Codes for Compression and Error Detection

Paulo E. D. Pinto[1]

[1]Instituto de Matemática e Estatística
Universidade do Estado do Rio de Janeiro (UERJ)
Rua São Francisco Xavier, 524, Rio de Janeiro – RJ – 20550-900
Rio de Janeiro – RJ – Brazil

pauloedp@ime.uerj.br

Abstract. *This article summarizes the main results about codes that combine the data compression power of Huffman trees with the error detection mechanisms proposed by Hamming. Two methods are described: the even trees method and the Hamming-Huffman trees method. Most of the works are the result of a close collaboration with Jayme Luiz Szwarcfiter.*

1. A Challenge in Coding Theory

In 2001, I was finally able to start my doctorate at COPPE/UFRJ. As a computer professional, I have had two careers. In the industry, I worked at Petrobras for 34 years. First, as a software developer, and then as a manager of technological innovations in software. At UERJ, I am an assistant professor in the computer science department since 1984 working in the area of Algorithms and Data Structures.

At that time, I met Jayme L. Szwarcfiter at a street market in Urca, a neighborhood in which we both used to live. In this meeting, I asked him for supervision on a doctorate thesis, which he promptly accepted. I started my doctorate having Jayme and Fabio Protti as advisors. After finishing the courses, Jayme offered me the challenge of working with a code that could integrate the tasks of compression and error detection. Interestingly, this subject was not Jayme's main line of research, Graph Theory, but it was quite suitable for me, since I had little familiarity with graphs. This decision is already a revelation of the versatility and generosity of this advisor, concerned with the possibilities of his advisees.

In the next section, I present some basic concepts related to my doctoral work and, following that, the main results that were obtained.

2. Huffman-like codes

Huffman codes [A. Huffman 1951] is one of the most traditional methods to compress data. An important aspect of these codes is the possibility of handling encodings of variable sizes and obtaining a set of prefix-free encodings. A great number of extensions and variations of the classical Huffman codes have been described through time.

On the other hand, Hamming formulated algorithms for the construction of error-detecting codes. Further, Hamming [Hamming 1986] posed the problem of describing an algorithm that would combine the advantages of Huffman codes with the noise protection of Hamming codes. The idea is to define a prefix code in which the encodings contain redundancies that allow the detection of certain kinds of errors. This is equivalent to forbidding some encodings which, when present in the reception, would signal an error.

Cadernos do IME - Série Informática
e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2022.70590

Such a code is a Hamming-Huffman code and its representing binary tree is a Hamming-Huffman tree. In a Huffman tree, all leaves correspond to encodings. In a Hamming-Huffman tree, there are encoding leaves and error leaves. Hitting an error leaf in the decoding process indicates the presence of an error. The problem posed by Hamming was to devise an efficient method to create that kind of tree. Figure 1 depicts an example of a Hamming-Huffman tree for five symbols, given by Hamming [Hamming 1986] in his book, page 76. The black leaves are error leaves used for error detection.

Possible advantages of this idea to other methods of compression with error detection would be a smaller encoded message, a higher capability of error detection, a faster encoding or decoding process, and a faster signaling of occurrence of errors.

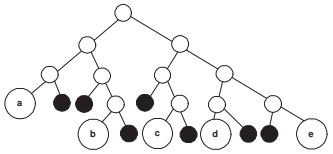

Figure 1. A Hamming-Huffman tree for 5 symbols.

After experimenting for some time with Hamming's idea, whose implementation proved to be very difficult, remaining an open problem, I ended up with an alternative simpler proposal, but of the same nature, the even trees. Even trees are also Huffman-like trees, also having error leaves, and with the basic property that all encodings have even parity. Figure 2 is an example of an optimal even tree for 11 symbols.

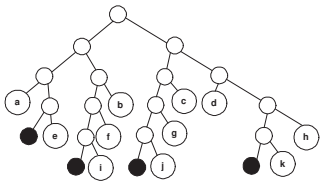

Figure 2. An optimal even tree for 11 symbols.

3. Main results

The initial results of the doctoral work are published in [Pinto et al. 2003], [Pinto et al. 2004], and [Pinto et al. 2005]. It can be observed that the most modern notation for even trees was not yet used in those publications. They addressed data compression in the situation of uniform frequencies for the symbols. A complete mathematical formulation was developed to characterize the optimal even trees for the described situation.

The production of a text with such a mathematical rigor, under the competent supervision of my advisors, meant growth for me as a researcher in exact sciences. Until then, I used an approach more focused on experimentation and empiricism in my scientific productions.

The next stage of the doctorate was to consider the general case of data compression, in the presence of arbitrary frequencies of symbols. An exact algorithm of complexity $O(n^3)$ was developed to create an optimal even tree for n symbols and also an

approximation algorithm of complexity $O(n \log n)$. It obtains even trees whose cost is at most $\frac{7}{6}$ of that produced by the corresponding Huffman tree. That is, the method created was very good, as the marginal effort of adding error detection capability to the Huffman tree is very low. Such results are published in [Pinto et al. 2009] and [Pinto et al. 2012].

4. Further results

Some time after finishing my doctorate, I supervised the master's thesis of Moysés S. Sampaio. He resumed the study of Hamming-Huffman trees. One of the results of this thesis was to show that the problem of creating Hamming-Huffman trees for n symbols is related to that of finding, in the hypercube $Q_{\lceil \log n \rceil + 1}$, an independent set with n vertices having minimum neighborhood. Part of this thesis was published in [Faria et al. 2016].

The recent resumption of this work, now with the participation of Jayme, obtained a relevant finding for the case of uniform frequencies, which is to guarantee that the optimal trees have leaves at a maximum of four different levels. This fact contributed to the solution of the problem of creating optimal Hamming-Huffman trees when the number of distinct levels is 1 or 2 and the symbol frequencies are uniform. This constitutes an advance towards the complete solution of the problem, which remains open. The result will be published in [Lin et al. 2022].

5. Acknowledgements

In this article, I described part of the work that I have carried out together with Jayme L. Szwarcfiter. I have already mentioned the richness of this relationship to me, for the competence, friendship, and generosity of Jayme. In recent years, we have had a closer collaboration, as he has the position of a visiting researcher at UERJ, starting in 2016. Since then, we have been working on two books, mentored students, shared some courses, and written several articles. In other words, I have been very privileged to have had many opportunities to produce together with Jayme and this article is to sincerely thank him for everything.

References

A. Huffman, D. (1951). A method for the construction of minimum redundancy codes. *Proceedings of the IRE*, 40:1098–1101.

Faria, L., Oliveira, F. S., Pinto, P. E. D., and Jr., M. S. S. (2016). On the minimum neighborhood of independent sets in the n-cube. *Matemática Contemporanêa*, 44:1 – 10.

Hamming, R. W. (1986). *Coding and Information Theory*. Prentice-Hall.

Lin, M. C., Oliveira, F. S., Pinto, P. E. D., JR, M. S. S., and Szwarcfiter, J. L. (2022). Restricted hamming-huffman trees. *Rairo - Theoretical Informatics and Applications*, aceito para publicação.

Pinto, P. E. D., Protti, F., and Szwarcfiter, J. L. (2003). Compactação de dados com cetecção de erros. *Anais do XXXV Simpósio Brasileiro de Pesquisa Operacional (SBPO 2003)*, 1:2463–2472.

Pinto, P. E. D., Protti, F., and Szwarcfiter, J. L. (2004). A Huffman-based error detecting code. *Proc. of the Experimental and Efficient Algorithms (WEA'2004), Lecture Notes in Computer Science*, 3059(6):446–457.

Pinto, P. E. D., Protti, F., and Szwarcfiter, J. L. (2005). Parity codes. *Rairo - Theoretical Informatics and Applications*, 39:263 – 278.

Pinto, P. E. D., Protti, F., and Szwarcfiter, J. L. (2009). Exact and experimental algorithms for a huffman-based error detecting code. *Lecture Notes in Computer Science*, 5532:311 – 324.

Pinto, P. E. D., Protti, F., and Szwarcfiter, J. L. (2012). Exact and approximation algorithms for error-detecting even codes. *Theoretical Computer Science*, 440-441:60 – 72.