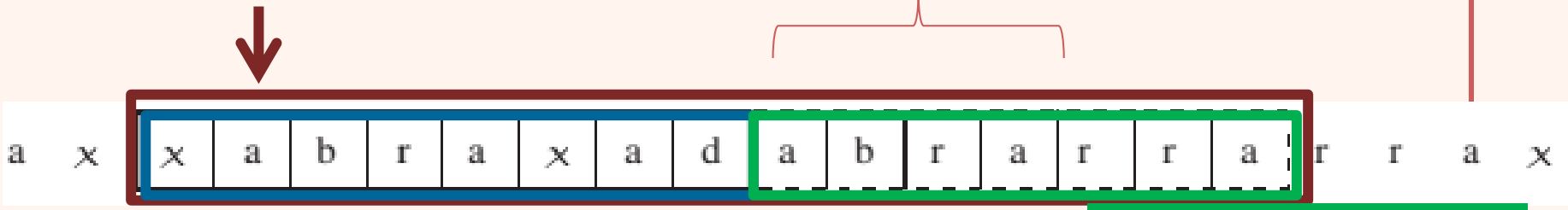
The LZ77 Approach

Jacob Ziv and Abraham Lempel

Sliding window

Match pointer

Length of match



Search buffer

offset

Look ahead buffer

لازم است کدگذاری شوند

کدگذاری شده

$< o,\ l,\ c >$

o: offset

l: length of macth

c: codeword of the symbol in the lookahed buffer following the match>

شیوه‌ی کدگذاری

Ziv, J. and A. Lempel (1977). "A universal algorithm for sequential data compression." IEEE Transactions on information theory 23(3): 337-343.

مثال- کدگذاری

طول پنجره: ۱۳ طول بافر پیش‌بینی: ۶

رشته‌ی زیر را در نظر بگیرید:

$$\cdots cabracadabrarrarrad \cdots$$

cabraca dabrar $<0,0,C(d)>$

abracad abrarr $<7,4,C(r)>$

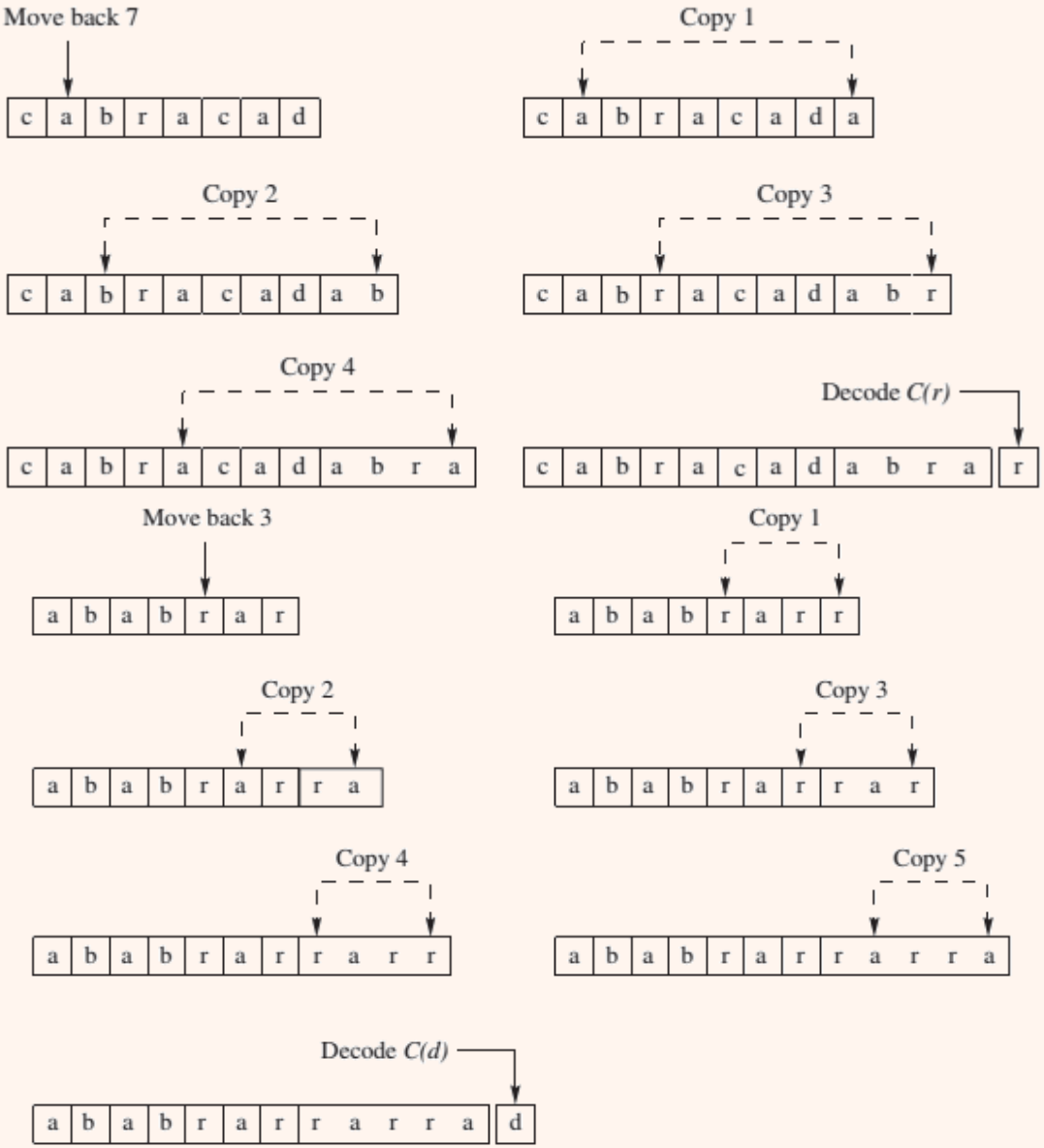
adabrar rarrad $<3,5,C(d)>$

کد به دست آمده بر اساس LZ77

cabraca $<0,0,C(d)><7,4,C(r)><3,5,C(d)>$

مثال- کدگشایی

cabraca <0,0,C(*d*)><7,4,C(*r*)><3,5,C(*d*)>



LZ77

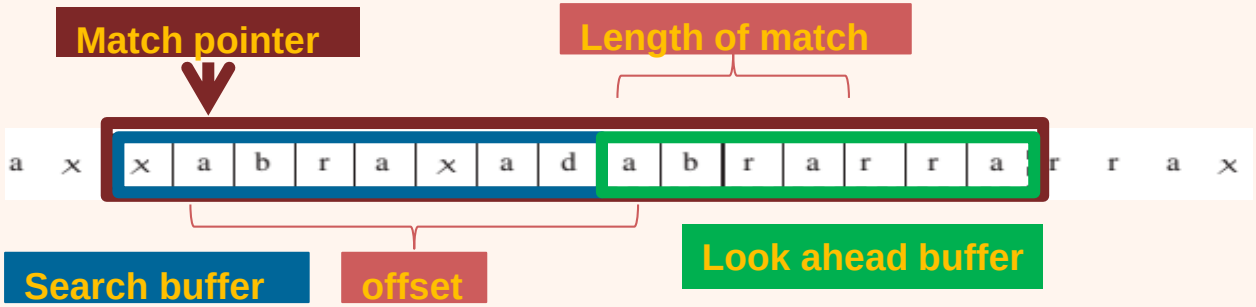
- چنانچه تکرار داده‌ها با تناوبی بیشتر از طول پنجره رخ دهد، LZ77 نه تنها کارایی خود را از دست می‌دهد، بلکه به جای فشرده‌سازی بر حجم داده‌ی کدشده می‌افزاید.

مثال

- ویرایش‌های بهبود یافته‌ی متعددی از LZ77 مطرح شده است.
  - بیشتر بر مبنای شیوه‌ی کدگذاری سه‌تایی‌ها هستند.
  - بسیاری از فشرده‌سازها نظیر ZIP از این شیوه‌ها استفاده می‌کنند.

تمرین

- سه‌تایی به دست آمده برای نمایش با طول ثابت $< o,\ l,\ c >$ به چند بیت نیاز دارد؟



$$\lceil \log_2 S \rceil + \lceil \log_2 W \rceil + \lceil \log_2 A \rceil$$

اندازه‌ی پنجره‌ی جستجو

اندازه‌ی پنجره

تعداد حروف الفبا

LZ78

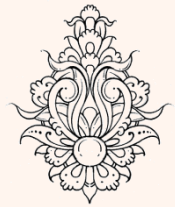
- LZ78 نیز راهی برای چیرگی بر مشکلات LZ77 است. در این شیوه واژه‌نامه به صورت مشابه در کدگذار و کدگشا ساخته می‌شود.
- در این شیوه داده‌ها به صورت زیر کد می‌شوند:

$$< i, c >$$

شاخص طولانی‌ترین رشته‌ی مشابه

کد نماد بعدی

- در صورتی نیافتن هیچ مشابهی در واژه‌نامه از شاخص «0» استفاده می‌شود.



مثال

wabba wabba wabba wabba woo woo woo

در ابتدا واژه‌نامه خالی است و چند نماد ابتدایی با شاخص 0 مشخص می‌شوند.

$<0,C(w)>$ $<0,C(a)>$ $<0,C(b)>$

واژه‌نامه

Index	Entry
1	*w*
2	*a*
3	*b*

چهارمین نماد «*b*» در واژه‌نامه وجود دارد، نماد بعدی را به آن الحاق کنیم: «*ba*» و به صورت زیر کد می‌شود:

$<3,C(a)>$

ادامه‌ی مثال

و به همین ترتیب دنباله‌نمادهای دیگر کد شده و هر چه پیش رویم نمادهای طولانی‌تری انتخاب می‌شوند. در پایان خواهیم داشت:

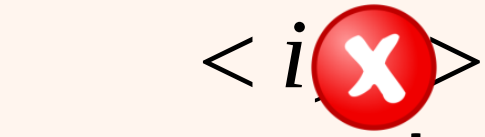
Encoder Output	Dictionary Index	Dictionary Entry
$\langle 0, C(w)\rangle$	01	w
$\langle 0, C(a)\rangle$	02	a
$\langle 0, C(b)\rangle$	03	b
$\langle 3, C(a)\rangle$	04	ba
$\langle 0, C(\not b)\rangle$	05	$\not b$
$\langle 1, C(a)\rangle$	06	wa
$\langle 3, C(b)\rangle$	07	bb
$\langle 2, C(\not b)\rangle$	08	$a\not b$
$\langle 6, C(b)\rangle$	09	wab
$\langle 4, C(\not b)\rangle$	10	$ba\not b$
$\langle 9, C(b)\rangle$	11	$wabb$
$\langle 8, C(w)\rangle$	12	$a\not bw$
$\langle 0, C(o)\rangle$	13	o
$\langle 13, C(\not b)\rangle$	14	$o\not b$
$\langle 1, C(o)\rangle$	15	wo
$\langle 14, C(w)\rangle$	16	$o\not bw$
$\langle 13, C(o)\rangle$	17	oo

wabba wabba wabba wabba woo woo woo

LZW

a modification by Terry Welch

- در این شیوه تنها بخش شاخص ارسال می‌شود.



- واژه‌نامه‌ی ابتدایی باید شامل تمام الفبا باشد.
- در صورت وجود ترکیب‌های طولانی در دنباله‌ی داده‌ها این ترکیبات به واژه‌نامه افزوده می‌شود.

مثال-کدگذاری

wabba wabba wabba wabba woo woo woo

الفبای کد منبع شامل حروف زیر است:

$$\{\not b, a, b, o, w\}$$

در نتیجه واژه‌نامه‌ی اولیه به صورت زیر خواهد بود:

Index	Entry
1	$\not b$
2	a
3	b
4	o
5	w

در کدگذار ابتدا با «w» مواجه می‌شود که در واژه‌نامه موجود است، از این رو «wa» انتخاب می‌شود و به عنوان ششمین مدخل به واژه‌نامه افزوده می‌شود.

مثال کدگذاری

سپس از حرف «a» ادامه می‌دهیم و «ab» را به واژه‌نامه می‌افزاییم.

همین روند را ادامه می‌دهیم تا به اولین حرف از دومین کلمه برسیم، تا اینجا رشته‌ی کد شده عبارتست از :

Index	Entry
01	⊬
02	*a*
03	*b*
04	*o*
05	*w*
06	*wa*
07	*ab*
08	*bb*
09	*ba*
10	*a*⊬
11	⊬ w
12	*w*…

wabba wabba wabba wabba woo woo woo

مثال- کدگذاری

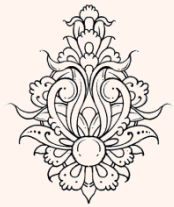
در کدگذار ابتدا با «*w*» مواجه می‌شود که در واژه‌نامه موجود است، از این رو «*wa*» انتخاب می‌شود، که آن‌هم در کدگذار موجود است. در نتیجه «*wab*» انتخاب می‌شود و به واژه‌نامه افزوده می‌وشود.

wabba wabba wabba wabba woo woo woo

کدگذار 6 را در ادامه‌ی دنباله ارسال خواهد کرد.

در نهایت واژه‌نامه و رشته‌ی کد شده به صورت زیر خواهند بود:

Index	Entry	Index	Entry
01	*ϕ*	14	*aϕ w*
02	*a*	15	*wabb*
03	*b*	16	*baϕ*
04	*o*	17	*ϕ wa*
05	*w*	18	*abb*
06	*wa*	19	*baϕ w*
07	*ab*	20	*wo*
08	*bb*	21	*oo*
09	*ba*	22	*oϕ*
10	*aϕ*	23	*ϕ wo*
11	*ϕ w*	24	*ooϕ*
12	*wab*	25	*ϕ woo*
13	*bba*		



5 2 3 3 2 1 6 8 10 12 9 11 7 16 5 4 4 11 21 23 4

مثال-کدگشایی

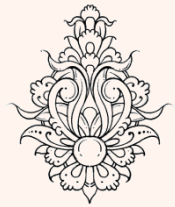
5 2 3 3 2 1 6 8 10 12 9 11 7 16 5 4 4 11 21 23 4

Index	Entry
1	␣
2	*a*
3	*b*
4	*o*
5	*w*

همزمان با کدگشایی واژه‌نامه نیز تشکیل می‌شود:

wabba 6 8 10 12 9 11 7 16 5 4 4 11 21 23 4

Index	Entry
01	␣
02	*a*
03	*b*
04	*o*
05	*w*
06	*wa*
07	*ab*
08	*bb*
09	*ba*
10	*a*␣
11	␣ . .



wabba wabba wabba wabba woo woo woo

مثال-کدگذاری

$abababababababababab\cdots$

الفبای کد منبع شامل حروف زیر است:

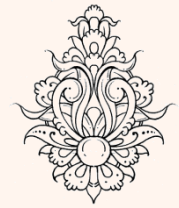
$\{a, b\}$

در نتیجه واژه‌نامه‌ی اولیه به صورت زیر خواهد بود:

Index	Entry
1	*a*
2	*b*

1 2 3 5 ⋯

Index	Entry
1	*a*
2	*b*
3	*ab*
4	*ba*
5	*aba*
6	*abab*
7	*b . . .*



مثال-کدگشایی

1 2 3 5 ⋯

Index	Entry
1	*a*
2	*b*

ab 3 5 ⋯

Index	Entry
1	*a*
2	*b*
3	*ab*
4	*b* . . .

*abab*5 ⋯

Index	Entry
1	*a*
2	*b*
3	*ab*
4	*ba*
5	*a* . . .

واژه‌نامه ناقص

LZW نیاز به یک سیستم مدیریت استثنا هم دارد!

کاربردها

- از الگوریتم‌های مبتنی بر LZ، در عمل فراوان مورد استفاده قرار گرفته‌اند. به عنوان مثال

LZW

– دستور compress در سیستم عامل UNIX

- اندازه‌ی واژه‌نامه وفقی است و از ۵۱۲ شروع می‌شود(نه بیت برای شاخص).
- با پر شدن واژه‌نامه، اندازه‌ی آن دو برابر می‌شود.
- وقتی به واژه‌نامه حداکثر انداره رسید(شانزده بیت برای شاخص)، عملا به یک روش ایستا تبدیل می‌شود.
- در این حالت نرخ فشردگی بررسی شده و چنان‌چه از یک حد آستانه کم‌تر بود، واژه‌نامه تخلیه شده و از نو ساخته می‌شود.

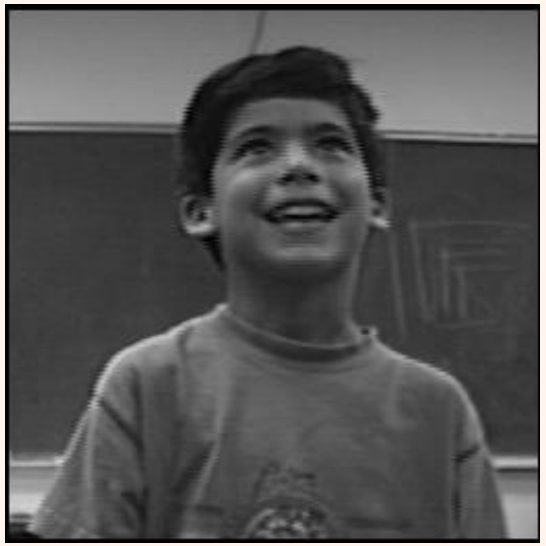
256×256

تصاویر نمونه

Earth

Omaha

Sensin



Sena

کاربردها(ادامه...)

- فرمت فشرده‌سازی GIF

Graphics Interchange Format

LZW

  - شبیه به compress است.
  - برای تصاویر طبیعی چندان مناسب نیست.

Image	GIF	Arithmetic Coding of Pixel Values	Arithmetic Coding of Pixel Differences
Sena	51,085	53,431	31,847
Sensin	60,649	58,306	37,126
Earth	34,276	38,248	32,137
Omaha	61,580	56,061	51,393

- فرمت فشرده‌سازی PNG

Portable Network Graphics

LZ77

  - با توجه به حق امتیاز LZW به عنوان جایگزین GIF مطرح شد.
- فرمت فشرده‌سازی V.42 bis در ارتباطات تلفنی

LZW