

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

HOÀNG VĂN SÁNG

TIẾP CẬN MÃ HUFFMAN THEO TÀN SUẤT VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

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

HOÀNG VĂN SÁNG

TIẾP CẬN MÃ HUFFMAN THEO TÀN SUẤT VÀ ỨNG DỤNG

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGSTSKH NGUYỄN XUÂN HUY

Thái Nguyên - 2015

LỜI CAM ĐOAN

Học viên xin cam đoan luận văn là công trình nghiên cứu của riêng cá nhân tôi, kết quả của luận văn hoàn toàn là kết quả của tự bản thân tôi tìm hiểu, nghiên cứu dưới sự hướng dẫn của giáo viên hướng dẫn *PGS.TSKH Nguyễn Xuân Huy*.

Học viên hoàn toàn chịu trách nhiệm về tính pháp lý quá trình nghiên cứu khoa học của luận văn này.

Thái Nguyên, tháng 10 năm 2015

Học viên

Hoàng Văn Sáng

LỜI CẢM ƠN

Số hóa bởi Trung tâm Học liệu - ĐHTN
<http://www.lrc.tnu.edu.vn/>

Lời đầu tiên, học viên xin gửi lời biết ơn sâu sắc đến PGS.TS Nguyễn Xuân Huy người đã tận tình hướng dẫn, chỉ bảo, giúp đỡ học viên trong suốt quá trình làm luận văn.

Học viên cũng xin gửi lời cảm ơn đến các thầy cô giáo trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên, các thầy cô Viện Công nghệ thông tin đã truyền đạt những kiến thức và giúp đỡ học viên trong suốt quá trình học tập.

Học viên cũng xin gửi lời cảm ơn tới Ban giám đốc Trung tâm Tin học - nơi học viên công tác - đã tạo điều kiện thuận lợi cho học viên tham gia khóa học và trong quá trình hoàn thành luận văn.

Và học viên cũng xin gửi lời cảm ơn tới các đồng nghiệp, gia đình và bạn bè, những người đã ủng hộ, động viên tạo mọi điều kiện giúp đỡ để học viên có được kết quả như ngày hôm nay.

Thái Nguyên, tháng 10 năm 2015

Học viên

Hoàng Văn Sáng

MỤC LỤC

Trang

Số hóa bởi Trung tâm Học liệu - ĐHTN
<http://www.lrc.tnu.edu.vn/>

LỜI CAM ĐOAN	i
MỤC LỤC	ii
DANH MỤC CÁC HÌNH	v
MỞ ĐẦU	1
1. Đặt vấn đề	1
2. Đối tượng và phạm vi nghiên cứu	1
3. Hướng nghiên cứu của đề tài	2
4. Phương pháp nghiên cứu	2
5. Ý nghĩa khoa học của đề tài.....	2
6. Bố cục của luận văn.....	3
CHƯƠNG 1: TỔNG QUAN VỀ NÉN DỮ LIỆU	4
1.1. Vấn đề về nén dữ liệu	4
1.2. Bài toán nén dữ liệu	6
1.3. Phân loại chương trình nén	7
1.4. Chất lượng của thuật toán nén dữ liệu.....	8
1.5. Vấn đề giải nén	9
1.6. Nén không tổn hao và nén tổn hao	11
1.7. Nén không tổn hao.....	11
1.8. Nén tổn hao.....	12
1.9. Đơn vị đo đặc tính nén.....	13
CHƯƠNG 2: TỔNG QUAN VỀ MÃ NÉN HUFFMAN	16
2.1. Mã tiền tố.....	16
2.2. Biểu diễn mã tiền tố trên cây nhị phân	17
2.3. Quy trình nén dữ liệu theo mã Huffman.....	19

2.3.1 Giới thiệu về mã Huffman.....	19
2.3.2. Phương pháp mã hóa Mã hóa Huffman.....	20
2.3.3 Tính chất cây Huffman.....	21
2.3.4. Thuật toán tạo mã Huffman.....	21
2.3.5. Giải mã thuật toán Huffman.....	31
2.4. Xây dựng và cải tiến thuật toán.....	32
CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH NÉN SỬ DỤNG.....	37
PHƯƠNG PHÁP MÃ HÓA HUFFMAN.....	37
3.1. Cấu trúc chương trình.....	37
3.2. Các thuật toán nhóm một.....	37
3.2.1. Thuật toán A1: Nén Huffman.....	37
3.2.2. Thuật toán A2: Dựng cây Huffman.....	38
3.2.3. Thuật toán A3: Huffman code.....	39
3.2.4. Thuật toán A4: Giải mã Huffman.....	41
3.3. Giới thiệu chương trình.....	43
3.4. Kết quả kiểm thử chương trình.....	46
KẾT LUẬN.....	48
TÀI LIỆU THAM KHẢO.....	50
Chương trình chi tiết thực hiện giải thuật bằng ngôn ngữ Dev-C++.....	51

DANH MỤC CÁC HÌNH

	Trang
<i>Hình 1. 1: Quy trình nén dữ liệu.....</i>	4
<i>Hình 1. 2: Bộ nén và bộ giải nén.</i>	10
<i>Hình 1. 3: Bộ mã hóa và bộ giải mã.....</i>	10
<i>Hình 1. 4: Những thuật toán nén không hao tổn.....</i>	11
<i>Hình 1. 5: Các thuật toán nén tổn hao</i>	12
<i>Hình 2. 1: Sắp xếp danh sách các kí tự (ví dụ 1).....</i>	22
<i>Hình 2. 2: Xây dựng cây Huffman (ví dụ 1)</i>	25
<i>Hình 2. 3: Cây Huffman điền đầy đủ thành phần (ví dụ 1)</i>	25
<i>Hình 2. 4: Sắp xếp danh sách các kí tự (ví dụ 2).....</i>	26
<i>Hình 2. 5: Xây dựng Cây Huffman (ví dụ 2)</i>	29
<i>Hình 2. 6: Cây Huffman điền đầy đủ thành phần (ví dụ 2)</i>	30

MỞ ĐẦU

1. Đặt vấn đề

Một trong những chức năng chính của máy tính là xử lý dữ liệu và lưu trữ. 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 tệp trở nên khá lớn. Những vấn đề trên nảy sinh ra khái niệm nén dữ liệu. Nén dữ liệu là quá trình làm giảm lượng thông tin “dư thừa” trong dữ liệu gốc, do vậy lượng thông tin thu được sau nén thường nhỏ hơn dữ liệu gốc rất nhiều. Nén dữ liệu là giải pháp hợp lý nhất nhằm mục đích giảm chi phí cho người sử dụng. Một trong những kỹ thuật thường dùng trong nén dữ liệu là xác định tần suất xuất hiện của các đối tượng trong file nguồn. Đối tượng nào xuất hiện càng nhiều thì mã càng ngắn. Heuristics này được sử dụng trong các thuật toán nén Huffman. Với mỗi văn bản cho trước, thuật toán đếm số lần xuất hiện của từng kí tự trong văn bản sau đó xây dựng cây Huffman cuối cùng từ cây Huffman sinh ra mã nén. Nếu ta có thể xác định tần suất xuất hiện của từng chữ cái trong một ngôn ngữ cho trước thì chương trình không phải thực hiện thủ tục đếm.

Vì vậy, “Tiếp cận mã Huffman theo tần suất và ứng dụng” được học viên chọn làm luận văn tốt nghiệp của mình.

2. Đối tượng và phạm vi nghiên cứu

- **Đối tượng:**

- Các phần mềm nén dữ liệu.

- Các thuật toán nén dữ liệu.
- Các phương pháp mã hóa Huffman.
- Hệ thống phần mềm nén dữ liệu.

- **Phạm vi:**

- Các khái niệm của ký tự mã hóa, các thuật toán của nén văn bản. Kiến trúc, chức năng và các thành phần của nén dữ liệu cụ thể cho bài toán nén sử dụng phương pháp mã hóa Huffman.
- Các chức năng chính và quy trình thực thi của bài toán nén dữ liệu.
- Hệ thống chương trình cho bài toán nén dữ liệu.

3. Hướng nghiên cứu của đề tài

- Tìm hiểu tổng quan về nén dữ liệu và nghiên cứu một thuật toán nén cụ thể.
- Tìm hiểu bài toán nén dữ liệu, tiến hành phân tích.
- Thu thập các số liệu có liên quan.
- Phân tích, đánh giá thông qua các số liệu thu thập được.
- Cài đặt thực nghiệm.

4. Phương pháp nghiên cứu

- Phương pháp nghiên cứu lý thuyết: Tổng hợp tài liệu, hệ thống lại các kiến thức, tìm hiểu các khái niệm, thuật toán sử dụng trong đề tài.
- Phương pháp quy nạp toán học.
- Phương pháp đối sánh.
- Phương pháp trao đổi khoa học, lấy ý kiến chuyên gia.

5. Ý nghĩa khoa học của đề tài

- Sau khi thực hiện đề tài học viên có thêm hiểu biết về công trình khoa học và nghiên cứu.

- Đề tài có thể đóng góp các phương pháp và kỹ thuật tối ưu hóa trong lĩnh vực mã hóa.

6. Bộ cục của luận văn

Bộ cục của luận văn bao gồm 3 chương:

Chương 1: Tổng quan về nén dữ liệu

Chương này học viên sẽ trình bày tổng quan về đề tài, khái niệm nén dữ liệu, các phương pháp nén dữ liệu, phân loại và đánh giá chất lượng của chương trình nén.

Chương 2: Tổng quan về mã nén Huffman

Chương này học viên sẽ trình bày mã tiền tố, cách biểu diễn mã tiền tố trên cây nhị phân, quy trình nén dữ liệu theo mã Huffman, thuật toán tạo mã Huffman và giải mã Huffman.

Chương 3: Xây dựng chương trình nén sử dụng phương pháp mã hóa Huffman

Dựa vào cơ sở lý thuyết của thuật toán được trình bày ở chương 2, trong chương này, học viên trình bày ứng dụng trong thực tế và cài đặt chương trình thực nghiệm cụ thể.