Um algoritmo aproximativo para a construção de árvores alfa-pares

Moysés da Silva Sampaio Júnior
Universidade do Estado do Rio de Janeiro - UERJ
20550-013, Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro, RJ
e-mail: moyses.sampaio@gmail.com

Paulo Eustáquio Duarte Pinto
Universidade do Estado do Rio de Janeiro - UERJ
20550-013, Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro, RJ
e-mail: pauloedp@ime.uerj.br

RESUMO

Inspirados em uma proposta de Hamming, de combinar compactação com detecção de erros, Pinto et all propuseram e desenvolveram métodos para a construção de árvores pares, que são árvores no estilo das árvores de Huffman, com a particularidade de que todas as codificações têm número par de bits 1. Essas árvores têm capacidade de detectar grande quantidade de erros introduzidos em uma mensagem compactada. Um código par pode ser representado em uma árvore binária onde existem folhas referentes a símbolos e folhas de erro. Mais recentemente foi proposta, pelos mesmos autores, uma extensão das árvores pares, as árvores alfa-pares, onde a capacidade de detecção de erros é ampliada, com base em um parâmetro alfa, que varia de 0 a 1. Já existem algoritmos para construção de árvores alfa-pares ótimas. Entretanto esses algoritmos têm complexidade $O(n^3)$ em tempo e $O(n^2)$ em espaço, o que inviabiliza seu uso para grandes volumes de dados. O presente artigo descreve um algoritmo aproximativo para a construção de árvores 1-pares quase ótimas a partir das árvores de Huffman.

A árvore de Huffman é percorrida por níveis. Sempre que são encontradas folhas em dado nível, é tomada uma decisão de alteração na árvore, considerando a soma das frequências das folhas do nível atual. Se essa soma for igual ou superior à soma das frequências das folhas dos níveis abaixo, um novo nível é criado na árvore e todos os nós do nível atual são removidos para nós de codificação par do nível criado. Os nós ímpares desse nível criado são transformados em folhas de erro. O processo continua no nível seguinte ao nível criado. Caso contrário, todas as folhas passam a ter dois nós filhos. Os símbolos das folhas são removidos para os filhos de codificação par. Os filhos ímpares são ocupados por subárvores removidas dos níveis abaixo, escolhidas por ordem decrescente da soma das frequências das folhas de cada subárvore. O processo prossegue a partir do nível seguinte.

Os resultados obtidos até o presente momento são os seguintes:

1. O algoritmo é 2-aproximativo em relação à construção da árvore de Huffman.

2. A complexidade do algoritmo é $O(n^2)$ em tempo e $O(n)$ em espaço. Espera-se em breve baixar a complexidade de tempo para $O(n \log n)$, o melhor que se pode ter.

3. Em experimentos práticos, envolvendo até 16000 símbolos, o aumento de custo das árvores aproximadas obtidas aumentou apenas 28% em relação à árvore de Huffman, quando os símbolos eram palavras da língua portuguesa, que tendem a ter uma distribuição Zipfiana. Quando se trabalhou com frequências aleatórias, o custo aumentou, no máximo, em 10%. A comparação com árvores 1-pares ótimas ainda não foi feita, mas só melhora esses resultados.

Palavras-chave*:* Algoritmos,Compactação, Detecção de Erros.
Área principal: TEL&SI.