

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

ĐỖ QUỲNH ANH

**NGHIÊN CỨU THUẬT TOÁN KNUTH-MORRIS-PRATT
VÀ ỨNG DỤNG**

**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: ĐỖ TRUNG TUẤN

Thái Nguyên – 2014

MỤC LỤC

MỤC LỤC	1
DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT	5
DANH MỤC CÁC HÌNH VẼ VÀ CÁC BẢNG	6
MỞ ĐẦU	7
CHƯƠNG 1. SO KHỚP CHUỖI	10
1.1. Khái niệm so khớp chuỗi	10
1.2. Lịch sử phát triển.....	11
1.3. Các cách tiếp cận.....	12
1.4. Ứng dụng của so khớp chuỗi.....	12
1.5. Các dạng so khớp chuỗi	13
1.5.1. So khớp đơn mẫu	13
1.5.2. So khớp đa mẫu.....	14
1.5.3. So mẫu mở rộng.....	15
1.5.4. So khớp chính xác.....	16
1.5.5. So khớp xấp xỉ	17
1.5.5.1. Phát biểu bài toán	17
1.5.5.2. Các tiếp cận so khớp xấp xỉ.....	18
1.5.5.3. Độ tương tự giữa hai xâu.....	19
1.5. Một số thuật toán so mẫu	20
1.5.1. Thuật toán Brute Force	20
1.5.2. Thuật toán Karp-Rabin	21
1.5.3. Thuật toán BM (Boyer- Moor)	24
1.5.4. Các thuật toán khác.....	27
1.6. Khớp chuỗi với otomat hữu hạn.....	28
1.6.1. Otomat hữu hạn.....	28
1.6.1.1. Ôtômát hữu hạn đơn định DFA.....	29
1.6.1.2. Ôtômát hữu hạn không đơn định NFA.....	33
1.6.2. Otomat khớp chuỗi.....	36
1.6.2.1. Giới thiệu	36
1.6.2.2. Thuật toán xây dựng Otomat so khớp chuỗi	38
1.7. Kết luận chương	40
CHƯƠNG 2. THUẬT TOÁN SO KHỚP CHUỖI KNUTH-MORRIS-PRATT.....	41
2.1. Thuật toán KMP	41

2.1.1. Giới thiệu thuật toán	41
2.1.2. Bảng so sánh một phần	45
2.1.3. Độ phức tạp của thuật toán KMP	47
2.2. Thuật toán KMP mờ	48
2.2.1. Otomat so mẫu	48
2.2.2. Thuật toán	49
2.2.2.1 Thuật toán tạo lập TFuzz	49
2.2.2.2. Thuật toán tìm kiếm mẫu dựa vào bảng TFuzz	51
2.2.3. So sánh KMP và thuật toán KMP mờ	52
2.3. Thuật toán KMP - BM mờ	53
2.3.1. Ý tưởng của thuật toán	53
2.4.2. Otomat mờ so mẫu	55
2.3.2.1. Giới thiệu	55
2.3.2.2. Hoạt động của otomat mờ so mẫu	55
2.3.3. Thuật toán tìm kiếm	56
2.4. Kết luận chương	57
CHƯƠNG 3. ỨNG DỤNG THUẬT TOÁN KMP TRONG TÌM KIẾM THÔNG TIN TRÊN VĂN BẢN	58
3.1. Bài toán tìm kiếm mẫu trên văn bản	58
3.1.1. Tìm kiếm mẫu	58
3.1.2. Tìm kiếm thông tin	59
3.1.2.1 Giới thiệu	59
3.1.2.2 Các mô hình tìm kiếm thông tin thường sử dụng	61
3.2. Mã nguồn mở Lucene	64
3.2.1. Giới thiệu	64
3.2.2. Các bước sử dụng Lucene	66
3.3. Ứng dụng tìm kiếm thông tin trên văn bản	67
3.4. Cài đặt chương trình thử nghiệm	68
3.4.1. Giải pháp, công nghệ sử dụng	68
3.4.2. Nội dung chương trình	68
3.4.3. Kết quả thực nghiệm	71
3.4.3.1. Giao diện chính của chương trình	72
3.4.3.2. Kết quả thử nghiệm của chương trình khi tìm kiếm với từ khóa “Văn bản”	72
3.5. Kết luận chương 3	73

KẾT LUẬN	74
TÀI LIỆU THAM KHẢO	76

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

BM	Thuật toán Boyer - Moore
DFA	Deterministic Finite Automata - Ôtômát hữu hạn đơn định
DOC	Document
FA	Finite Automata - Ôtômát hữu hạn
HTML	HyperText Markup Language
IDF	Inverse document frequency - Tần suất tài liệu ngược
KMP	KNUTH-MORRIS-PRATT
LAN	Local area network
NFA	Nondeterministic Finite Automata - Ôtômát hữu hạn không đơn định
TF	Term frequency - Tần suất từ

DANH MỤC CÁC HÌNH VẼ VÀ CÁC BẢNG

Hình 1.1. Sơ đồ chuyển của một DFA	30
Hình 1.2. Mô tả một DFA	31
Bảng 1.1. Ví dụ hàm chuyển δ của DFA	32
Hình 1.3. Sơ đồ của một NFA.....	34
Hình 1.4. Di chuyển chuỗi	35
Bảng 1.2. Ví dụ hàm chuyển trạng thái δ của NFA	35
Hình 1.5. Ví dụ so khớp chuỗi	37
Hình 1.6. Ví dụ otomat so khớp chuỗi	38
Bảng 2.1. Bảng so sánh một phần.....	46
Bảng 2.2. Thí dụ khác	46
Bảng 2.3. Trường hợp mẫu xấu nhất với thuật toán KMP.....	47
Bảng 2.4. Bảng next	51
Bảng 2.5. Bảng TFuzz.....	51
Bảng 2.6. Minh họa thí dụ.....	52
Hình 2.1. Dịch chuyển con trỏ trên mẫu	52
Bảng 2.7. Kết quả tìm sự xuất hiện mẫu P trong tập S theo KMP và tiếp cận mờ	53
Hình 2.2. Ý tưởng chung của thuật toán KMP-BM mờ.....	55
Hình 3.1. Mô hình biểu diễn và so sánh thông tin	60
Hình 3.2. Mô hình không gian vec tơ	62
Bảng 3.1. Tính điểm số	64
Hình 3.3. Mô hình đánh chỉ mục của Lucene	65
Hình 3.4. Mô hình ứng dụng tìm kiếm thông tin văn bản	68
Hình 3.5. Giao diện chính của chương trình.....	72
Hình 3.6. Kết quả tìm kiếm của chương trình.....	73

MỞ ĐẦU

1. Lý do chọn đề tài

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.

Cùng với sự phổ biến của công nghệ thông tin, số lượng các tài liệu điện tử cũng gia tăng từng ngày. Đến nay, số lượng các tài liệu được lưu trữ lên đến hàng tỷ trang. Trong khi đó, nhu cầu khai thác trong kho tài liệu khổng lồ này để tìm kiếm những thông tin cần thiết đang là nhu cầu thường ngày và thiết thực của người sử dụng. Tuy nhiên, một trong những khó khăn con người gặp phải trong việc khai thác thông tin là khả năng tìm chính xác thông tin họ cần trong kho tài liệu. Để trợ giúp công việc này, các hệ thống tìm kiếm đã lần lượt được phát triển nhằm phục vụ cho nhu cầu tìm kiếm của người sử dụng.

Những hệ thống tìm kiếm bắt đầu phát triển và đưa vào ứng dụng, phổ biến là các hệ thống tìm kiếm theo từ khóa. Nhiều hệ thống hoạt động hiệu quả trên Internet như Google, Bing, Yahoo!... Tuy nhiên, phần lớn các công cụ tìm kiếm này là những sản phẩm thương mại và mã nguồn được giữ bí mật. Hoặc các hệ thống tìm kiếm trên máy cá nhân như Windows Search, Google Desktop... đã đáp ứng phần nào nhu cầu của người sử dụng, miễn phí cho cá nhân, tuy nhiên cũng chỉ đáp ứng được trên phạm vi nhỏ và mới chỉ dừng lại ở mức độ tìm kiếm từ khóa theo tiêu đề và phần tóm tắt.

Có một cách tiếp cận hiệu quả để giải quyết vấn đề này là thực hiện việc

so khớp và tìm kiếm toàn văn. Một trong những thuật toán so khớp chuỗi kinh điển là thuật toán KMP. Có thể nói, KMP là một thuật toán mới mẻ ít được sử dụng tại Việt Nam trong việc quản lý, lưu trữ và xử lý lượng dữ liệu lớn nhưng rất hiệu quả và chính xác. Dựa trên hướng tiếp cận đó và sự hướng dẫn của giáo viên, tôi mạnh dạn nhận đề tài “So khớp chuỗi và thuật toán Knuth-Morris-Pratt”.

2. Đối tượng và phạm vi nghiên cứu

- Các khái niệm so khớp chuỗi.
- Các khái niệm thuật toán so khớp chuỗi KMP.
- Một số ứng dụng trong thuật toán KMP.

3. Hướng nghiên cứu của đề tài

- Nghiên cứu tìm kiếm Knuth–Morris–Pratt và ứng dụng trong việc tìm kiếm thông tin trên văn bản.
- Nghiên cứu giải pháp công nghệ cài đặt chương trình thử nghiệm.

4. Những nội dung chính

Luận văn được trình bày trong 3 chương, có phần mở đầu, phần kết luận, phần mục lục, phần tài liệu tham khảo. Luận văn được chia làm ba chương với nội dung cơ bản như sau:

- Chương 1: Trình bày khái niệm về so khớp chuỗi, các hướng tiếp cận, các dạng so khớp và một số thuật toán so mẫu.
- Chương 2: Trình bày về thuật toán KMP, thuật toán KMP mờ và thuật toán KMP-BM mờ.
- Chương 3: Trình bày về bài toán tìm kiếm thông tin trên văn bản và tiến hành cài đặt thử nghiệm chương trình.

5. Phương pháp nghiên cứu

Tổng hợp các tài liệu đã được công bố về thuật toán tìm kiếm thông tin,

khai phá dữ liệu, đặc biệt các kết quả nghiên cứu liên quan đến thuật toán tìm kiếm thông tin.

Thực nghiệm thuật toán tìm kiếm KMP với dữ liệu mẫu. Nhận xét, đánh giá kết quả thử nghiệm.

6. Ý nghĩa khoa học của đề tài

Luận văn nghiên cứu kỹ thuật, thuật toán tìm kiếm thông tin là cơ sở hỗ trợ cho công tác dự báo, lập kế hoạch, quy hoạch, phân tích dữ liệu quản lý, chuyên môn, nghiệp vụ.

CHƯƠNG 1. SO KHỚP CHUỖI

1.1. Khái niệm so khớp chuỗi

So khớp chuỗi là một kỹ thuật đóng vai trò nền tảng trong lĩnh vực xử lý văn bản. Hầu như tất cả các trình soạn thảo và xử lý văn bản đều cần phải có một cơ chế để so khớp các chuỗi trong tài liệu hiện tại. Việc tích hợp các thuật toán so khớp chuỗi là một trong những khâu cơ bản được sử dụng trong việc triển khai phần mềm và được thực hiện trên hầu hết các hệ điều hành.

Mặc dù hiện nay dữ liệu được lưu trữ dưới nhiều hình thức khác nhau, nhưng văn bản vẫn là hình thức chủ yếu để lưu trữ và trao đổi thông tin. Trong nhiều lĩnh vực như so khớp, trích chọn thông tin, tin sinh học..., một lượng lớn dữ liệu thường được lưu trữ trong các tập tin tuyến tính. Hơn nữa khối lượng dữ liệu thu thập được tăng lên rất nhanh nên đòi hỏi phải có các thuật toán xử lý và so khớp dữ liệu văn bản hiệu quả

So khớp chuỗi là việc so sánh một hoặc nhiều chuỗi (thường được gọi là mẫu hoặc Pattern) với văn bản để tìm vị trí và số lần xuất hiện của chuỗi đó trong văn bản.

Ta hình thức hoá bài toán so khớp chuỗi như sau: coi văn bản là một mảng $T[1..n]$ có chiều dài n và khuôn mẫu là một mảng $P[1..m]$ có chiều dài m ; các thành phần của T và P là các ký tự được rút từ một bảng chữ cái hữu hạn Σ . Ví dụ, ta có thể có $\Sigma = \{0,1\}$ hoặc $\Sigma = \{a,b,\dots,z\}$. Các mảng ký tự P và T thường được gọi là các chuỗi ký tự. Ta nói rằng một chuỗi w là tiền tố (hậu tố) của một chuỗi x , ký hiệu là $w \subset x$ ($w \supset x$), nếu $x = wy$ ($x = yw$), với y là một chuỗi nào đó. Để ngắn gọn, ta kí hiệu P_k để thể hiện tiền tố k - ký tự $P[1..k]$ của khuôn mẫu $P[1..m]$. Ta nói rằng khuôn mẫu P xảy ra với khoá chuyển s trong văn bản T (hoặc, theo tương đương, nói rằng khuôn mẫu P xảy ra bắt đầu tại vị trí $s + 1$ trong văn bản T) nếu $0 \leq s \leq n-m$ và $T[s + 1..s + m] = P[1..m]$ (nghĩa là, nếu $T[s+j] = P[j]$, với $1 \leq j \leq m$). Bài toán so khớp chuỗi là bài toán tìm tất cả các