Para compreender técnicas ou algoritmos de compressão, é preciso conhecer a base teórica em que tais técnicas são desenvolvidas. Nessa seção serão apresentados conceitos de compressão de dados, fundamentos teóricos de teoria da informação, que servem como base para os algoritmos sem perda. Dos algoritmos sem perda, serão exibidas também técnicas de compressão sem perdas com código de tamanho variável, técnicas de dicionário e técnicas baseadas em contexto.

Subsection 2.1

Conceitos básicos de Teoria da Informação

//Trocar de lugar?(Técnicas de compressão ou algoritmos de compressão são divididas em dois algoritmos, um de compressão, onde dada um entrada X, o algoritmo de compressão gera uma saída Xc que precisa de menos bits para ser representada. E o algoritmo de reconstrução, que tem como entrada Xc e geram Y de saída. Quando Y é idêntico à entrada inicial X, esse esquema de compressão é considerado sem perdas, se não, é chamado de compressão com perdas. Algoritmos sem perda são utilizados em casos onde a informação é não-tolerante a diferenças entre os dados originais e os dados reconstruídos. Nos casos onde são aceitáveis diferenças entre os dados originais e os dados reconstruídos, técnicas com perdas podem ser utilizadas para reduzir o tamanho de Xc. Algoritmos sem perda têm como base conceitos em teoria da informação.)

A idéia de quantizar informação é apresentada por Claude Elwood Shannon, como auto-informação. Dado um evento A, um conjunto de resultados de um experimento qualquer. Sendo P(A), a probabilidade do evento A ocorrer. Então a auto-informação associada com tal evento é expressa por:

i(A) = log b (1/P(A)) = - log b P(A)

Onde a base b do logaritmo representa a unidade do evento A, no caso de b = 2, a unidade é bits, em b = e, a unidade é natural e em b = 10 a unidade é decimal. Sabendo que log (1) = 0 e –log(x) cresce quando x diminui, a auto informação i então cresce quando a probabilidade de um evento é pequena. Ou seja, em casos onde um evento pouco provável acontece, a informação associada com esse evento é mais alta do que em eventos que são mais prováveis de acontecer. Essa noção é importante e será base de alguns esquemas que serão apresentados. //cite book

Um fator que deriva da noção de auto-informação é a entropia, que é dado um conjunto de eventos independentes Ai, de um experimento S, a média da auto-informação associada com esse evento é dada por:

H= sum(P(Ai)i(Ai)) = -sum(P(Ai)logb(P(Ai)))

Tal quantidade é chamada de entropia associada ao experimento. No caso em que S é uma fonte que possui símbolos Ai de um alfabeto qualquer A(A diferente) e b = 2, a entropia representa a média de símbolos binários que são necessários para codificar essa fonte. //cite book

Outra quantidade importante para a compressão sem perdas é a entropia da fonte, que como mostrado por Shannon //Cite Shannon representa o melhor desempenho que uma técnica de compressão sem perda pode alcançar. A entropia da fonte é dada por:

H(S) = lim 1/n Gn

Onde:

Gn =...

Normalmente, a entropia é estimada, já que não é possível conhecer a entropia de uma fonte física. Então tal estimativa depende dos pressupostos que são feitos sobre a estrutura de uma fonte, conhecer a estrutura dos dados de uma fonte, pode ajudar a reduzir a entropia estimada a fim de chegar o mais próximo possível da entropia da fonte. Os pressupostos são separados em modelos matemáticos, e é possível obter modelos baseados na análise empírica das estatísticas dos dados. E uma boa modelagem tende a trazer melhores resultados em esquemas de compressão.//cite book