ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT & TRUYỀN THÔNG

PHẠM THU HƯỜNG

NÉN VĂN BẢN TIẾNG VIỆT THEO HUFFMAN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2013

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT & TRUYỀN THÔNG

PHẠM THU HƯỜNG

NÉN VĂN BẢN TIẾNG VIỆT THEO HUFFMAN

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS Nguyễn Hữu Điển

Thái Nguyên - 2013

LỜI CẢM ƠN

Để đạt được kết quả ngày hôm nay l à một sự cố gắng, nỗ lực rất lớn của bản thân, cũng như sự giúp đỡ nhiệt tình của quý thầy cô, bạn bè để tôi hoàn thành luận văn này.

Tôi xin trân trọng cảm ơn:

- PGS.TS Nguyễn Hữu Điển – Giám đốc Trung tâm tính toán hiệu năng cao Trường Đại học khoa học tự nhiên Hà Nội.
- Các thầy cô trong hội đồng phản biện.

Cuối cùng tôi xin chân thành cảm ơn các thầy, cô, các bạn đã động viên và giúp đỡ tôi trong thời gian làm luận văn.

Xin trân trọng cám ơn quý thầy cô, các bạn!.

DANH MỤC CÁC HÌNH

Hình 1. Quy trình nén dữ liệu 3
Hình 2. Xây dựng cây nhị phân từ bảng mã không ở dạng tiền tố 8
Hình 3. Sắp xếp danh sách các ký tự 20
Hình 4. Xây dựng cây Huffman 22
Hình 5. Cây Huffman điền đầy đủ thành phần 22
Hình 6. Một trường hợp xây dựng khác 23
Hình 7. Lưu đồ giải mã 24
Hình 8. Ý tưởng xây dựng cây theo phương pháp Shannon – Fano 26
Hình 9. Xây dựng cây theo phương pháo Shannon-Fano 27
Hinh 10. Mã hóa bằng phương pháp Huffman động 31
Hình 11. Giải mã bằng phương pháp Huffman động 33
Hình 12. Quá trình thực hiện nén bằng LZ 43
Hình 13. Sơ đồ nén LZ78 47
Hình 14. Sơ đồ giải nén LZ78 48
Hình 15. Sơ đồ nén LZW 51
Hình 16. Sơ đồ giải nén LZW 54
Hình 17. Phương pháp MTF (tốt) 57
Hình 18. Phương pháp MTF (xấu) 57
Hình 19. Phương pháp BW tìm chuỗi sau mã hóa 59
Hình 20. Hai cách tìm chuỗi gốc 60
Hình 21. Giao diện chương trình 62

MỤC LỤC

LỜI CẢM ƠN ..... **iii**
DANH MỤC CÁC HÌNH ..... **iv**
MỞ ĐẦU **1**
1. Đặt vấn đề ***1***
2. Đối tượng và phạm vi nghiên cứu ***1***
2.1. Đối tượng ***1***
2.2. Phạm vi ***2***
3. Hướng nghiên cứu của đề tài ***2***
4. Phương pháp nghiên cứu ***2***
5. Ý nghĩa khoa học của luận văn ***2***
CHƯƠNG 1: TỔNG QUAN VỀ CÔNG NGHỆ NÉN DỮ LIỆU **3**
1.1. Sơ lược về nén dữ liệu ***3***
1.1.1. Khái niệm về nén dữ liệu ***3***
1.1.2. Những vấn đề phải giải quyết trong nén dữ liệu ***4***
1.1.3. Phân loại chương trình nén ***5***
1.1.4. Đánh giá chất lượng của chương trình nén ***6***
1.2. Mã nén dữ liệu ***7***
1.2.1. Định nghĩa mã hoá ***7***
1.2.2. Các khái niệm về ký tự mã hóa ***8***
1.2.3. Mã tổng và mã phân tách ***13***
1.2.4. Định lý mã nén ***18***
CHƯƠNG 2. MỘT SỐ MÃ NÉN CƠ BẢN **21**
2.1. Mã hóa Huffman (Huffman coding) ***21***
2.1.1. Phương pháp mã hóa ***21***
2.1.2. Thuật toán tạo mã Huffman ***21***
2.1.3. Giải mã thuật toán Huffman : ***25***
2.2. Mã hóa Huffman động (Adaptive Huffman coding) ***31***
2.2.1. Phương pháp mã hóa: ***31***
2.2.2. Thuật toán nén ***31***
2.2.3. Thuật toán giải nén ***33***
2.3. Thuật toán xử lý sự lặp lại của xâu (RLE) ***36***

2.3.1. Phương pháp: *36*
2.3.2. Thuật toán tạo mã *36*
2.3.3. Quá trình giải mã *36*
2.4. Mã hóa kiểu từ điển (Dictionary -based compression)..... 39
2.4.1. Nguyên lý LZ *39*
2.4.2.Từ điển *40*
2.4.3. Quá trình thực hiện khi nén bằng mã LZ..... *41*
2.4.4. Các thuật toán nén LZ..... *42*
2.5. Một số phương pháp biến đổi (transform) 54
2.5.1. Phương pháp đẩy về phía trước (Move to front): *54*
2.5.2. Phương pháp Burrows – Wheeler (BW): *56*
CHƯƠNG 3. XÂY DỰNG CHƯƠNG TRÌNH NÉN TIẾNG VIỆT SỬ DỤNG PHƯƠNG PHÁP MÃ HÓA HUFFMAN..... 59
3.1. Bộ gõ Tiếng việt..... 59
3.2. Quy ước biểu diễn ký tự tiếng Việt. 59
3.3. Chuẩn dấu Tiếng việt..... 60
3.3.1. Unicode..... *60*
3.3.2. TCVN3 ..... *60*
3.3.3. VNI *60*
3.4. Phương pháp mã hóa Huffman 60
3.5. Giới thiệu chương trình 61
3.5.1. Hướng dẫn sử dụng..... *62*
3.5.2. Kết quả kiểm thử chương trình *64*
KẾT LUẬN 65
TÀI LIỆU THAM KHẢO 66
PHỤ LỤC 67

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.

Như chúng ta cũng đã biết tiếng Việt là một ngôn thuộc hệ thống chữ cái Latin sử dụng nhiều dấu đi kèm với nguyên âm. Với bảng mã ASCII 8 bit sử dụng phổ biến trên máy tính, chúng ta có thể mã hóa 256 ký tự. Để đưa tiếng Việt vào máy tính, các phần mềm tiếng Việt hiện nay sử dụng một trong hai phương pháp mã hóa : mã dựng sẵn hoặc mã tổ hợp để xây dựng trang mã ký tự tiếng Việt. Bảng mã phổ biến nhất chúng ta thường sử dụng là bảng mã Unicode để thể hiện tiếng Việt. Nhưng bảng mã Unicode yêu cầu 16 bit để thể hiện một ký tự điều này sẽ dẫn đến sự lãng phí và dư thừa dữ liệu.

Vì vậy, " Nén văn bản tiếng Việt theo Huffman " được em chọn làm luận văn tốt nghiệp của mình.

2. Đối tượng và phạm vi nghiên cứu

2.1. Đối tượng

- Các phần mềm nén dữ liệu;
- Các thuật toán nén dữ liệu;
- Các phương pháp mã hóa tiếng Việt;
- Hệ thống phần mềm nén dữ liệu từ đó ứng dụng vào để nén dữ liệu cho tiếng Việt.

2.2. 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 văn bản Tiếng việt 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;

Vì thời gian có hạn, trong khuôn khổ một luận văn tốt nghiệp cao học, việc giải quyết bài toán nén dữ liệu chỉ giới hạn ở một vài thuật toán nén cổ điển.

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

- Nghiên cứu các tài liệu và viết tổng quan;
- Phương pháp khảo sát thực tế;
- Phương pháp phân tích và đánh giá các thuật toán;
- Nghiên cứu triển khai thuật toán và thử nghiệm hệ thống.

5. Ý nghĩa khoa học của luận văn

- Bản thân hiểu sâu hơn và áp dụng được các thuật toán của nén dữ liệu vào thực tế;
- Triển khai một số thuật toán nén dữ liệu qua đó ứng dụng chính là phương pháp mã hóa Huffman vào tiếng Việt;
- Xây dựng được chương trình nén dữ liệu dành cho tiếng Việt trên máy tính.

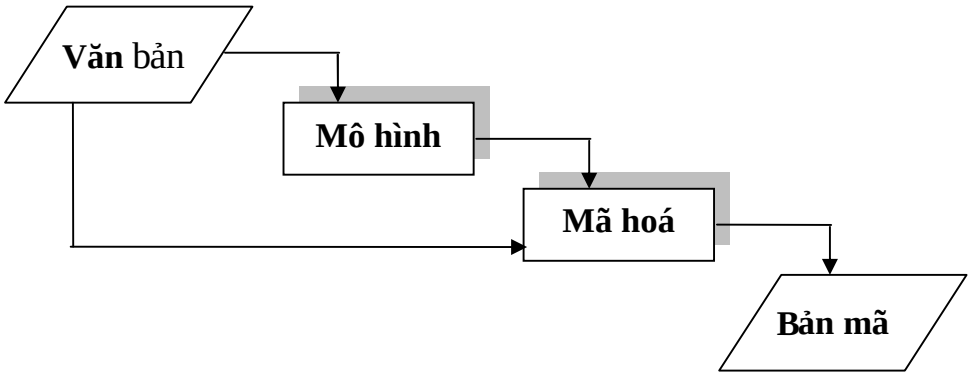
CHƯƠNG 1: TỔNG QUAN VỀ CÔNG NGHỆ NÉN DỮ LIỆU

1.1. Sơ lược về nén dữ liệu

1.1.1. Khái niệm về nén dữ liệu

Nén là một quá trình giảm lượng không gian cần thiết để biểu diễn cùng một lượng thông tin cho trước. Người ta còn gọi nén là biến đổi một luồng ký hiệu thành một luồng các từ mã.

Quá trình nén như sau:



Hình 1. Quy trình nén dữ liệu

Trong đó:

- Văn bản là văn bản ban đầu cần nén.
- Mô hình là tập hợp các chữ cái cùng quy tắc được sử dụng để xử lý các chữ cái vào và đưa ra các từ mã. Một mô hình sẽ xác định chính xác xác suất xuất hiện của từng chữ cái và một bộ mã sẽ tạo ra các từ mã dựa trên xác suất đó.
- Mã hoá là chỉ quá trình thay thế các chữ cái trong văn bản ban đầu bằng các từ mã tương ứng để đưa ra bản mã chính xác.

Như vậy, quá trình nén diễn ra như sau: quá trình mô hình căn cứ vào văn bản cần nén sẽ tạo ra các từ mã. Sau đó, từ bộ từ mã vừa tạo được và văn bản ban đầu quá trình nén sẽ đưa ra bản mã.

Mã hoá và mô hình là hai giai đoạn hoàn toàn khác nhau vì trong giai đoạn mô hình có rất nhiều cách để xử lý các chữ cái của văn bản mà cùng sử dụng một phương pháp xây dựng mã để cho ra các từ mã.

Nếu bản mã có kích thước nhỏ hơn văn bản thì phương pháp nén đó có hiệu quả.

Ví dụ:

Chúng ta sử dụng cùng phương pháp mã Huffman cho hai mô hình khác nhau:

- Mô hình 1: dựa trên xác suất độc lập của từng chữ cái xuất hiện bất kì trong văn bản.
- Mô hình 2: cần tính được xác suất phụ thuộc dựa trên những chữ cái nhận được lúc đó trong văn bản.

Do mô hình khác nhau nên cùng sử dụng mã Huffman để đưa ra từ mã nhưng hiệu quả nén của chúng rất khác nhau.

Tuy nhiên, chúng ta vẫn quen dùng từ mã hoá để chỉ cho cả quá trình nén văn bản mặc dù đó chỉ là một giai đoạn của một quá trình nén.

Người ta thường mã hoá thông qua các từ mã của một bảng chữ cái nào đó.

Có thể có nhiều thuật toán nén dữ liệu khác nhau. Mỗi thuật toán có một kiểu dữ liệu nhất định và cùng một số modem có đặc điểm nén thích ứng có nghĩa là chúng có khả năng chọn thuật toán nén thích hợp phụ thuộc vào kiểu dữ liệu cần nén. Trong số các cách mã thì cách nào mã ngắn hơn chúng ta nói là nó nén tốt hơn (so với cách mã khác).

1.1.2. Những vấn đề phải giải quyết trong nén dữ liệu

Mục tiêu của nén dữ liệu là đưa ra thuật toán để giảm thiểu sự lãng phí từ những phần giống nhau khi biểu diễn dữ liệu. Thông thường một chương trình nén cần quan tâm đến khả năng nó có thể cắt triết được nhiều hay ít dung lượng của dữ liệu sau khi nén, điều này còn phụ thuộc vào thuật toán, và điều kiện đầu vào của dữ liệu. Như vậy những dữ liệu đầu vào khác nhau có thể đòi hỏi những thuật toán khác nhau để nhận diện nhậy nhất những chỗ tiêu sài lãng phí bộ nhớ. Cần nhận ra