ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT & TRUYỀN THÔNG

Nguyễn Thành Trung

NÉN DỮ LIỆU KẾT HỢP VỚI CÁC PHƯƠNG PHÁP BIẾN ĐỔI SƠ BỘ DỮ LIỆU

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số : 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS Bùi Văn Thanh

Thái Nguyên, năm 2013

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn này được thực hiện bằng công sức của mình theo sự hướng dẫn của TS Bùi Văn Thanh, không sao chép từ công trình khác. Mọi thông tin tham khảo đều được trích dẫn đầy đủ. Nếu có gì gian dối tôi xin chịu hoàn toàn trách nhiệm.

Thái Nguyên, tháng 9 năm 2013

Học viên

Nguyễn Thành Trung

LỜI CẢM ƠN

Tôi xin chân thành nói lời cảm ơn Thầy giáo TS Bùi Văn Thanh, người đã tận tình giúp đỡ hướng dẫn tôi trong suốt quá trình thực hiện luận văn cùng với những kinh nghiệm quý báu trong nghiên cứu khoa học cũng như cuộc sống từ Thầy.

Tôi chân thành cảm ơn Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên, Khoa Công nghệ Thông tin đã tạo điều kiện tốt nhất cho tôi được học tập và nghiên cứu. Xin cảm ơn quý Thầy giáo, cô giáo đã tận tình giảng dạy, giúp đỡ và hướng dẫn tôi trong suốt khóa học.

Cảm ơn các bạn đồng nghiệp đã động viên đóng góp ý kiến cho luận văn của tôi.

Mặc dù đã cố gắng hết sức cùng với sự tận tâm của thầy giáo hướng dẫn, song do trình độ còn hạn chế nên Luận văn khó tránh khỏi những thiếu sót. Tôi rất mong nhận được sự thông cảm và góp ý của quý thầy cô và các bạn.

Thái Nguyên, tháng 9 năm 2013

Học viên

Nguyễn Thành Trung

MỤC LỤC

Trang

DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

TT	Từ viết tắt	Viết đầy đủ
1	ARC	ARithmetic Coding
2	De_ARC	Decoder ARithmetic Coding
3	ASCII	American Standard Code for Information Interchange
4	BIT	BInary digiT
5	BPS	Bits Per Second
6	BWT	Burrow-Wheeler Transform
7	InvBWT	Invert Burrow-Wheeler Transform
8	EC	Entropy Coding
9	LZW	Lempel-Ziv-Welch
10	MTF	Move-To-Front
11	InvMTF	Invert Move-To-Front
12	RLE	Run Length Encoding

DANH MỤC CÁC HÌNH VẼ & BẢNG BIỂU

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại CNTT, nhu cầu trao đổi dữ liệu ngày một tăng, dữ liệu cần chia sẻ, trao đổi có dung lượng ngày một lớn hơn, phức tạp hơn và đa dạng hơn. Việc nén dữ liệu nhằm mục đích làm giảm kích thước của dữ liệu gốc giúp cho việc xử lý dữ liệu nhanh hơn (sao chép, di chuyển, tải lên, tải xuống,…).

Việc nén dữ liệu là tất yếu cần thiết do hai lý do chính sau đây. Thứ nhất là do lưu trữ: dữ liệu sau khi nén có dung lượng nhỏ hơn, do vậy cần ít không gian lưu trữ hơn. Thứ hai là giảm thiểu thời gian truyền: dữ liệu sau khi nén có dung lượng nhỏ hơn nên được truyền nhanh hơn.

Do vậy, cần có các thuật toán nén dữ liệu hiệu quả hơn và nhanh hơn vẫn luôn luôn tăng, nhất là với các ứng dụng trực tuyến (hình ảnh, âm thanh). Thực tế là các thuật toán nén dữ liệu đã được liên tục phát triển kể từ khi giới hạn lý thuyết về hiệu suất nén dữ liệu (được định nghĩa như là tỷ lệ các ký tự mã trên một ký tự nguồn) đã được chứng minh một cách chặt chẽ dựa trên lý thuyết thông tin.

Nhiều kỹ thuật nén (không mất thông tin) đã được phát triển như nhóm các phương pháp với tên gọi mã hóa entropy bao gồm mã số học và mã Huffman. Sau đó, hàng loạt các kỹ thuật mới ra đời để cải tiến các kỹ thuật trên như: mã hóa RLE (Run Length Encoding), LZW (Lempel-Ziv-Welch),...

Nhưng nhìn chung không có kỹ thuật nén nào có thể áp dụng một cách hiệu quả với tất cả các loại dữ liệu. Có những thuật toán cho hiệu suất nén cao, nhưng cài đặt phức tạp và mất nhiều thời gian nén cũng như giải nén. Chính vì vậy, để đạt được hiệu suất nén cao với thuật toán tương đối đơn giản và thời gian nén/giải nén chấp nhận được cần phối hợp các thuật toán biến đổi sơ bộ dữ liệu vào để chuyển dữ liệu cần nén sang dạng dữ liệu thích hợp, sau đó áp dụng một thuật toán nén phù hợp để tăng hiệu suất nén.

Với ý tưởng trên, chúng ta có thể sử dụng kỹ thuật biến đổi Burrow-Wheeler (BWT) kết hợp với kỹ thuật Move-To-Front (MTF) để xử lý sơ bộ dữ liệu cần nén với mục đích thay đổi tính chất thống kê của dữ liệu sao cho việc áp dụng các thuật

toán nén trở lên dễ dàng hơn, sau đó áp dụng phương pháp nén số học hoặc nén Huffman để được kết quả mong muốn.

2. Nhiệm vụ nghiên cứu

- Tìm hiểu tổng quan về nén dữ liệu, kỹ thuật biến đổi dữ liệu Burrows-Wheeler, kỹ thuật biến đổi dữ liệu Move-To-Front, thuật toán nén số học và thuật toán nén Huffman.

- Xây dựng ứng dụng thử nghiệm phối hợp kỹ thuật biến đổi Burrows-Wheeler và kỹ thuật biến đổi Move-To-Front để sơ chế dữ liệu trước khi sử dụng phương pháp nén số học hoặc nén Huffman và áp dụng trên tất cả các dữ liệu.

3. Đối tượng và phạm vi nghiên cứu

- Tổng quan về nén dữ liệu.
- Tìm hiểu kỹ thuật biến đổi dữ liệu Burrows-Wheeler.
- Tìm hiểu kỹ thuật biến đổi dữ liệu Move-To-Front.
- Tìm hiểu thuật toán nén số học và nén Huffman.
- Cài đặt thử nghiệm chương trình ứng dụng nén số học hoặc nén Huffman kết hợp với Burrows-Wheeler và Move-To-Front.

4. Phương pháp nghiên cứu

Sử dụng các phương pháp nghiên cứu chính sau:

- Phương pháp nghiên cứu lý thuyết.
- Phương pháp thực nghiệm.
- Phương pháp trao đổi khoa học, lấy ý kiến chuyên gia.

5. Ý nghĩa khoa học của đề tài

- Giúp tìm hiểu, đánh giá khái quát về nén dữ liệu. Nén = mô hình + mã hóa.
- Xây dựng được chương trình nén phục vụ cho công tác lưu trữ tại đơn vị đang công tác.