

Flavia Medeiros dos Anjos

**Reorganização e Compressão
de Dados Sísmicos**

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Programa de Pós-Graduação em
Informática

Rio de Janeiro, agosto de 2007



Flavia Medeiros dos Anjos

Reorganização e Compressão de Dados Sísmicos

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientadores: Eduardo Sany Laber
Pedro Mário Cruz e Silva

Rio de Janeiro
Agosto de 2007



Flavia Medeiros dos Anjos

Reorganização e Compressão de Dados Sísmicos

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Eduardo Sany Laber

Orientador

Departamento de Informática - PUC-Rio

Pedro Mário Cruz e Silva

Co-orientador

TeCGraf - PUC-Rio

Marcelo Gattass

Departamento de Informática - PUC-Rio

Carlos Alves da Cunha Filho

Petrobras

Prof. José Eugenio Leal

Coordenador Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 9 de agosto de 2007

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, da autora e do orientador.

Flavia Medeiros dos Anjos

Graduou-se em Engenharia de Computação pela Pontifícia Universidade Católica do Rio de Janeiro, onde continuou seus estudos no programa de Mestrado em Informática. Sempre com interesse nas áreas de algoritmos e computação gráfica participou de projetos no laboratório LEARN e desde 2003 atua com desenvolvimento e pesquisa no laboratório TeCGraf.

Ficha Catalográfica

Anjos, Flavia Medeiros dos

Reorganização e compressão de dados sísmicos / Flavia Medeiros dos Anjos ; orientadores: Eduardo Sany Laber, Pedro Mário Cruz e Silva. – 2007.

78 f. : il. (col.) ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Inclui bibliografia

1. Informática – Teses. 2. Dados sísmicos 3. Sub-volumes. 4. Transferência disco-memória. 5. Agrupamento. 6. Compressão. 7. Reorganização. 8. K-medianas. I. Laber, Eduardo Sany. II. Silva, Pedro Mário Cruz e III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

às minhas avós Helena e Guilhermina e ao meu avô Hermano, que tenha
encontrado sua paz.

Agradecimentos

À minha família que me apoiou desde o início.

Ao meu orientador Eduardo Laber que acreditou e depositou sua confiança em mim.

Aos colegas e amigos do TeCGraf e da Petrobras, em especial à equipe do v3o2, que me ajudaram e forneceram recursos para o desenvolvimento do trabalho.

À CNPq pelo apoio financeiro.

Aos meus amigos que sempre torceram pelo meu sucesso.

A todos aqueles que me acompanharam ao longo desta trajetória.

Resumo

dos Anjos, Flavia Medeiros; Laber, Eduardo Sany. **Reorganização e Compressão de Dados Sísmicos**. Rio de Janeiro, 2007. 78p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Dados sísmicos, utilizados principalmente na indústria de petróleo, costumam apresentar dimensões de dezenas de gigabytes e em alguns casos, centenas. Este trabalho apresenta propostas de manipulação destes dados que ajudem a contornar problemas enfrentados por aplicativos de processamento e interpretação sísmica ao trabalhar com arquivos deste porte. As propostas se baseiam em reorganização e compressão. O conhecimento do formato de utilização dos dados permite reestruturar seu armazenamento diminuindo o tempo gasto com a transferência entre o disco e a memória em até 90%. A compressão é utilizada para diminuir o espaço necessário para armazenamento. Para dados desta natureza os melhores resultados, em taxa de redução, são das técnicas de compressão com perda, entre elas as compressões por agrupamento. Neste trabalho apresentamos um algoritmo que minimiza o erro médio do agrupamento uma vez que o número de grupos tenha sido determinado. Em qualquer método desta categoria o grau de erro e a taxa de compressão obtidos dependem do número de grupos. Os dados sísmicos possuem uma coerência espacial que pode ser aproveitada para melhorar a compressão dos mesmos. Combinando-se agrupamento e o aproveitamento da coerência espacial conseguimos comprimir os dados com taxas variando de 7% a 25% dependendo do erro associado. Um novo formato é proposto utilizando a reorganização e a compressão em conjunto.

Palavras-chave

dados sísmicos; sub-volumes; transferência disco-memória; agrupamento; compressão; reorganização; k-medianas

Abstract

dos Anjos, Flavia Medeiros; Laber, Eduardo Sany. **Reorganization e Compression of Seismic Data**. Rio de Janeiro, 2007. 78p. MSc Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Seismic data, used mainly in the petroleum industry, commonly present sizes of tens of gigabyte, and, in some cases, hundreds. This work presents propositions for manipulating these data in order to help overcoming the problems that application for seismic processing and interpretation face while dealing with file of such magnitude. The propositions are based on reorganization and compression. The knowledge of the format in which the data will be used allows us to restructure storage reducing disc-memory transference time up to 90%. Compression is used to save storage space. For data of such nature, best results in terms of compression rates come from techniques associated to information loss, being clustering one of them. In this work we present an algorithm for minimizing the cost of clustering a set of data for a pre-determined number of clusters. Seismic data have space coherence that can be used to improve their compression. Combining clustering with the use of space coherence we were able to compress sets of data with rates from 7% to 25% depending on the error associated. A new file format is proposed using reorganization and compression together.

Keywords

seismic data, sub-volume, disc-memory transference, clustering, compression, reorganization, k-medians

Sumário

1 Introdução	14
1.1. Um novo formato para armazenar dados sísmicos	15
1.2. Compressão de dados sísmicos	17
1.3. Organização da dissertação	19
2 Uma proposta para reorganização de dados sísmicos	20
2.1. Seleção e tempo de leitura	21
2.2. Estratégia de multi-fatias	23
2.3. Estratégia de sub-volumes	25
2.3.1. Algoritmo de leitura de uma fatia	27
2.3.2. Experimentos computacionais	28
2.3.3. Otimizações	30
2.3.4. Paralelepípedos	33
2.4. Observações finais	39
3 Uma proposta para compressão de dados sísmicos	41
3.1. Agrupamento uniforme	42
3.2. Agrupamento com minimização do erro médio	43
3.2.1. Algoritmo	44
3.2.2. Implementação não recursiva	45
3.2.3. Implementação recursiva	47
3.2.4. Otimizando o tempo de execução	49
3.3. Amostragem	51
3.4. Comparação do agrupamento uniforme com K-medianas	54
3.5. Transformando o dado através da diferença lateral	57
3.6. Curva de Hilbert	60
3.7. Predição por casamento parcial - PPM	63
3.7.1. Algoritmo	64
3.7.2. Experimentos computacionais	66
3.8. Observações finais	68
4 Conclusão	71

4.1. União entre reorganização e compressão	72
4.2. Trabalhos futuros	75
5 Referências bibliográficas :	77

Lista de figuras

Figura 1: Aplicativo sendo utilizado para ver fatias de um dado.	15
Figura 2: Aplicativo de visualização volumétrica sendo utilizado para ver amostras cujos valores encontram-se entre um intervalo escolhido.	16
Figura 3: Índices de cada amostra.	20
Figura 4: Exemplo das etapas do procedimento de leitura atual.	21
Figura 5: Estratégia de multi-fatias.	24
Figura 6: Gráficos mostrando o tempo total de leitura em função do tamanho do bloco.	24
Figura 7: Gráficos mostrando o tempo de leitura de um bloco e o tempo de seleção das amostras no bloco lido em função do tamanho do bloco.	24
Figura 8: Exemplo da indexação de acordo com a posição.	26
Figura 9: Exemplo de leitura utilizando sub-volumes	27
Figura 10: Gráfico mostrando o tempo médio da leitura de uma fatia em função do fator de subdivisão.	28
Figura 11: Gráfico mostrando o tempo máximo, entre as três direções, para a leitura de uma fatia em função do fator de subdivisão.	29
Figura 12: Gráfico mostrando o tempo máximo, entre as três direções, em função do tamanho da aresta do sub-volume para três volumes diferentes.	30
Figura 13: Um mesmo volume sub-dividido de duas maneiras diferentes.	34
Figura 14: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo z é superior as dimensão dos eixos x e y .	34
Figura 15: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo y é superior as dimensão dos eixos x e z .	35
Figura 16: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo x é superior as dimensão dos eixos y e z .	35
Figura 17: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo z é superior a dimensão do eixo y ; que por sua vez é superior a dimensão do eixo x .	36

- Figura 18: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo z é superior a dimensão do eixo x ; que por sua vez é superior a dimensão do eixo y . 36
- Figura 19: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo y é superior a dimensão do eixo z ; que por sua vez é superior a dimensão do eixo x . 37
- Figura 20: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo y é superior a dimensão do eixo x ; que por sua vez é superior a dimensão do eixo z . 37
- Figura 21: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo x é superior a dimensão do eixo z ; que por sua vez é superior a dimensão do eixo y . 38
- Figura 22: Gráficos mostrando os tempos de leitura de fatias dos tipos $x=cte$, $y=cte$ e $z=cte$ para as duas estratégias de divisão em um volume cuja dimensão do eixo x é superior a dimensão do eixo y ; que por sua vez é superior a dimensão do eixo z . 38
- Figura 23: Exemplo de horizonte sísmico cortado por duas fatias. 39
- Figura 24: Histograma de dois dados sísmicos. 43
- Figura 25: Gráficos comparando os tempos de execução dos algoritmos não recursivo e recursivo em função do tamanho do conjunto de amostras com um número fixo de representantes. 48
- Figura 26: Gráficos comparando os tempos de execução dos algoritmos não recursivo e recursivo em função do número de grupos para um tamanho fixo do conjunto de amostras. 48
- Figura 28: Gráficos comparando os tempos de execução dos algoritmos com e sem o corte no espaço de busca em função do tamanho do conjunto de amostras com um número fixo de representantes. 51
- Figura 29: Gráficos comparando os tempos de execução dos algoritmos com e sem o corte no espaço de busca em função do número de grupos para um tamanho fixo do conjunto de amostras. 51
- Figura 30: Gráficos mostrando, para dois arquivos, o erro do agrupamento do dado todo em função do tamanho da amostragem utilizada para

determinar os representantes.	52
Figura 31: Os gráficos comparam o erro do agrupamento do dado todo com o erro de agrupamento da sub-amostragem utilizando o mesmo grupo de representantes em função do tamanho da amostragem.	53
Figura 32: Coerência horizontal maior que coerência vertical .	58
Figura 33: Exemplo de aplicação da diferença lateral.	58
Figura 34: Histogramas dos resultados do agrupamento para diferentes técnicas e arquivos.	59
Figura 35: Exemplos de curva de Hilbert.	61
Figura 36: Exemplo da utilização de curvas de Hilbert para cobrir áreas de qualquer tamanho.	62

Lista de tabelas

Tabela 1: Exemplo de dimensões e tamanho de arquivos sísmicos.	14
Tabela 2: Ambiente de testes.	21
Tabela 3: Testes de leitura de fatias de dados armazenados na forma tradicional.	22
Tabela 4: Comparação entre os tempos de leitura das estratégias de sub-volumes com e sem as otimizações para $x=cte$ e $y=cte$ num volume de dimensões 600x600x600	31
Tabela 5: Comparação entre os tempos de leitura das estratégias de sub-volumes com e sem as otimizações para $x=cte$ e $y=cte$ num volume de dimensões 800x800x800	31
Tabela 6: Comparação entre os tempos de leitura das estratégias de sub-volumes com e sem as otimizações para $x=cte$ e $y=cte$ num volume de dimensões 1.000x1.000x1.000	32
Tabela 7: Comparação entre os tempos de leitura originais e com a estratégia eleita melhor para o volume e máquina utilizados.	32
Tabela 8: Média dos tempos de leitura de diversas fatias em volumes não cúbicos.	33
Tabela 9: Número de ocorrências dos de cada símbolo do texto do exemplo.	55
Tabela 10: Comparação entre agrupamento por intervalos uniformes e medianas utilizando os critérios de erro médio e entropia.	56
Tabela 11: Comparação entre o número de grupos que cada técnica precisa para atingir níveis semelhantes de erro médio e entropia.	56
Tabela 12: Entropia de 3 dados após a aplicação de diferentes agrupamento e após aplicação da diferença lateral no resultado do agrupamento.	60
Tabela 13: Comparação, para três dados, entre a entropia após calcular a diferença lateral e a entropia após calcular a diferença seguindo a curva de Hilbert.	63
Tabela 14: Entropia do PPM com contexto de até 1 símbolo.	66
Tabela 16: Resultados da compressão com código aritmético em comparação com as estimativas da Entropia no Dado C.	70
Tabela 17: Resultados da compressão com código aritmético em comparação com as estimativas da Entropia no Dado F.	70