

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

MỞ ĐẦU	1
Chương 1. TÌM KIẾM MẪU TRONG VĂN BẢN THEO CÁCH TIẾP CẬN OTOMAT MỜ	5
1.1. Tổng quan về tìm kiếm mẫu trên văn bản	5
1.1.1 Giới thiệu chung về vấn đề tìm kiếm văn bản	5
1.1.2. Các dạng tìm kiếm và các kết quả nghiên cứu	7
1.1.2.1. Tìm đơn mẫu	7
1.1.2.2. Tìm đa mẫu.....	8
1.1.2.3. Tìm mẫu mở rộng	9
1.1.2.4. Tìm kiếm xấp xỉ.....	10
1.1.2.4.1. Phát biểu bài toán	10
1.1.2.4.2. Các tiếp cận tìm kiếm xấp xỉ	11
1.1.2.4.3. Độ tương tự giữa hai xâu	12
1.1.3. Tìm kiếm trong văn bản nén và mã hoá	14
1.2. Hệ mờ.....	15
1.3. Ý tưởng chung của tiếp cận otomat mờ.....	15
1.4. Khái niệm otomat mờ	17
1.5. Một số thuật toán so mẫu.....	18
1.5.1. Thuật toán KMP (Knuth- Morris- Pratt)	18
1.5.2. Thuật toán BM (Boyer- Moor).....	22
1.6. Kết luận chương 1	26
Chương 2. BÀI TOÁN SO MẪU THEO CÁCH TIẾP CẬN OTOMAT MỜ	27
2.1. Bài toán so mẫu chính xác	27
2.1.1. Phát biểu bài toán	27
2.1.2. Độ mờ của mô hình	27

2.1.3. Thuật toán KMP mờ	28
2.1.3.1. Otomat so mẫu.....	28
2.1.3.2. Tính đúng đắn của thuật toán	29
2.1.3.3. Thuật toán.....	29
2.1.3.4. So sánh KMP và thuật toán KMP mờ	32
2.1.4. Thuật toán KMP - BM mờ.....	33
2.1.4.1. Ý tưởng của thuật toán.....	33
2.1.4.2. Otomat mờ so mẫu.....	35
2.1.4.3. Thuật toán 2.4.....	37
2.2. Bài toán so mẫu xấp xỉ.....	38
2.2.1. Đặt vấn đề.....	38
2.2.2. Bài toán	39
2.2.3. Độ tương tự dựa trên độ dài khúc con chung của hai xâu.....	40
2.2.3.1. Phát biểu bài toán.....	40
2.2.3.2. Otomat so mẫu.....	42
2.2.4. Độ gần tựa ngữ nghĩa.....	43
2.2.4.1. Ý tưởng về độ gần	43
2.2.4.2. Thuật toán sơ bộ tính độ gần.....	44
2.2.4.2.1. Ý tưởng	44
2.2.4.2.2. Thuật toán chi tiết	44
2.2.4.3. Giải thích độ mờ của mô hình	45
2.3. Kết luận chương 2	46
Chương 3. TÌM KIẾM MẪU TRONG VĂN BẢN NÉN VÀ MÃ	
HOÁ	47
3.1. Tiếp cận tìm kiếm tổng quát trên văn bản nén và mã hoá.....	47
3.2. Tìm kiếm trên văn bản nén	50
3.2.1. Các mô hình nén văn bản.....	50

3.2.2. Thuật toán tìm kiếm trên dữ liệu nén dạng text	50
3.3. Tìm kiếm trên văn bản mã hóa.....	55
3.3.1. Tìm kiếm trên văn bản mã hóa dạng khối kí tự	55
3.3.2. Mã đàn hồi.....	55
3.3.3. Tìm kiếm trên văn bản mã hóa bởi mã đàn hồi	58
3.3.3.1. Ý tưởng chung	58
3.3.3.2. Phương pháp đánh giá độ mờ xuất hiện mẫu trên văn bản mã hóa.....	59
3.3.3.2.1. Bài toán	59
3.3.3.2.2. Mô tả phương pháp.....	59
3.3.3.2.3. Chi tiết hóa các otomat trong thuật toán	60
3.3.3.2.4. Thuật toán tìm kiếm mẫu dựa trên otomat	61
3.3.4. Tìm kiếm trên văn bản mã hóa hai tầng	63
3.4. Kết luận chương 3	64
KẾT LUẬN.....	65
TÀI LIỆU THAM KHẢO.....	67

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Các ký hiệu

ε	Xâu rỗng
w_i	Ký tự thứ i của chuỗi w
$w(f, d)$	Xâu con (hay khúc con) độ dài f của chuỗi 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 $\text{pref}_t(w)$	Khúc đầu độ dài t của chuỗi w
$\text{suf}_t(w)$	Khúc cuối độ dài t của chuỗi w
$ A $	Lực lượng của tập A

Các chữ viết tắt

NFA	Otomat đa định hữu hạn
-----	------------------------

DANH MỤC CÁC HÌNH VẼ

Hình 1.1. Ý nghĩa của mảng next	19
Hình 1.2. Ý nghĩa của mảng next tại vị trí $m + 1$	19
Hình 2.1. Dịch chuyển con trỏ trên mẫu	32
Hình 2.2. Ý tưởng chung của thuật toán KMP-BM mờ	35
Hình 2.3. Một ví dụ với các khối độ dài $t = 3$	44
Hình 2.4. Tập mờ mô tả độ gần tựa ngữ nghĩa của mẫu P so với xâu đích S.....	45
Hình 3.1. Phương pháp so mẫu trên miền nén có sử dụng otomat mờ ..	48
Hình 3.2. Phương pháp so mẫu không giải mã	49
Hình 3.3. Queue trước (a) và sau (b) khi thực hiện thủ tục Decompress	52
Hình 3.4. Queue trước (a) và sau (b) bước nhảy $n2'$	53
Hình 3.5. Đồ thị xây dựng khái niệm tích đàn hồi	56
Hình 3.6. Đồ thị xác định mã đàn hồi	58
Hình 2.7. Quá trình mã hóa hai tầng	64
Hình 2.8. Quá trình giải mã hai tầng.....	64
Hình 2.9. Quá trình tìm kiếm mẫu trên văn bản mã hóa hai tầng	64

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 lý 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.