ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN THỊ PHƯỢNG

KỸ THUẬT NÉN DỮ LIỆU BURROW WHEELER VÀ CÁC CẢI TIẾN

LUẬN VĂN THẠC SỸ: KHOA HỌC MÁY TÍNH

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN THỊ PHƯỢNG

KỸ THUẬT NÉN DỮ LIỆU BURROW WHEELER VÀ CÁC CẢI TIẾN

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ CHUYÊN NGÀNH: 60 80 01

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

HƯỚNG DẪN KHOA HỌC:
PGS TSKH NGUYỄN XUÂN HUY

Thái Nguyên - 2011

LỜI CAM ĐOAN

Tôi xin cam đoan: Luận văn "*Kỹ thuật nén dữ liệu Burrow Wheeler và các cải tiến*" là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nghiên cứu trong luận văn được sử dụng trung thực và có nguồn trích dẫn.

- -

LỜI CẢM ƠN

Em xin gửi lời cảm ơn sâu sắc tới thầy PGS TSKH Nguyễn Xuân Huy - Viện Công nghệ thông tin, người đã gợi mở và định hướng cho em tìm hiểu về lĩnh vực giấu tin trong ảnh. Thầy đã hết lòng giúp đỡ, tạo điều kiện cho em nghiên cứu và hoàn thành luận văn này.

Em xin cảm ơn các thầy cô trong Viện Công nghệ thông tin, các thầy cô giáo khoa Công nghệ thông tin ĐH Thái nguyên, đã giảng dạy và giúp đỡ em trong hai năm học qua.

Cuối cùng tôi xin cảm ơn tới gia đình, các bạn cùng lớp và các bạn đồng nghiệp đã giúp đỡ, động viên, cùng nghiên cứu, đóng góp ý kiến, chia sẻ kinh nghiệmvới tôi trong suốt quá trình học tập và làm luận văn!

Thái Nguyên - 2011

Nguyễn Thị Phượng

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Chữ viết tắt	Diễn giải	Ý nghĩa
AWFC	Advanced Weighted Frenquency Count	Đếm trọng số tần số cao cấp
BWCA	Burrows-Wheeler Compression Algorithm	Thuật toán nén Burrows-Wheeler
BWT	Burrows Wheeler Transform	Chuyển đổi Burrows Wheeler
DC	Distance Coding	Mã hóa khoảng cách
EC	Entropy Coding	Mã hóa Entropy
GST	Global Structure Transformation	Chuyển đổi cấu trúc tổng thể
IF	Inversion Frequencies	Sự đảo ngược tần số
IFC	Inversion Frequencies Count	Đếm gia tăng tần số
LUA	List Update Algorithm	Thuật toán cập nhật danh sách
MTF	Move To Front	Di chuyển lên phía trước
RLE	Run Length Encoding	Mã hóa loạt dài
RLE – BIT	Run Length Encoding - BIT	Thuật toán RLE – BIT
RLE – EXP	Run Length Encoding - EXP	Thuật toán RLE – EXPBIT
RLE0	Run Length Encoding 0	Mã hóa chuyển đổi run 0
RMB	RLE Mantissia Buffer	Luồng dữ liệu riêng biệt
SIF	Sort Inversion Frequencies	Sự đảo ngược tần số có sắp xếp
WFC	Weighted Frequency Count	Đếm trọng số tần số

DANH SÁCH BẢNG BIỂU

DANH SÁCH HÌNH ẢNH

MỞ ĐẦU

1. Lý do chọn đề tài

Một trong các chức năng chính của máy tính là xử lý và lưu trữ dữ liệu. Bên cạnh việc xử lý nhanh người ta còn quan tâm đến việc lưu trữ được nhiều dữ liệu nhưng lại tiết kiệm được vùng nhớ và giảm chi phí lưu trữ. Về mặt lý thuyết thì các thiết bị lưu trữ là không có giới hạn, nhưng ngày nay do nhu cầu xử lý nhiều tập tin, nhiều loại dữ liệu trong cùng một tệp do vậy mà kích thước tập tin trở nên khá lớn.

Trong nhiều năm gần đây, mạng máy tính đã trở nên khá phổ biến trên thế giới. Sự ra đời của mạng đã thực hiện ước mơ chinh phục khoảng cách của con người. Những lợi ích mà mạng cung cấp rất đa dạng và phong phú trên các lĩnh vực khác nhau của toàn xã hội như cung cấp, trao đổi thông tin giữa các máy tính, giữa máy chủ với server hoặc giữa các server với nhau. Điều này dẫn đến phải làm thế nào để giảm thiểu thời gian, chi phí sử dụng để trao đổi dữ liệu trên mạng. Nó đồng nghĩa với việc bên cạnh nâng cao chất lượng của các thiết bị truyền dữ liệu trên mạng thì mặt khác chúng ta phải nghĩ ra một phương pháp để sao cho việc truyền dữ liệu có hiệu quả hơn.

Tất cả những vấn đề trên nảy sinh ra nhu cầu nén dữ liệu với mong muốn thu gọn kích thước các tập tin làm cho thông tin chiếm không gian trên đĩa là ít nhất. Nó là một trong những kỹ thuật quyết định đến cuộc cách mạng đa phương tiện kỹ th uật số đang diễn ra trong nhiều thập kỷ.

Trong một quá trình phát triển lâu dài, nhiều kỹ thuật nén dữ liệu đã ra đời và được chia làm 2 nhóm kỹ thuật chính là nén tổn hao và nén không tổn hao. Nén không tổn hao là kỹ thuật nén mà sau đó ta có thể khôi phục lại chính xác dữ liệu ban đầu. Nén tổn hao là kỹ thuật nén mà sau khi nén ta không thể khôi phục lại chính xác dữ liệu ban đầu của nó. Nén không tổn hao rất được mong đợi vì tỉ lệ nén cũng như tốc độ nén cao của nó. Tuy nhiên, kỹ thuật này chỉ được dùng cho nén âm thanh, hình ảnh bởi cách thức mà các hệ thống thị giác và thính giác của con người làm việc có thể chấp nhận được. Với dữ liệu gốc của nguồn là rất quan trọng mà ta không thể để mất bất kỳ chi tiết nào dụ như hình ảnh y tế , văn bản, các hình ảnh được bảo vệ vì lý do

pháp lý, một số file khả thi của máy tính … thì ta không thể sử dụng kỹ thuật nén tổn hao.

Nhiều kỹ thuật nén không tổn hao đã ra đời như: Phương pháp 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, CD, MTF, LZW... Và gần đây là kỹ thuật nén dữ liệu Burrows Wheeler được công bố bởi Burrows và Wheerler năm 1994. Trong vòng hơn một thập kỷ qua, thuật toán nén Burrows Wheeler [6] đã trở thành một trong những công cụ then chốt trong lĩnh vực nén dữ liệu nói chung. Lý do thành công của nó là tốc độ nén cao kết hợp với tỷ lệ nén tốt. Nhiều cải tiến của kỹ thuật này cũng đã được trình bày. Hiện nay, nén dữ liệu đặc biệt là kỹ thuật nén không tổn hao Burrows Wheeler đang là vấn đề quan tâm rất lớn của các cá nhân, tổ chức, trường học, viện nghiên cứu... trên thế giới.

Chính vì vậy, tôi đã chọn đề tài "***Kỹ thuật nén dữ liệu Burrows Wheeler và các cải tiến***".

Cấu trúc của luận văn được chia làm 3 chương. **Chương 1: Tổng quan về nén dữ liệu.** Trình bày các khái niệm cơ bản như "dữ liệu", "Nén dữ liệu"... Các phương pháp nén dữ liệu...**Chương 2: Kỹ thuật nén dữ liệu Burrows Wheeler và các cải tiến.** Trong chương này trình bày cách làm việc của chuyển đổi Burrows Wheeler, kỹ thuật nén dữ liệu Burrows Wheeler và các cải tiến với kỹ thuật nén này. **Chương 3: Cài đặt thử nghiệm**. Áp dụng chuyển đổi BWT tiến hành một kỹ thuật nén số học kết hợp với BWT và MTF trên đối tượng là tệp văn bản. Xây dựng chương trình thử nghiệm áp dụng thuật toán nén số học kết hợp với BWT và MTF.

2. Mục tiêu của đề tài

Luận văn tập trung tìm hiểu kỹ thuật nén BWT và các cải tiến của kỹ thuật. Cuối cùng là cài đặt thử nghiệm chương trình nén số học kết hợp với BWT và MTF.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng và phạm vi nghiên cứu của luận văn tập trung vào kỹ thuật nén dữ liệu Burrows Wheeler và các cải tiến. Từ đó xây dựng chương trình ứng dụng nén số học kết hợp với Burrows Wheeler và MTF, sử dụng phương pháp Quicksort để sắp xếp và áp dụng trên đối tượng dữ liệu là tệp văn bản.