ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TÔ PHƯỚC SANG – CHÂU HUỲNH NHƯ**

**TÌM HIỂU VÀ ỨNG DỤNG GIẢI THUẬT DI TRUYỀN**

**VÀO BÀI TOÁN TÌM KIẾM VĂN BẢN**

**KHOÁ LUẬN TỐT NGHIỆP**

**NGÀNH: CÔNG NGHỆ THÔNG TIN**

**TRÌNH ĐỘ ĐÀO TẠO: ĐẠI HỌC**

NGƯỜI HƯỚNG DẪN: T.S. HUỲNH MINH TRÍ

**TP. HỒ CHÍ MINH, THÁNG 12 NĂM 2016**

**LỜI CAM ĐOAN**

*Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, các số liệu và kết quả nghiên cứu nêu trong luận văn là trung thực, được các đồng tác giả cho phép sử dụng và chưa từng được công bố trong bất kì một công trình nào khác*

**TP.Hồ Chí Minh, tháng 12 năm 2016**

*Tác giả khoá luận*

**Châu Huỳnh Như**

**Tô Phước Sang**

**LỜI CẢM ƠN**

*Để hoàn thành khoá luận này, chúng em xin chân thành tỏ lòng biết ơn và niềm tri ân sâu sắc đến T.S. Huỳnh Minh Trí , người đã trực tiếp hướng dẫn tận tình giúp đỡ chúng em có thể hoàn thành tốt nhất bài khoá luận này.*

*Chúng em cũng chân thành gửi lời cảm ơn đến với quý thầy cô trong khoa Công Nghệ Thông Tin, trường Đại Học Sài Gòn đã tận tình truyền đạt những kiến thức quý báu trong suốt quá trình học tập, với vốn kiến thức ấy, không chỉ là nền tảng cho quá trình nghiên cứu khoá luận mà còn là hành trang vững chắc để chúng em bước vào đời một cách tự tin nhất.*

*Trong quá trình nghiên cứu cũng như hoàn thành bài khoá luận này, khó có thể tránh khỏi sai sót, rất mong các quý thầy cô bỏ qua, đồng thời chúng em cũng rất mong nhận được ý kiến đóng góp của các quý thầy cô để chúng em có thêm nhiều kinh nghiệm và có một bài khoá luận hoàn hảo nhất.*

*Cuối cùng, chúng em kính chúc quý thầy cô dồi dào sức khoẻ, tràn đầy niềm tin để thực hiện sứ mệnh cao đẹp của mình và đạt được nhiều công tốt đẹp trong công việc cũng như trong cuộc sống.*

*Chúng em xin chân thành cảm ơn!*

**TP.Hồ Chí Minh, tháng 12 năm 2016**

*Tác giả khoá luận*

**Châu Huỳnh Như**

**Tô Phước Sang**

**Lời mở đầu**

Thuật giải di truyền, cũng như các thuật toán tiến hoa nói chung, hình thành dựa trên quan niệm cho rằng, quá trình tiến hóa tự nhiên là hoàn hảo nhất, hợp lý nhất, và tự nó đã mang tính tối ưu. Quan niệm này có thể được xem là một tiên đề đúng, không chứng minh được, nhưng phù hợp với thực tế khách quan. Quá trình tiến hóa thể hiện tính tối ưu ở chổ, thế hệ sau bao giờ cũng tốt hơn, phát triển hơn và hoàn thiện hơn thế hệ trước.

Hiện nay, thuật toán di truyền cùng với logic mờ được ứng dụng rất rộng rãi trong các lĩnh vực tương đối phức tạp. Thuật toán di truyền kết hợp logic mờ đã chứng tỏ được hiệu quả của nó trong các vấn đề khó giải quyết bằng các phương pháp thông thường hay các phương pháp cố 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, thuật giải di truyền (Genetic Algorithm) đã trở thành đề tài nghiên cứu thú vị và mang đến nhiều ứng dụng thực tiễn ngày nay.

Thuật giải di truyền giải bài toán bằng cách mô phỏng lại và áp dụng các quy luật trong quá trình tiến hóa của tự nhiên. Phương pháp này không đảm bảo đưa ra lời giải tối ưu mà chỉ cố gắng tìm kiếm các lời giải tương đối tối ưu.

Các thuật toán tìm kiếm không có thông tin cổ điển, vét cạn sử dụng phương pháp tìm kiếm đơn giản và trực quan. Hay những phương pháp tìm kiếm có thông tin, sử dụng cấu trúc của không gian tìm kiếm giúp giảm thời gian tìm kiếm. Tuy nhiên những phương pháp này chỉ có tác dụng tốt khi áp dụng đối với những không gian tìm kiếm nhỏ. Trong thực tiễn, có rất nhiều bài toán tìm kiếm với không gian tìm kiếm lớn. Vì vậy, đòi hỏi cần có những thuật toán chất lượng hơn, và sử dụng trí tuệ nhân tạo là rất cần thiết trong giải quyết những bài toán có không gian tìm kiếm lớn. Thuật giải di truyền (genetic algorithm) là một trong những thuật giải đáp ứng được yêu cầu đó.

Hiện nay, nhiều hệ điều hành đã cung cấp công cụ hỗ trợ tìm kiếm các file văn bản chủ yếu dựa trên 2 cách là: tìm theo từ khoá tên file và tìm theo từ khoá nội dung trong file. Mặc dù đã được hỗ trợ tốt như vậy nhưng những phương pháp này vẫn còn nhiều hạn chế, chẳng hạn như chỉ tìm được các file văn bản có chính xác từ khoá chứa trong nội dung văn bản, như vậy sẽ rất khó khăn cho người dùng nếu họ không nhớ chính xác từ khoá, hơn nữa phương pháp này chưa chỉ ra được vị trí cụ thể của từ khoá trong đoạn văn bản và tần suất xuất hiện của chúng.

Vì lẽ đó bài toán tìm kiếm văn bản là bài toán rất thiết thực đang được nhiều người quan tâm, vấn đề cấp thiết đặt ra là giải quyết bài toán tìm kiếm văn bản sao cho hiệu quả, đáp ứng được nhu cầu của người sử dụng. Khoá luận này định hướng nghiên cứu sử dụng giải thuật di truyền tìm trong file văn bản các đoạn văn bản giống hoặc gần giống với mẫu (từ khoá) cần tìm kiếm.

Với mục tiêu đó, nhóm chúng em lựa chọn đề tài nghiên cứu của khoá luận là “Bài toán tìm kiếm văn bản sử dụng giải thuật di truyền”. Đây là hướng tiếp cận khá mới đối với bài toán này, hy vọng rằng kết quả đạt được sẽ có hiệu quả đáng kể so với các phương pháp tìm kiếm khác.

**TÓM TẮT NỘI DUNG KHOÁ LUẬN**

**Chương 1**

*Giới Thiệu Giải Thuật Di Truyền*

* 1. **Tổng quan về giải thuật di truyền**
* Giới thiệu sơ lược về giải thuật di truyền
* Các tính chất của giải thuật di truyền
  1. **Cấu trúc giải thuật di truyền**

Giới thiệu sơ lược về cấu trúc của một giải thuật di truyền.

* 1. **Các thao tác cơ bản**
* Các thao tác được áp dụng để giải một bài toán tối ưu bằng giải thuật di truyền
* Các khái niệm về mã hoá, phát sinh quần thể, hàm đánh giá, hàm thích nghi, các toán tử di truyền, các tham số,…
  1. **Các bước qua trọng trong việc áp dụng giải thuật di truyền**

Chỉ ra các bước quan trọng để giải quyết bài toàn bằng giải thuật di truyền.

**Chương 2**

*Bài Toán Tìm Kiếm Văn Bản Với Giải Thuật Di Truyền*

**2.1. Yêu cầu đặt ra bài toán tìm kiếm văn bản**

Đặt vấn đề cho bài toán tìm kiếm văn bản.

**2.2. Xây dựng hàm tìm kiếm**

Cách xây dựng và triển khai hàm tìm kiếm.

**2.3. Phát biểu bài toán tìm kiếm văn bản theo hướng tiếp cận di truyền**

Phát biểu bài toán đặt vấn đề theo hướng di truyền.

**2.4. Tìm độ dài xâu con chung lớn nhất bằng quy hoạch động**

Định nghĩa, công thức và cách cài đặt một xâu con chung.

**2.5. Áp dụng giải thuật di truyền**

Sử dụng các khái niệm, định nghĩa, các thao tác áp dụng giải thuật di truyền vào bài toán

**Chương 3**

*Ứng Dụng Giải Thuật Di Truyền*

*Cho Bài Toán Tìm Kiếm Văn Bản*

**3.1. Giới thiệu chung**

Giới thiệu sơ lược về ứng dụng.

**3.2. Một số ví dụ thử nghiệm**

Trình bày bài toán, đưa ra các tham số và tiến hành chạy thử nghiệm trên ứng dụng đã được giới thiệu và đưa ra nhận xét.

**Chương 4**

*Tổng Kết*

Tổng kết, đánh giá, đưa ra kết luận, kiến nghị hướng phát triển, những gì đã đạt được và hạn chế của khoá luận.

**TÀI LIỆU THAM KHẢO**

**PHỤ LỤC**