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

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

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 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.

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, KPM 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 thoải 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 $\sum$. Ví dụ, ta có thể có $\sum = \{0,1\}$ hoặc $\sum = \{a,b,....,z\}$. 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 + i trong văn bản T) nếu $0 \le s \le n-m$ và T[s + 1..s + m] = P[1..m] (nghĩa là, nếu T[s+j] = P[j], với $1 \le j \le m$). Bài toán so khớp chuỗi là bài toán tìm tất cả các