ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỖ THỊ HẠNH

TÌM KIẾM MỜ VÀ ỨNG DỤNG TÌM KIẾM THÔNG TIN TRONG CÁC VĂN BẢN NÉN

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 35 01

LUẬN VĂN THẠC SĨ

Người hướng dẫn: **PGS.TS. ĐOÀN VĂN BAN**

Thái Nguyên - 2009

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỖ THỊ HẠNH

TÌM KIẾM MỜ VÀ ỨNG DỤNG TÌM KIẾM THÔNG TIN TRONG CÁC VĂN BẢN NÉN

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 35 01

LUẬN VĂN THẠC SĨ

Người hướng dẫn: **PGS.TS. ĐOÀN VĂN BAN**

Thái Nguyên - 2009

LỜI CẢM ƠN

Em xin chân thành cảm ơn các thầy, cô khoa Công nghệ thông tin trường Đại học Thái Nguyên đã tạo điều kiện giúp đỡ và truyền đạt cho em những kiến thức về chuyên ngành và những kiến thức xã hội.

Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến PGS.TS. Đoàn Văn Ban - Viện Khoa học Công nghệ Việt Nam. Thầy đã trực tiếp hướng dẫn và giúp đỡ em hoàn thành luận văn. Mặc dù, trong quá trình làm luận văn em đã gặp nhiều khó khăn nhưng thầy luôn động viên, chia sẻ, đó là nguồn động lực lớn giúp em vượt qua. Thầy chính là tấm gương cho em trong công tác giảng dạy, nghiên cứu khoa học, cũng như trong cuộc sống. Em xin cảm ơn thầy.

Em không quên sự động viên, khích lệ của gia đình, bạn bè và những người thân đã giúp đỡ em vượt qua mọi khó khăn để em hoàn thành khoá học.

Em xin chân thành cảm ơn!

Thái Nguyên, tháng 11 năm 2009

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Các ký hiệu

ε	Xâu rỗng
w_i	Ký tự thứ i của xâu w
w(f, d)	Xâu con (hay khúc con) độ dài f của xâu w, kết thúc ở vị trí d trên w
$w1 \leq_s w2$	Xâu w1 là khúc đuôi của w2
$w1 \leq_{ls} w2$	Xâu w1 là khúc đuôi dài nhất của w2
w(t) hoặc $pref_t(w)$	Khúc đầu độ dài t của xâu w
$suf_t(w)$	Khúc cuối độ dài t của xâu w
|A|	Lực lượng của tập A

Các chữ viết tắt

NFA	Otomat đa định hữu hạn

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Bộ não của con người có thể xử lý thông tin ở hai mức:

- Mức định lượng (chính xác)
- Mức định tính (không chính xác, bất định, mơ hồ, không chắc chắn, nhập nhằng, không rõ ràng, mờ)

Tính thông minh trong quá trình xử lý thông tin thể hiện ở khả năng xử lý thông tin định tính. Đây là điều mà thế hệ máy tính hiện nay đang hướng tới.

Máy tính ngày nay đã được sử dụng trong hầu hết các lĩnh vực và đã góp phần quan trọng vào việc thúc đẩy sự phát triển kinh tế, xã hội, khoa học kỹ thuật, … Máy tính ra đời nhằm phục vụ cho những mục đích nhất định của con người. Với tất cả sự xử lý của máy tính để lấy thông tin hữu ích và trong quá trình xử lí đó một vấn đề đặc biệt quan trọng là tìm kiếm thông tin với khối lượng lớn, độ chính xác cao, thời gian nhanh nhất.

Tìm kiếm thông tin thì bài toán đóng vai trò quan trọng là bài toán so mẫu, với mẫu có thể ở bất kỳ kiểu dữ liệu nào, từ văn bản đến các loại dữ liệu đa phương tiện khác (ảnh, video, âm thanh, …). Trên thực tế có rất nhiều ứng dụng tìm kiếm thông tin như: công cụ tìm kiếm của các hệ điều hành, khai phá web trên Internet, ...

Để tìm kiếm thông tin thì cần phải xem thông tin đó lưu trữ dưới dạng dữ liệu nào? Dữ liệu được lưu trữ dưới nhiều dạng, song phổ biến nhất vẫn là dạng text nên chúng tôi chọn đề tài này cụ thể là tìm kiếm văn bản text. Tìm kiếm văn bản text nếu như những văn bản có khối lượng lớn thì có thể mất nhiều thời gian với những thuật toán kinh điển. Vậy đặt ra vấn đề tìm kiếm văn bản nhưng ở dạng nén sẽ nhanh hơn.