ISSN: 2526-1371

Abordagem Entrópica de Codificação de Fonte usando compressão de dados

Kaio Pablo Gomes[1], **Eduardo El Akkari Sallum**[2], **Everton Leonardo Skeika**[3]

[1]Departamento de Informática
Universidade Tecnológica Federal do Paraná (UTFPR) – Ponta Grossa, PR – Brazil

[2]Departamento de Informática
Universidade Tecnológica Federal do Paraná (UTFPR) – Ponta Grossa, PR – Brazil

[3]Departamento de Informática
Universidade Tecnológica Federal do Paraná (UTFPR) – Ponta Grossa, PR – Brazil

kaiopablo@live.com, esallum@gmail.com, everton_skeika@hotmail.com

Abstract. *This paper presents an entropic coding approach for the transmission of information between a source and a destination interconnected under a LAN network architecture. In this scenario, an entropic font coding system using Huffman, with variable code size based on unequivocal frequencies of appearance of English letters, was proposed. To make this change in the form of coding, it is suggested to insert pre-coding and pre-decoding steps in the local network data transmission model. Huffman coding results show that in a 1460 textit bytes data packet, it is possible to send up to 2768 symbols, whereas currently with the ASCII system it can only send up to 1460. In the scenario considered, the methodology has 1.89 times more symbols to contribute to channel efficiency. Also, the efficiency of the Huffman coding obtained was superior to an alternative entropic approach evaluated by the Shannon-Fano technique, 99.19% against 99.16%, respectively.*

Resumo. *Este trabalho apresenta uma abordagem de codificação entrópica para a transmissão de informação entre uma fonte e um destino interligados sob uma arquitetura de rede LAN. Neste cenário, foi proposto um sistema de codificação entrópica de fonte utilizando Huffman, com tamanho de código variável baseado em frequências não-equiprováveis de aparição das letras da língua inglesa. Para realizar essa mudança na forma de codificar, sugere-se a inserção de etapas de pré-codificação e pré-decodificação no modelo de transmissão de dados em redes locais. Os resultados da codificação de Huffman demonstram que em um pacote de dados de 1460 bytes, é possível enviar até 2768 símbolos, enquanto que atualmente com o sistema ASCII pode ser enviado apenas até 1460. No cenário considerado, a abordagem comporta 1,89 vezes mais símbolos de modo a contribuir com a eficiência do canal. Também, a eficiência da codificação de Huffman obtida foi superior em relação a uma abordagem entrópica alternativa avaliada por meio da técnica de Shannon-Fano, 99,19% contra 99,16%, respectivamente.*

Keywords: Entropic coding; Local Area Networks; Huffman; ASCII; Data Transmission; Channel Efficiency; Shannon-Fano.

Palavras-chave: Codificação Entrópica; Redes Locais; Huffman; ASCII; Transmissão de Dados; Eficiência do Canal; Shannon-Fano.

1. Introdução

Os sistemas de redes locais estão fundamentados no padrão *Ethernet*. Nesse modelo de transmissão de dados, a comunicação é por meio de pacotes de dados denominados quadros ou *frames*, segundo o modelo OSI. Dentre outras especificações da tecnologia *Ethernet*, determina-se a técnica de codificação da fonte e do canal. A codificação da fonte é dada por meio da tabela *ASCII* para a representação dos símbolos transmitidos ou recebidos, enquanto que a do canal é por meio do modelo *Manchester* [Tanenbaum et al. 2003].

A eficiência de um canal está associada com a quantidade de dados que podem ser enviados em uma transmissão de mensagem [Forouzan 2009]. Ao possibilitar maior agregação de informação em um único pacote de dados, aumenta-se a eficiência do canal levando em consideração um cenário ideal. Nesse cenário, não considera outros fatores que afetam a comunicação de dados como: tempo de atraso, colisões de pacotes, dentre outros.

Neste trabalho, para permitir que mais informações possam ser agregadas em um único pacote, propõe-se uma abordagem entrópica de codificação de fonte usando compressão de dados. O objetivo principal é apresentar uma abordagem capaz de proporcionar maior eficiência para o canal considerando a transmissão de dados em redes locais.

2. Abordagem Proposta para Transmissão de Dados em Redes Locais

A abordagem proposta neste trabalho visa melhorar a comunicação entre a fonte e o destino utilizando uma codificação entrópica. Analisando o padrão dos sistemas de redes locais, a fonte utiliza a tabela ASCII para codificar as mensagens. No entanto, essa técnica de codificação é fundamentada no princípio de geração de códigos de tamanho fixo, sendo qualquer símbolo da tabela equiprovável.

A análise estatística representada pela Tabela 1 mostra a frequência das letras da língua inglesa [Zim 1962] . Conforme os valores identificados, as letras do alfabeto nesse idioma assim como em outros não são equiprováveis. Ao considerar um sistema de codificação em que não conta com a possibilidade de existir diferentes probabilidade, afeta-se a eficiência do processo de transmissão.

A ideia principal da nova abordagem que está sendo proposta é utilizar uma quantidade variável de informação por símbolo. Sendo assim, um processo entrópico de codificação é adotado. No entanto, conhecendo-se apenas as probabilidades de ocorrência das letras de um alfabeto, este trabalho visa codificar apenas tais tipos de caracteres. Ainda, considera-se letras pertencentes ao alfabeto inglês, uma vez que as frequências de utilização das mesmas segue a análise mostrada na Tabela 1.

Para conceder essa melhoria na transmissão de dados, a fonte passa a utilizar um componente adicional chamado pré-codificação. O componente inserido é responsável

Tabela 1. Tabela da Esquerda

Letra	Frequência
e	12.702%
t	9.056%
a	8.167%
o	7.507%
i	6.966%
n	6.749%
s	6.327%
h	6.094%
r	5.987%
d	4.253%
l	4.025%
c	2.782%
u	2.758%

Tabela 2. Tabela da Direita

Letra	Frequência
m	2.406%
w	2.360%
f	2.228%
g	2.015%
y	1.974%
p	1.929%
b	1.492%
v	0.978%
k	0.772%
j	0.153%
x	0.150%
q	0.095%
z	0.074%

por permitir um processo de codificação alternativo para a fonte. Tanto a codificação por meio da tabela ASCII, como a entrópica de Huffman, são possíveis uma vez que verifica-se inicialmente os símbolos que formam a mensagem a ser enviada. Toda vez que uma letra é identificada, utiliza-se o algoritmo de Huffman para codificar tal caractere. Por outro lado, qualquer outro símbolo que não seja uma letra do alfabeto adotado, mapeia-se em código ASCII respeitando a mesma estrutura já existente nos sistemas de redes locais.

Além da fonte, o destino precisa sofrer um ajuste necessário para conseguir lidar com a abordagem entrópica. Por isto, é necessário adicionar um pré-decodificador, o qual é acionado quando uma mensagem composta apenas de letras é recebida. Essa mensagem não chega a ser decodificada pela maneira usual, consultando a tabela ASCII. O pré-decodificador passa a consultar uma nova tabela referente ao mapeamento das letras em códigos Huffman.

As mensagens que antes seriam enviadas por uma única transmissão de dados podem depender de mais de um envio de pacote de dados. Essa relação de dependência está vinculada com o processo entrópico proposto, que modela apenas símbolos referentes a letras. Não é possível misturar a codificação de Huffman e ASCII, uma vez que o processo de decodificação não saberia interpretar os dados transmitidos. Sendo assim, uma mensagem pode ser quebrada em um ou mais envios de dados, dependendo inclusive da taxa de MTU (*Maximum Transmission Unit*) que é seguida no padrão *Ethernet*.

2.1. Compressão de Dados

A abordagem entrópica adotada para a fonte, permite uma codificação mais eficiente por meio de compressão de dados. A informação representada em cada símbolo varia de acordo com a probabilidade de ocorrência do mesmo. Para esse trabalho, duas técnicas de compressão de dados foram estudadas e consideradas: algoritmo de Shannon-Fano e algoritmo de Huffman. Apesar de as duas técnicas de compressão poderem apresentar resultados similares em alguns contextos, o funcionamento delas são diferentes na forma como

cada técnica manipula uma árvore de codificação [Kodituwakku and Amarasinghe 2010]. A eficiência da codificação e da fonte utilizando essas técnicas de compressão de dados podem variar de acordo com o contexto da aplicação, inclusive entre ambas as técnicas os resultados podem apresentar uma diferença considerável. A Tabela 3 apresenta a aplicação do algoritmo de Shannon-Fano para esse conjunto de símbolos, respeitando as suas probabilidades de ocorrências identificadas pela Tabela 1.

Para compararmos os resultados obtidos pela codificação de Shannon-Fano, utilizamos a codificação de Huffman. A Tabela 4 mostra o resultado da codificação de Huffman. Nota-se que ambos os algoritmos apresentaram resultados mais eficientes do que o sistema ASCII, em termos de quantidade de *bits*, para representar cada letra do alfabeto inglês.

Tabela 3. Codificação Shannon

Shannon-Fano		
Símbolo	**Probabilidade**	**Codificação**
e	12.702%	000
t	9.056%	001
a	8.167%	0100
o	7.507%	0101
i	6.966%	0110
n	6.749%	0111
s	6.327%	1000
h	6.094%	1001
r	5.987%	1010
d	4.253%	1011
l	4.025%	11000
c	2.782%	11001
u	2.758%	11010
m	2.406%	11011
w	2.360%	11100
f	2.228%	111010
g	2.015%	111011
y	1.974%	111100
p	1.929%	111101
b	1.492%	111110
v	0.978%	1111110
k	0.772%	11111110
j	0.153%	1111111100
x	0.150%	1111111101
q	0.095%	1111111110
z	0.074%	1111111111

Tabela 4. Codificação Huffman

Huffman		
Símbolo	**Probabilidade**	**Codificação**
e	12.702%	100
t	9.056%	000
a	8.167%	1110
o	7.507%	1101
i	6.966%	1011
n	6.749%	1010
s	6.327%	0111
h	6.094%	0110
r	5.987%	0101
d	4.253%	11111
l	4.025%	11110
c	2.782%	01001
u	2.758%	01000
m	2.406%	00111
w	2.360%	00110
f	2.228%	00101
g	2.015%	110011
y	1.974%	110010
p	1.929%	110001
b	1.492%	110000
v	0.978%	001000
k	0.772%	0010011
j	0.153%	001001011
x	0.150%	001001010
q	0.095%	001001001
z	0.074%	001001000

Ao avaliar as duas codificações geradas, percebe-se que ambos os sistemas de

codificação apresentam resultados diferentes para o contexto da abordagem proposta. Nesse cenário de aplicação, para criar a tabela de Shannon-Fano precisa de 148 *bits*, enquanto que a de Huffman precisa de 143 *bits*. Esse valor é a soma dos *bits* de cada símbolo, nesse caso as letras de a-z. O comprimento médio de *bits* por símbolo é 4,21977 *bits* para a codificação Shannon, enquanto que utiliza-se 4,21854 *bits* para a de Huffman. Considerando que é a entropia da fonte $H(X)$ é de 4,18 *bits* por mensagem, calcula-se a relação entre a entropia e comprimento médio dos símbolos. A eficiência da codificação de Shannon é de 99,16%, enquanto que a de Huffman é 99,19%. Como a eficiência de Huffman apresentou um melhor desempenho, a abordagem proposta utiliza a codificação Huffman para a sua abordagem entrópica de compressão de dados.

2.2. Avaliação da Abordagem Proposta

A abordagem proposta foi avaliada por meio da análise dos resultados implicados pela utilização de compressão de dados pela fonte. É feita uma comparação entre a quantidade de dados enviados por pacote em sistemas de redes locais. A técnica de compressão adotada pela abordagem é a de Huffman. O gráfico ilustrado pela Figura 1 mostra a comparação entre a abordagem proposta e os sistemas de redes locais que utilizam a tabela ASCII.

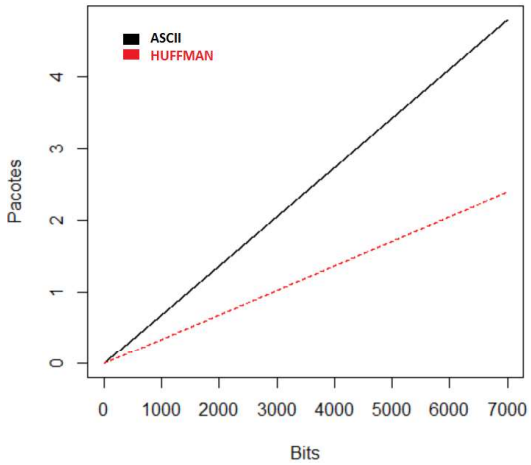

Figura 1. Gráfico comparativo de pacotes por *bits* do Huffman com o ASCII

O gráfico presente na Figura 1 mostra que o uso da abordagem proposta possibilita enviar aproximadamente 1,98 vezes mais símbolos em cada pacote de dados. O comprimento médio dos símbolos com a codificação Huffman é de 4,21854 *bits*, enquanto que o tamanho médio com o código ASCII é de 8 *bits*. Considerando o cenário de redes locais com *Ethernet*, o tamanho máximo de dados que um pacote de dados comporta é 1460 bytes ou 11680 *bits*. Em apenas um único pacote com dados comprimidos pela abordagem proposta é possível enviar aproximadamente até 2768 símbolos. No entanto, ao utilizar dados codificados em ASCII pode-se enviar apenas 1460 símbolos.

Por meio desta avaliação, identifica-se que a aplicação de uma abordagem entrópica de compressão de dados possibilita o aumento da eficiência do canal. Essa

melhoria é possível uma vez que a fonte passa a codificar símbolos com menos dados, e consequentemente uma maior quantidade de símbolos podem ser enviados pelo canal. Ressalta-se que os símbolos comprimidos pela abordagem proposta são apenas as letras do alfabeto inglês, qualquer outro caractere tais como números, pontos, dentre outro não são considerados. Para realizar a transmissão desses outros símbolos, a execução de transmissão considerando codificação ASCII é adotada.

3. Conclusão

De acordo com os resultados obtidos na avaliação da abordagem sugerida, a compressão de dados pode aumentar a eficiência do canal. Conhecendo a probabilidade de ocorrência dos símbolos codificados pela fonte, torna-se possível uma abordagem entrópica na qual dados podem ser transmitidos de forma comprimida. A escolha da codificação, nesse trabalho, é feita com base na comparação da eficiência entre dois algoritmos de compressão de dados: Huffman e Shannon-Fano.

A abordagem sugere que a fonte realize uma combinação dos sistemas de codificação ASCII e Huffman. No contexto aplicado, considera-se a utilização de Huffman para a transmissão de símbolos referentes as letras do alfabeto inglês, enquanto que o sistema ASCII codifica todos seus caracteres com exceção das letras. Para possibilitar essa mudança nos sistemas de rede local, um processo de pré-codificação e pré-decodificação é proposto.

A principal vantagem associada a aplicação da abordagem proposta é a possibilidade de enviar mais informações em um único pacote de dados. No entanto, outros fatores associados ao impacto de adicionar mecanismos de codificação e decodificação de código Huffman pode afetar a eficiência da transmissão de dados.

A extensão dos estudos da abordagem proposta pode ser feita por meio de um trabalho futuro relacionado a aplicação da abordagem em cenários reais considerando o tempo de atraso dos pacotes, confiabilidade e outros parâmetros avaliativos. A eficiência da codificação e a redução do número de transmissões de dados, são parâmetros que podem ser aprimorados ao utilizar um único sistema de representação de símbolos nas fonte. Um estudo para avaliar a probabilidade de ocorrência de símbolos em uma transmissão pode proporcionar a utilização de um único sistema de codificação.

Referências

Forouzan, B. A. (2009). *Comunicação de dados e redes de computadores*. AMGH Editora.

Kodituwakku, S. and Amarasinghe, U. (2010). Comparison of lossless data compression algorithms for text data. *Indian journal of computer science and engineering*, 1(4):416–425.

Tanenbaum, A. S. et al. (2003). Computer networks, 4-th edition. *ed: Prentice Hall*.

Zim, H. S. (1962). Codes and secret writing (abridged edition). *Scholastic Book Services, fourth printing*.