Pontifícia Universidade Católica
do Rio de Janeiro

Bruno Tenório Ávila

Compressão de números naturais, seqüência de bits e grafos

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática

Orientador: Prof. Eduardo Sany Laber

Rio de Janeiro
Setembro de 2011

Pontifícia Universidade Católica
do Rio de Janeiro

Bruno Tenório Ávila

Compressão de números naturais, seqüência de bits e grafos

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Eduardo Sany Laber
Orientador
Departamento de Informática — PUC–Rio

Prof. Marco Antônio Casanova
Departamento de Informática — PUC–Rio

Prof. Weiler Alves Finamore
CETUC — PUC–Rio

Prof. Claudson Ferreira Bornstein
Departamento de Ciências da Computação — UFRJ

Prof. Artur Alves Pessoa
Departamento de Engenharia de Produção — UFF

Prof. José Eugênio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 16 de Setembro de 2011

Bruno Tenório Ávila

Bacharelado em Ciências da Computação, UFPE – 2004.
Mestrado em Engenharia Elétrica, UFPE – 2006.

Ficha Catalográfica

Ávila, Bruno Tenório

Compressão de números naturais, seqüência de bits e grafos / Bruno Tenório Ávila; orientador: Eduardo Sany Laber. – 2011.

v., 100 f ; il. ; 30 cm.

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.

Inclui bibliografia.

1. Informática – Teses. 2. Compressão de dados. 3. Seqüência de bits. I. Laber, Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD 004

Agradecimentos

Expresso aqui meus agradecimentos às pessoas que ajudaram no desenvolvimento desta tese. Em especial agradeço:

– à minha família, que sempre me apoiou incondicionalmente em todos os momentos;

– ao CNPq pelo suporte financeiro através de uma bolsa de estudos;

– à banca examinadora composta pelos professores Marco Antônio Casanova (PUC–Rio), Weiler Alves Finamore (PUC–Rio), Claudson Ferreira Bornstein (UFRJ) e Artur Alves Pessoa (UFF);

– aos meus amigos que compartilharam as dificuldades nas disciplinas e no desenvolvimento desta tese. Em especial: Gustavo Moreira e Luiz Albuquerque.

Resumo

Ávila, Bruno Tenório; Laber, Eduardo Sany. **Compressão de números naturais, seqüência de bits e grafos**. Rio de Janeiro, 2011. 100p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Esta tese aborda os problemas de compressão para os seguintes tipos de dados: seqüência de bits e grafos web. Para o problema de compressão de seqüência de bits, demonstramos a relação entre algoritmos de intercalação e codificadores de fonte binária. Em seguida, mostramos que os algoritmos de intercalação binária (Hwang e Lin, 1972), recursivo (Dudzinski, 1981) e probabilístico (Vega, 1993), geram respectivamente os codificadores de entropia baseado em comprimentos de carreiras codificados com o código de Rice, o codificador de intercalação binária (Moffat, 2000) e o codificador de Rice aleatório, na qual é um novo variante do código de Rice. Para o problema de compressão de grafos web, propomos uma nova representação compacta para grafos web, intitulada *árvore-w*, construída especificamente para memória externa (disco), sendo a primeira nesse gênero. Propomos também um novo tipo de layout projetado especificamente para grafos web, intitulado *layout escalado*. Além disso, mostramos como construir um layout *cache-oblivious* para explorar a hierarquia de memórias, sendo a primeira desse tipo. Apresentamos vários tipos de consultas que podem ser executadas e é a primeira representação a suportar execução de consulta de leitura aleatória em lote e a otimização de consultas avançadas, inclusive em memória principal. Por fim, executamos uma série de experimentos que mostra que a árvore-w apresenta taxas de compressão e de tempo de execução competitivas com outras representações compactas em memória principal. Assim, demonstramos empiricamente a viabilidade de uma representação compacta para memória externa na prática, contrariando a afirmação de vários pesquisadores (Suel, 2001) (Buehrer, 2008).

Palavras–chave

Compressão de dados; Seqüência de bits; Grafos web.

Abstract

Ávila, Bruno Tenório; Laber, Eduardo Sany (Advisor). **Compression of natural numbers, sequence of bits and graphs**. Rio de Janeiro, 2011. 100p. DSc Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

This thesis addresses the problems of compression for the following data types: numbers, sequence of bits and webgraphs. For the problem of compression of a sequence of bits, we demonstrate the relationship between merge algorithms and binary source coders. Then, we show that the algorithms binary merge (Hwang and Lin, 1972), recursive merge (Dudzinski, 1981) and probabilistic merge (Vega, 1993), generate respectively an entropy coder based runlengths encoded with the Rice code, the interpolative binary coder (Moffat, 2000) and the random Rice coder, which is a new variant of the Rice code. For the problem of webgraph compression, we propose a new compact representation for webgraphs, entitled *w-tree*, built specifically for external memory (disk), being the first one in this genre. We also propose a new type of layout designed specifically for webgraphs, entitled *scaled layout*. In addition, we show how to build a *cache-oblivious* layout to explore the hierarchy of memories, being the first of its kind. We offer several types of queries that can be performed and it is the first representation to support batched random read query execution and advanced query optimization, including in main memory. Finally, we performed a series of experiments showing that the w-tree provides compression rates and running times competitive with other compact representations for main memory. Therefore, we demonstrate empirically the feasibility of a compact representation for external memory in practice, contrary to the assertion of several researchers (Suel, 2001) (Buehrer, 2008).

Keywords

Data compression; Sequence of bits; Webgraphs.

Sumário

Lista de figuras

Lista de tabelas