

UNIVERSIDADE FEDERAL DO AMAZONAS
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INOVAÇÃO CIENTÍFICA

UM ESTUDO SOBRE TÉCNICAS DE COMPRESSÃO DE DADOS
DE BAIXO CUSTO COMPUTACIONAL PARA OTIMIZAÇÃO DE
COMUNICAÇÃO ASSÍNCRONA EM PLATAFORMAS MÓVEIS

Bolsista: André Luís Serudo da Silva, CNPq

MANAUS

2011

UNIVERSIDADE FEDERAL DO AMAZONAS
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INOVAÇÃO CIENTÍFICA

RELATÓRIO FINAL

PIB-E/0020/2010

UM ESTUDO SOBRE TÉCNICAS DE COMPRESSÃO DE DADOS
DE BAIXO CUSTO COMPUTACIONAL PARA OTIMIZAÇÃO DE
COMUNICAÇÃO ASSÍNCRONA EM PLATAFORMAS MÓVEIS

Bolsista: André Luís Serudo da Silva

Orientador: Prof. Dr. César Augusto Viana Melo

MANAUS

2011

RESUMO

Recentemente, podemos visualizar o uso crescente das mensagens de texto como um dos principais métodos de comunicação entre usuários de dispositivos móveis. E com isso, temos novos campos de perspectivas a serem trabalhados para a sua utilização, tais como: monitoramento de pessoas e patrimônios. Nesses novos contextos, a restrição quanto ao tamanho de carga útil de cada mensagem é um impedimento a ser considerado. Por outro lado, o emprego de técnicas de compressão de dados de baixo custo operacional, porém se destacada eficiência, representa uma alternativa para o uso efetivo da comunicação via SMS na construção de aplicações mais sofisticadas. Este trabalho teve como objetivo a identificação de técnicas de compressão de baixo custo operacional, assim otimizando o processo de comunicação de sistemas de monitoramento geográfico entre plataformas móveis. Com isso é essencial a primeiro passo, o processo de levantamento das técnicas existentes para que se possa ser feita uma análise de suas características e após escolhida-a, uma avaliação de sua efetividade e capacidade de otimização para que se tenha resultados mais sólidos. Sendo assim, dentre os métodos de compressão estatísticos, optou-se pela utilização da compressão de Huffman. Método o qual utiliza a probabilidade de ocorrência de determinada informação dentro do conjunto de dados a ser comprimido, gerando assim uma sequência binária que representará a informação. Tendo como base de entrada uma série contendo informações de localização geográfica de um móvel obtida a partir de um sistema GPS, e através dela, amostras foram tiradas e submetidas a compressão. Diferentes saídas foram obtidas, de acordo com a mudança em que a composição da amostra era submetida, procurando sempre ter como medida de controle o limite de caracteres suportados em uma mensagem SMS. A solução baseada na compressão de Huffman, não mostrou resultados expressivos. No entanto, após a realiza-

ção de otimizações, envolvendo tal técnica verificou-se altas taxas de compressão, em sua versão modificada viabilizando-a como a técnica de compressão a ser implementada.

Sumário

1. INTRODUÇÃO.....	5
2. REVISÃO BIBLIOGRÁFICA.....	6
2.1 Tecnologias Utilizadas.....	6
2.2 Compressão de dados.....	7
3. MÉTODOS UTILIZADOS.....	9
4. RESULTADOS E DISCUSSÕES.....	11
5. CONCLUSÕES.....	16
6. FONTES E REFERÊNCIAS BIBLIOGRÁFICAS.....	17
7. CRONOGRAMA EXECUTADO.....	18

1. INTRODUÇÃO

Entre as várias técnicas de utilização da tecnologia de comunicação celular-móvel, o envio de mensagens de texto, torpedos ou SMS, pode ser apontado como uma das aplicações de maior sucesso entre os usuários. O relatório[1] publicado pelo CETIC.br aponta que 55% dos usuários da telefonia móvel faz uso de SMS para se comunicar. Esse número está próximo de 75% dos usuários quando se considera a faixa etária de 16 a 24 anos. Ainda que amplamente utilizado, o serviço de SMS impõe restrições aos seus usuários, sendo a principal delas o número de caracteres que cada mensagem pode carregar em sua carga útil. Na grande maioria dos prestadores desse serviço esse número não excede 160 caracteres. No entanto, essa restrição tem sido superada em parte pela criatividade dos usuários que utilizam, em geral, uma linguagem fonética para escrever seus textos, conseguindo assim reduzir o tamanho das palavras, e com isso otimizar a carga útil por mensagem enviada.

Essa utilização massiva abre perspectivas para a sua utilização em outros contextos, como por exemplo: monitoramento de pessoas e patrimônios. Aonde que o usuário fica sabendo via SMS, a localização do patrimônio a ser monitorado. Nesse novo contexto, a restrição quanto a carga útil de cada mensagem é algo a ser considerado. Neste trabalho estudou-se técnicas de compressão de baixo custo operacional que permitem a otimização da carga útil transportada em mensagens SMS. Avaliou-se em tal cenário a utilização do algoritmo de Huffman.

Este relatório está organizado da seguinte forma: Na seção 2 mostra-se a remissão bibliográfica realizada para confecção deste trabalho. Na seção 3, os méto-

dos em que foi conduzido o desenvolvimento da pesquisa. Na seção 4, os resultados em que este resultou e suas conclusões, na seção 5. Contidos na seção 6 e 7, temos respectivamente, as referências bibliográfica e o cronograma de execução.

2. REVISÃO BIBLIOGRÁFICA

2.1 Tecnologias Utilizadas

Para desenvolvimento da aplicação de teste neste projeto, foram usadas as seguintes tecnologias web: HTML, CSS e Javascript. E cada uma dessas tecnologias, desempenhou um papel específico na construção da aplicação. O HTML (Hyper Text Markup Language), é uma linguagem com a qual se definem páginas web. Realiza-se isso através de marcadores(tags) que são usadas para separar texto e outros elementos da página. Cada marcador realiza uma função específica, mais temos três marcadores básicos iniciais: o html, responsável por definir o início e o fim do arquivo html elaborado; o head, que funciona como um cabeçalho onde podemos definir padrões e também aonde podemos utilizar esse espaço para incluir códigos de outras linguagens de programação; e por fim o body, o conteúdo em si da página. Aqui se faz a verdadeira manipulação do conteúdo disposto na página a ser construída.

O CSS (Cascading Style Sheet) é uma linguagem usada para alterar estilos em linguagens de marcação, tais como html e xml. Através de seu uso, pode-se realizar a criação de estilos em marcadores específicos e até em páginas inteiras, criando com isso a modelagem do que pode-se chamar de design da página e componentes que a compõem. Como linguagem de programação escolhida, temos o Javascript como uma das principais linguagens web atuais. Concebida como uma linguagem script que adota paradigmas de orientação a objetos, é uma linguagem que atualmente está bem forte no mercado. Além de apresentar recursos e possibilidades que se assemelham a outras linguagens mais comumente usadas, tais como C++ e Java.

Trabalhar no desenvolvimento para dispositivos móveis, implica em produzir softwares para dispositivos com baixo recurso computacional, quando comparado a capacidade dos atuais laptops e computadores. Com isso, há um processo maior de delimitação quanto ao desenvolvimento. Já que softwares que fazem uso de um uso excessivo do processador implica em um consumo mais prolongado de energia. Gasto que poderia ser evitado. Para solucionar esse problema, já existem plataformas desenvolvidas que auxiliam o desenvolvedor nas mais diversas linguagens de programação, podemos citar: Java ME, plataforma Symbian e Android. Essas plataformas foram propostas por diferentes fornecedores, e as aplicações desenvolvidas para elas são utilizadas em aparelhos de mais de um fabricante.

Neste projeto, utilizou-se a plataforma Symbian, e suas aplicações foram desenvolvidas usando a tecnologia WRT(Web Run Time), que integra serviços da web 2.0 com serviços da plataforma móvel. Essa tecnologia, disponibilizada pela Nokia à comunidade de desenvolvedores, possui um elemento de software conhecido como engine webKit. Essa engine permite a interpretação do conjunto de códigos HTML, CSS e Javascript, que geralmente são usados somente em páginas web, mas aqui são interpretados pela plataforma como uma aplicação móvel. Ainda disponível a comunidade, temos a possibilidade da obtenção de documentação e uma interface de programação de aplicações (APIs), escritas em Javascript.

2.2 Compressão de dados

Temos que a compressão de dados tem como objetivo a representação de um conjunto de informações em menor escala, possibilitando assim um melhor aproveitamento do espaço a ser utilizado. Para isso, algoritmos de compressão, avaliam a informação e interpretam as variações dos caracteres e determinar a melhor forma de reduzir a sua utilização.

A respeito dos algoritmos de compressão, uma forma de dividi-los é quanto a sua operação, temos: os métodos estatísticos, tais como: método Huffman e o método aritmético; e os métodos baseados em dicionários: como o LZ77 e o LZ78. Para o projeto, verificou-se que como o modelo estático sendo a melhor alternativa. Já que estamos tratando da probabilidade de um ambiente para dispositivos móveis, precisamos ser cientes de que a capacidade computacional do mesmo é reduzida. Ainda a princípio temos somente os métodos originais, só que temos que fazer uma avaliação perante a capacidade de adaptação e otimização desse possível código. E para isso, a codificação Huffman corresponde ou pelo menos, é a resposta mais próxima ao melhor cenário.

A codificação Huffman, proposta por David Huffman em 1952, permite a atribuição de um conjunto de informações a ser comprimido em uma sequência de números de código binário. Através da probabilidade de ocorrência de determinado símbolo no conjunto de dados, é adotado como seleção para a atribuição de um valor em código binário de tamanho variável. A codificação se dá através da construção de uma árvore ordenada a partir dos símbolos e suas frequências.

De modo que ao fim da compressão, cada símbolo que compõe a informação inicial, tem um representante codificado, correspondente ao percurso da raiz até a folha em que o símbolo corresponde na árvore. E com o conjunto dessas representações, podemos montar a informação comprimida de saída.

3. MÉTODOS UTILIZADOS

Tendo conhecimento das linguagens de desenvolvimento padrão em mãos, é posto em caráter de decisão, qual delas tende-se a uma melhor relação. De modo que, a escolha da linguagem a qual será implementada o projeto é de importante caráter, já que podemos ter o fator complexidade associado ao futuro da pesquisa. Então, propõe-se exemplos de linguagens e é escolhida a que já se tem um conhecimento prévio. Mediante a escolha da linguagem a qual seria trabalhada, inicia-se o estudo dos métodos de compressão de dados. Primeiro, estuda-se os conceitos básicos. Definição, características e exemplos são conceitos com os quais há um trabalho de compreensão mais aprofundado. A literatura existente sobre esse assunto é vasta, e a princípio foca-se nos artigos científicos relacionados a essa temática. Principalmente aos que apresentam a princípio, uma contextualização e descrição de como todo o processo funciona. De modo com que após visualização de todos os conceitos iniciais, partimos para o estudo das principais técnicas de compressão mais comumente utilizadas.

Com as principais técnicas catalogadas, analisamos o custo operacional de cada uma delas assim como a sua complexidade, e paralelamente, são identificados todos os conceitos por trás de cada método e o seu grau de desenvolvimento, visando com isso, saber se será necessário o estudo de assuntos que ainda não são previamente conhecidos. Além de que, mais fatores são postos em estudo, tais como: capacidade de otimização, custo operacional. Obtém-se então, a finalização de todo o

levantamento bibliográfico do assunto e inicia-se somente o processo de avaliação até que se entre em consenso e consiga a definição de um método a ser utilizado.

Método escolhido. E com a literatura pesquisada em mãos, com intuito de auxílio, inicia-se o processo de desenvolvimento da aplicação com a qual se fará uso do método escolhido. Dividimos o desenvolvimento da aplicação em problemas menores, para que o trabalho fique mais flexível e o tratamento de erros seja o mais eficiente possível, procurando primeiro desenvolver o método original. Com isso, obtemos um ponto de partida para as nossas futuras comparações e/ou análises.

Para futura composição da base para realizar o estudo de caso, obtém uma amostra foi capturada através de um objeto móvel em um cenário livre de tráfego. Nessas amostras, são capturados os campos a compor as futuras mensagens, são estes: latitude, longitude, número de satélites utilizados, velocidade do satélite vertical, velocidade do satélite horizontal e data. Após o término da implementação do método escolhido, a amostra composta dos campos anteriormente citados, são postos sobre compressão. E assim, são traçados diretrizes de comparações, tais como: quantidade de caracteres de entrada/saída, taxas de compressão e quantidade de coordenadas por mensagens. Com isso, um estudo de caso inicial é adotado, como parâmetro comparativo inicial. Catalogando os resultados obtidos após traçadas as diretrizes de situação, de modo que nesse processo, obteve-se uma melhor visualização dos cenários do processo inicial ao pós-otimizado. O estudo de caso final é então posto em prática, e é composto de análise conjunta a gráfico que são gerados a partir dos dados obtidos dos experimentos. Com isso, novos rumos de desenvolvimento são propostos de maneira sempre a buscar a otimização de resultados e melhoria.

4. RESULTADOS E DISCUSSÕES

Inicialmente, realizou-se experimentos com o método de Huffman normal aplicado a base de dados, porém restritos ao tamanho máximo de uma mensagem SMS que é de 160 caracteres. O método utilizado funciona de maneira satisfatória, obtendo-se uma compressão razoável. Porém, faz-se necessário envio da representação junto a árvore que é usada como base, a quantidade de caracteres usada para isso, acaba excedendo de o limite de caracteres. De modo que, foi preciso re-planejar a maneira como o método poderia ser executado. Já que o resultado obtido, era uma cadeia de zeros e um, o dicionário que identificava os significados dos mesmos. Já que a problemática não era o método em si, e sim o resultado de saída que era gerado, a solução seria procurar uma forma de trabalhar com o grande contingente de informação a qual o código estava gerando.

A solução para tal problema foi encontrada nos princípios por trás da funcionalidade da tabela ASC, que codifica caracteres utilizando oito bits. Os códigos representam texto em computadores, de maneira que um número binário constituídos de até oito algarismos possa representar algum texto dentro da tabela. Com isso, foi proposto o agrupamento de oito em oito da codificação final resultante do método e a transformação dessa codificação binária em um resultado decimal, e com isso obtendo-se o texto correspondente a ele na tabela ASC.

Mais testes podem ser realizados e partindo do primeiro cenário, temos a visualização de uma melhora, mas trabalhando com todos os campos, temos uma árvore de representacao de maior tamanho, encurtando assim a carga útil da

mensagem. Gerando para a primeira composição, verifica-se pouco ganho em comprimir a mensagem. Diante disso, reduzimos um campo da amostra, possuindo agora 5 campos. E assim, obtêm-se resultados um pouco mais satisfatórios porém não tão conclusivos. Após a mudança, são mandadas em média de um conjunto de três a quatro coordenadas por mensagem que em relação ao teste anterior têm-se uma melhora mas em termos de taxa de compressão temos um valor não muito relevante.

De novo, é feita a redução de campos de composição da mensagem, fazendo testes para mensagens com quatro e três campos. E para as duas variações, obtêm-se a mesma quantidade de coordenadas enviadas por mensagem, quatro. Figuras 1 e 2 abaixo mostram o valor das taxas aplicados sobre diferentes arranjos.

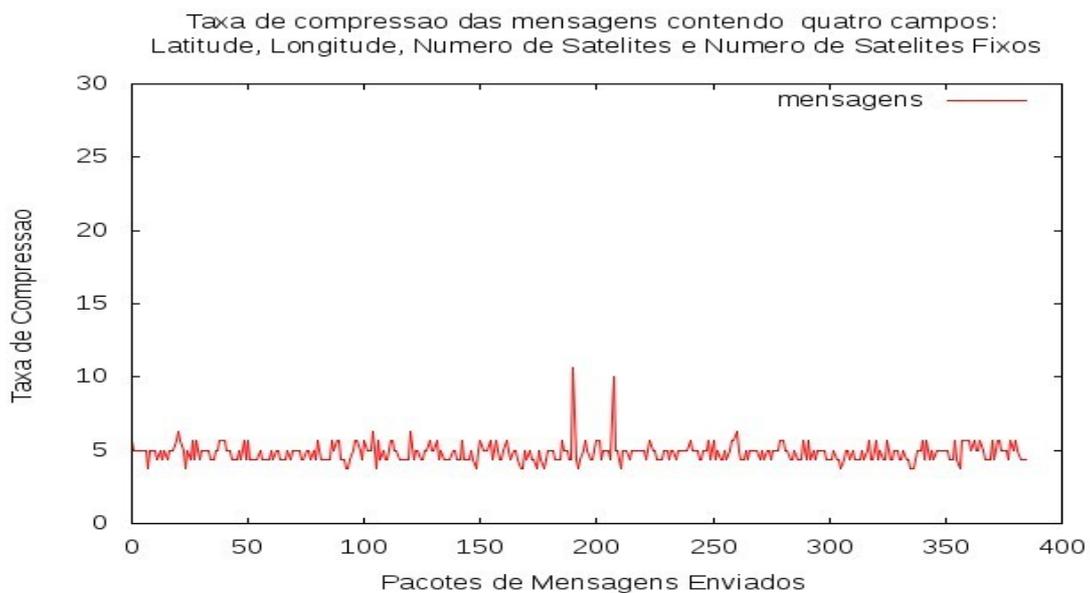


Figura 1 – Taxa de Compressão das mensagens contendo quatro campos: Latitude, Longitude, Número de Satélites e Número de Satélites Fixos

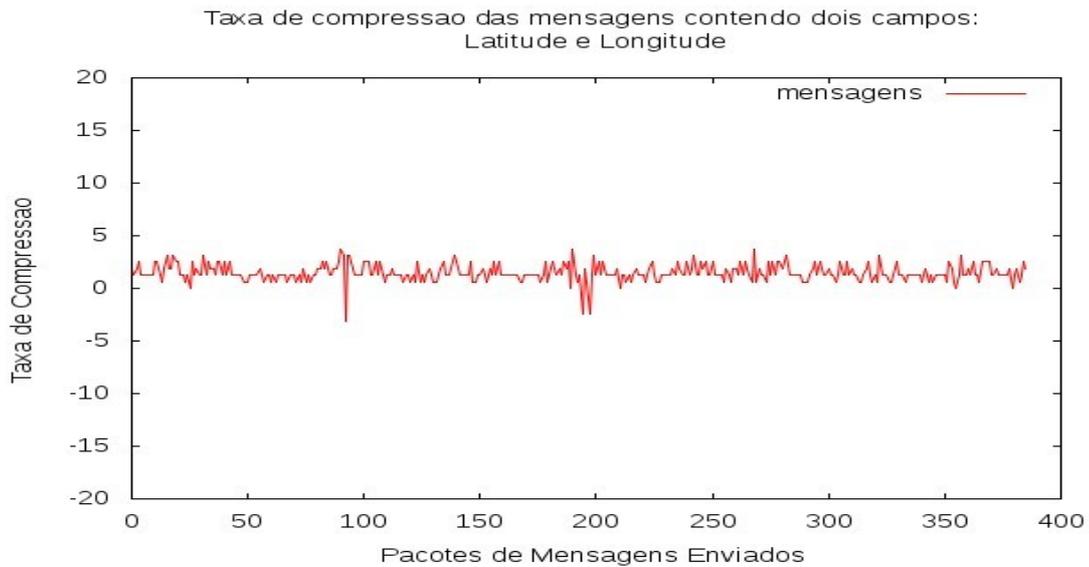


Figura 2 – Taxa de Compressão das mensagens contendo dois campos: Latitude e Longitude
 eficiente enviar as coordenadas sozinhas sem a necessidade de se efetuar a compressão.

Analisando a composição das mensagens comprimidas, temos que o espaço em que a representação da árvore se faz presente, retira grande parte da carga útil da mensagem. Então, como reduzir o espaço da representação da árvore, possibilitando assim o aumento de número de coordenadas nas mensagens e aumentando a carga útil delas?

A solução encontrada, estava no conceito inicial do método. Pois, sabendo que o Huffman trabalha com a probabilidade de incidência de um símbolo no conjunto de informações, e tendo conhecimento de que ao passar do tempo o conjunto delas a ser manipulado não sofre alterações de maneiras drásticas, nós não teríamos mudança nos símbolos. Logo, eles sempre existirão em todos os conjuntos, claro que uns mais que outros. Então, a partir do nosso melhor cenário, mensagens contendo latitude e longitude, trabalharíamos em sua otimização. Com o uso de dois campos, limitamos a abrangência dos símbolos utilizados a números, a vírgula e o hífen. Sabendo que em todo o conjunto de coordenadas, esses valores são sempre fixos podendo ocorrer apenas em lugares diferentes, determina-se que a primeira

mensagem tem que ser capaz de ter todo o universo de símbolos existentes nas informações. Assim, a árvore de representação é gerada apenas uma vez e é capturada pelo receptor, fazendo com que as futuras coordenadas a serem recebidas possam vir comprimidas.

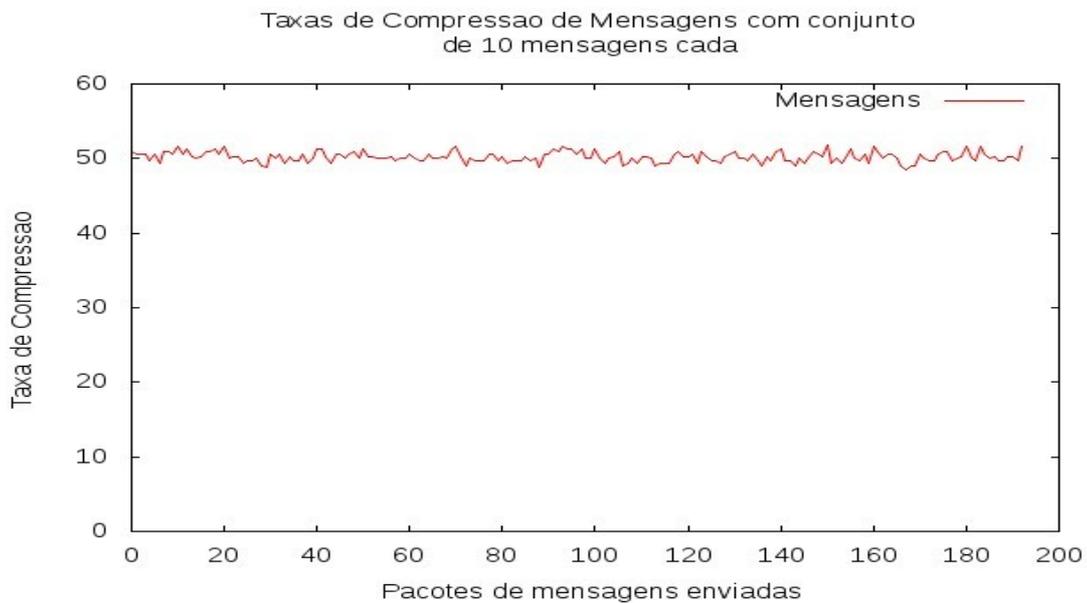


Figura 3 - Taxa de Compressão de mensagens com conjunto de dez coordenadas cada.

Os resultados obtidos foram realmente satisfatórios. Após a mudança, pode-se mandar de nove a dez coordenadas por mensagem. Essas dez mensagens, que estariam representadas por 320 caracteres, com a aplicação do algoritmo são agora representadas por um conjunto de no máximo 160 caracteres, representando uma taxa de compressão de 50% aproximadamente.

Através do gráfico acima, pode-se ter a comparação entre os dois cenários abordado no projeto, a compressão com 5 coordenadas por mensagem e a compressão de 10 coordenadas por mensagem, mandando a árvore de representação na primeira mensagem. O novo método mostrou uma elevada taxa de compressão.

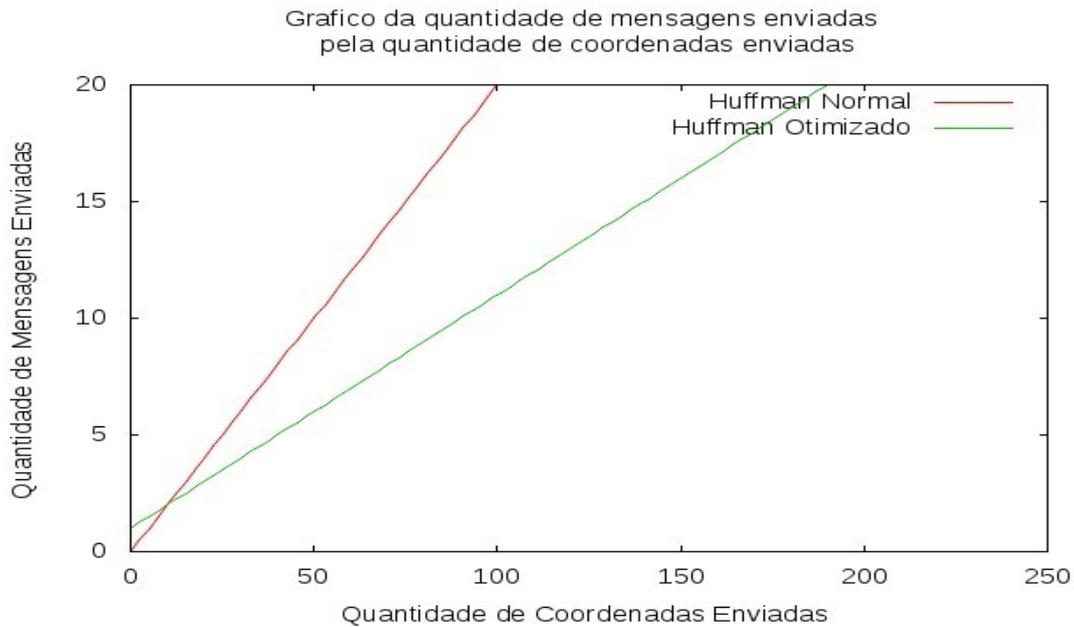


Figura 4 - Gráfico da quantidade de mensagens enviadas pela quantidade de coordenadas enviadas

Em sua primeira mensagem, temos o envio de somente a árvore da representação dos elementos, e a partir dessa árvore, as outras codificações serão trabalhadas. Então, a segunda mensagem enviada contém 10 coordenadas, com a terceira mensagem mais 10 são enviadas. De maneira que ao prazo de envio de 10 mensagens, temos 90 coordenadas enviadas de um lado, enquanto o ultimo metodo testado com apenas 50 coordenadas. A figura 4 mostra o número de mensagens enviadas para transferências do conjunto de coordenadas com as duas abordagens.

5. CONCLUSÕES

Visando atender a crescente utilização da mensagem SMS como ferramenta da comunicação entre dispositivos moveis, a possibilidade de se atender a essa tendência que conjunta a novos ramos de estudo, como a

monitoração geográfica, necessita de técnicas que acompanhem esse desenvolvimento. Já que cada vez mais, temos um número maior de informações circulando por entre esses dispositivos. Como técnica de compressão, o método Huffman se mostrou eficaz o suficiente para ser adotado em sistemas de comunicação assíncrona. Além de se mostrar flexível a otimizações, de maneira que a medida em que ele é construído, ele pode ser usado como base para uma solução adaptável e que funcione diferente. Considerando o cenário da comunicação entre dispositivos móveis, a solução se mostra eficaz.

De modo geral, a efetividade da compressão é garantida mas a compressão com uso da árvore de representação a ser enviada antes mesmo das coordenadas serem iniciadas, apresentaram valores de taxa de compressão mais elevados. Ou seja, a quantidade de mensagens em que as coordenadas são enviadas representam um valor mais expressivo ao que seria adotado a partir do método de compressão normal. Têm-se como recomendações a serem usadas no futuro, o estudo de como o método passa a reagir para cenários aonde a localização passa a ser diferente ou até quantas informações seriam possíveis passar através de mensagens SMS fazendo uso da aplicação que foi gerada a partir do método escolhido.

6. FONTES E REFERÊNCIAS BIBLIOGRÁFICAS

1] Pesquisa Sobre o uso das tecnologias da informação e da Comunicação no Brasil, Comitê gestor da Internet no Brasil, Disponível em www.cgi.br. Acessado em Abril de 2010.

[2] Grandes Desafios da Computação no Brasil, SBC. Disponível em www.sbc.org.br, Acessado em Abril de 2010.

[3] <http://www.w3school.org>, Acessado em Abril de 2010.

[4] Handbook of data compression, Salomon, D., Motta, G., 5th Ed. Springer.

[5] MAVAM, Acision Monitor for VAS, November, 2009. Disponível em www.telecom.com.br. Acessado em Abri de 2010.

7. CRONOGRAMA EXECUTADO

Nº	Descrição	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
----	-----------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

