ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN VŨ MINH

THUẬT TOÁN DI TRUYỀN VÀ MỘT SỐ ỨNG DỤNG VỚI LỚP CÁC BÀI TOÁN NP

LUẬN VĂN THẠC SỸ CÔNG NGHỆ THÔNG TIN

Chuyên ngành: Khoa học máy tính

Mã số chuyên ngành: 60.48.01

Người hướng dẫn khoa học: **TS. Vũ Vinh Quang**

Thái Nguyên - 2012

LỜI CẢM ƠN

Trong thời gian hai năm của chương trình đào tạo thạc sỹ, trong đó gần một nửa thời gian dành cho các môn học, thời gian còn lại dành cho việc lựa chọn đề tài, giáo viên hướng dẫn, tập trung vào nghiên cứu, viết, chỉnh sửa và hoàn thiện đề tài. Với quỹ thời gian như vậy và với vị trí công việc đang phải đảm nhận, không riêng bản thân em mà hầu hết các sinh viên cao học muốn hoàn thành tốt luận văn của mình trước hết đều phải có sự sắp xếp thời gian hợp lý, có sự tập trung học tập và nghiên cứu với tinh thần nghiêm túc, nỗ lực hết mình; tiếp đến cần có sự ủng hộ về tinh thần, sự giúp đỡ về chuyên môn một trong những điều kiện không thể thiếu quyết định đến việc thành công của đề tài.

Để hoàn thành được đề tài này trước tiên em xin gửi lời cảm ơn đến thầy giáo hướng dẫn **TS. Vũ Vinh Quang**, người đã có những định hướng cho em về nội dung và hướng phát triển của đề tài, người đã có những đóng góp quý báu cho em về những vấn đề chuyên môn của đề tài, giúp em tháo gỡ kịp thời những vướng mắc trong quá trình làm luận văn.

Em cũng xin cám ơn các thầy cô giáo Trường Đại học Công nghệ thông tin và Truyền thông cũng như bạn bè cùng lớp đã có những ý kiến đóng góp bổ sung cho đề tài luận văn của em. Xin cảm ơn gia đình, người thân cũng như đồng nghiệp luôn quan tâm, ủng hộ hỗ trợ về mặt tinh thần trong suốt thời gian từ khi nhận đề tài đến khi hoàn thiện đề tài này.

Em xin hứa sẽ cố gắng hơn nữa, tự trau dồi bản thân, tích cực nâng cao năng lực chuyên môn của mình để sau khi hoàn thành đề tài này sẽ có hướng tập trung nghiên cứu sâu hơn, không ngừng hoàn thiện hơn nữa đề tài của mình để có những ứng dụng thực tiễn cao trong thực tế.

Thái Nguyên, tháng 8 năm 2012

Sinh viên

Trần Vũ Minh

MỤC LỤC

Lời cam đoan.....i
Lời cảm ơn.....ii
Mục lục.....iii
Danh mục các ký hiệu, các chữ viết tắt vi
Danh mục các bảng.....vii
Danh mục các hình.....viii
LỜI MỞ ĐẦU 1
CHƯƠNG 1 3
GIẢI THUẬT DI TRUYỀN 3
1.1 Giới thiệu về GA 3
1.2 Các khái niệm cơ bản 4
1.2.1 Cá thể, nhiễm sắc thể.....4
1.2.2 Quần thể.....4
1.2.3 Chọn lọc (Selection).....4
1.2.4 Lai ghép (Cross-over).....5
1.2.5 Đột biến (Mutation).....5
1.3 Mô hình GA 5
1.4 Các tham số của GA.....7
1.4.1 Kích thước quần thể 7
1.4.2 Xác suất lai ghép 7
1.4.3 Xác suất đột biến 7
1.5 Cơ chế thực hiện GA.....8
1.5.1 Mã hóa.....8
1.5.2 Khởi tạo quần thể ban đầu.....9
1.5.3 Xác định hàm thích nghi 9
1.5.4 Cơ chế lựa chọn 10
1.5.5 Các toán tử di truyền 11
1.6. Thuật toán di truyền kinh điển 13

1.6.1. Mã hóa 13
1.6.2. Toán tử chọn lọc 13
1.6.3. Toán tử lai ghép 14
1.6.4. Toán tử đột biến 16
1.6.5. Thuật toán di truyền mã hóa số thực (RCGA) 18
CHƯƠNG 2 25
CƠ SỞ TOÁN HỌC CỦA GIẢI THUẬT DI TRUYỀN 25
2.1. Định lý sơ đồ của Holland 25
2.1.1. Một số khái niệm 25
2.1.2. Định lý sơ đồ (Holland 1975) 26
2.2. Mô hình Markov của GA 27
2.2.1. Tính Markov 28
2.2.2. Xích Markov trong GA 29
2.2.3. Sự hội tụ của thuật toán di truyền 29
CHƯƠNG 3 32
GIẢI THUẬT DI TRUYỀN ĐỐI VỚI MỘT SỐ BÀI TOÁN THUỘC LỚP NP
3.1. Khái niệm về lớp các bài toán NP 32
3.2. Thuật toán di truyền với bài toán TSP 33
3.2.1 Giới thiệu bài toán 33
3.2.2 Mô tả bài toán 34
3.2.3 Giải thuật GA đối với bài toán TSP 36
3.3 Thuật toán GA giải bài toán TSP 39
3.3.1 Biểu diễn NST 39
3.3.2 Khởi tạo quần thể ban đầu 39
3.3.3 Chọn hàm thích nghi 39
3.3.4 Các toán tử di truyền 39
3.3.5 Toán tử đột biến 39
3.4. Thuật toán di truyền với bài toán tách từ trong văn bản 48
3.4.1 Một số thuật toán tách từ tiếng Việt hiện nay 50

3.4.2 Công cụ tách từ dùng GA.....52
3.4.3 Công cụ Opensource tách từ tiếng việt 59
KẾT LUẬN.....67
TÀI LIỆU THAM KHẢO.....68
NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN.....69
NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN.....70

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

GA – Genetic Algorithm: giải thuật di truyền

TSP - Travelling Salesman Problems: bài toán người du lịch

EC - Evolutionary computation: tính toán tiến hóa

EP - Evolutionary Programming: quy hoạch tiến hóa

ES - Evolutionary Strategies: các chiến lược tiến hóa

GP - Genetic Programming: lập trình di truyền

CS - Classifier Systems: các hệ thống phân loại

NST – nhiễm sắc thể

Selection: chọn lọc

Cross-over: lai ghép

Mutation: đột biến

Reproduction: sinh sản

pop-size: kích cỡ quần thể

RCGA: thuật toán di truyền mã hóa số thực

BLX-α - Blend Crossover: lai ghép BLX-α

CMX - Center of Mass Crossover: lai ghép CMX

NP-hard: bài toán NP khó

NP-complete: bài toán NP đầy đủ

WFST - Weighted finit-state Transducer: mô hình mạng chuyển dịch trạng thái hữu hạn có trọng số

IGATEC - Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese: Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật toán di truyền

df - document frequency: tần số tài liệu

fitness: độ thích nghi

DANH MỤC CÁC BẢNG

Bảng 1: Các tham số điều khiển hoạt động của thuật giải di truyền

Bảng 2. Thống kê độ dài từ trong từ điển

Bảng 3. Tham số thực hiện GA

Bảng 4. Gói vn.hus.mim, tokenizer và các gói con

DANH MỤC CÁC HÌNH

Hình 1: Sơ đồ mô tả GA

Hình 2: Lai ghép CMX

Hình 3: Phân bố của x_j^{ci}

Hình 4: Toán tử lai ghép SX

Hình 5: Sự phân lớp các bài toán

Hình 6: Giao diện chương trình TSP

Hình 7: Giao diện nhập dữ liệu chương trình TSP

Hình 8: Giao diện kết quả chương trình TSP

Hình 9. Biểu diễn cá thể bằng các bit 0,1

Hình 10. Thang tỉ lệ phát sinh loại từ

Hình 11. Quá trình lai ghép

Hình 12. Quá trình đột biến

Hình 13. Quá trình sinh sản

Hình 14. Quá trình chọn cá thể

Hình 15. Giao diện chính vnToolkit 3.0.0

Hình 16. Kết quả tách từ

Hình 17. Kết quả thống kê từ

Hình 18. Kết quả gỡ rối tách từ

Hình 19. Kết quả tách câu

Hình 20. Kết quả gán nhãn

Hình 21. Bộ dán nhãn được sử dụng

LỜI MỞ ĐẦU

Hiện nay trong ngành khoa học máy tính, việc tìm kiếm lời giải tối ưu cho các bài toán là vấn đề luôn được các nhà khoa học đặc biệt quan tâm. Mục đích chính của các thuật toán tìm kiếm lời giải là tìm ra lời giải tối ưu cho bài toán trong thời gian nhỏ nhất. Các thuật toán như tìm kiếm không có thông tin, vét cạn (tìm kiếm trên danh sách, trên cây hoặc đồ thị) hoặc các thuật toán tìm kiếm có thông tin được sử dụng nhiều trong không gian tìm kiếm nhỏ. Đối với không gian tìm kiếm lớn, việc tìm kiếm các lời giải tối ưu cho bài toán gặp nhiều khó khăn. Do đó, cần thiết phải có những thuật giải tốt và sử dụng kỹ thuật trí tuệ nhân tạo khi giải quyết các bài toán có không gian tìm kiếm lớn. Thuật giải di truyền (Genetic Algorithm - GA) là một trong những kỹ thuật tìm kiếm lời giải tối ưu đã đáp ứng được yêu cầu của nhiều bài toán và ứng dụng. Cùng với logic mờ, GA được ứng dụng rất rộng rãi trong các lĩnh vực phức tạp. Sự kết hợp giữa GA và logic mờ đã chứng tỏ được hiệu quả trong các vấn đề khó mà trước đây thường được giải quyết bằng các phương pháp thông thường hay các phương pháp cổ điển, nhất là trong các bài toán cần có sự lượng giá, đánh giá sự tối ưu của kết quả thu được. Chính vì vậy, GA đã trở thành một trong những đề tài nghiên cứu thu hút được nhiều sự quan tâm và hiện nay đã và đang đem đến rất nhiều ứng dụng trong thực tiễn.

Xuất phát từ thuyết tiến hóa muôn loài của Darwin, GA là một kỹ thuật chung giúp giải quyết vấn đề bài toán bằng cách mô phỏng sự tiến hóa của con người hay của sinh vật nói chung trong những điều kiện được qui định sẵn của môi trường. GA là một thuật giải và mục tiêu của GA không nhằm đưa ra lời giải chính xác tối ưu mà là đưa ra lời giải tương đối tối ưu.

John Holland (1975) và Goldberg (1989) đã đề xuất và phát triển GA, là thuật giải tìm kiếm dựa trên cơ chế chọn lọc và di truyền tự nhiên. Thuật giải này sử dụng các nguyên lý di truyền về sự thích nghi và sự sống các cá thể thích nghi nhất trong tự nhiên.

Ngày nay, GA được ứng dụng khá nhiều trong các lĩnh vực như khoa học, kinh doanh và giải trí. Đầu tiên phải kể đến là các bài toán tối ưu bao gồm: tối ưu số và tối ưu tổ hợp; đã sử dụng GA để tìm lời giải như là bài toán người du lịch (Travelling Salesman Problems - TSP).

Một ứng dụng khác cũng đang được ứng dụng rộng rãi của GA là giải quyết vấn đề bùng nổ về lượng thông tin trên mạng internet bao gồm: thư viện điện tử, thông tin điện tử... dẫn đến phát sinh một số lượng lớn văn bản với tốc độ tăng chóng mặt. Vấn đề làm sao để tổ chức và tìm kiếm một lượng thông tin lớn như vậy một cách có hiệu quả? GA hiện đang được ứng dụng hiệu quả trong việc phân loại thông tin phục vụ cho việc tìm kiếm văn bản.

Với những lý do trên, em chọn đề tài: "***Thuật toán di truyền và một số ứng dụng với lớp các bài toán NP***" làm luận văn tốt nghiệp.

Nội dung chính của luận văn gồm 3 chương:

Chương 1 trình bày các khái niệm cơ bản, mô hình, các tham số cơ bản, các phép toán, cơ chế thực hiện tổng quát của thuật toán di truyền, thuật toán di truyền mã hóa số thực.

Chương 2 trình bày cơ sở toán học về sự hội tụ của thuật toán di tuyền thông qua mô hình Markov và định lý sơ đồ của Holland.

Chương 3 trình bày hai nội dung chính:

+ Giới thiệu bài toán người du lịch (Travelling Salesman Problems – TSP) là một trong những bài toán thuộc lớp NP và phương pháp giải bài toán này bằng thuật toán di truyền.

+ Giới thiệu về bài toán tách từ trong văn bản, ứng dụng của GA đối với bài toán tách từ trong văn bản thông qua bộ công cụ tách từ dùng thuật giải di truyền ***vnToolkit 3.0***.

Các kết quả lý thuyết về bài toán TSP và bài toán tách từ trong văn bản đã được kiểm nghiệm thông qua các chương trình thực nghiệm viết trên nền ngôn ngữ C# và Java.