ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN&TRUYỀN THÔNG

Nguyễn Thị Phương Thảo

MỘT HỌ THUẬT TOÁN ĐỐI SÁNH MẪU CHÍNH XÁC NHANH SSABS - TVSBS - FQS VÀ THỰC NGHIỆM

Chuyên ngành: **Khoa học máy tính**

Mã số: **60. 48. 01. 01**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS Hà Quang Thụy

Thái Nguyên - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan:

Những nội dung trong luận văn này là do tôi thực hiện dưới sự hướng dẫn trực tiếp của thầy giáo hướng dẫn PGS TS. Hà Quang Thụy.

Mọi tham khảo trong luận văn đều được trích dẫn rõ ràng tác giả, tên công trình, thời gian, địa điểm công bố.

Tôi xin cam đoan luận văn không phải là sản phẩm sao chép của bất kỳ tài liệu khoa học nào.

Học viên

Nguyễn Thị Phương Thảo

LỜI CẢM ƠN

Đầu tiên tôi xin gửi lời cảm ơn sâu sắc nhất tới PGS. TS Hà Quang Thụy người hướng dẫn khoa học, đã tận tình chỉ bảo, giúp đỡ tôi thực hiện luận văn. Tôi cũng xin lời lời cám ơn trân thành tới PGS. TS. Nguyễn Trí Thành và các anh chị em Phòng Thí nghiệm Khoa học dữ liệu và Công nghệ Tri thức, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã giúp đỡ và tạo điều kiện hỗ trợ tôi.

Tôi xin cảm ơn các thầy cô trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã giảng dạy và truyền đạt kiến thức cho tôi.

Tôi xin trân thành cảm ơn Ban giám hiệu trường Cao đẳng nghề Phú Thọ và các đồng nghiệp trong khoa Công nghệ thông tin đã tạo mọi điều kiện giúp đỡ tôi hoàn thành nhiệm vụ học tập.

Cuối cùng, tôi xin cảm ơn những người thân và các bạn bè chia sẻ, giúp đỡ tôi hoàn thành luận văn này.

Mặc dù đã hết sức cố gắng hoàn thành luận văn với tất cả sự nỗ lực của bản thân, nhưng luận văn vẫn còn những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý Thầy, Cô và bạn bè đồng nghiệp.

Tôi xin chân thành cảm ơn!

Việt Trì, ngày 10 tháng 09 năm 2015

Nguyễn Thị Phương Thảo

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

MP70	Morris-Pratt
BM	Boyer-Moore
CLRS01	CLRS01
BDM	CR94
ACR99	Bakward-Oracle
BDM	BNDM
QS	Quick Search
KMP	Knuth-Morris-Pratt
brBc	Berry-Ravindran

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

Đối sánh xâu chính xác (*exac string matching*, sau đây gọi tắt là "*sánh xâu chính xác*"), còn được gọi là sánh mẫu chính xác (*exac pattern matching*) là bài toán tìm ra tất cả sự xuất hiện của một xâu *p* cho trước trong một văn bản *t*, trong đó *p*, *t* đều là các xâu văn bản theo một bảng chữ cái; *p* được gọi là "mẫu" (*pattern*) còn *t* là được gọi là "văn bản đích" (*target text*) [3, 8]. Một ví dụ gần gũi là cho một truy vấn *p* và một trang web *t*, kiểm tra xem *p* có xuất hiện trong nội dung của *t* hay không.

Theo Simone Faro và Thierry Lecroq [8], trong 40 năm gần đây, đối sánh xâu là một trong những bài toán được nghiên cứu rộng rãi nhất trong khoa học máy tính, chủ yếu vì các ứng dụng trực tiếp của nó cho rất nhiều lĩnh vực khác nhau như xử lý văn bản - hình ảnh - tín hiệu (*text, image and signal processing*), phân tích và nhận dạng giọng nói (*speech analysis and recognition*), truy hồi thông tin (*information retrieval*), nén dữ liệu (*data compression*), sinh học và hóa học tính toán (*computational biology and chemistry*). Bài toán sánh xâu chính xác trực tuyến (*onlineexac string matching*) nhận được sự quan tâm rất lớn của cộng đồng nghiên cứu. Trong thập kỷ 2001-2010, hơn 50 thuật toán mới được đưa ra, mở rộng thêm số lượng khoảng 40 thuật toán đã có từ trước năm 2000 [2]. Hệ thống các thuật toán này đã được phân tích, đánh giá công phu [5, 6, 7].

Theo Simone Faro và Thierry Lecroq, nhóm các thuật toán sánh mẫu nhanh có nguồn gốc từ thuật toán QS [9] đã chứng tỏ được lợi thế, đặc biệt khi mẫu đối sánh có độ dài ngắn. Chính vì lý do đó, luận văn này định hướng nghiên cứu một số thuật toán sánh mẫu chính xác nhanh có nguồn gốc từ thuật toán QS, tập trung vào họ thuật toán SSABS [8] – TVSBS [4] - FQS [3] do các thuật toán này đã tỏ ra ưu việt trong bài toán sánh mẫu ngắn. Họ thuật toán này là ở nhánh thuật toán khác với các thuật toán trong [1], hơn nữa, thuật toán FQS là mới được công bố vào năm 2014.

Nội dung chủ yếu của luận văn là nghiên cứu, phân tích chi tiết các thuật toán sánh mẫu SSABS – TVSBS - FQS, khai thác công cụ [11] để tiến hành thực nghiệm. Nội dung chính của luận văn gồm phần mở đầu, bốn chương nội dung, phần kết luận. Nội dung của bốn chương nội dung được giới thiệu sơ bộ như sau:

Chương 1. Giới thiệu chung về thuật toán sánh mẫu trình bày các khái niệm và đặc trưng của bài toán sánh mẫu, các ứng dụng của sánh mẫu, khái quát về các thuật toán sánh mẫu chính xác nhanh.

Chương 2. Họ thuật toán sánh mẫu chính xác nhanh SSABS -TVSBS- FQS giới thiệu về một lớp thuật toán sánh mẫu chính xác nhanh, trình bày và phân tích họ thuật toán SSABS-TVSBS- FQS. Đồng thời, các bước tiến hóa và hiệu suất của ba thuật toán này cũng được giới thiệu.

Chương 3. Chương trình thực nghiệm họ thuật toán đối sánh mẫu chính xác nhanh với bộ công cụ Smart.

Phần kết luận tổng kết các kết quả chính cũng như các hạn chế của luận văn, đồng thời, ý tưởng về các nghiên cứu tiếp theo cũng được giới thiệu.