

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN THỊ THANH

ỨNG DỤNG GIẢI THUẬT DI TRUYỀN VÀO PHÂN LOẠI
TÀI LIỆU DẠNG VĂN BẢN

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn “**Ứng dụng giải thuật di truyền vào phân loại tài liệu dạng văn bản**” là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của **PGS.TS. Bùi Thế Hồng**. Toàn bộ phần mềm do chính tôi lập trình và kiểm thử. Tôi xin chịu trách nhiệm về lời cam đoan của mình.

Các số liệu và thông tin sử dụng trong luận văn này hoàn toàn là trung thực.

Tác giả

Trần Thị Thanh

MỤC LỤC

MỤC LỤC.....	i
DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT.....	vi
DANH MỤC CÁC BẢNG.....	vii
DANH MỤC CÁC HÌNH VẼ.....	viii
MỞ ĐẦU.....	1
CHƯƠNG 1: TÌM HIỂU VỀ KHAI PHÁ DỮ LIỆU.....	3
1.1 Giới thiệu chung.....	3
1.1.1. Giới thiệu.....	3
1.1.2. Khái niệm.....	3
1.1.3. Đặc điểm của bài toán khai phá dữ liệu.....	4
1.2. Quá trình khám phá tri thức trong cơ sở dữ liệu.....	6
1.2.1. Gom dữ liệu.....	7
1.2.2. Trích lọc dữ liệu.....	7
1.2.3. Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu.....	8
1.2.4. Chuyển đổi dữ liệu.....	9
1.2.5. Khai phá dữ liệu - Phát hiện và trích mẫu dữ liệu.....	9
1.2.6. Đánh giá kết quả mẫu.....	10
1.3. Khái quát các kỹ thuật khai phá dữ liệu.....	10
1.3.1. Kỹ thuật khai phá dữ liệu dự đoán.....	10
1.3.1.1. Phân lớp dữ liệu.....	10
1.3.1.2. Hồi quy.....	12
1.3.2. Kỹ thuật khai phá dữ liệu mô tả.....	13
1.3.2.1 Phân cụm dữ liệu.....	13
1.3.2.2. Tóm tắt.....	14

1.3.3. So sánh các tiếp cận khai phá dữ liệu: phân cụm - phân lớp	14
1.3.4. Ứng dụng phân cụm	15
1.3.5. Ví dụ	15
1.4. Ý nghĩa thực tiễn và tình hình ứng dụng	17
1.4.1. Ý nghĩa thực tiễn	17
1.4.2. Tình hình ứng dụng	18
CHƯƠNG 2: TÌM HIỂU VỀ THUẬT GIẢI DI TRUYỀN	19
2.1. Tổng quan về giải thuật di truyền	19
2.1.1. Giới thiệu	19
2.1.2. Các tính chất quan trọng của giải thuật di truyền.....	20
2.1.3. Cơ sở sinh học của giải thuật di truyền	21
2.1.4. Sơ đồ thực hiện giải thuật di truyền	21
2.1.5. Ứng dụng	24
2.2. Các khái niệm chung về giải thuật di truyền	24
2.2.1. Chuỗi nhiễm sắc thể	24
2.2.2. Các cá thể	25
2.2.3. Phương pháp mã hóa	25
2.2.4. Quần thể	25
2.2.5. Hàm thích nghi	26
2.2.6. Lai ghép, đột biến, tái sinh và chọn lọc.....	26
2.3. Các phép toán di truyền.	27
2.3.1. Mã hóa	27
2.3.1.1 Mã hóa nhị phân.....	27
2.3.1.2 Mã hóa hoán vị	28

2.3.1.3 Mã hóa giá trị.....	28
2.3.1.4 Mã hóa theo cây	28
2.3.2. Quá trình lai ghép.....	29
2.3.2.1. Lai ghép giá trị thực.....	29
2.3.2.2. Lai ghép giá trị nhị phân.....	31
2.3.3. Đột biến	32
2.3.3.1. Đột biến các giá trị thực.....	32
2.3.3.2 Đột biến các giá trị nhị phân.....	33
2.3.4. Phép chọn lọc	33
2.3.4.1. Phương pháp chọn lọc dùng bánh xe Roulette	33
2.3.4.2. Phương pháp chọn lọc Stochastic Universal Sampling.....	34
2.3.4.3. Phương pháp chọn lọc địa phương	35
2.3.4.4. Phương pháp lựa chọn loại bỏ	36
2.4. Các tham số của thuật giải di truyền.....	36
2.4.1. Kích thước quần thể	36
2.4.2. Xác suất lai giống.....	37
2.4.3. Xác suất đột biến	37
2.4.4. Số lượng thế hệ.....	38
CHƯƠNG 3: ỨNG DỤNG GIẢI THUẬT DI TRUYỀN VÀO PHÂN LOẠI.....	39
TÀI LIỆU DẠNG VĂN BẢN	39
3.1. Phân loại văn bản.....	39
3.1.1. Khái niệm	39
3.1.2. Quá trình phân loại văn bản	39
3.2. Giới thiệu bài toán phân loại văn bản.....	41

3.3. Các phương pháp biểu diễn văn bản.....	41
3.3.1. Mô hình không gian vector (Vector Space Model - VSM).....	41
3.3.2. Mô hình BOOLEAN	43
3.3.3. Mô hình tần suất	44
3.3.3.1. Phương pháp dựa trên tần số thuật ngữ (TF)	44
3.3.3.2. Phương pháp dựa trên nghịch đảo tần số văn bản (TDF)	45
3.3.3.3. Phương pháp $TF \times IDF$	45
3.3.4. Phương pháp xử lý vector thưa	46
3.3.5 Mô hình đồ thị	46
3.4. Các thuật toán phân loại văn bản.....	48
3.4.1. Bộ phân loại Vector hỗ trợ (SVM).....	48
3.4.2. Phân loại văn bản và SVM	53
3.4.3. Thuật toán k-NN (k-Nearest Neighbor)	60
3.5. Giải thuật di truyền phân loại văn bản.....	62
3.5.1. Lựa chọn mô hình biểu diễn văn bản	62
3.5.1.1. Biểu diễn vector của văn bản	63
3.5.1.2. Phép tính độ tương tự giữa hai vector.....	63
3.5.1.3. Vector trọng tâm của một nhóm văn bản.....	63
3.5.1.4. Phép tính độ tương tự giữa hai nhóm văn bản.....	63
3.5.2. Phương án tách thuật ngữ.....	64
3.5.2.1. Đối với các ngôn ngữ đơn âm tiết (single-term)	64
3.5.2.2. Đối với các ngôn ngữ đa âm tiết (multi-term)	64
3.5.2.3. Loại nhiễu	65
3.5.2.4. Mã hóa ký tự	66

3.5.2.5. Tách từ khóa.....	66
3.5.2.6. Loại từ dừng (Stop Words)	66
3.5.2.7. Thống kê từ khóa.....	66
3.5.3. Sử dụng thuật giải di truyền trích chọn từ khóa.....	67
3.5.3.1. Giới thiệu	67
3.5.3.2. Độ thích hợp của từ khóa	67
3.5.3.3. Ứng dụng giải thuật di truyền để tối ưu hóa độ thích nghi của từ khóa	69
3.6. Cài đặt và thử nghiệm chương trình	69
KẾT LUẬN	73
TÀI LIỆU THAM KHẢO.....	74

DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT

Các từ viết tắt	Nghĩa tiếng anh	Nghĩa tiếng việt
KDD	Knowledge Discovery and Data Mining	Kỹ thuật phát hiện tri thức và khai phá dữ liệu
VSM	Vector Space Model	Mô hình không gian vector
VC	Vapnik-Chervonenkis	Kích thước VC
SVM	Support Vector Machine	Bộ phân loại Vector hỗ trợ
RBF	Radial Basis Functions	Bộ phân loại chức năng
SMO	Sequential Minimal Optimization	Tối ưu hóa tuần tự cực tiểu
TF	term frequency	Tần suất từ
k-NN	k-Nearest Neighbor	Thuật toán k-NN
WFST	Weighted Finite State Transducer	Mô hình WFST kết hợp mạng Noron
SW	Stop Words	Loại từ dừng

DANH MỤC CÁC BẢNG

Bảng 2.1: Biểu diễn cá thể trước và sau đột biến.....	33
Bảng 2.2: Độ thích nghi và xác suất của cá thể	34
Bảng 3.1: Vector biểu diễn văn bản 1 và văn bản 2 theo tần suất xuất hiện	43
Bảng 3.2: Vector Boolean biểu diễn văn bản 1.....	44
Bảng 3.3: Các tham số tối ưu tương ứng với mỗi số lượng đặc trưng.....	58
Bảng 3.4: Độ chính xác phân loại trên mỗi lớp và trên toàn bộ	58
Bảng 3.5: Một số từ dừng trong tiếng Việt	66

DANH MỤC CÁC HÌNH VẼ

Hình 1.1: Quá trình khám phá tri thức	7
Hình 1.2: Các đường biên phân loại đối với một lát giềng gần nhất	11
Hình 1.3: Đường biên phân loại học bởi phân loại không tuyến tính.....	12
Hình 1.4: Một hồi quy tuyến tính đơn giản với tập dữ liệu vay nợ	12
Hình 1.5: Một phép phân cụm đơn giản của tập dữ liệu vào 3 cụm.....	14
Hình 1.6: Phân cụm các điểm trong không gian	15
Hình 1.7: Phân cụm các ngôi nhà dựa vào khoảng cách địa lý.....	16
Hình 2.1: Giải quyết vấn đề bằng giải thuật di truyền.	20
Hình 2.2: Sơ đồ giải thuật di truyền.	22
Hình 2.3: Nguyên tắc thực hiện lai ghép chéo	31
Hình 2.4: Nguyên tắc thực hiện lai ghép đa điểm.....	32
Hình 2.5: Ảnh hưởng của quá trình đột biến	32
Hình 2.6: Quá trình chọn lọc cá thể bằng phương pháp bánh xe Roulette	34
Hình 2.7: Quá trình chọn lọc cá thể bằng phương pháp Stochastic Universal Sampling.....	35
Hình 2.8: Mô tả các lân cận của cá thể	35
Hình 2.9: Mô tả các lân cận của cá thể	36
Hình 3.1: Các bước nhỏ trong quá trình đánh chỉ số	40
Hình 3.2: Biểu diễn các vector văn bản trong không gian chỉ có 2 thuật ngữ.....	42
Hình 3.3: Đồ thị biểu diễn văn bản	47
Hình 3.4. Đồ thị đồng hiện của văn bản.....	48
Hình 3.5. Mặt phẳng tách các mẫu dương khỏi các mẫu âm.....	49
Hình 3.8: Minh họa việc khoanh vùng k văn bản gần nhất với $k = 5$	60
Hình 3.9: Mô hình tách từ khoá từ văn bản thô	65
Hình 3.10: Giao diện chương trình chính	70
Hình 3.11: Thực hiện phân tách từng văn bản định dạng txt.....	70
Hình 3.12: Quá trình loại bỏ các stop word có trong từng văn bản.....	70
Hình 3.13:Thực hiện học phân lớp thể thao và pháp luật	71