webir

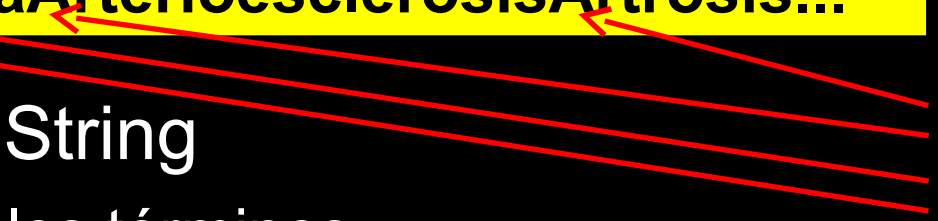
Clase 6

Temas

- Compresión de índices - vocabulario
- Compresión de índices – lista de postings

Compresión de Vocabulario

... ArbolArteriaArterioesclerosisArtrosis...



- Diccionario en un sólo String
  - Lista continua de todos los términos
  - Reducción de hasta 60%
- Por Bloques

... 5Arbol7Arteria17Arterioesclerosis8Artrosis...

... 9Elemental8Elemento8Eliminar...

...

... 5Zorro6Zorzal8Zozobrar...

- Codificación Frontal – Front Coding
  - 11.2 → 5.9

... 8Automata8Automate9automatic10Automation...

... 8Automat*a1¤e2¤ic3¤ion...

Compresión de la Lista de Postings REUTERS-RCV1

- 800 000 documentos
  - $\log_2$ 800 000 = 20 bits para docID
- 200 tokens por documento
- Tokens de 6 caracteres = 6 bytes
- 100 000 000 postings

- Tamaño de la colección
  - 800 000 * 200 * 6 = 960 MB
- Tamaño del conjunto de postings (sin comprimir)
  - 100 000 000 * 20/8 = 250 MB

Compresión de la Lista de Postings

- Ejemplo Reuters-RCV1
  - 800 000 documentos (~ 20 bits)
- ¿Menos de 20 bits para docID (por documento)?
- Dependiendo de la frecuencia de aparición de los términos en los documentos es la magnitud del gap

the	docIDs	...	283042	283043	283044	
	gaps			1	1	
computer	docIDs	...	283047	283154	283159	
	gaps			107	5	
zenith	docIDs	252000			500100	
	gaps	252000	2481000			

Compresión de la Lista de Postings

- Ejemplo Reuters-RCV1
  - 800 000 documentos (~ 20 bits)
- ¿Menos de 20 bits para docID (por documento)?
- Dependiendo de la frecuencia de aparición de los términos en los documentos es la magnitud del gap

the	docIDs	...	283042	283043	283044	
	gaps			1	1	
computer	docIDs	...	283047	283154	283159	
	gaps			107	5	
zenith	docIDs	252000			500100	
	gaps	252000	2481000			

Compresión de la Lista de Postings

- ¿Menos de 20 bits para docID (por documento)?
- Los términos frecuentes aparecen en muchos documentos
  - Es más “económico” representar el espacio entre documentos < 20 bits

the	docIDs	...	283042	283043	283044	
	gaps			1	1	
computer	docIDs	...	283047	283154	283159	
	gaps			107	5	
zenith	docIDs	252000			500100	
	gaps	252000	2481000			

Compresión de la Lista de Postings

- ¿Menos de 20 bits para docID (por documento)?
- Los términos poco frecuentes aparecen en pocos documentos
  - No es tan “económico” representar el espacio entre documentos ~ 20 bits

the	docIDs	...	283042	283043	283044	
	gaps			1	1	
computer	docIDs	...	283047	283154	283159	
	gaps			107	5	
zenith	docIDs	252000			500100	
	gaps	252000	2481000			

Compresión de la Lista de Postings

- Representación de largo variable!
- Compresión
  - Por bytes
  - Por bits

Código Variable en Bytes- Código VB

- Número variable de bytes (8 bits)
- El primer bit de cada byte indica “continuación”
  - 1 si es el último byte
  - 0 en caso contrario
- Se concatenan 7 bits de cada byte
- Se obtiene número binario

docID	... 824	829	2125406...
gap	5	214577	
Código VB	10000101	00001101 00001100 10110001	

Código Variable en bytes - Código VB

- REUTERS-RCV1
  - Se reduce de 250 MB a 116 MB
  - Más del 50%
- Se puede aplicar otras unidades
  - 32 bits
  - 16 bits
  - 4 bits
- Espacio ahorrado/tiempo de decodificación

Código Variable en bits – Código Vb

- Número variable de bits
- Códigos γ y δ

Código Variable en bits – Código Vb

- Código unario
- Código variable en bits más sencillo
- Ineficiente

0	0
1	10
2	110
3	1110
4	11110
9	1111111110

Código Variable en bits – Código Vb

- Cantidad de gaps G
  - $1 <= G <= 2^n$
  - Asumiendo que son equiprobables
- Código óptimo usa n bits
  - En particular $g=2^n$, no se puede codificar con menos de $\log_2 2^n$ bits = n bits
- Acercarse al óptimo de n bits
  - Se debe distinguir un gap de otro = desperdicio

Código Variable en bits – Código Vb

- Representación del código γ
  - *largo* y *offset*
- *offset*
  - Es el número en representación binaria sin el 1 más significativo
  - 13 → 1101, *offset* = 101
- *largo*
  - Largo del *offset* en código unario + 0
  - 13 → 1101, *offset* = *101* y *largo* = 1110
- Código γ para 13 es 1110*101*

Código Variable en bits – Código Vb

- Código γ

- Leer el *largo* hasta el 1er 0
- Leer el *offset*
- Agregar el 1 que se eliminó

- Código γ para 13 es 1110101
  - *largo* = 1110 → *offset* = 101 → número es 1101 = 13

Número	Código γ
0	¿Porqué?
1	0
2	10,0
3	10,1
4	110,00
9	1110,001
13	1110,101
24	11110,1000
511	111111110,11111111

Código Variable en bits – Código Vb

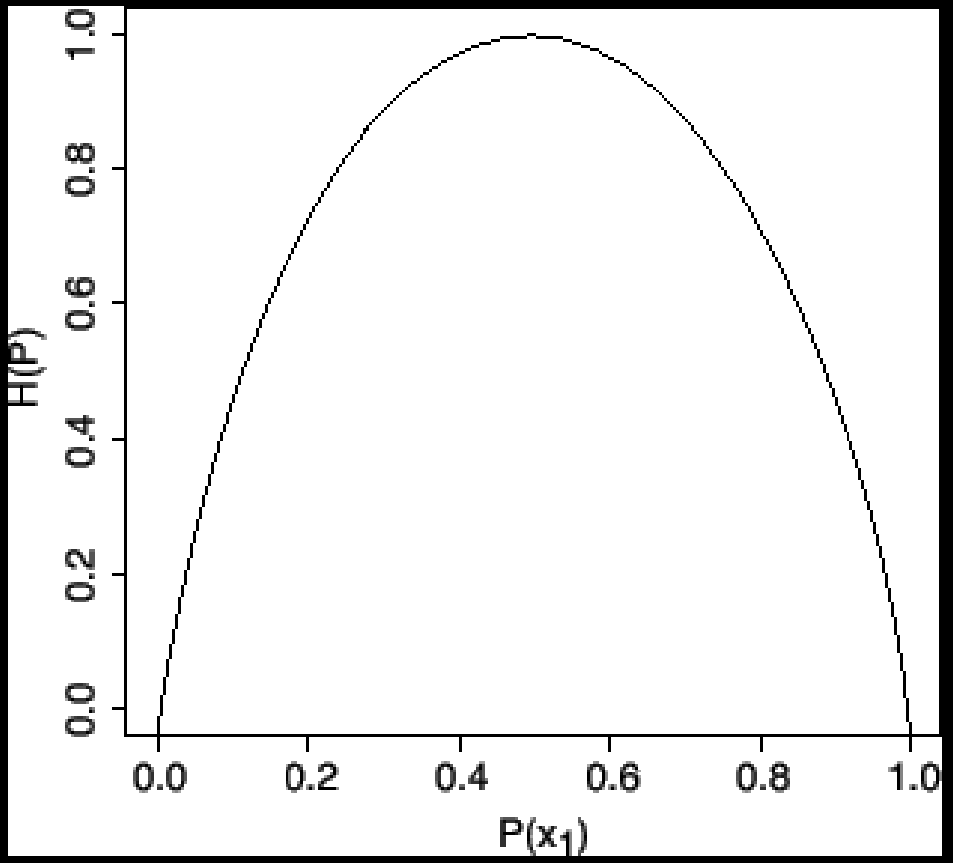
- Cantidad de gaps G
  - $1 <= G <= 2^n$
- Código γ
  - *offset* es $\lfloor \log_2 G \rfloor$
  - *largo* es $\lfloor \log_2 G \rfloor + 1$
  - Total $2 * \lfloor \log_2 G \rfloor + 1$
  - Diferencia del óptimo por constante multiplicativa = 2
- ¿Equiprobables?

Código Variable - Entropía

- Propiedades de la codificación en base a una distribución de probabilidad P de los datos a codificar (en este caso gaps) se determina mediante
  - Entropía $$H(P) = -\sum_{x \in X} P(x) \log_2 P(x)$$
  - *X* es el conjunto de números a codificar $$\sum_{x \in X} P(x) = 1$$
  - H(P) es una cota inferior para el largo esperado de la codificación óptima (optimalidad del código)
  - Código γ constante multiplicativa = 3, sin importar P → codificación Universal

Código Variable - Entropía

- Dos eventos x_1 y x_2
- X = $\{x_1, x_2\}$
- H(P) = 1, máxima
  - $P(x_1) = P(x_2) = 0.5$
  - Máxima incertidumbre
- H(P) = 0, mínima
  - $P(x_1) = 1$ y $P(x_2) = 0$ o viceversa
  - Mínima incertidumbre



Código Variable en bits – Código Vb

- Código γ – Ventajas
  - Universalidad
  - Sin prefijos comunes – no se necesitan delimitadores
  - Sin parámetros que ajustar
- REUTERS-RCV1
  - Se reduce de 250 MB a 101 MB

Código Variable en bits – Código Vb

- Código γ – Desventajas
  - Decodificación costosa
  - No siempre coinciden las unidades con palabras/unidades en memoria
  - Las operaciones de bajo nivel se aplican a las unidades de memoria
- Tiempo vs espacio

Código Variable en bits – Código Vb

- Código δ (para mejorar rep. del largo en γ)
  - *largo* y *offset*
- Optimizar la representación del *largo*
- Codificar el largo con código γ

- Código γ para 7 es 10,0,11
  - Código γ para *largo* 2 → 10,0
  - *offset* = 11 permanece igual

Código Variable en bits – Código Vb

Method	Bits per gap			
	Bible	GNUBib	Comact	TREC
Unary	262	909	487	1918
Gamma γ	6.51	5.68	4.48	6.63
Delta δ	6.23	5.08	4.35	6.38

- ¿Para qué rango de valores es más corto el código δ que el código γ?

Códigos de Compresión

- Tabla anterior:
  - I.H. Witten, A. Moffat and T.C. Bell, “Managing Gigabytes”, Morgan Kaufmann, 1999
- Compresión de documentos de texto con posibilidades de búsqueda de palabras o frases:
  - de Moura, Edleno Silva, Gonzalo Navarro, Nivio Ziviani, and Ricardo Baeza-Yates, “Fast and flexible word searching on compressed text”, 2000
  - Brisaboa, Nieves R., Antonio Fariña, Gonzalo Navarro, and José R. Paramá, “Lightweight natural language text compression”, 2007