TRƯỜNG ĐẠI HỌC THỦY LỢI

****

ĐỀ CƯƠNG
ĐỀ ÁN THẠC SĨ

ĐỀ TÀI:

ỨNG DỤNG THUẬT TOÁN DI TRUYỀN TRONG TÌM KIẾM VĂN BẢN

Học viên cao học	**: Nguyễn Trọng Anh**
Lớp	**: 31CNTT21**
Mã số học viên	**: 238217501**
Chuyên ngành	**: Công nghệ Thông tin**
Mã số	**: 8480201**
Người HD khoa học	**: TS. Trần Mạnh Tuấn**
	TS. Lê Minh Tuấn
Bộ môn quản lý	**: Hệ thống thông tin**

Hà Nội, 2025

ĐĂNG KÝ ĐỀ TÀI ĐỀ ÁN THẠC SĨ

I. HỌC VIÊN CAO HỌC

1. Họ và tên: **Nguyễn Trọng Anh**
2. Sinh ngày: 29/10/1977
3. Học viên lớp cao học: 31CNTT21
4. Chuyên ngành: Công nghệ Thông tin Mã số: 8480201
5. Cơ quan công tác: Công ty TNHH CMC Global
6. Điện thoại: 0912.099.254 Email: anhnguyen.gck@gmail.com

II. THÔNG TIN VỀ NGƯỜI HƯỚNG DẪN

1. Họ và tên : **Trần Mạnh Tuấn**
   - Học hàm, học vị: Tiến sĩ
   - Chuyên ngành: Cơ sở Toán học cho hệ thống thông tin
   - Đơn vị công tác: Bộ môn HTTT, Khoa CNTT, Trường ĐH Thuỷ Lợi
   - Địa chỉ: Phòng 205-C1, Trường Đại học Thuỷ lợi
   - Điện thoại: 098.366.8841
2. Họ và tên : **Lê Minh Tuấn**
   - Học hàm, học vị: Tiến sĩ
   - Ngành: Công nghệ thông tin
   - Đơn vị công tác: Khoa Công nghệ và Kỹ thuật, Trường ĐH Công Đoàn
   - Địa chỉ: số 169 Tây Sơn – Đống Đa – Hà Nội
   - Điện thoại: 0937.862.789

III. THÔNG TIN VỀ ĐỀ TÀI

1. Tên đề tài

Ứng dụng thuật toán di truyền trong tìm kiếm văn bản.

2. Bộ môn quản lý

- Bộ môn Hệ thống thông tin – Khoa Công nghệ Thông tin.

3. Nội dung, phương pháp nghiên cứu và các kết quả dự kiến đạt được

3.1. Nội dung nghiên cứu

- Nghiên cứu cơ sở lý thuyết về thuật toán di truyền.

- Nghiên cứu ứng dụng tính tương đồng trong việc tìm kiếm văn bản.

- Ứng dụng và sử dụng thuật toán để cài đặt module tìm kiếm văn bản trong C#.

3.2. Phương pháp nghiên cứu

- Tối ưu hóa một hàm cụ thể, tìm kiếm giải pháp cho một bài toán tối ưu, hoặc cải thiện hiệu suất của thuật toán di truyền.

- Nghiên cứu và tổng hợp các tài liệu liên quan đến thuật toán di truyền, bao gồm nguyên lý hoạt động, các thuật toán đã được phát triển, và ứng dụng thực tiễn.

- Phương pháp kế thừa: Dựa trên các công trình nghiên cứu đã được công bố liên quan đề tài trong và ngoài nước, các luận văn, luận án của các tác giả đã được bảo vệ và đánh giá tại các hội đồng Khoa học;

- Phương pháp thực nghiệm: Trên cơ sở phân tích, tổng hợp, kế thừa tác giả áp dụng và thực nghiệm các phương pháp học sâu với bài toán của đề tài.

3.3. Kết quả dự kiến đạt được

- Chọn một bài toán thực tế để áp dụng thuật toán di truyền (cụ thể: ứng dụng thuật toán đề tìm kiếm văn bản).

- Sử dụng ngôn ngữ lập trình C# để triển khai thuật toán di truyền, bao gồm các bước khởi tạo, đánh giá, chọn lọc, lai ghép, và đột biến.

- Khi cài đặt và ứng dụng thuật toán di truyền tác giả cho phép thay đổi các tham số như kích thước population, tỷ lệ đột biến, và số thế hệ, ... trong chương trình ứng dụng.

4. Những yêu cầu thực hiện luận văn (nếu có):

- Đảm bảo tiến độ thực hiện các nội dung trong đề cương.
- Thường xuyên trao đổi với GVHD.
- Nộp luận văn đúng hạn.

5. Các công việc thực hiện có liên quan đến luận văn

- ***Các môn học chính học viên đã học và dự kiến lựa chọn học có liên quan đến đề tài:*** Học máy nâng cao, Khai phá dữ liệu nâng cao, Dữ liệu lớn, Hệ thống thông minh nâng cao.

Hà Nội, ngày 20 tháng 05 năm 2025

NGƯỜI ĐĂNG KÝ

Nguyễn Trọng Anh

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

Từ viết tắt	**Nội dung đầy đủ**
GA	Genetic Algorithm
GVHD	Giảng viên hướng dẫn

MỞ ĐẦU

1. Giới thiệu

- **Nghiên cứu Giải Thuật Di Truyền**:
  - Tìm hiểu các nguyên lý hoạt động của GA, bao gồm cách thức chọn lọc, lai ghép và đột biến.
  - Phân tích các tham số quan trọng như kích thước population, tỷ lệ đột biến, và cách đánh giá độ phù hợp (fitness).
- **Bài Toán Tìm Kiếm Văn Bản**:
  - Xác định các yêu cầu của bài toán tìm kiếm văn bản, chẳng hạn như tìm kiếm chuỗi từ khóa trong một tập hợp tài liệu lớn.
  - Đánh giá các phương pháp tìm kiếm hiện tại và xác định những hạn chế mà GA có thể khắc phục.
- **Ứng Dụng Giải Thuật Di Truyền**:
  - Thiết kế một mô hình sử dụng GA để tối ưu hóa quá trình tìm kiếm văn bản.
  - Sử dụng độ tương đồng cosine hoặc các chỉ số tương tự để đánh giá độ phù hợp của các tài liệu so với từ khóa tìm kiếm.

2. Đối tượng và phạm vi nghiên cứu

a. Đối tượng nghiên cứu - Xây Dựng Mô Hình GA

- Chromosome: Mỗi cá thể (chromosome) đại diện cho một truy vấn tìm kiếm. Các gene có thể là các từ khóa hoặc cụm từ trong văn bản.
- Hàm Đánh Giá: Sử dụng độ tương đồng cosine để tính toán độ phù hợp của tài liệu với truy vấn tìm kiếm.

b. *Quy Trình Tìm Kiếm:*

- Khởi tạo population của các truy vấn tìm kiếm ngẫu nhiên.
- Đánh giá độ phù hợp của từng truy vấn với các tài liệu trong cơ sở dữ liệu.
- Chọn lọc các truy vấn tốt nhất, thực hiện lai ghép và đột biến để tạo ra thế hệ mới.
- Lặp lại cho đến khi đạt được độ phù hợp mong muốn hoặc đến số thế hệ tối đa.

3. Xây dựng phần mềm Tìm Kiếm Văn Bản

- Giao Diện Người Dùng:
    - Thiết kế một giao diện đơn giản cho phép người dùng nhập từ khóa tìm kiếm và chọn tập tài liệu cần tìm kiếm.
- Chức Năng Tìm Kiếm:
    - Cài đặt thuật toán GA để xử lý các truy vấn người dùng và tìm kiếm trong cơ sở dữ liệu.
    - Hiển thị kết quả tìm kiếm một cách rõ ràng và nhanh chóng.
- Kiểm Tra và Đánh Giá:
    - Thực hiện kiểm tra với nhiều tập dữ liệu khác nhau để đánh giá hiệu suất của phần mềm.
    - Đo lường độ chính xác và tốc độ tìm kiếm, so sánh với các phương pháp tìm kiếm truyền thống.

4. Cách tiếp cận và phương pháp nghiên cứu

a. Cách tiếp cận

- Tiếp cận kế thừa: Sử dụng các tài liệu, bài báo chuyên đề nghiên cứu có sẵn.

- Trao đổi với GVHD về các vấn đề nghiên cứu để xây dựng mô hình.

- Tiếp cận hệ thống: Xây dựng mô hình cải thiện tham số dựa trên kết quả thu thập từ các mô hình học máy.

b. Phương pháp nghiên cứu

- Phương pháp khảo sát, phân tích: Khảo sát, phân tích các bài toán sử dụng mô hình học sâu, học máy. Từ đó thu thập, phân tích, đánh giá dữ liệu liên quan.

- Phương pháp thu thập, phân tích, tổng hợp tài liệu liên quan bài toán

- Phương pháp kế thừa: Dựa trên các công trình nghiên cứu đã được công bố liên quan đề tài trong và ngoài nước, các nghiên cứu luận văn, luận án của các tác giả đã được bảo vệ và đánh giá tại các hội đồng Khoa học;

- Phương pháp thực nghiệm: Trên cơ sở phân tích, tổng hợp, kế thừa tác giả áp dụng và thực nghiệm các phương pháp học sâu với bài toán của đề tài.

5. Kết quả dự kiến đạt được

- Thu thập, tổng hợp dữ liệu cho bài toán.

- Phần mềm đã hoạt động đúng và cho ra kết quả chính xác theo mong muốn, giúp người dùng tìm kiếm văn bản một cách nhanh chóng và hiệu quả.

- Kết quả tìm kiếm được hiển thị rõ ràng, cho phép người dùng dễ dàng truy cập vào các tài liệu liên quan.

- Báo cáo đề án hoàn chỉnh.

6. Nội dung của đề án

PHẦN I: MỞ ĐẦU

1. Giới thiệu bài toán
2. Mục tiêu nghiên cứu
3. Đối tượng và phạm vi nghiên cứu
4. Phương pháp nghiên cứu
5. Kết cấu đề án

PHẦN II: NỘI DUNG NGHIÊN CỨU

1. Giới Thiệu Về Giải Thuật Di Truyền

- **Khái niệm**: Giới thiệu về GA, lịch sử phát triển và ứng dụng.
- **Nguyên lý hoạt động**: Mô tả các bước cơ bản của GA, bao gồm khởi tạo population, đánh giá độ phù hợp, chọn lọc, lai ghép và đột biến.

2. Bài Toán Tìm Kiếm Văn Bản

- **Định nghĩa bài toán**: Giới thiệu về bài toán tìm kiếm văn bản, mục tiêu và tầm quan trọng.
- **Các phương pháp hiện tại**: Phân tích các phương pháp tìm kiếm truyền thống (như TF-IDF, Vector Space Model) và hạn chế của chúng.

3. Sử Dụng Giải Thuật Di Truyền Trong Tìm Kiếm Văn Bản

- **Mô hình hóa bài toán**: Thiết kế mô hình GA cho bài toán tìm kiếm văn bản, xác định cách mã hóa các truy vấn.
- **Hàm đánh giá**: Xây dựng hàm đánh giá độ phù hợp dựa trên độ tương đồng cosine hoặc các chỉ số khác.

4. Thiết Kế và Triển Khai Phần Mềm

- **Giao diện người dùng**: Thiết kế giao diện cho phần mềm tìm kiếm văn bản.

- **Cài đặt thuật toán GA**: Triển khai GA trong mã nguồn để xử lý truy vấn tìm kiếm.

5. Thực Nghiệm và Đánh Giá

- **Phương pháp thực nghiệm**: Mô tả cách thức thực hiện thử nghiệm, tiêu chí đánh giá (độ chính xác, tốc độ).
- **Kết quả**: Trình bày và phân tích kết quả thu được từ phần mềm, so sánh với các phương pháp khác.

6. Kết Luận và Đề Xuất Hướng Phát Triển

- **Tóm tắt kết quả nghiên cứu**: Đánh giá hiệu quả của GA trong tìm kiếm văn bản.
- **Hướng phát triển**: Đề xuất những cải tiến có thể thực hiện cho thuật toán và phần mềm trong tương lai.

PHẦN III: CÀI ĐẶT VÀ THỰC NGHIỆM

1. Cài Đặt Môi Trường Phát Triển

- **Chọn Ngôn Ngữ Lập Trình**: Lựa chọn ngôn ngữ lập trình phù hợp (ví dụ: C#) để triển khai thuật toán.
- **Cài Đặt Thư Viện**: Cài đặt các thư viện cần thiết cho việc xử lý văn bản và triển khai GA, chẳng hạn như:
    - **C#**: GeneticSharp hoặc các thư viện tương tự.

2. Xây Dựng Cấu Trúc Dữ Liệu

- **Tổ Chức Tài Liệu**: Tạo một cơ sở dữ liệu hoặc tập tin chứa các tài liệu văn bản để tìm kiếm. Đảm bảo rằng dữ liệu được định dạng đúng (ví dụ: JSON, CSV).
- **Mã Hóa Truy Vấn**: Thiết kế cách mã hóa truy vấn tìm kiếm thành các chromosome trong GA. Mỗi chromosome có thể chứa các từ khóa hoặc cụm từ.

3. Triển Khai Giải Thuật Di Truyền

- **Khởi Tạo Population**: Tạo một population ban đầu với các truy vấn ngẫu nhiên.

- **Hàm Đánh Giá**: Xây dựng hàm tính độ phù hợp cho mỗi truy vấn dựa trên độ tương đồng cosine hoặc các chỉ số tương tự.
- **Quy Trình GA**:
    - Chọn lọc: Sử dụng phương pháp chọn lọc (như roulette wheel selection).
    - Lai ghép: Áp dụng kỹ thuật lai ghép để tạo ra các truy vấn mới.
    - Đột biến: Thực hiện các phép đột biến để duy trì tính đa dạng.

4. Chạy Thí Nghiệm

- **Thực Hiện Nhiều Chạy Thí Nghiệm**: Chạy GA với các tham số khác nhau (kích thước population, tỷ lệ đột biến, số thế hệ) để tìm ra cấu hình tối ưu.
- **Ghi Nhận Kết Quả**: Lưu trữ kết quả của các thí nghiệm, bao gồm độ chính xác và thời gian tìm kiếm cho từng cấu hình.

5. Phân Tích Kết Quả

- **So Sánh Hiệu Suất**: So sánh kết quả tìm kiếm của GA với các phương pháp tìm kiếm truyền thống (như TF-IDF) trên cùng một tập dữ liệu.
- **Đánh Giá Độ Chính Xác**: Tính toán và phân tích độ chính xác tìm kiếm (precision, recall) để đánh giá hiệu quả của hệ thống.

6. Báo Cáo Kết Quả

- **Trình Bày Kết Quả**: Sử dụng biểu đồ, bảng, hoặc đồ họa để trình bày kết quả một cách trực quan.
- **Thảo Luận**: Phân tích các kết quả, nêu rõ những điểm mạnh và điểm yếu của GA trong bài toán tìm kiếm văn bản.

PHẦN KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận
2. Hướng phát triển

TÀI LIỆU THAM KHẢO

[1] Nguyễn, V. Q. (2009). Bài toán tìm kiếm văn bản sử dụng giải thuật di truyền (Doctoral dissertation, Đại học Thái Nguyên).

[2] Sahu, A., Sinha, S., & Banka, H. (2024). Fuzzy inference system using genetic algorithm and pattern search for predicting roof fall rate in underground coal mines. International Journal of Coal Science & Technology, 11(1), 1.

[3] Challagundla, B. C., & Challagundla, S. (2024). Dynamic Adaptation and Synergistic Integration of Genetic Algorithms and Deep Learning in Advanced Natural Language Processing.

KẾ HOẠCH THỰC HIỆN

STT	Thời gian	Nội dung công việc	Kết quả dự kiến đạt được
1	4 tuần	Xác định rõ mục tiêu nghiên cứu và các câu hỏi chính cần trả lời	Hiểu về thuật toán di truyền
2	8 tuần	Đọc và tổng hợp các tài liệu về giải thuật di truyền và bài toán tìm kiếm văn bản..	Nghiên cứu được các phương pháp tìm kiếm văn bản hiện tại và xác định điểm mạnh, điểm yếu của chúng
3	6 tuần	Thiết kế mô hình giải thuật di truyền cho bài toán tìm kiếm văn bản.	Xác định và xây dựng hàm đánh giá độ phù hợp cho các truy vấn tìm kiếm.
4	4 tuần	Hoàn thiện mô hình. Viết báo cáo luận văn.	Báo cáo luận văn thạc sĩ.
5	2 tuần	Kiểm tra, chỉnh sửa báo cáo. Chuẩn bị tài liệu bảo vệ.	Báo cáo luận văn thạc sĩ hoàn chỉnh. Slide bảo vệ.
6	1 tuần	Bảo vệ luận văn.	Hoàn thành bảo vệ luận văn thạc sĩ.

Hà Nội, ngày 02 tháng 05 năm 2025

Người viết Đề cương

Nguyễn Trọng Anh

Ý KIẾN CỦA NGƯỜI HƯỚNG DẪN:

.....

.....

.....

.....

.....

.....

Ý KIẾN CỦA BỘ MÔN:

.....

.....

.....

.....

.....

Ý KIẾN CỦA HỘI ĐỒNG KHOA HỌC:

.....

.....

.....

.....

.....

.....