Sobre a Eficiência da Codificação de Huffman Binária em Extensões de Fontes de Informação

Diocleciano D. Neto†, Everton B. Lacerda† e Daniel C. Cunha†

***Resumo*— Este trabalho investiga, de forma preliminar, o comportamento da eficiência de códigos de Huffman binários aplicados a fontes sem memória com três símbolos em seu alfabeto e suas extensões. Foi observado que, de forma semelhante ao caso em que os alfabetos de fonte e código possuem a mesma cardinalidade, nem sempre há ganho na eficiência da codificação de Huffman ao se aumentar a ordem da extensão da fonte, quando se usa códigos binários mapeando fontes com alfabeto de três símbolos.**

***Palavras-Chave*— Teoria da informação, codificação de fonte, eficiência de códigos, códigos de Huffman.**

***Abstract*— This paper investigates, in a preliminary way, the behavior of the efficiency of binary Huffman codes applied to memoryless sources with three symbols in its alphabet and its extensions. It was observed that, similarly to the case where source and code alphabets have the same cardinality, there is not always gain in efficiency of the Huffman coding by increasing the order of the source extension, when using binary codes to map sources with alphabet with three symbols.**

***Keywords*— Information theory, source coding, coding efficiency, Huffman coding.**

I. Introdução

A importância da Codificação de Fonte se relaciona diretamente com aspectos fundamentais da capacidade e desempenho de um sistema de comunicações [1]. Considerando as tendências atuais, as quais demandam cada vez mais comunicação entre as pessoas por meio de seus diversos dispositivos eletrônicos, codificar a informação de forma eficiente se tornou um requisito indispensável para se obter sucesso em aplicações de comunicações.

Um dos resultados conhecidos da Teoria de Codificação de Fonte (conhecido como Primeiro Teorema de Shannon) [1], [2] estabelece que a eficiência de um código tende ao seu valor máximo à medida que se aumenta a ordem da extensão da fonte de informação. Isso significa que é possível aumentar a eficiência de códigos por meio da codificação do agrupamento de símbolos da fonte original. No entanto, foi mostrado em [3] que ao se utilizar a codificação de Huffman [4] para certas distribuições de probabilidade dos símbolos da fonte, existem casos em que o aumento da ordem da extensão causa a diminuição da eficiência, quando eficiências de fontes estendidas de ordens consecutivas são comparadas. Tais casos são denominados de "irregularidades".

Dessa forma, foram apresentadas em [3] duas contribuições principais: (i) uma análise do comportamento da eficiência de códigos de Huffman binários para extensões de fontes binárias de informação sem memória em função da distribuição de probabilidades da fonte original e da ordem da extensão (com maiores detalhes para o caso da fonte estendida de segunda ordem); e (ii) uma avaliação similar da eficiência de códigos de Huffman ternários para extensões de fontes sem memória com três símbolos. Não se abordou o caso de se utilizar códigos binários para mapear fontes cujo alfabeto possui mais de dois símbolos e suas extensões. Face ao exposto, este trabalho investiga, de forma preliminar, o comportamento da eficiência de códigos de Huffman binários aplicados a fontes sem memória com três símbolos em seu alfabeto e suas extensões. Na Seção II, são definidos conceitos relacionados à eficiência de um código de fonte. Na Seção III, resultados preliminares são apresentados e, por fim, conclusões são realizadas na Seção IV.

† Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife-PE, Brasil. E-mails: [ddn2, ebl3, dcunha]@cin.ufpe.br.

II. Eficiência de um Código

Considere uma fonte discreta sem memória com alfabeto $S = \{s_1, \ldots, s_k, \ldots, s_q\}$, com probabilidades de ocorrência $\{p_1, \ldots, p_k, \ldots, p_q\}$, respectivamente. A entropia por símbolo da fonte, denotada por $H(S)$, vale $H(S) = -\Sigma p_k \log(p_k)$. Ao codificarmos essa fonte (por meio de codificação de Huffman), cada símbolo s_k terá uma palavra-código associada de comprimento l_k, em que palavras-código mais curtas são atribuídas aos símbolos mais prováveis, enquanto palavras-código mais longas são alocadas aos símbolos menos prováveis. Dessa forma, dado que cada palavra-código de comprimento l_k mapeia um símbolo de probabilidade p_k, define-se o comprimento médio do código como $L = \Sigma p_k l_k$. Uma vez que o Primeiro Teorema de Shannon estabelece que $H(S) \leq L$, define-se a eficiência de um código, denotada por η, como $\eta = H(S)/L \leq 1$ [2].

A partir da fonte S, é possível obter fontes estendidas (sem memória) de ordem n, denotadas por S^n, assumindo que os alfabetos dessas fontes possuem símbolos que são formados por blocos de n símbolos da fonte S. As probabilidades dos símbolos das fontes estendidas são obtidas por meio do produto das probabilidades dos símbolos s_k que compõem o bloco. Obtida a fonte estendida S^n (alfabeto e distribuição de probabilidades), aplica-se a codificação de Huffman da mesma forma que no caso da fonte original. A eficiência do código que mapeia a fonte estendida S^n é então dada por $\eta_n = H(S^n)/L_n$, em que $H(S^n) = nH(S)$ é a entropia da fonte S^n e L_n é o comprimento médio do código associado à fonte S^n.

III. Resultados Numéricos

Inicialmente, implementou-se um algoritmo para a reprodução dos resultados obtidos em [3]. A partir daí, a

investigação foi ampliada considerando o uso de códigos de Huffman binários para extensões de fontes com três símbolos, conforme já mencionado anteriormente. A implementação do algoritmo foi realizada no *software* MATLAB [5].

A Fig. 1 mostra a ocorrência das "irregularidades" na eficiência da codificação de Huffman binária para uma fonte ternária, $S = \{A, B, C\}$, com probabilidades $\{p_A, p_B, p_C\}$, quando a extensão de ordem n é codificada em relação à codificação da extensão de ordem $(n-1)$. As perdas de eficiência apresentadas correspondem à diferença entre as eficiências das extensões em questão. A codificação da extensão da fonte foi realizada da segunda até a quarta ordem. Os conjuntos de probabilidades dos símbolos da fonte S foram identificados por um número de 1 a 833, em que os valores de p_A estão no intervalo $[0{,}01 - 0{,}33]$, os de p_B no intervalo $[0{,}01 - 0{,}49]$ e os valores de p_C são obtidos em decorrência da expressão $p_C = 1 - p_A - p_B$. Os conjuntos de probabilidades foram dispostos em ordem crescente dos valores de p_A, em que para cada valor de p_A fixado, variou-se p_B e p_C conforme já explicado. Nota-se na Fig. 1, que não houve "irregularidades" na eficiência quando se realizou a codificação da extensão de segunda ordem da fonte S. Por outro lado, percebe-se que "irregularidades" ocorreram para as codificações das extensões de maior ordem, com magnitude de perda na eficiência, em grande parte, menor que 0,02.

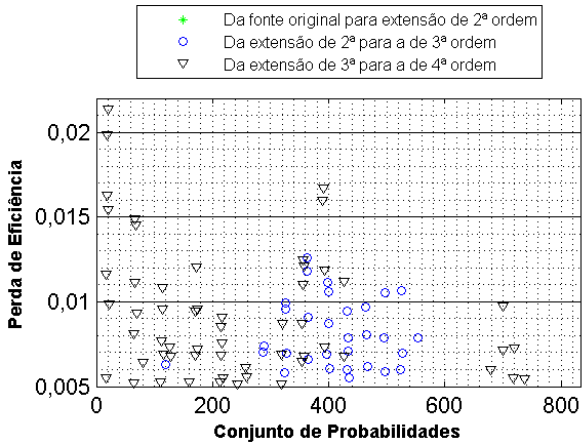

Fig. 1: Perda acima de $0{,}5\%$ da eficiência da codificação de Huffman binária entre as extensões da fonte $S = \{A, B, C\}$, em função das probabilidades $\{p_A, p_B, p_C\}$. Cada conjunto de probabilidades $\{p_A, p_B, p_C\}$ está identificado por um número (de 1 a 833).

A Fig. 2 ilustra a eficiência da codificação de Huffman binária para duas extensões de uma fonte $S = \{A, B, C\}$ com probabilidades $\{p_A, p_B, p_C\}$, quais sejam, a de segunda e a de terceira ordem. Na referida figura, p_A é denotada por "Prob. do 1º Símbolo" e seus valores estão no intervalo $[0{,}01 - 0{,}33]$, enquanto p_B é denotada por "Prob. do 2º Símbolo" e seus valores estão no intervalo $[0{,}01 - 0{,}49]$. Para as duas extensões apresentadas é possível observar regiões de "irregularidades", ou seja, trechos em que a eficiência do código da fonte estendida de segunda ordem supera a eficiência do código da fonte estendida de terceira ordem. Essas regiões são evidenciadas pelos trechos em que a superfície em linhas azuis está acima da superfície em linhas pretas.

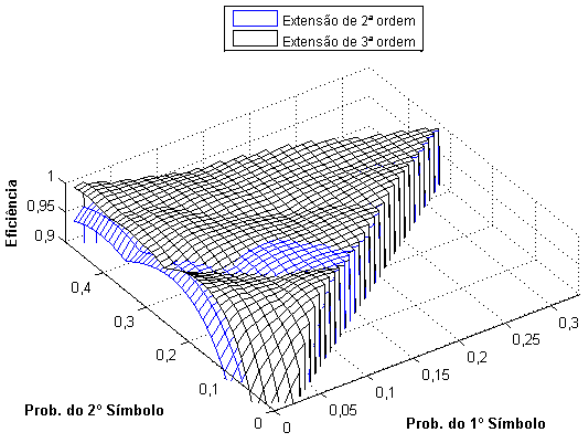

Fig. 2: Eficiência da codificação de Huffman binária de extensões de uma fonte com três símbolos em função da distribuição de probabilidades $\{p_A, p_B, p_C\}$.

IV. Conclusões

Este trabalho investigou, de forma preliminar, o comportamento da eficiência de códigos de Huffman binários aplicados a fontes sem memória com três símbolos em seu alfabeto e suas extensões. Observou-se que, de forma semelhante ao caso em que os alfabetos de fonte e código possuem a mesma cardinalidade, nem sempre há ganho na eficiência da codificação de Huffman ao se aumentar a ordem da extensão de uma fonte, quando se usam códigos binários mapeando fontes com alfabeto de três símbolos. Este fato não nos permite generalizar este comportamento para códigos r-ários e fontes q-árias, juntamente com suas extensões. As ideias apresentadas podem ser utilizadas na compressção de dados de forma direta, visto que é comum nesse tipo de aplicação, a diferença de cardinalidade entre a fonte de informação e o código. Como continuidade do trabalho, espera-se realizar uma análise mais detalhada dessa questão, como, por exemplo, aquela que foi realizada em [6], para obter condições que possibilitem garantir um ganho na eficiência à medida que a codificação das extensões são realizadas ou obter funções analíticas que descrevam o comportamento da codificação. Também existe a possibilidade de se estudar o comportamento específico das irregularidades em função da extensão da fonte.

Referências

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 2006.

[2] C. E. Shannon, "The mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379-423, 1948.

[3] P. M. Fenwick, "Huffman code efficiencies for extensions of sources," *IEEE Trans. Commun.*, vol. 43, n. 2/3/4, pp. 163-165, 1995.

[4] D. A. Huffman, "A method for the construction of minimum redundancy codes," *Proc. of the IRE*, vol. 40, pp. 1098-1101, 1952.

[5] D. Hanselman and B. Littlefield, "*MATLAB 6: Curso Completo*,". Prentice-Hall, 2003.

[6] J. Cheng, "On the expected codeword length per symbol of optimal prefix codes for extended sources," *IEEE Trans. Inf. Theory*, vol. 55, pp. 1692-1695, 2009.