ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

NGÂN HOÀNG MỸ LINH

BÀI TOÁN ĐỐI SÁNH MẪU SỬ DỤNG GIẢI THUẬT DI TRUYỀN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

NGÂN HOÀNG MỸ LINH

BÀI TOÁN ĐỐI SÁNH MẪU SỬ DỤNG GIẢI THUẬT DI TRUYỀN

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: TS. VŨ MẠNH XUÂN

THÁI NGUYÊN - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này của tự bản thân tôi tìm hiểu, nghiên cứu dưới sự hướng dẫn của TS Vũ Mạnh Xuân. Các chương trình thực nghiệm do chính bản thân tôi lập trình, các kết quả là hoàn toàn trung thực. Các tài liệu tham khảo được trích dẫn và chú thích đầy đủ.

TÁC GIẢ LUẬN VĂN

Ngân Hoàng Mỹ Linh

LỜI CẢM ƠN

Tôi xin bày tỏ lời cảm ơn chân thành tới tập thể các thầy cô giáo Viện công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam, các thầy cô giáo Trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên đã dạy dỗ chúng tôi trong suốt quá trình học tập chương trình cao học tại trường.

Đặc biệt tôi xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo TS Vũ Mạnh Xuân đã quan tâm, định hướng và đưa ra những góp ý, gợi ý, chỉnh sửa quý báu cho tôi trong quá trình làm luận văn tốt nghiệp.

Cuối cùng, tôi xin chân thành cảm ơn các bạn bè đồng nghiệp, gia đình và người thân đã quan tâm, giúp đỡ và chia sẻ với tôi trong suốt quá trình làm luận văn tốt nghiệp.

Thái Nguyên, tháng 08 năm 2015

Ngân Hoàng Mỹ Linh

MỤC LỤC

DANH MỤC THUẬT NGỮ, TỪ VIẾT TẮT, KÍ HIỆU

GA	Giải thuật di truyền
NST	Nhiễm sắc thể
Population	Quần thể
Pattern matching	Đối sánh mẫu
TSP	Bài toán người bán hàng

DANH MỤC CÁC HÌNH VẼ

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

Hiện nay, cùng với sự phát triển không ngừng của ngành khoa học máy tính chính là việc hệ thống thông tin được lưu trữ ngày càng đồ sộ. Đối với một kho thông tin lớn như vậy, việc người dùng muốn tra cứu, truy vấn dữ liệu cũng ngày càng khó khăn hơn. Bên cạnh đó, khi lượng thông tin phát triển quá nhiều, việc tổ chức, quản lí chúng để làm sao kiểm soát được việc bùng nổ thông tin cũng là một trong những vấn đề cần quan tâm của các nhà quản lí. Hiện nay đã có rất nhiều công cụ truy vấn có thể hỗ trợ cho người dùng phần nào trong việc tìm kiếm:

* Công cụ tìm kiếm của wikipedia: Chỉ tìm ra tên tựa bài của văn bản nào trùng hợp với từ khóa.

* Công cụ tìm kiếm của phần mềm ứng dụng Microsoft word: Công cụ FIND cho phép người dùng tìm kiếm cụm từ nội bên trong một hồ sơ, văn bản.

* Công cụ tìm kiếm của hệ điều hành Microsoft Windows và Adobe Reader: Cả hai công cụ này cho phép tìm kiếm các hồ sơ có chứa từ khóa trong một hồ sơ, một thư mục hay trong các ổ đĩa của máy tính.

Tuy nhiên, các công cụ trên vẫn tồn tại những hạn chế nhất định.Trong khi đó, công việc tìm kiếm, truy vấn dữ liệu làm sao để nhanh chóng và hiệu quả vẫn đang là một vấn đề cấp thiết đang được rất nhiều người dùng quan tâm. Các thông tin được lưu trữ trên máy tính tuy lớn nhưng đa số đều được lưu dưới dạng văn bản, và mặc dù có rất nhiều công cụ tìm kiếm nhưng cơ chế chung của chúng vẫn là dựa trên phương pháp sử dụng chuỗi. Đối sánh mẫu (pattern matching) là một bài toán quan trọng trong việc hỗ trợ tìm kiếm văn bản được áp dụng để tìm một xâu khớp với mẫu trong văn bản hoặc tìm các văn bản có chứa mẫu.

Giải thuật di truyền (GA – Genetic Algorithms) là một kỹ thuật cơ bản của tính toán mềm nhằm tìm kiếm giải pháp thích hợp cho các bài toán tối ưu tổ hợp, nó vận dụng các nguyên lý của tiến hóa như lai ghép, đột biến, chọn lọc. Ngày nay,

giải thuật di truyền được ứng dụng rộng rãi trên mọi lĩnh vực như tin sinh học, khoa học máy tính, trí tuệ nhân tạo, tài chính và một số ngành khác.

Đề tài **"Bài toán đối sánh mẫu sử dụng giải thuật di truyền"** nhằm mục đích nghiên cứu bài toán đối sánh mẫu, giải thuật di truyền và ứng dụng của giải thuật di truyền trong đối sánh mẫu và tìm kiếm văn bản.

Ngoài phần mở đầu và kết luận, luận văn gồm có 3 chương:

- Chương 1: Một số thuật toán đối sánh mẫu
- Chương 2: Giới thiệu về giải thuật di truyền
- Chương 3: Bài toán đối sánh mẫu sử dụng giải thuật di truyền

Phương pháp nghiên cứu

Trong luận văn, học viên đã sử dụng các phương pháp nghiên cứu chính sau:

- Phương pháp nghiên cứu lý thuyết: Tìm tòi, tổng hợp tài liệu, hệ thống lại các kiến thức, tìm hiểu các khái niệm, thuật toán sử dụng trong luận văn.
- Lập trình thử nghiệm: Luận văn sử dụng ngôn ngữ lập trình là Visual Studio C# 2012 để viết chương trình thử nghiệm.
- Các phương pháp so sánh.

CHƯƠNG 1
MỘT SỐ THUẬT TOÁN ĐỐI SÁNH MẪU

Chương này giới thiệu và phát biểu bài toán đối sánh mẫu, tìm hiểu một số thuật toán đã và đang được sử dụng để giải bài toán đối sánh mẫu.

1.1. Giới thiệu về bài toán đối sánh mẫu

Trong khoa học máy tính, đối sánh mẫu là hành động kiểm tra xem một trình tự các kí tự có hiện diện trong một xâu cho trước hay không. Ngược lại với nhận dạng mẫu, đối sánh mẫu thường có sự chính xác hơn. Dạng phổ biến nhất của bài toán đối sánh mẫu là: Cho trước nguồn tìm kiếm là một tập D các văn bản, cho một câu hỏi dạng văn bản q (thường là một từ, một xâu văn bản ngắn), hãy tìm tất cả các văn bản thuộc D mà có chứa q. Trong nhiều trường hợp (chẳng hạn, tìm kiếm thông qua máy tìm kiếm) q còn được gọi là "truy vấn" và bài toán còn có tên gọi là "tìm kiếm theo truy vấn". Để tìm được các văn bản có chứa văn bản truy vấn q, hệ thống tìm kiếm cần phải kiểm tra văn bản truy vấn q có là một xâu con của các văn bản thuộc tập D hay không (sánh mẫu) và đưa ra các văn bản đáp ứng. Trong nhiều trường hợp, bài toán còn đòi hỏi tìm tất cả các vị trí của các xâu con trong văn bản trùng với q. Đồng thời, điều kiện tìm kiếm có thể được làm "xấp xỉ" theo nghĩa văn bản kết quả có thể không cần chứa q mà chỉ cần "liên quan" tới q, nghĩa là có xâu con trong văn bản xấp xỉ q. Có thể thấy, các máy tìm kiếm sử dụng cả cơ chế tìm kiếm xấp xỉ khi mà văn bản kết quả tìm kiếm không chứa hoàn toàn chính xác văn bản truy vấn .[6]

1.2. Phát biểu bài toán

Đối sánh mẫu là một bài toán cơ bản trong xử lý văn bản, bài toán yêu cầu tìm ra một hoặc nhiều vị trí xuất hiện của mẫu q trên một văn bản S. Mẫu q và văn bản S là các chuỗi có độ dài M và N ($M \leq N$); q và S là các xâu ký tự trên cùng một bảng chữ cái Σ có δ ký tự. Bài toán sánh mẫu tổng quát được phát biểu như sau: