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

**LÊ VĂN NINH**

# NÉN DỮ LIỆU THEO KỸ THUẬT MOVE – TO - FRONT

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

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

**Thái Nguyên - 2011**

## *Lời cảm ơn!*

Em xin chân thành cảm ơn các thầy giáo trong Viện công nghệ thông tin, Viện khoa học và công nghệ Việt Nam; các thầy giáo, cô giáo trường Đại học CNTT & TT - Đại học Thái Nguyên đã tạo điều kiện, giúp đỡ em hoàn thành luận văn này.

Đặc biệt, em xin chân thành cảm ơn PGS.TSKH Nguyễn Xuân Huy, thầy giáo đã giảng dạy và trực tiếp hướng dẫn trong suốt quá trình nghiên cứu và hoàn thành luận văn.

Dù đã có nhiều cố gắng, nhưng chắc chắn luận văn sẽ không tránh khỏi những thiếu sót và hạn chế. Vì vậy, rất mong được sự góp ý, chỉ dẫn của các thầy, cô giáo, bạn bè và đồng nghiệp.

Trân trọng cảm ơn!

*Thái Nguyên, tháng 11 năm 2011*

**Lê Văn Ninh**

| DANH MỤC TỪ VIẾT TẮT | |
|---|---|
| Move – To – Front (Di chuyển lên phía trước) | MTF |
| Borrows – Wheeler Tranform (Chuyển đổi BW) | BWT |
| Run length (Mã hóa Run Length) | RUL |
| Inversion Frequencies (Sự đảo ngược tần số) | *IF* |
| Distance Coding (Mã hóa khoảng cách) | *DC* |
| Weighted Frequency Count (Đếm trọng số tần số) | *WFC* |
| Increased Frequency Count (Đếm gia tăng tần số) | IFC |
| Local to Global Transform (Chuyển đổi từ cục bộ đến tổng thể ) | LGT |

DANH MỤC HÌNH VẼ *PAGE*

**MỤC LỤC** *PAGE*

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Trong các lĩnh vực của công nghệ thông tin và viễn thông hiện nay, việc truyền tải tin tức là công việc xảy ra thường xuyên. Tuy nhiên, thông tin được truyền tải đi thường rất lớn, điều này gây khó khăn cho công việc truyền tải như: gây tốn kém tài nguyên mạng, tiêu phí khả năng của hệ thống,.. Để giải quyết vấn đề đó, các thuật toán nén dữ liệu đã được ra đời.

Các kỹ thuật nén được nhúng ngày càng nhiều trong phần mềm và đã trở thành yêu cầu chung cho hầu hết phần mềm ứng dụng như một lĩnh vực nghiên cứu quan trọng và tích cực trong khoa học máy tính.

Trong kỹ thuật truyền tin nối tiếp, do các bit dữ liệu được truyền đi nối tiếp, lại bị giới hạn về dải thông của kênh truyền và giới hạn về các chuẩn ghép nối... nên tốc độ truyền tin tương đối chậm. Nén dữ liệu trước khi truyền đi là một trong các phương pháp nhằm tăng tốc độ truyền dữ liệu. Nguyên tắc của nén dữ liệu là quá trình mã hóa thông tin dùng ít bit hơn so với thông tin chưa được mã hóa bằng cách dùng một hoặc kết hợp các phương pháp nào đó. Dựa theo nguyên tắc này giúp tránh các hiện tượng kênh truyền bị quá tải và việc truyền tin trở nên kinh tế hơn.

Mặc dù các chương trình nén dữ liệu thường sử dụng kết hợp nhiều thuật toán có độ phức tạp khác nhau nhằm đạt được hiệu quả cao nhất cho dữ liệu được nén để đáp ứng yêu cầu đặt ra. Nhưng nhìn chung không thể có phương pháp nén tổng quát nào cho kết quả tốt đối với tất cả các loại tập tin. Kỹ thuật nén tập tin thường được áp dụng cho các tập tin văn bản (Trong đó có một số kí tự nào đó có xác suất xuất hiện nhiều hơn các kí tự khác), các tập tin ảnh bitmap (Mà có thể có những mảng đồng nhất), các tập tin dùng để biểu diễn âm thanh dưới dạng số hoá và các tín hiệu tương tự  khác (các tín hiệu này có thể có các mẫu được lặp lại nhiều lần). Đối với các tập tin nhị phân như tập tin chương trình thì sau khi nén cũng không tiết kiệm được

nhiều. Ngoài ra, trong một số trường hợp để nâng cao hệ số nén người ta có thể bỏ bớt một số thông tin của tập tin (Ví dụ như kỹ thật nén ảnh JPEG).

Nén dữ liệu theo kỹ thuật Move-To-Front (MTF) là một trong những kỹ thuật nén dữ liệu được thiết kế để cải tiến hiệu quả của kỹ thuật nén mã hóa *entropy*. Nó được sử dụng sau kỹ thuật chuyển đổi Burrows-Wheeler để xếp hạng các biểu tượng theo tần số tương quan của chúng. Mục đích là để đạt được một hiệu suất nén tốt hơn cho mã hóa entropy.

Xuất phát từ ý tưởng đó, tôi đã lựa chọn đề tài "Nén dữ liệu theo kỹ thuật Move – To – Front".

**2. Đối tượng nghiên cứu:**

- Kỹ thuật nén dữ liệu Move-To-Front

**3. Phạm vi nghiên cứu:**

- Tìm hiểu tổng quan về nén dữ liệu
- Nén dữ liệu Burrows-Wheeler
- Nén dữ liệu theo kỹ thuật Move – to – front

**4. Mục tiêu nghiên cứu**

Luận văn tập trung nghiên cứu, đánh giá về nén dữ liệu theo kỹ thuật Move-To-Front. Vận dụng nén dữ liệu trong một số lĩnh vực đặc thù.

**5. Ý nghĩa khoa học của đề tài**

- Giúp tìm hiểu, đánh giá khái quát về nén dữ liệu theo kỹ thuật MTF.
- Vận dụng được phương pháp nén dữ liệu theo kỹ thuật MTF trong một số lĩnh vực đặc thù.

**6. 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 thống kê
- Phương pháp trao đổi khoa học, lấy ý kiến chuyên gia.

# **Chương I.** TỔNG QUAN VỀ NÉN DỮ LIỆU

## 1.1. Giới tiệu

*Nén dữ liệu* trong ngữ cảnh khoa học máy tính là khoa học để biểu diễn thông tin dưới dạng thu gọn. Nói cách khác nén dữ liệu là việc thực hiện thu gọn kích thước các tập tin hoặc làm cho thông tin lưu trữ chiếm không gian lưu trữ ít nhất. Có nhiều cách để thực hiện điều này tùy vào từng đối tượng cụ thể.

Nén dữ liệu đã trở thành yêu cầu chung cho hầu hết phần mềm ứng dụng như một lĩnh vực nghiên cứu quan trọng và tích cực trong khoa học máy tính. Nếu không có các kỹ thuật nén, Internet sẽ không bao giờ phát triển, TV kỹ thuật số , các kỹ thuật tru yền thông di động hoặc truyền thông video đã được phát triển trên thực tế.

Các lĩnh vực ứng dụng có liên quan và đươc thúc đẩy bởi nén dữ liệu gồm có:

- Các hệ thống truyền thông cá nhân: Fax, thư thoại (voice mail) và điện thoại.
- Các hệ thống máy tính: Cấu trúc bộ nhớ, đĩa và băng.
- Tính toán di động.
- Các hệ thống máy tính phân tán.
- Mạng máy tính, đặc biệt là Internet.
- Sự phát triển đa phương tiện, hình ảnh, xử lý tín hiệu.
- Lưu trữ hình ảnh và hội nghị truyền hình.
- Ti vi kỹ thuật số và truyền hình vệ tinh.
- ...

Nhiều vấn đề trên thực tế đã thúc đẩy nhiều nghiên cứu khác nhau về nén dữ liệu. Tương tự, nghiên cứu về nén dữ liệu cũng đã dựa trên hay được

kích thích bởi các lĩnh vực mới khác. Một phần do phạm vi ứng dụng rộng rãi của nó, nén dữ liệu bao trùm nhiều ngành khoa học và có thể được tìm thấy trong nhiều lĩnh vực khác nhau như: Lý thuyết thông tin; Lý thuyết mã hóa; Mạng máy tính và viễn thông; Xử lý tín hiệu kỹ thuật số; Xử lý ảnh; Đa phương tiện; Bảo mật máy tính...

Trong nén dữ liệu, từ *dữ liệu* có nghĩa là thông tin ở dạng kỹ thuật số mà những chương trình máy tính hoạt động và *nén,* có nghĩa là quá trình loại bỏ dư thừa trong dữ liệu. Cụm từ "*nén dữ liệu*", có nghĩa là đưa ra các kỹ thuật hay cụ thể hơn là thiết kế những thuật toán hiệu quả nhằm để:

- Biểu diễn dữ liệu theo dạng mà chứa ít dư thừa.
- Loại bỏ dư thừa trong dữ liệu.
- Cài đặt thuật toán nén và giải nén.

### 1.1.1. Một số vấn đề về Nén dữ liệu

Một vấn đề nén liên quan đến việc tìm một thuật toán hiệu quả để loại bỏ dư thừa khác nhau từ một kiểu dữ liệu nhất định. Ví dụ cho một xâu *s*, câu hỏi là *dãy các biểu tượng có thể thay thế mà chiếm ít không gian lưu trữ là dãy nào*? *Giải pháp* cho vấn đề nén là thuật toán nén nhằm đưa ra dãy các biểu tượng chứa ít số lượng bit hơn , cộng với các thuật toán giải nén để phục hồi xâu gốc.

Vậy số lượng bit ít hơn là bao nhiêu ? Điều đó phụ thuộc vào những thuật toán nhưng nó cũng phụ thuộc vào sự dư thừa có thể chiết ra từ dữ liệu gốc là bao nhiêu. Dữ liệu khác nhau có thể yêu cầu những kỹ thuật khác nhau để xác định dư thừa và loại bỏ dư thừa trong dữ liệu.

Không có giải pháp nào phù hợp cho tất cả vấn đề nén dữ liệu . Theo các nghiên cứu về nén dữ liệu , ta chủ yếu phải phân tích những đặc tính của dữ liệu đã được nén và hy vọng đưa ra một số mô hình để đạt được sự biểu