

Classificação e Pesquisa de Dados

Aula 25
Compressão de Arquivos: Métodos de Compressão

UFRGS

INF01124

Compressão de Arquivos

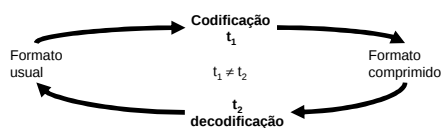
◆Propósito

- Redução do número de bits necessário para a representação dos dados, sem perda de informação

◆Motivação

- armazenamento com economia de espaço
- transmissão com economia de tempo

Compressão de dados



◆Codificação

- item → tabela → código

◆Decodificação

- código → tabela → item

Compactação ou compressão?



◆ Compactação:

- consiste na busca por regiões do arquivo que não contêm dados e posterior recuperação desses espaços perdidos. Os espaços vazios são provocados, p. ex., pela eliminação de registros

◆ Compressão:

- envolve a codificação da informação de modo que o arquivo tome menos espaço. Algumas técnicas são gerais e outras são específicas p/ certos tipos de dados: voz, imagem ou texto.

Considerações sobre a implementação

- ◆ Os custos de processamento para a *compressão* e *expansão* podem ser significativos
- ◆ A relação entre a quantidade de recuperações e a quantidade de alterações deve ser avaliada para permitir a ponderação dos tempos de compressão e expansão dos métodos quando da sua escolha
- ◆ Os mecanismos de compressão e expansão podem ser implementados por:
 - Software
 - Hardware

Métodos de compressão

- ◆ Recodificação com códigos de comprimento fixo
- ◆ Supressão de caracteres repetidos
- ◆ Codificação de itens
- ◆ Compressão de seqüências
- ◆ Códigos de comprimento variável
 - Código de Huffman
 - Pike

Supressão de caracteres repetidos

◆ Marcas de supressão

< marca, número >

Quantidade de caracteres suprimidos

Identifica a supressão de um caracter

a Zeros
b Brancos

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0	0	0	0	0	5	3	J	ô		L	i	m	a					2	5	0	0	0	0

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
a	5	5	3	J	ô		L	i	m	a	b	4	2	5	a	4

Supressão de
zeros e brancos

Codificação de itens

◆ Alguns itens ocorrem com frequências significativamente maiores do que os demais

■ Programas fonte

• Begin, integer, end, if, then, ...

■ Cadastro de nomes

• José, Maria, João

■ Textos em computação

• Sistema, linguagem, algoritmo, dado, ...

Como reduzir o espaço de armazenamento desses itens ?



Codificação de itens

A redução do espaço de armazenamento pode ser obtida substituindo-se os itens mais freqüentes por **códigos** !

◆ Tipos de codificação

- Códigos simples
- Códigos compostos

Códigos simples

◆ Estratégia de codificação

- Códigos de comprimento fixo
 - Ex.: 8 bits
- Representação dos 254 nomes e sobrenomes mais freqüentes pelos códigos de 1 a 254
- 0 : indica início de trecho não codificado
- 255 : indica supressão de espaços
- Nomes por extenso são precedidos por um par <0, n>, onde **n** é igual ao número de caracteres do nome

Códigos simples

◆ Codificação

- Item → tabela → código
- Exige eficiência no acesso à tabela
- Exige a decomposição do item

◆ Decodificação

- Código → tabela → item
- Código é o próprio endereço de acesso à tabela
- Procedimento mais rápido que a codificação

Eficiência no acesso à tabela

◆ Tabela residente na memória durante o uso dos dados

◆ Como selecionar os itens da tabela?

- Amostra significativa de itens ordenados por frequência

Códigos simples

Exemplos:

José Antônio Silva ... (30)

31	7	77	255	11
----	---	----	-----	----

1 Espaço após cada nome

João Augusto Silva

27	0	7	A	U	G	U	S	T	O	77	255	11
----	---	---	---	---	---	---	---	---	---	----	-----	----

→ Tamanho do trecho não codificado

Código	Nome
1	Abílio
2	Ana
...	...
7	Antônio
...	...
27	João
...	...
31	José
...	...
57	Maria
...	...
77	Silva
...	...
254	Zuleika

Códigos compostos

Este método de compressão baseia-se em códigos de comprimento variável. Neste caso, os códigos têm 4, 8 ou 12 bits.

Diferentemente de outros métodos, são codificados, simultaneamente, caracteres e palavras, conforme descreve o autor Pike, 1981.

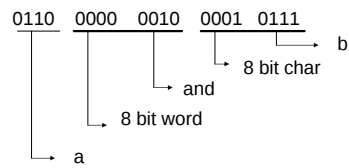
Códigos compostos (Pike)

Tabela criada c/ as frequências de caracteres e palavras observadas em textos em inglês.

Código de 4 bits	Caracteres com 8 bits	Palavras com 8 bits	12 bits
0 8 bit word	u	the	67 caracteres
1 8 bit chars	p	of	189 palavras
2 12 bits	g	and	
3 end of line	m	to	
4 space	h	in	
5 e	y	that	
6 a	f	it	
7 i	b	is	
8 s	v	for	
9 r	w	be	
10 t	k	was	
11 n	x	as	
12 o	j	you	
13 e	q	with	
14 d	,	he	
15 c	.	on	

Códigos compostos

Exemplo:



Códigos compostos

0000 0101 0000 0111 0110 1011

?

Códigos compostos

0000 0101 0000 0111 0110 1011 = 24 bits
that is a n = 64 bits

Percentual de redução = 62,5%

Compressão de seqüências ordenadas

Este método explora entradas ordenadas, com prefixos comuns, como aquelas que aparecem em *dicionários* e *índices*.

Compressão de seqüências

Exemplo de uma seqüência de palavras:

SEPARAR	0, SEPARAR
SEPARATA	6, TA
SEPARATISMO	7, ISMO
SEPARATISTA	9, TA
SEPARATIVO	8, VO
SEPARATORIO	7, ORIO
SEPARATRIZ	7, RIZ
SEPARAVEL	6, VEL

Sequência original Sequência Comprimida