Tópicos em Recuperação de Informação[1]

Nivio Ziviani

[1]Conjunto de transparências elaborado por Nivio Ziviani e Patrícia Correia

Compressão de Índices

- Arquivos invertidos são amplamente usados para indexar grandes arquivos de texto.

- O tamanho de um arquivo invertido pode ser reduzido usando compressão nas listas invertidas.

- Números dos documentos em ordem ascendente: seqüência de *d-gaps* (a diferença entre d_{k+1} e d_k) entre os números de documentos.

- Processamento é seqüencial desde início da lista: números originais dos documentos podem ser obtidos através do cálculo da soma dos *d-gaps*.

  Ex.: Suponha que um termo ocorra em 8 documentos:

  $$<8;\ 3, 5, 20, 21, 23, 76, 77, 78>$$

  Calculando-se os *gaps* entre os documentos:

  $$<8;\ 3, 2, 15, 1, 2, 53, 1, 1>$$

Compressão de Índices

- Ambos os métodos de representação das listas necessitam de $\lceil \log N \rceil$ bits por apontador, onde N é o número de documentos da coleção.

- Entretanto, *d-gaps* podem ser codificados em muito menos do que $\lceil \log N \rceil$ bits.

- Os *gaps* são pequenos para palavras freqüentes e grandes para palavras infreqüentes, a compressão pode ser obtida codificando-se pequenos valores com códigos menores
  - Por exemplo, o código unário, no qual um inteiro x é codificado como $x - 1$ bits 1 e seguido de um bit 0.

    Ex.: O código unário para o inteiro 3 é 110.

Métodos para Compressão de Índices

Métodos locais

Modelo de compressão para cada lista associada a um termo é ajustado de acordo com algum parâmetro local. Por exemplo, a freqüência do termo.

Métodos globais

Listas são comprimidas usando um modelo comum de compressão

- Não-parametrizados: Unário, Elias-δ e Elias-γ.

- Parametrizados: Golomb
    - envolve algum parâmetro que pode ser ajustado à distribuição dos *d-gaps*
    - exige duas passadas no texto.

Métodos para Compressão de Índices

Método	**Referência**
Métodos Globais	
Não parametrizados	
Unário	
Binário	
γ	Elias (1975); Bentley and Yao (1976)
δ	Elias (1975); Bentley and Yao (1976)
Parametrizados	
Bernoulli	Golomb(1966);
	Gallager and Van Voorhis (1975)
Observed frequency	
Métodos Locais	
Bernoulli	Witten, Bell, and Nevill (1992);
	Bookstein, Klein, and Raita (1992)
Skewed Bernoulli	Teuhola (1978); Moffat and Zobel (1992)
Hyperbolic	Schuegraf (1976)
Observed frequency	
Batched frequency	Moffat and Zobel (1992)
Interpolative	Moffat and Zobel (1996)

Codificação Unária

Para a codificação de um inteiro x usando o método unário repete-se $(x-1)$ bits com valor 1, em seguida acrescenta-se um bit com o valor 0.

1	0
2	10
3	110
4	1110
5	11110
6	111110
7	1111110
8	11111110
9	111111110
10	1111111110

Codificação de Elias-γ

O número é dividido em duas partes:

- Primeira parte:
  - código unário de $1 + \lfloor \log x \rfloor$ bits.
- Segunda parte:
  - tamanho do código em bits: $\lfloor \log x \rfloor$ bits.
  - codificando o número $x - 2^{\lfloor \log x \rfloor}$ em binário.

Ex.: $x = 5$

$1 + \lfloor \log 5 \rfloor = 3$ (110 em unário)

$5 - 2^{\lfloor \log 5 \rfloor} = 5 - 2^2 = 1$ (01 em binário de tamanho 2)

portanto o código de 5 é 11001.

Codificação de Elias-δ

O número é dividido em duas partes:

- Primeira parte:
  - código Elias-γ de $1 + \lfloor \log x \rfloor$ bits.
- Segunda parte:
  - tamanho do código em bits: $\lfloor \log x \rfloor$ bits.
  - codificando o número $x - 2^{\lfloor \log x \rfloor}$ em binário.

Ex.: $x = 5$

$1 + \lfloor \log 5 \rfloor = 3$ (101 em Elias-γ)

$5 - 2^{\lfloor \log 5 \rfloor} = 5 - 2^2 = 1$ (01 em binário de tamanho 2)

portanto o código de 5 é 10101.

Codificação de Golomb

Solomon Golomb, USC (1966):

- Definir o valor do parâmetro b.

- $(q+1)$ em unário, onde $q = \lfloor (x-1)/b \rfloor$.

- $r = (x-1) - q \times b$ codificado em binário com $\lfloor \log b \rfloor$ ou $\lceil \log b \rceil$ bits.

- Valores de r com $\lfloor \log b \rfloor$ bits começam com 0.

- Valores de r com $\lceil \log b \rceil$ bits começam com 1.

Ex.: $b = 3$ e $x = 7$

$q = \lfloor (7-1)/3 \rfloor = 2.$ Então $(q+1)$ em unário é 110

$r = (7-1) - 2 \times 3 = 0$

representação de 7 é 1100.

Exemplos de Códigos para Inteiros

Gap x	Unary	Elias-γ	Elias-δ	Golomb ($b = 3$)
1	0	0	0	00
2	10	100	1000	010
3	110	101	1001	011
4	1110	11000	10100	100
5	11110	11001	10101	1010
6	111110	11010	10110	1011
7	1111110	11011	10111	1100
8	11111110	1110000	11000000	11010
9	111111110	1110001	11000001	11011
10	1111111110	1110010	11000010	11100

- O código Elias-δ para um inteiro arbitrário x requer $1 + 2\lfloor \log \log 2x \rfloor + \lfloor \log x \rfloor$ bits.

- Para pequenos valores de x os códigos de Elias-γ são menores que os de Elias-δ. Entretanto, no limite, como x torna-se grande, a situação é revertida.