

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

VÕ DUY THANH

**NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT
HỌC BÁN GIÁM SÁT VÀO LĨNH VỰC
PHÂN LOẠI VĂN BẢN TIẾNG VIỆT**

LUẬN ÁN TIẾN SĨ KỸ THUẬT

Đà Nẵng - 2017

BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

VÕ DUY THANH

**NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT
HỌC BÁN GIÁM SÁT VÀO LĨNH VỰC
PHÂN LOẠI VĂN BẢN TIẾNG VIỆT**

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số : 62 48 01 01

LUẬN ÁN TIẾN SĨ KỸ THUẬT

Người hướng dẫn khoa học:

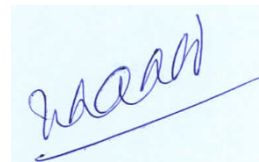
- 1. PGS.TS. Võ Trung Hùng**
- 2. PGS.TS. Đoàn Văn Ban**

Đà Nẵng - 2017

LỜI CAM ĐOAN

Tôi tên là Võ Duy Thanh. Tôi xin cam đoan đây là công trình nghiên cứu do tôi thực hiện. Các nội dung và kết quả nghiên cứu được trình bày trong Luận án là trung thực và chưa được công bố bởi bất kỳ tác giả nào hay trong bất kỳ công trình khoa học nào khác.

Tác giả Luận án



Võ Duy Thanh

MỤC LỤC

| | |
|--|------------|
| LỜI CAM ĐOAN..... | i |
| MỤC LỤC..... | ii |
| DANH MỤC CÁC TỪ VIẾT TẮT..... | vi |
| DANH MỤC HÌNH VẼ | vii |
| DANH MỤC BẢNG..... | ix |
| MỞ ĐẦU | 1 |
| Chương 1. NGHIÊN CỨU TỔNG QUAN | 9 |
| 1.1. Học máy | 9 |
| 1.1.1. Khái niệm | 9 |
| 1.1.2. Ứng dụng của học máy..... | 9 |
| 1.1.3. Các dạng dữ liệu trong học máy..... | 10 |
| 1.2. Các phương pháp học máy | 12 |
| 1.2.1. Học có giám sát | 12 |
| 1.2.2. Học không giám sát..... | 13 |
| 1.2.3. Học bán giám sát | 14 |
| 1.2.4. Học tăng cường | 14 |
| 1.2.5. Học sâu | 15 |
| 1.3. Các phương pháp học bán giám sát..... | 16 |
| 1.3.1. Một số phương pháp học bán giám sát..... | 18 |
| 1.3.2. Thuật toán học có giám sát SVM và bán giám sát SVM..... | 28 |
| 1.3.3. SVM trong phân loại văn bản..... | 33 |

| | |
|--|-----------|
| 1.3.4. Bán giám sát SVM và phân loại trang Web | 34 |
| 1.3.5. Thuật toán phân loại văn bản..... | 35 |
| 1.4. Bài toán phân loại văn bản | 37 |
| 1.4.1. Văn bản..... | 37 |
| 1.4.2. Biểu diễn văn bản bằng véc tơ đặc trưng | 37 |
| 1.4.3. Phân loại văn bản..... | 40 |
| 1.5. Đề xuất nghiên cứu..... | 42 |
| 1.6. Tiểu kết chương..... | 45 |
| Chương 2. XÂY DỰNG KHO DỮ LIỆU | 46 |
| 2.1. Giới thiệu kho dữ liệu phân loại văn bản tiếng Việt | 46 |
| 2.2. Tổng quan về kho dữ liệu..... | 47 |
| 2.2.1. Khái niệm kho dữ liệu | 47 |
| 2.2.2. Đặc điểm của kho dữ liệu | 48 |
| 2.2.3. Mục đích của kho dữ liệu | 49 |
| 2.2.4. Kiến trúc kho dữ liệu..... | 50 |
| 2.3. Phân tích yêu cầu..... | 52 |
| 2.3.1. Xây dựng kho | 52 |
| 2.3.2. Khai thác kho..... | 54 |
| 2.3.3. Cập nhật kho..... | 55 |
| 2.4. Phân tích và đặc tả dữ liệu..... | 55 |
| 2.5. Giải pháp xây dựng kho | 56 |
| 2.5.1. Đề xuất mô hình tổng quát | 56 |
| 2.5.2. Quá trình xây dựng kho dữ liệu..... | 56 |

| | |
|--|-----------|
| 2.5.3. Quy trình của chương trình phân loại văn bản | 57 |
| 2.5.4. Sử dụng thuật toán Naïve Bayes để phân loại văn bản | 62 |
| 2.5.5. Định dạng đầu ra của dữ liệu trong kho | 64 |
| 2.6. Kết quả kho dữ liệu thử nghiệm và đánh giá..... | 65 |
| 2.6.1. Kết quả kho dữ liệu thử nghiệm | 65 |
| 2.6.2. Đánh giá kho dữ liệu | 66 |
| 2.7. Tiểu kết chương..... | 66 |
| Chương 3. PHÂN LOẠI VĂN BẢN DỰA TRÊN MÔ HÌNH CỤ LY TRẮC ĐỊA | 67 |
| 3.1. Mô hình cụ ly trắc địa trên máy véc tơ hỗ trợ..... | 67 |
| 3.1.1. Mô hình cụ ly trắc địa | 67 |
| 3.1.2. Kỹ thuật phân cụm đa dạng sử dụng cụ ly trắc địa | 71 |
| 3.1.3. Phương pháp tính toán cụ ly trắc địa..... | 72 |
| 3.1.4. Hàm nhân trong máy véc tơ hỗ trợ sử dụng cụ ly trắc địa | 74 |
| 3.2. Phương pháp phân loại văn bản dựa trên mô hình cụ ly trắc địa | 75 |
| 3.3. Thực nghiệm phân loại văn bản dựa trên mô hình cụ ly trắc địa | 76 |
| 3.3.1. Phát triển chương trình ứng dụng..... | 76 |
| 3.3.2. Chuẩn bị dữ liệu | 76 |
| 3.3.3. Triển khai chương trình..... | 78 |
| 3.3.4. Kết quả thực nghiệm | 80 |
| 3.4. Tiểu kết chương..... | 84 |
| Chương 4. RÚT GỌN SỐ CHIỀU VEC TƠ DỰA TRÊN ĐỒ THỊ DENDROGRAM..... | 86 |

| | |
|--|------------|
| 4.1. Giới thiệu..... | 86 |
| 4.1.1. Định nghĩa đồ thị Dendrogram..... | 86 |
| 4.1.2. Giải pháp đề xuất..... | 86 |
| 4.2. Xây dựng đồ thị Dendrogram từ dữ liệu Wikipedia..... | 91 |
| 4.2.1. Thuật toán xử lý Wikipedia..... | 91 |
| 4.2.2. Thuật toán xử lý từ điển | 92 |
| 4.2.3. Thuật toán tính toán ma trận P tần số xuất hiện chung | 93 |
| 4.2.4. Thuật toán xây dựng đồ thị Dendrogram | 94 |
| 4.2.5. Triển khai phân cụm..... | 95 |
| 4.2.6. Thử nghiệm | 96 |
| 4.3. Áp dụng véc tơ rút gọn vào phân loại văn bản..... | 102 |
| 4.3.1. Dữ liệu đầu vào | 102 |
| 4.3.2. Kết quả thực nghiệm | 102 |
| 4.4. Tiểu kết chương..... | 107 |
| KẾT LUẬN..... | 108 |
| CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ..... | 111 |
| TÀI LIỆU THAM KHẢO | 112 |

DANH MỤC CÁC TỪ VIẾT TẮT

| | |
|--------|---|
| ANN | Artificial Neural Network (Mạng nơ ron nhân tạo) |
| CRFs | Conditional Random Fields (Trường điều kiện ngẫu nhiên) |
| DM | Data Marts (Kho dữ liệu chủ đề) |
| DWH | Data WareHouse (Kho dữ liệu) |
| GD | Geodesic Distance (Cự li trắc địa) |
| IDF | Inverse Document Frequency (Tần số nghịch đảo văn bản) |
| IID | Independently and Identically Distributed (phân phối độc lập và phân bố tương tự) |
| ISOMAP | Isometric Feature Mapping (Lập bản đồ đặc trưng đều Metric) |
| KNN | K - Nearest Neighbor (K - Láng giềng gần nhất) |
| LDA | Linear Discriminant Analysis (Phân tích biệt thức tuyến tính) |
| MAP | Maximum A Posteriori (Hậu nghiệm cực đại) |
| MDA | Multiple Discriminant Analysis (Phân tích đa biệt thức) |
| MDP | Markov decision process (Quy trình quyết định Markov) |
| MEM | Maximum Entropy Markov Model (Mô hình Markov cực đại hóa entropy) |
| NB | Naïve Bayes |
| NLP | Natural Language Processing (Xử lý ngôn ngữ tự nhiên) |
| OLAP | On line Analytical Processing (Phân tích xử lý trực tuyến) |
| PCA | Principal Component Analysis (Phân tích thành phần chính) |
| POS | Part – of – Speech tagging (Gán nhãn từ loại) |
| SVM | Support vector Machine (Máy véc tơ hỗ trợ) |
| S3VM | Semi-Supervised Support Vector Machine (bán giám sát dựa trên máy véc tơ hỗ trợ) |
| TF | Term frequency (tần suất của từ) |
| RBF | Radial Basis Functions (Hàm cơ sở Radial) |
| VC | Vapnik-Chervonenkis (Khoảng cách VC) |

DANH MỤC HÌNH VẼ

| | | |
|----------|---|----|
| Hình 1.1 | Siêu phẳng cực đại | 21 |
| Hình 1.2 | Biểu diễn trực quan của thiết lập Self-training | 23 |
| Hình 1.3 | Sơ đồ biểu diễn trực quan thiết lập Co-training | 25 |
| Hình 1.4 | Siêu mặt tối ưu và biên | 29 |
| Hình 1.5 | Véc tơ đặc trưng biểu diễn văn bản mẫu | 39 |
| Hình 1.6 | Mô hình tổng quát của hệ thống phân loại văn bản | 42 |
| Hình 1.7 | Mô hình phân lớp văn bản | 43 |
| Hình 1.8 | Mô hình đề xuất phân lớp văn bản sử dụng Self-training | 44 |
| Hình 2.1 | Kiến trúc DWH cơ bản | 50 |
| Hình 2.2 | Kiến trúc DWH với khu vực xử lý | 51 |
| Hình 2.3 | Mô hình đề xuất tổng quát kho dữ liệu | 56 |
| Hình 2.4 | Quy trình phân loại văn bản | 58 |
| Hình 2.5 | Mô hình không gian véc tơ 3 chiều | 61 |
| Hình 3.1 | Cự ly Euclid và cự ly trắc địa | 68 |
| Hình 3.2 | Mô hình đề xuất | 68 |
| Hình 3.3 | Mô hình đề xuất phân loại văn bản dựa trên cự ly trắc địa | 76 |
| Hình 3.4 | Giá trị trung bình và độ lệch chuẩn của tỷ lệ phân loại | 84 |
| Hình 4.1 | Đồ thị Dendrogram | 86 |
| Hình 4.2 | Ví dụ về đồ thị Dendrogram | 90 |
| Hình 4.3 | Lưu đồ thuật toán xử lý tập tin dữ liệu Wikipedia | 92 |

| | | |
|-----------|---|-----|
| Hình 4.4 | Sơ đồ thuật toán xử lý từ điển | 93 |
| Hình 4.5 | Ví dụ cho việc cắt đồ thị Dendrogram, kết quả nhận được 3 cụm | 97 |
| Hình 4.6 | Số lượng cặp từ theo tần số xuất hiện chung | 98 |
| Hình 4.7 | Số lượng nhóm phụ thuộc phân cụm trên đồ thị Dendrogram | 99 |
| Hình 4.8 | Kết quả phân cụm với Dendrogram | 99 |
| Hình 4.9 | Một ví dụ khác thể hiện những từ liên quan đến âm nhạc | 100 |
| Hình 4.10 | Một ví dụ đồ thị Dendrogram cho các từ | 100 |
| Hình 4.11 | Ví dụ đồ thị Dendrogram cho các từ thuộc chủ đề y học | 101 |
| Hình 4.12 | Dung lượng lưu trữ véc tơ phụ thuộc vào số lượng từ | 104 |
| Hình 4.13 | Đồ thị thể hiện thời gian gán nhãn của 5 lần huấn luyện | 105 |
| Hình 4.14 | Thời gian phân loại văn bản trung bình của 5 lần huấn luyện | 105 |
| Hình 4.15 | Đồ thị thể hiện độ phân loại của 5 lần HL theo tỷ lệ phân cụm | 106 |
| Hình 4.16 | Đồ thị thể hiện sự thay đổi của kết quả theo tỷ lệ phân loại | 107 |

DANH MỤC BẢNG

| | | |
|-----------|--|-----|
| Bảng 2.1 | Dữ liệu thô tải về | 53 |
| Bảng 2.2 | Dữ liệu huấn luyện | 63 |
| Bảng 2.3 | Kết quả kho dữ liệu thử nghiệm | 65 |
| Bảng 3.1 | Thống kê số tập tin trong kho dữ liệu | 77 |
| Bảng 3.2 | Kết quả phân loại lần 1 sử dụng SVM | 80 |
| Bảng 3.3 | Kết quả phân loại lần 1 sử dụng SVM với mô hình cự ly trắc địa | 81 |
| Bảng 3.4 | Kết quả phân loại lần 2 sử dụng SVM | 81 |
| Bảng 3.5 | Kết quả phân loại lần 2 sử dụng SVM với mô hình cự ly trắc địa | 81 |
| Bảng 3.6 | Kết quả phân loại lần 3 sử dụng SVM | 82 |
| Bảng 3.7 | Kết quả phân loại lần 3 sử dụng SVM với mô hình cự ly trắc địa | 82 |
| Bảng 3.8 | Kết quả phân loại lần 4 sử dụng SVM | 82 |
| Bảng 3.9 | Kết quả phân loại lần 4 sử dụng SVM với mô hình cự ly trắc địa | 83 |
| Bảng 3.10 | Kết quả phân loại lần 5 sử dụng SVM | 83 |
| Bảng 3.11 | Kết quả phân loại lần 5 sử dụng SVM với mô hình cự ly trắc địa | 83 |
| Bảng 4.1 | Dữ liệu huấn luyện, kiểm thử | 102 |

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, cùng với sự phát triển chung của khoa học kỹ thuật và công nghệ thông tin đã đem đến cho con người khả năng tiếp cận với tri thức khoa học một cách nhanh chóng, cụ thể như: thư viện điện tử, cổng thông tin điện tử, báo mạng, các ứng dụng tìm kiếm, ..., đã giúp con người thuận tiện hơn trong việc trao đổi, cập nhật thông tin trên toàn cầu thông qua mạng Internet.

Tuy nhiên, với quá trình trao đổi và cập nhật thông tin một cách nhanh chóng, khối lượng thông tin được lưu trữ (dưới dạng tài liệu số) ngày càng tăng nên gặp phải khó khăn trong việc sắp xếp phân loại. Phân loại văn bản là một bước quan trọng nhằm giúp cho việc xử lý hiệu quả hơn. Nếu thực hiện quá trình phân loại bằng thủ công sẽ tốn nhiều thời gian và chi phí. Vì vậy, thực hiện việc phân loại tự động văn bản số hiện nay là một vấn đề cấp thiết.

Để phân loại văn bản số, nhiều phương pháp phân loại đã được đề xuất dựa trên mô hình không gian véc tơ. Từ mô hình này, các mô hình xác suất được xây dựng thông qua việc học máy nhằm mục đích phân loại văn bản tự động. Máy véc tơ hỗ trợ (SVM) là một trong những công cụ phân loại tự động hữu hiệu, là dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó SVM là một thuật toán phân loại nhị phân và được các nhà nghiên cứu trong lĩnh vực học máy đánh giá cao. Tuy nhiên, để áp dụng vào phân loại văn bản số tiếng Việt, việc sử dụng mô hình không gian véc tơ thường không đem lại hiệu quả cao bởi ngôn ngữ tiếng Việt khá phức tạp. Rất nhiều từ đồng âm khác nghĩa và cũng có rất nhiều từ khi so sánh trên ký tự thì khác nhau hoàn toàn nhưng lại có cùng ý nghĩa khi phân loại. Điều này dẫn đến trong không gian véc tơ, hai văn bản chứa các từ đồng âm khác nghĩa sẽ có khoảng cách nhỏ cho dù nội dung hoàn toàn khác nhau. Điều này dẫn đến việc phân loại không thành công.

Mặt khác, khi phát triển các ứng dụng dựa trên học máy thì kho dữ liệu huấn luyện đóng một vai trò quan trọng. Khối lượng và chất lượng dữ liệu sử dụng để

huấn luyện hệ thống nhằm tạo ra một mô hình tốt có ý nghĩa vô cùng quan trọng, quyết định đến chất lượng của hệ thống. Tuy nhiên, đối với tiếng Việt, các kho dữ liệu phục vụ cho việc phát triển các ứng dụng phân loại văn bản dựa trên học máy chưa có nhiều. Vì vậy, sử dụng phương pháp học bán giám sát để không cần lượng dữ liệu lớn đã xác định nhãn (đã xác định tên loại dữ liệu) khi phân loại là phù hợp với các ngôn ngữ mà kho ngữ liệu còn hạn chế.

Phân loại văn bản tự động là gán các nhãn phân loại lên một văn bản mới dựa trên mức độ tương tự của văn bản đó so với các văn bản đã được gán nhãn trong tập huấn luyện. Nhiều kỹ thuật máy học và khai phá dữ liệu đã được áp dụng vào bài toán phân loại văn bản, chẳng hạn: phương pháp quyết định dựa vào Naive Bayes, cây quyết định, k-láng giềng gần nhất, mạng nơ ron, ... Trong những nghiên cứu gần đây, phương pháp phân loại văn bản sử dụng Máy véc tơ hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực phân loại. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn.

Trong những năm gần đây vấn đề phân loại văn bản tiếng Việt được nhiều cơ sở nghiên cứu trong cả nước quan tâm. Một số công trình nghiên cứu cũng đạt được những kết quả khả quan. Các hướng tiếp cận bài toán phân loại văn bản đã được nghiên cứu bao gồm: hướng tiếp cận bài toán phân loại bằng lý thuyết đồ thị, cách tiếp cận sử dụng lý thuyết tập thô, cách tiếp cận thống kê, cách tiếp cận sử dụng phương pháp học không giám sát và đánh chỉ mục. Nhìn chung, những cách tiếp cận này đều cho kết quả chấp nhận được. Tuy nhiên SVM chưa được áp dụng một cách có hiệu quả vào phân loại văn bản tiếng Việt. Vì vậy với mục đích xây dựng mô hình và cải tiến phương pháp máy véc tơ hỗ trợ SVM để nâng cao hiệu quả phân loại văn bản tiếng Việt là một công việc cấp thiết.

Chính vì vậy, trong luận án này, tập trung nghiên cứu các phương pháp mới nhằm phân loại văn bản tiếng Việt hiệu quả hơn dựa trên kỹ thuật học bán giám sát.

2. Tổng quan tình hình nghiên cứu

Trong khoa học máy tính, hiện nay có một số thuật toán phân loại văn bản thực

hiện đạt kết quả khi được xây dựng trên các tập mẫu huấn luyện lớn. Nhưng trong thực tế thì điều kiện này hết sức khó vì các tập mẫu huấn luyện thường được gán nhãn bởi con người nên đòi hỏi phải mất nhiều thời gian và công sức. Trong khi đó dữ liệu chưa gán nhãn thì rất phong phú. Đối với các bài toán phân loại văn bản nói trên trở nên phổ biến hơn. Chính vậy, việc xem xét các thuật toán phân loại văn bản chỉ cần ít dữ liệu được gán nhãn, sử dụng nguồn dữ liệu lớn chưa gán nhãn đã nhận được sự quan tâm của nhiều nhà khoa học trong và ngoài nước. Học bán giám sát là một lớp kỹ thuật học máy kết hợp việc sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn trong huấn luyện. Số lượng của dữ liệu gán nhãn thường là rất ít so với số lượng của dữ liệu chưa gán nhãn, bởi vì việc gán nhãn cho các mục dữ liệu đòi hỏi chi phí về thời gian rất lớn. Nhiều nhà nghiên cứu trong lĩnh vực học máy đã thấy rằng dữ liệu không có nhãn, khi dùng kết hợp với một số lượng nhỏ dữ liệu có nhãn, có thể đưa ra được những cải tiến đáng kể trong việc học chính xác.

a. Tình hình nghiên cứu trên thế giới

Trước năm 2005, đã có một số công trình nghiên cứu và đã đề xuất một số thuật toán phục vụ học bán giám sát áp dụng giới hạn trong một số lĩnh vực [10], [14], [15], [69]. Nhưng các nghiên cứu này chưa đầy đủ và chưa khái quát được bài toán học bán giám sát. Trong những năm gần đây, đã có nhiều công trình nghiên cứu về tổng quan học bán giám sát như [5], [11], [12], [83], [93], [95]. Một số nghiên cứu khác tập trung chủ yếu trên: Học bán giám sát dựa trên máy véc tơ hỗ trợ [8], [30], [32], [33], [49], [71], [79], [94] hoặc học bán giám sát với sự trợ giúp cây Bayes [10], [37]; Phân loại bán giám sát với quá trình xử lý hồi quy Gauss [57]; Phân nhóm dựa trên khoảng cách đo đặc trên nhiều mặt khác nhau Gaussian cho việc tự động phân loại các gián đoạn [4], [28]. Đây là những phương pháp có hiệu quả và được áp dụng trong thực tế.

b. Tình hình nghiên cứu trong nước

Việc nghiên cứu ứng dụng kỹ thuật học bán giám sát vào các bài toán trong xử lý ngôn ngữ tự nhiên như phân loại văn bản, dịch thống kê, hỏi đáp tự động, ... là rất phù hợp. Tuy nhiên, hiện tại các nghiên cứu trong nước chủ yếu sử dụng kỹ thuật

n-grams [72], [73] trong việc giải quyết các bài toán này mà chưa ứng dụng nhiều kỹ thuật học bán giám sát. Ở Việt Nam, kỹ thuật học bán giám sát được tập trung nghiên cứu trong lĩnh vực tin sinh học (phân loại gen, protein) và chưa được phổ biến rộng rãi.

Trong lĩnh vực xử lý văn bản tiếng Việt, một số kết quả nghiên cứu như: Gán nhãn từ loại; Tách từ tiếng Việt [22]; Phân loại văn bản tiếng Việt dựa trên tập thô [24], [59]; Rút trích và tóm tắt nội dung trang Web tiếng Việt [35]; Phân loại nội dung tài liệu web tiếng Việt [82]; Nghiên cứu ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét đến ngữ cảnh [87]; Nghiên cứu gom cụm đồ thị và ứng dụng vào việc rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận [26]; Nghiên cứu độ tương đồng ngữ nghĩa giữa hai câu và áp dụng vào bài toán sử dụng tóm tắt văn bản để đánh giá chất lượng phân cụm dữ liệu trên máy tìm kiếm VNSEN [78], [81]; Nghiên cứu ứng dụng tập phổ biến tối đại vào bài toán tóm tắt văn bản hỗ trợ phân lớp văn bản dựa trên SVM [29], [80]; Nghiên cứu ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét ngữ nghĩa [25]; Phân loại văn bản tiếng Việt dựa trên mô hình chủ đề [9]; Phương án xây dựng tập mẫu cho bài toán phân lớp văn bản tiếng Việt, nguyên lý, giải thuật, thử nghiệm và đánh giá kết quả [23].

3. Mục tiêu nghiên cứu

Mục tiêu chung của đề tài là nghiên cứu ứng dụng kỹ thuật học bán giám sát vào phân loại văn bản tiếng Việt.

Mục tiêu cụ thể như sau:

- Hệ thống hoá và phân tích đầy đủ các vấn đề liên quan đến phân loại văn bản gồm: các mô hình phân loại, các phương pháp, kỹ thuật học có giám sát, học không có giám sát, học bán giám sát và học tăng cường;
- Đề xuất được các giải pháp nhằm cải tiến phương pháp phân loại văn bản tiếng Việt để cải thiện chất lượng phân loại cả về kết quả lẫn tốc độ xử lý;
- Tạo ra được kho dữ liệu và các công cụ phục vụ phân loại văn bản tiếng Việt.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài gồm:

- Kỹ thuật học bán giám sát;
- Các thuật toán phân loại, phân cụm dữ liệu trong cơ sở dữ liệu có cấu trúc và bán cấu trúc, phương pháp tách từ, tách câu trong các loại văn bản.
- Một số hệ thống phân loại văn bản hiện có.

Chúng tôi giới hạn phạm vi nghiên cứu trong luận án này gồm:

- Chỉ nghiên cứu một số kỹ thuật học bán giám sát dựa trên SVM, phân loại bán giám sát với quá trình xử lý hồi quy Gauss, phân loại học bán giám sát sử dụng hàm nhân, kỹ thuật nhân đồ thị sử dụng phép biến đổi phổ, phương pháp cự ly trắc địa kết hợp với máy véc tơ hỗ trợ, thuật toán tìm đường đi ngắn nhất trong mô hình cự ly trắc địa để xây dựng ma trận nhân trong SVM, phương pháp rút gọn số chiều véc tơ, gom cụm từ;
- Chỉ tập trung cho phân loại văn bản tiếng Việt.

5. Nội dung nghiên cứu

Để đạt được mục tiêu đề ra, nội dung nghiên cứu của luận án gồm:

- Xác định một hàm hoặc một phương thức cho phép phân loại hiệu quả các lớp dữ liệu (thường là hai lớp);
- Đưa ra dự đoán lớp cho những dữ liệu chưa biết nhãn;
- Nghiên cứu sự ảnh hưởng của số lượng dữ liệu chưa biết nhãn đến kết quả của thuật toán;
- Xây dựng các phần mềm thử nghiệm phân loại văn bản tiếng Việt.

6. Phương pháp nghiên cứu

- **Phương pháp tài liệu:** Nghiên cứu các tài liệu có liên quan đến các nội dung nghiên cứu như: học máy, học bán giám sát, phân loại văn bản, phân loại văn bản tiếng Việt, cự ly trắc địa, đồ thị Dendrogram.

- **Phương pháp thực nghiệm:** Nghiên cứu đánh giá thực nghiệm từng mô hình, phương pháp phân loại văn bản, từ đó so sánh, đánh giá với mô hình, phương pháp được đề xuất. Xây dựng chương trình phân loại văn bản, chương trình rút gọn số chiều véc tơ, gom cụm từ.

- **Phương pháp chuyên gia:** Lấy ý kiến các chuyên gia về phương pháp lấy ý kiến, các giải pháp đề xuất và khảo sát ý kiến của người sử dụng.

7. Đóng góp chính của luận án

Luận án tiến sĩ này có những đóng góp chính như sau:

1) *Đề xuất được một giải pháp mới trong phân loại văn bản dựa trên mô hình trắc địa và lý thuyết đồ thị.* Tất cả các nghiên cứu trước đây về phân loại văn bản đều sử dụng khoảng cách Euclid để đo mức độ gần nhau giữa các văn bản khi thực hiện gom cụm, xây dựng mô hình ngôn ngữ hoặc phân loại văn bản. Về mặt hình học, khoảng cách Euclid dựa trên đo khoảng cách theo đường chim bay (nối 2 điểm mà không tính đến mặt cong phân bố các điểm) nên chưa thể hiện chính xác mức độ gần nhau thực tế của các điểm. Mô hình trắc địa sử dụng hệ tương quan ngắn nhất (trong phân loại văn bản là mức độ gần nhau giữa các văn bản) để tính khoảng cách giữa hai điểm, khoảng cách tính trên mặt cong phân bố các điểm. Khoảng cách này được gọi là cự ly trắc địa và khác với khoảng cách Euclid. Về mặt mô hình toán học, khi xây dựng được một mô hình đường trắc địa hợp lý và tính khoảng cách các điểm dựa trên cự ly trắc địa thì việc phân loại văn bản tự động sẽ chính xác hơn. Vấn đề khó khăn nhất khi áp dụng mô hình trắc địa là việc tính toán phức tạp hơn trên không gian Euclid và làm thế nào để xác định khoảng cách giữa tất cả các điểm phân bố trên các mặt cong của mô hình trắc địa. Vấn đề này được luận án giải quyết thông qua việc áp dụng lý thuyết đồ thị. Mỗi một điểm trên mô hình trắc địa được xem như một đỉnh đồ thị và luận án xác lập một đường đi từ một đỉnh đến các đỉnh khác theo thứ tự khoảng cách giữa chúng. Cách tính này dẫn đến một ưu điểm nổi bật của mô hình trắc địa kết hợp với lý thuyết đồ thị là cho phép phân loại văn bản (thực chất là phân chia các điểm/đỉnh đồ thị) thành nhiều loại/nhóm thay vì chỉ phân ra hai loại (dựa trên phân lớp nhị phân) như các phương pháp cũ dựa trên cự ly Euclid. Giải pháp mà luận án đề xuất đã được kiểm chứng và cho kết quả phân loại tốt hơn so với các phương pháp sử dụng cự ly Euclid. Ngoài ra, giải pháp này có thể được áp dụng sang các ứng dụng khác mà trong đó có tính đến yếu tố khoảng cách giữa các điểm trong không gian nhiều chiều. Kết quả có một công trình công bố tại Hội thảo quốc tế ISDA 2014, IEEJ catalog, ISSN: 2150-7996.

2) Đề xuất được một giải pháp mới để rút gọn số chiều của véc tơ biểu diễn văn bản dựa trên đồ thị Dendrogram. Phương pháp biểu diễn văn bản được sử dụng phổ biến hiện nay là sử dụng véc tơ, trong đó mỗi từ (hoặc tần số xuất hiện từ đó trong văn bản) là một phần tử của véc tơ. Vì vậy, số chiều của véc tơ biểu diễn văn bản là rất lớn. Do số chiều véc tơ rất lớn nên nếu áp dụng cự ly đường trắc địa sẽ có ảnh hưởng lớn đến tốc độ xử lý. Để giải quyết vấn đề này, luận án đề xuất giải pháp tiếp theo là rút gọn số chiều véc tơ bằng phương pháp phân cụm các từ dựa trên đồ thị Dendrogram. Ý tưởng của đề xuất này là sử dụng Từ điển Bách khoa toàn thư Wikipedia và đồ thị Dendrogram nhằm mục đích phân cụm từ tiếng Việt dựa trên tần suất xuất hiện đồng thời của các từ trên các văn bản và trên cơ sở đó rút gọn số chiều véc tơ thuộc tính của văn bản (hợp nhất các phần tử gần nhau trên đồ thị Dendrogram). Việc áp dụng không gian véc tơ đã được rút gọn sẽ giúp giảm số chiều véc tơ biểu diễn văn bản và qua đó tiết kiệm thời gian phân loại văn bản tiếng Việt mà vẫn đảm bảo tỉ lệ phân loại đúng ở mức cao. Giải pháp rút gọn số chiều véc tơ này không phải chỉ áp dụng cho phân loại văn bản mà có thể áp dụng cho tất cả các ứng dụng khác có biểu diễn văn bản bằng véc tơ như xác định mức độ giống nhau giữa các văn bản, nhận dạng ngôn ngữ, ... Kết quả có một công trình công bố tại Hội thảo quốc tế ACIS 2014, ISBN: 978-4-88686-7.

Bên cạnh hai đóng góp chính trên, luận án cũng đã xây dựng được kho dữ liệu phục vụ phân loại văn bản tiếng Việt. Đóng góp này không có nhiều ý nghĩa về mặt khoa học nhưng có ý nghĩa thực tiễn rất cao vì kho dữ liệu ngôn ngữ là cơ sở để thực hiện các nghiên cứu thực nghiệm liên quan đến xử lý ngôn ngữ. Đối với các ngôn ngữ như tiếng Anh, Pháp, Tây Ban Nha, Nhật, ... người ta đã xây dựng các kho dữ liệu ngôn ngữ (là các văn bản trong một ngôn ngữ cụ thể đã được tiền xử lý như gán nhãn, tách từ, gán nhãn từ loại, ...) để phục vụ triển khai các thử nghiệm và đánh giá kết quả. Tuy nhiên, đối với tiếng Việt, người ta chưa xây dựng hoặc chưa công bố các kho dữ liệu ngôn ngữ như vậy để cộng đồng các nhà khoa học sử dụng. Trong luận án này, đã tạo ra một kho dữ liệu với số lượng 5027 văn bản đã được tiền xử lý và gán nhãn với 5 chủ đề khác nhau. Luận án đã sử dụng kho dữ liệu này

cho tất cả các thử nghiệm về phân loại văn bản và đánh giá kết quả đạt được cho các phương pháp khác nhau trên cùng một tập dữ liệu.

8. Bố cục của luận án

Nội dung chính của luận án được trình bày trong 4 chương:

Chương 1: Nghiên cứu tổng quan

Trình bày các kết quả nghiên cứu tổng quan liên quan đến học máy, các phương pháp học máy, phân lớp dữ liệu và phân loại văn bản, nghiên cứu ứng dụng kỹ thuật học bán giám sát vào lĩnh vực phân loại văn bản tiếng Việt. Trên cơ sở nghiên cứu, đánh giá các vấn đề còn tồn tại, đề xuất những nội dung nghiên cứu trình bày trong các chương tiếp theo.

Chương 2. Xây dựng kho dữ liệu

Trình bày các vấn đề cơ bản về kho dữ liệu như: giới thiệu về kho dữ liệu, phân tích, đặc tả dữ liệu, đưa ra giải pháp xây dựng kho, phân tích thiết kế kho dữ liệu, đồng thời thiết kế cơ sở dữ liệu cho kho để phân loại văn bản tiếng Việt. Mục đích xây dựng kho dữ liệu ở chương này là để phục vụ huấn luyện và kiểm thử cho thực nghiệm ở các chương sau.

Chương 3: Phân loại dựa trên mô hình cự ly đường trắc địa

Trình bày kết quả nghiên cứu phân loại văn bản dựa trên mô hình cự ly trắc địa. Nội dung chủ yếu liên quan đến các khái niệm mô hình cự ly trắc địa, xây dựng mô hình cự ly trắc địa dựa trên máy véc tơ hỗ trợ, thuật toán tính cự ly trắc địa và ứng dụng để xây dựng phần mềm phân loại văn bản tiếng Việt.

Chương 4: Rút gọn số chiều véc tơ dựa trên đồ thị Dendrogram

Giới thiệu kết quả nghiên cứu về đồ thị Dendrogram, kết hợp mô hình đồ thị Dendrogram và dữ liệu Từ điển Bách khoa toàn thư Wikipedia để thực hiện phân cụm từ và áp dụng để rút gọn số chiều của véc tơ trong quá trình phân loại văn bản tiếng Việt.

Chương 1. NGHIÊN CỨU TỔNG QUAN

Trong chương này giới thiệu các kết quả nghiên cứu tổng quan liên quan đến học máy, phân loại văn bản và đề xuất các vấn đề nghiên cứu. Nội dung chính trình bày các khái niệm, phương pháp và kỹ thuật sử dụng trong học máy. Tiếp theo, trình bày về phân loại văn bản, cách biểu diễn văn bản bằng véc tơ và các phương pháp phân loại văn bản đang sử dụng phổ biến hiện nay. Trên cơ sở đó, đề xuất những vấn đề nghiên cứu trong các chương tiếp theo.

1.1. Học máy

1.1.1. Khái niệm

Học máy [52] là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc phát triển các kỹ thuật cho phép các máy tính có thể "học" [69], [88]. Học máy là lĩnh vực liên quan nhiều đến thống kê do cả hai lĩnh vực đều tập trung vào việc nghiên cứu để phân tích dữ liệu. Tuy nhiên, học máy có sự khác biệt với thống kê, học máy tập trung vào nghiên cứu sự phức tạp của các giải thuật trong quá trình tính toán, xử lý dữ liệu. Trên thực tế, có nhiều bài toán suy luận được xếp loại là bài toán NP- khó, vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ để có thể xử lý được lớp các bài toán nhị phân một cách tổng quát nhất.

Trên cơ sở đó, người ta phân loại học máy theo hai dạng:

- Học máy dựa trên quy nạp: Máy học phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụng được nguồn dữ liệu rất nhiều, sẵn có.
- Học máy dựa trên suy diễn: Máy học phân biệt các khái niệm dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ học máy.

1.1.2. Ứng dụng của học máy

Chúng ta đều biết khái niệm về việc xếp hạng trang web. Đó là quá trình gửi một câu truy vấn đến một công cụ tìm kiếm, sau đó sẽ được trả một danh sách các trang web có liên quan đến câu đã truy vấn theo một thứ tự nhất định. Để thực hiện được

chức năng này, một công cụ tìm kiếm phải “Biết” được những kiến thức về các trang phù hợp hay liên quan với truy vấn. Ví dụ về kết quả sắp xếp của các trang web được truy vấn bởi từ khóa “Học máy”. Những kiến thức như vậy có thể được tổng hợp từ nhiều nguồn khác nhau như: cấu trúc liên kết, nội dung hay tần số sử dụng của các trang web. Ngoài ra cũng có thể được kết hợp với cách xếp hạng thủ công để đưa ra kết quả xếp hạng tự động từ một câu truy vấn.

Tuy nhiên học máy chính là một sự lựa chọn tốt hơn cả cho việc thiết kế một công cụ tìm kiếm.

Việc sử dụng lọc cộng tác trong hệ thống tư vấn như ở các trang Amazon hay Netflix nhằm khuyến khích người sử dụng mua sách hay thuê phim cũng là một ví dụ minh họa cho một ứng dụng của học máy. Tương tự như việc xếp hạng trang web, chúng ta cần một bảng sắp xếp danh sách các ấn phẩm (sách hoặc phim) theo sở thích của khách hàng. Việc giải quyết những vấn đề trên một cách tự động là hết sức cấp bách nhằm mục đích tránh phỏng đoán sai cũng như tiết kiệm thời gian.

Một ứng dụng khác của học máy đó chính là dịch tự động cho văn bản. Thông thường, để dịch văn bản từ thứ tiếng này sang thứ tiếng khác, chúng ta cần phải hiểu rõ tất cả những quy tắc được quy định bởi các chuyên gia (nhà ngôn ngữ học) am hiểu cả hai ngôn ngữ mà chúng ta cần dịch. Đây là một việc làm khá phức tạp và tốn nhiều chi phí bởi vì chúng ta không thể thu thập được hết tất cả các quy tắc, cũng như không phải bất kì văn bản nào cũng tuân theo một quy tắc nhất định. Thay vào đó chúng ta có thể sử dụng một số bản dịch mẫu để học một cách tự động phương pháp dịch giữa hai ngôn ngữ. Nói cách khác, học máy chính là một công cụ tốt nhất để xây dựng một hệ thống phiên dịch tự động.

Có rất nhiều ứng dụng học máy khác, như sử dụng nhận dạng khuôn mặt để phục vụ các hệ thống điều khiển tự động hay bảo mật. Hệ thống cần phải học và tìm ra những đặc trưng tốt nhất trong việc nhận dạng mặt người. Đó chính là nhiệm vụ của học máy.

1.1.3. Các dạng dữ liệu trong học máy

Khi nói tới học máy, chúng ta không thể bỏ qua việc định dạng các thể loại của

dữ liệu. Việc định dạng dữ liệu giúp chúng ta có thể tìm ra những hướng giải quyết vấn đề mới nhờ vào việc sử dụng những kỹ thuật có chung kiểu dữ liệu. Ví dụ, trong xử lý ngôn ngữ tự nhiên, dữ liệu thường là những chuỗi ký tự. Sau đây là một số kiểu dữ liệu thường được sử dụng trong học máy.

- **Véc tơ:** là kiểu dữ liệu cơ bản nhất trong học máy. Nó thể hiện các đặc tính của một sự vật, sự việc trong một môi trường cụ thể. Mỗi văn bản được biểu diễn thành một véc tơ đặc trưng, mỗi thành phần là một từ khóa trong tập văn bản gốc và được gán một trọng số xác định dựa trên tần suất xuất hiện của các từ hay cụm từ trong văn bản, ...

- **Danh sách:** là danh sách các dữ liệu hoặc đặc tính được liệt kê của sự vật, sự việc. Khác với véc tơ đặc trưng, danh sách không nhất thiết phải liệt kê đầy đủ các thông số của đặc tính. Ví dụ, một bác sỹ không nhất thiết phải thực hiện đầy đủ tất cả các bước trong quy trình khám mà vẫn có thể xác định được bệnh nhân có khỏe mạnh hay không. Trong trường hợp này, chúng ta có thể sử dụng danh sách để tiết kiệm bộ nhớ máy tính.

- **Tập hợp:** là một tập hợp các dữ liệu, trong đó thứ tự của các phần tử dữ liệu không quan trọng, không ảnh hưởng đến kết quả của các thuật toán trong học máy và các phần tử thường không ảnh hưởng lẫn nhau.

- **Ma trận:** thường như là một bảng dữ liệu 2 chiều trong đó dữ liệu có thể được xác định khi và chỉ khi biết chính xác số hàng và số cột của dữ liệu đó. Ví dụ, biểu diễn văn bản, ta có thể chia văn bản thành n đoạn, mỗi đoạn biểu diễn bằng một véc tơ m đặc trưng, ta có n véc tơ. Xếp xếp các véc tơ này thành n cột và n hàng liên tiếp thành một ma trận.

- **Hình ảnh:** hình ảnh được hiểu như một mảng hai chiều, trong đó dữ liệu là các con số như cường độ ánh sáng, màu sắc, điểm ảnh (pixel) của ảnh được số hóa.

- **Video:** là một danh sách các hình ảnh chúng được biểu diễn bởi một mảng 3 chiều để thuận lợi trong việc tính toán, xử lý.

- **Cây hoặc đồ thị:** thể hiện các mối quan hệ giữa các dữ liệu với nhau thông qua các nút của cây hoặc các đỉnh của đồ thị. Biểu diễn văn bản bằng đồ thị, mỗi đồ thị

là một văn bản. Đỉnh của đồ thị có thể là câu, hoặc từ, hoặc kết hợp câu và từ. Cạnh nối giữa các đỉnh là vô hướng hoặc có hướng, thể hiện mối quan hệ trong đồ thị. Nhãn của đỉnh thường là tần số xuất hiện của đỉnh, còn nhãn của cạnh là tên mối liên kết khái niệm giữa hai đỉnh, hay tần số xuất hiện chung của hai đỉnh trong một phạm vi nào đó, hay tên vùng mà đỉnh xuất hiện.

- **Xâu ký tự:** là một chuỗi các ký tự. Thường được sử dụng trong xử lý ngôn ngữ tự nhiên và tin sinh học. Trong phân loại văn bản, phân tách chuỗi ký tự thành chuỗi các từ. Giai đoạn này có thể đơn giản hay phức tạp tùy theo ngôn ngữ và quan niệm về đơn vị từ vựng.

- **Cấu trúc sưu tập (Collection):** là cấu trúc có thể hỗn hợp của nhiều kiểu dữ liệu khác nhau khi thể hiện một đối tượng nào đó.

1.2. Các phương pháp học máy

1.2.1. Học có giám sát

Trong học có giám sát, tập dữ liệu huấn luyện gồm các mẫu đã gán nhãn hoặc có giá trị hàm đích đi kèm. Học có giám sát có thể giúp chúng ta phân loại một cách chính xác và phù hợp với mục đích của từng bài toán phân loại [12], [51], [89], [92]. Tuy nhiên để gán nhãn cho các dữ liệu trong tập huấn luyện cần đòi hỏi nhiều thời gian và chi phí cho việc gán nhãn. Học có giám sát là phương thức xây dựng mô hình phân loại được thể hiện thông qua các thành phần:

Tập huấn luyện: $L = \{ (x_1, y_1), \dots, (x_n, y_n) \}$, trong đó $x_i \in \mathbb{R}^d$ là véc tơ d chiều thể hiện các đặc tính của đối tượng thứ i và có nhãn là y_i .

Mục đích: gán nhãn cho các đối tượng $x \in X$ không biết trước.

Cho trước một mẫu bao gồm các cặp đối tượng - nhãn (x_i, y_i) , cần tìm ra mối quan hệ giữa các đối tượng và các nhãn. Mục đích là học để xác định được một phép ánh xạ từ X tới Y , khi cho trước một tập huấn luyện gồm các cặp (x_i, y_i) , trong đó $y_i \in Y$ gọi là các nhãn hoặc đích của các mẫu x_i .

Các bước xử lý: Để giải quyết một bài toán nào đó của học có giám sát, người ta phải xem xét nhiều bước khác nhau:

- **Xác định loại của các mẫu huấn luyện:** Trước khi làm bất cứ điều gì, người làm nhiệm vụ phân lớp nên quyết định loại dữ liệu nào sẽ được sử dụng làm mẫu. Chẳng hạn đó có thể là một kí tự viết tay đơn lẻ, tập hợp các từ viết tay, hay tập hợp một dòng chữ viết tay.

- **Thu thập tập huấn luyện:** Tập huấn luyện cần có độ bao phủ để chứa tất cả các đặc trưng của đối tượng. Vì thế, một tập dữ liệu mô tả thông tin các đối tượng đầu vào được thu thập và đầu ra tương ứng được thu thập, hoặc từ các chuyên gia hoặc từ việc đo đạc tính toán.

- **Xác định cách biểu diễn các đặc trưng đầu vào:** Sự chính xác của hàm chức năng phụ thuộc lớn vào cách các đối tượng đầu vào được biểu diễn. Thông thường, đối tượng đầu vào được chuyển đổi thành một véc tơ đặc trưng, chứa một số các đặc trưng nhằm mô tả cho đối tượng đó. Số lượng các đặc trưng không nên quá lớn, do sự bùng nổ tổ hợp, nhưng phải đủ lớn để dự đoán chính xác đầu ra. Xác định cấu trúc của hàm chức năng cần tìm và giải thuật học tương ứng.

- **Hoàn thiện thiết kế:** Các tham số của giải thuật học có thể được điều chỉnh bằng cách tối ưu hoá hiệu năng trên một tập con (gọi là tập kiểm chứng – validation set) của tập huấn luyện, hay thông qua kiểm chứng chéo (cross-validation).

Sau khi học và điều chỉnh tham số, hiệu năng của giải thuật có thể được đo đạc trên một tập kiểm thử độc lập với tập huấn luyện.

1.2.2. Học không giám sát

Là phương thức phân cụm mà tập huấn luyện không được gán nhãn trước [14]. Trong đó. Học không giám sát thì không đòi hỏi chi phí cho việc gán nhãn.

- **Tập huấn luyện:** $L = \{x_1, x_2, \dots, x_n\}$, trong đó $x_i \in R^d$ là véc tơ d chiều thể hiện các đặc tính của đối tượng thứ i .

- **Mục đích:** gán nhãn cho đối tượng x .

Biểu diễn toán học của phương pháp này như sau:

Đặt $X = (x_1, x_2, \dots, x_n)$ là tập hợp gồm n mẫu. Mục đích của học không giám sát là tìm ra một cấu trúc thông minh trên tập dữ liệu đó.

Từ đó, **học bán giám sát** có thể được xem là:

- Học giám sát cộng thêm dữ liệu chưa gán nhãn.
- Học không giám sát cộng thêm dữ liệu đã gán nhãn.

Học bán giám sát chính là cách học kết hợp sử dụng thông tin chứa trong cả dữ liệu chưa gán nhãn và tập dữ liệu huấn luyện. Các thuật toán học bán giám sát có nhiệm vụ chính là mở rộng tập các dữ liệu đã gán nhãn ban đầu. Hiệu quả của thuật toán phụ thuộc vào chất lượng của các mẫu đã gán nhãn được thêm vào ở mỗi vòng lặp và được đánh giá dựa trên hai tiêu chí:

- Các mẫu được thêm vào phải được gán nhãn một cách chính xác.
- Các mẫu được thêm vào phải mang lại thông tin hữu ích cho bộ phân lớp (hoặc dữ liệu huấn luyện).

1.2.3. Học bán giám sát

Học bán giám sát là kết hợp việc học cả dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn. Từ số lượng lớn các dữ liệu chưa gán nhãn, và một lượng nhỏ dữ liệu đã gán nhãn ban đầu (thường gọi là seed set) để xây dựng một bộ phân lớp dữ liệu tốt hơn.

Trong quá trình học như thế phương pháp học bán giám sát sẽ tận dụng được nhiều thông tin đa dạng của dữ liệu chưa gán nhãn, trong khi chỉ yêu cầu với một số lượng rất nhỏ các dữ liệu đã gán nhãn, vẫn thu được kết quả phân loại tốt.

Vấn đề được đặt ra là: Liệu các phương pháp học bán giám sát này có ích hay không? Hay chính xác hơn là, việc so sánh với phương pháp học giám sát chỉ sử dụng dữ liệu đã gán nhãn, ta có thể hy vọng vào sự chính xác của dự đoán khi xét thêm các điểm không gán nhãn [68], [74], [96], [98], [99].

Học bán giám sát được thể hiện thông qua các thành phần:

- **Tập huấn luyện:** $L = \{ (x_1, y_1), \dots, (x_k, y_k), x_{k+1}, \dots, x_n \}$, trong đó $x_i \in \mathbb{R}^d$ là véc tơ d chiều thể hiện các đặc tính của đối tượng thứ i . Với $i \in \{1, \dots, k\}$ là số thứ tự các nhãn đã gán nhãn và $i \in \{k+1, \dots, n\}$ là số thứ tự của các đối tượng chưa gán nhãn.

- **Mục đích:** gán nhãn cho các đối tượng $x_i, i \in \{k+1, \dots, n\}$ chưa gán nhãn trong tập huấn luyện hay đối tượng x ngoài tập huấn luyện.

Các phương pháp học bán giám sát sẽ được trình bày chi tiết hơn ở phần sau:

1.2.4. Học tăng cường

Học tăng cường là một lĩnh vực con của học máy, nghiên cứu cách thức một tác

tử trong một môi trường nên chọn thực hiện các hành động nào để cực đại hóa một khoản thưởng nào đó về lâu dài. Các thuật toán học tăng cường cố gắng tìm một chiến lược ánh xạ các trạng thái của thế giới tới các hành động mà tác tử nên chọn trong các trạng thái đó.

Học tăng cường là học với dữ liệu thường không được cho trước mà được tạo ra trong quá trình tương tác với môi trường. Mục tiêu của phương pháp này là tìm ra một sách lược lựa chọn hành động để cực tiểu hóa chi phí dài hạn nào đó. Quy trình động của môi trường và chi phí tối ưu thường không được biết trước nhưng có thể ước lượng được.

Môi trường thường được biểu diễn dưới dạng một quá trình quyết định Markov trạng thái hữu hạn và các thuật toán học tăng cường cho ngữ cảnh này có liên quan nhiều đến các kỹ thuật quy hoạch động. Các xác suất chuyển trạng thái và các xác suất thu lợi trong MDP thường là ngẫu nhiên nhưng lại tĩnh trong quá trình của bài toán.

Khác với học có giám sát, trong học tăng cường không có các cặp dữ liệu vào, kết quả đúng, các hành động gần tối ưu cũng không được đánh giá đúng sai một cách tường minh.

Học tăng cường đặc biệt thích hợp cho các bài toán có sự được mất giữa các khoản thưởng ngắn hạn và dài hạn. Học tăng cường đã được áp dụng thành công cho nhiều bài toán, trong đó có điều khiển rô bốt, điều vận thang máy, các trò chơi backgammon, cờ vua và các nhiệm vụ quyết định tuần tự khác.

1.2.5. Học sâu

Học sâu (Deep Learning) là một kỹ thuật học máy đang được nhiều nhà khoa học nghiên cứu. Kỹ thuật này nổi trội là do chúng thực hiện đồng thời hai việc: biểu diễn thông tin và học máy. Do đó, kỹ thuật này còn được gọi là học biểu diễn (representation learning) [20], [39], [91].

Học sâu dựa trên một tập hợp các thuật toán để mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng nhiều lớp xử lý với cấu trúc phức tạp, hoặc bằng cách khác bao gồm nhiều phép biến đổi phi tuyến, rộng hơn dựa trên đại diện học của dữ

liệu. Một quan sát (ví dụ như một hình ảnh) có thể được biểu diễn bằng nhiều cách như một véc tơ của các giá trị cường độ cho mỗi điểm ảnh, hoặc một cách trừu tượng hơn như là một tập hợp các cạnh, các khu vực hình dạng cụ thể, ...

Các nghiên cứu trong lĩnh vực này cố gắng tạo ra các mô hình để tìm hiểu các đại diện từ dữ liệu quy mô lớn không dán nhãn. Nhiều kiến trúc học sâu khác nhau như mạng nơ-ron sâu [6], mã mạng nơ-ron tích chập sâu và mạng nơ-ron tái phát đã được áp dụng cho các lĩnh vực như thị giác máy tính, tự động nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, nhận dạng âm thanh và tin sinh học.

Học sâu ảnh hưởng đến các lĩnh vực liên quan tới học máy và xử lý ngôn ngữ tự nhiên cũng không phải là ngoại lệ. Nhiều bài toán trong NLP như nhận dạng, dịch máy, POS Tagging, ... đã có được sự phát triển vượt trội nhờ học sâu.

1.3. Các phương pháp học bán giám sát

Trong khi xử lý các bài toán phân loại văn bản tự động ta thấy tồn tại một số lượng khổng lồ các dữ liệu văn bản trên Internet, thư điện tử, thư viện số, ... Các thuật toán học mang tính thống kê có thể được huấn luyện để phân lớp xấp xỉ các dữ liệu đó vào các chủ đề tương ứng.

Đã có một số thuật toán phân loại văn bản đã được sử dụng để phân lớp các bài báo [43], [47], phân lớp trang web [17], [70], tự động học thêm các sở thích về việc đọc của người dùng [44], [63], tự động sắp xếp thư điện tử [48], [67].

Tuy nhiên, trong thực tế các thuật toán này thường gặp phải những khó khăn như: Để xây dựng được thuật toán phân loại có độ tin cậy cao đòi hỏi phải có một số lượng lớn các mẫu dữ liệu huấn luyện (chính là các văn bản đã gán nhãn lớp tương ứng). Các dữ liệu huấn luyện đã gán nhãn này là rất ít và phải chi phí thời gian lớn vì dữ liệu này thường được thực hiện bởi con người, một tiến trình tốn thời gian và công sức.

Sự tồn tại của dữ liệu, trong thực tế dữ liệu thường tồn tại ở dạng trung gian: Không phải tất cả dữ liệu đều được gán nhãn cũng như không phải tất cả chúng đều chưa gán nhãn. Học bán giám sát là một phương pháp học sử dụng thông tin từ cả hai nguồn dữ liệu này.

Sự hiệu quả của học bán giám sát

Những năm gần đây có nhiều nghiên cứu về học bán giám sát. Những kết quả thực nghiệm cũng như lý thuyết đã chỉ ra rằng, sử dụng cách tiếp cận đánh giá khả năng giống nhau cực đại có thể cải tiến độ chính xác phân loại khi có thêm các dữ liệu chưa gán nhãn [60], [61]. Tuy nhiên, cũng có những nghiên cứu chỉ ra rằng, dữ liệu chưa gán nhãn có thể cải tiến độ chính xác phân loại hay không là phụ thuộc vào cấu trúc bài toán có phù hợp với giả thiết của mô hình hay không? Gần đây, Cozman [15] đã thực nghiệm trên dữ liệu giả hướng vào tìm hiểu giá trị của dữ liệu chưa gán nhãn. Thực nghiệm chỉ ra rằng, độ chính xác phân loại có thể giảm đi khi thêm vào ngày càng nhiều dữ liệu chưa gán nhãn. Nguyên nhân của sự giảm này là do sự không phù hợp giữa giả thiết của mô hình và phân phối dữ liệu thực tế.

Học bán giám sát để mang lại hiệu quả cần một điều kiện tiên quyết là: Phân phối các mẫu cần phát hiện phải phù hợp với bài toán phân loại. Về mặt công thức, các tri thức thu được từ dữ liệu chưa gán nhãn $p(x)$ phải mang lại thông tin hữu ích cho suy luận $p(x|y)$. Olivier Chapelle [12] đã đề xuất một giả thiết làm trơn, đó là hàm nhãn lớp ở vùng có mật độ cao thì trơn hơn ở vùng có mật độ thấp. Giả thiết được phát biểu như sau:

Giả thiết bán giám sát: Nếu hai điểm x_1, x_2 thuộc vùng có mật độ cao là gần nhau thì đầu ra tương ứng của chúng nên là y_1, y_2 .

Với giả thiết này, nếu hai điểm được liên kết với nhau bởi một đường dẫn trên vùng mật độ cao thì đầu ra của chúng nên gần nhau.

Đối với bài toán phân loại văn bản, được hình dung như sau: Dữ liệu chưa gán nhãn sẽ cung cấp thông tin về phân phối xác suất đồng thời của các từ khóa. Ví dụ với bài toán phân lớp trang web với hai lớp: trang chủ của một khoá học và không phải trang chủ của một khoá học. Ta coi trang chủ của một khoá học là hàm đích. Vì vậy, trang chủ của một khoá học sẽ là mẫu dương, và các trang còn lại là các mẫu âm.

Để có thể hiểu được bản chất của học bán giám sát, đầu tiên chúng ta cần hiểu thế nào là học giám sát và học không giám sát.

1.3.1. Một số phương pháp học bán giám sát

Trước khi quyết định lựa chọn phương pháp học cho một bài toán cụ thể cần phải xem xét các giả thiết của mô hình bài toán. Để sử dụng phương pháp học nào mà giả thiết của mô hình phù hợp với cấu trúc của bài toán [88]. Việc lựa chọn này có thể là khó trong thực tế, tuy nhiên. Nếu các lớp tạo ra dữ liệu có tính phân cụm cao thì sử dụng phương pháp cực đại hóa kỳ vọng (Expectation Maximization - EM) với mô hình trộn sinh có thể là một sự lựa chọn tốt; nếu các đặc trưng có sự phân chia tự nhiên thành hai tập thì phương pháp đồng huấn luyện (Co-training) có thể là một sự lựa chọn phù hợp; nếu hai mẫu dữ liệu với các đặc trưng tương tự nhau hướng tới thuộc về cùng một lớp thì có thể sử dụng các phương pháp dựa trên đồ thị; nếu các bộ phân loại giám sát được xây dựng từ trước là phức tạp và khó sửa đổi thì phương pháp tự huấn luyện (Self-training) sẽ là một lựa chọn ưu tiên.

Trước khi trình bày chi tiết hai phương pháp học Self-training và Co-training, chúng ta sẽ tìm hiểu một số phương pháp học bán giám sát điển hình khác:

Thuật toán cực đại kỳ vọng, thuật toán SVM truyền dẫn và thuật toán phân hoạch đồ thị quang phổ.

1.3.1.1. Thuật toán cực đại hóa kỳ vọng

Trong thống kê và học máy, thuật toán cực đại hóa kỳ vọng (EM -Expectation-Maximization) được sử dụng rộng rãi để giải bài toán tìm hợp lý cực đại hoặc hậu nghiệm cực đại (MAP) của một mô hình xác suất có các biến ẩn. EM sở dĩ được gọi vậy một phần do thuật toán này bao gồm việc thực hiện liên tiếp tại mỗi vòng lặp hai quá trình (E): tính kỳ vọng của hàm hợp lý của giá trị các biến ẩn dựa theo ước lượng đang có về các tham số của mô hình và (M): ước lượng tham số của mô hình để cực đại hóa giá trị của hàm tính được ở (E). Các giá trị tìm được ở (E) và (M) tại mỗi vòng lặp sẽ được dùng cho việc tính toán ở vòng lặp kế tiếp.

Trong thống kê học, nếu một mô hình xác suất có chứa các biến ẩn hoặc thiếu dữ liệu thì việc tính toán ước lượng của các tham số trở nên khó khăn hoặc không thực hiện được. Thật vậy, thông thường ta cần một trong hai đại lượng trên (biến ẩn và tham số) để ước lượng giá trị của cái còn lại.

Giải thuật EM cho ta một phương pháp giải quyết bài toán trên một lớp bài toán tương đối rộng. Nguyên lý của nó là tại mỗi bước (E) ta giả thiết rằng tham số đã biết và cố gắng ước lượng giá trị của biến ẩn này và dùng giá trị tìm được này ở bước (M) để tìm giá trị của các tham số. Ta có thể chứng minh được rằng tại mỗi vòng lặp, ta luôn tìm được kết quả tốt hơn của vòng lặp trước đó, vì thế EM luôn hội tụ về giá trị tối ưu (cục bộ).

Thuật toán EM là một thuật toán tổng quát đánh giá sự khả năng cực đại (ML – Maximum Likelihood) [2] mà dữ liệu không hoàn chỉnh hoặc hàm khả năng liên quan đến các biến ẩn [3], [61]. Ở đây, hai khái niệm “dữ liệu không hoàn chỉnh” và “biến ẩn” có liên quan đến nhau: Khi tồn tại biến ẩn, thì dữ liệu là không hoàn chỉnh vì ta không thể quan sát được giá trị của biến ẩn; tương tự như vậy khi dữ liệu là không hoàn chỉnh, ta cũng có thể liên tưởng đến một vài biến ẩn với dữ liệu thiếu. Thuật toán EM gồm hai bước lặp: Bước mong đợi (Expectation step) và bước cực đại (Maximization step). Khởi đầu, nó gán giá trị ngẫu nhiên cho tất cả các tham số của mô hình. Sau đó, tiến hành lặp hai bước lặp sau:

Bước mong đợi (E-step): tính toán khả năng mong muốn cho dữ liệu dựa trên các thiết lập tham số và dữ liệu không hoàn chỉnh.

Bước cực đại (M-step): Tính toán lại tất cả các tham số sử dụng tất cả các dữ liệu. Khi đó, ta sẽ có một tập các tham số mới.

Tiến trình tiếp tục cho đến khi hội tụ, ví dụ như đạt tới cực đại cục bộ. EM sử dụng phương pháp leo đồi, nên chỉ đảm bảo đạt được cực đại cục bộ. Khi tồn tại nhiều cực đại, việc đạt tới cực đại toàn cục hay không là phụ thuộc vào điểm bắt đầu leo đồi. Nếu ta bắt đầu từ một đồi đúng, ta sẽ có khả năng tìm được cực đại toàn cục. Tuy nhiên, việc tìm được đồi đúng thường là rất khó. Có hai giải pháp được đưa ra để giải quyết bài toán này:

Một là, chúng ta thử nhiều giá trị khởi đầu khác nhau, sau đó lựa chọn giải pháp có giá trị khả năng hội tụ lớn nhất.

Hai là, sử dụng mô hình đơn giản hơn để xác định giá trị khởi đầu cho các mô hình phức tạp.

Ý tưởng là: một mô hình đơn giản hơn sẽ giúp tìm được vùng tồn tại cực đại

toàn cục và ta bắt đầu bằng một giá trị trong vùng đó để tìm kiếm tối ưu chính xác khi sử dụng mô hình phức tạp hơn.

Thuật toán EM rất đơn giản, ít nhất là về mặt khái niệm. Nó được sử dụng hiệu quả nếu dữ liệu có tính phân cụm cao.

1.3.1.2. Học SVM truyền dẫn

Gồm nội dung cơ bản của học quy nạp và học truyền dẫn [41].

• **Học quy nạp:** Xét hàm f ánh xạ từ đầu vào x tới đầu ra y : $y = f(x)$ với $(y \in \{-1, 1\})$.

Học quy nạp sẽ dựa vào các dữ liệu huấn luyện có dạng $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ để tìm hàm f . Sau đó, ta sẽ sử dụng hàm f để dự đoán nhãn y_{n+1} cho các mẫu chưa gán nhãn x_{n+1} . Các vấn đề của phương pháp:

- Khó tập hợp các dữ liệu gán nhãn.
- Lấy các mẫu dữ liệu chưa gán nhãn thì dễ dàng.
- Các mẫu cần phân lớp là biết trước.
- Không quan tâm đến hàm phân lớp f .

Do vậy cần ứng dụng học theo kiểu truyền dẫn.

• **Học truyền dẫn:** Học truyền dẫn được Vapnik đề cập từ năm 1998. Một bộ học được gọi là truyền dẫn nếu bộ học chỉ xử lý trên dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn và không thể xử lý dữ liệu mà bộ học chưa biết. Cho trước một tập các mẫu gán nhãn $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ và một tập các dữ liệu chưa gán nhãn x'_1, x'_2, \dots, x'_m , mục đích là tìm các nhãn y'_1, y'_2, \dots, y'_m . Học truyền dẫn không cần thiết phải xây dựng hàm f , đầu ra sẽ là một véc tơ các nhãn lớp được xác định bằng việc chuyển thông tin từ dữ liệu gán nhãn sang dữ liệu chưa gán nhãn. Các phương pháp dựa trên đồ thị lúc đầu thường là truyền dẫn.

• **Phương pháp học SVM truyền dẫn:**

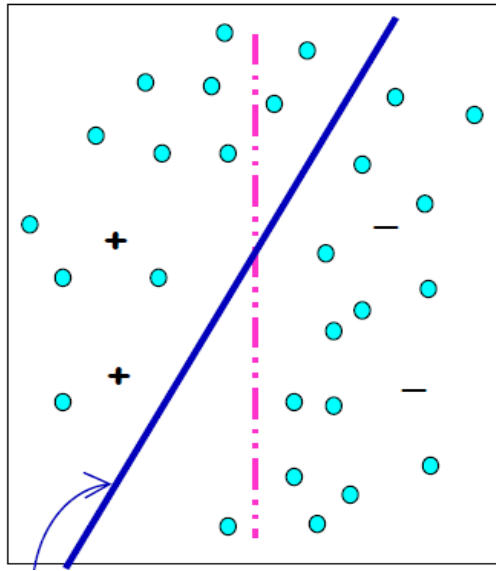
SVM truyền dẫn (TSVM) một dạng mở rộng của SVM chuẩn. Trong SVM chỉ có dữ liệu đã gán nhãn được sử dụng, mục đích là tìm siêu phẳng cực đại dựa trên các mẫu dữ liệu huấn luyện. Với TSVM, các điểm dữ liệu chưa gán nhãn cũng được sử dụng. Mục đích của TSVM là gán nhãn cho các điểm dữ liệu chưa gán nhãn để

cho biên tuyến tính có lề phân cách là lớn nhất trên cả dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn (hình 1.1).

Qui ước:

+, -: các mẫu dương, âm

● : các mẫu chưa gán nhãn



Hình 1.1 Siêu phẳng cực đại

Đường chấm chấm là kết quả của bộ phân lớp SVM quy nạp, đường liên tục chính là phân lớp SVM truyền dẫn.

1.3.1.3. Thuật toán Self-training [34]

a. Giới thiệu

Trong quá trình xử lý các bài toán phân loại văn bản tự động, chúng ta cần phải xử lý số lượng khổng lồ các dữ liệu văn bản trên Web, E-mail, tài liệu thư viện số, sách điện tử, ... Các thuật toán học mang tính thống kê có thể được huấn luyện để phân lớp xấp xỉ các dữ liệu đó vào các chủ đề tương ứng giúp cho việc lưu trữ, sắp xếp, quản lý, tra cứu, tìm kiếm được thuận tiện.

Phương pháp học bán giám sát đã được nghiên cứu phát triển trong những năm gần đây, nhất là từ khi xuất hiện các trang Web với số lượng thông tin ngày càng nhiều, các chủ đề ngày càng phong phú. Chúng ta có thể nêu lên quá trình phát triển của học bán giám sát trải qua các thuật toán được nghiên cứu như sau.

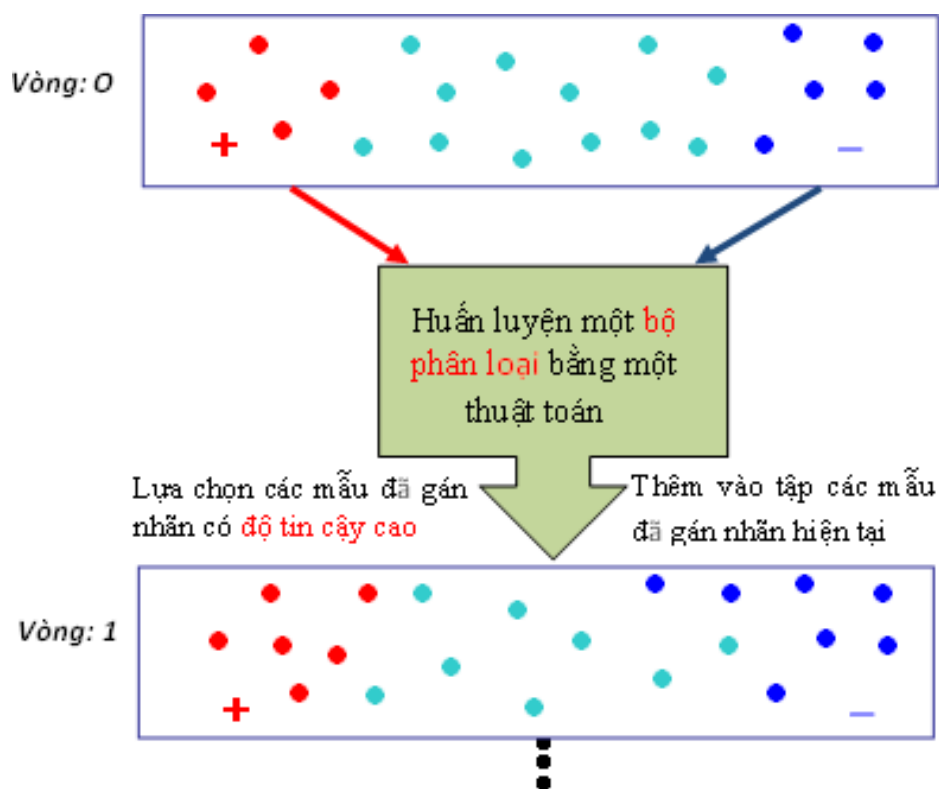
Cùng với số liệu lớn của dữ liệu chưa gán nhãn, các thành phần hỗn hợp có thể được nhận ra cùng với thuật toán EM. Chỉ cần một mẫu đơn đã gán nhãn cho mỗi thành phần để xác định hoàn toàn được mô hình hỗn hợp. Mô hình này được áp dụng thành công vào việc phân loại văn bản. Một biến thể khác của mô hình này chính là phương pháp Self-training. Cả 2 phương pháp này được sử dụng cách đây một thời gian dài. Chúng được sử dụng phổ biến vì dựa trên khái niệm đơn giản của chúng và sự dễ hiểu của thuật toán [16], [27], [54], [56].

Thuật toán Co-training là thuật toán học bán giám sát điển hình tiếp theo mà các nhà khoa học đầu tư nghiên cứu. Trong khi thuật toán Self-training là thuật toán mà khi có một sự phân loại lỗi thì có thể tăng cường thêm cho chính nó, thì thuật toán Co-training giảm bớt được lỗi tăng cường có thể xảy ra khi có một quá trình phân loại bị lỗi.

Cùng với quá trình phát triển và việc áp dụng phổ biến và sự tăng lên về chất lượng của thuật toán SVM (Máy véc tơ hỗ trợ - Support Vector Machine), SVM truyền dẫn (Transductive Support Vector Machine – TSVM) nổi bật lên như một SVM chuẩn mở rộng cho phương pháp học bán giám sát.

Gần đây các phương pháp học bán giám sát dựa trên đồ thị thu hút nhiều sự quan tâm của các nhà khoa học cũng như những người quan tâm đến lĩnh vực khai phá dữ liệu. Các phương pháp dựa trên đồ thị bắt đầu với một đồ thị mà các nút là các điểm dữ liệu gán nhãn và các điểm dữ liệu chưa gán nhãn qua các điểm nối phản ánh được sự giống nhau giữa các nút này. Có thể thấy học bán giám sát là một quá trình hoàn thiện dần các thuật toán để áp dụng giải quyết các bài toán thực tế.

Thuật toán về Self-training được đề xuất từ những năm 1960. Có thể nói rằng, ý tưởng đầu tiên về sử dụng dữ liệu chưa gán nhãn trong phân loại là thiết lập Self-training. Đó là thuật toán sử dụng vòng lặp nhiều lần một phương pháp học giám sát. Hình 1.2 biểu diễn trực quan của thiết lập Self-training.



Hình 1.2 Biểu diễn trực quan của thiết lập Self-training

Thuật toán Self-training đã được ứng dụng trong xử lý ngôn ngữ tự nhiên như phân tích cú pháp, dịch máy, chương trình môi (bootstrap)...

Thuật toán Self-training là kỹ thuật học bán giám sát được sử dụng rất phổ biến, với một bộ phân lớp ban đầu được huấn luyện bằng một số lượng nhỏ các dữ liệu đã gán nhãn. Sau đó, sử dụng bộ phân lớp này để gán nhãn các dữ liệu chưa gán nhãn. Các dữ liệu được gán nhãn có độ tin cậy cao (vượt trên một ngưỡng nào đó) và nhãn tương ứng của chúng được đưa vào tập huấn luyện (training set). Tiếp đó, bộ phân lớp được học lại trên tập huấn luyện mới ấy và thủ tục lặp tiếp tục. Ở mỗi vòng lặp, bộ học sẽ chuyển một vài các mẫu có độ tin cậy cao nhất sang tập dữ liệu huấn luyện cùng với các dự đoán phân lớp của chúng. Tên gọi Self-training xuất phát từ việc nó sử dụng dự đoán của chính nó để dạy chính nó.

b. Thuật toán

❖ **Mục đích:** Mở rộng tập các mẫu đã gán nhãn ban đầu bằng cách chỉ cần một bộ phân lớp với một khung nhìn của dữ liệu.

❖ **Giải thuật:**

Dữ liệu vào:

- L : là tập các dữ liệu đã gán nhãn
- U : là tập các dữ liệu chưa gán nhãn

Dữ liệu ra:

- Gán nhãn cho tập con U' của U có độ tin cậy cao nhất

Repeat

- Huấn luyện bộ phân lớp h trên tập dữ liệu huấn luyện L .
- Sử dụng h để phân lớp dữ liệu trong tập U .
- Tìm tập con U' của U có độ tin cậy cao nhất.
- $L + U' \rightarrow L$
- $U - U' \rightarrow U$

Until ($U = \emptyset$)

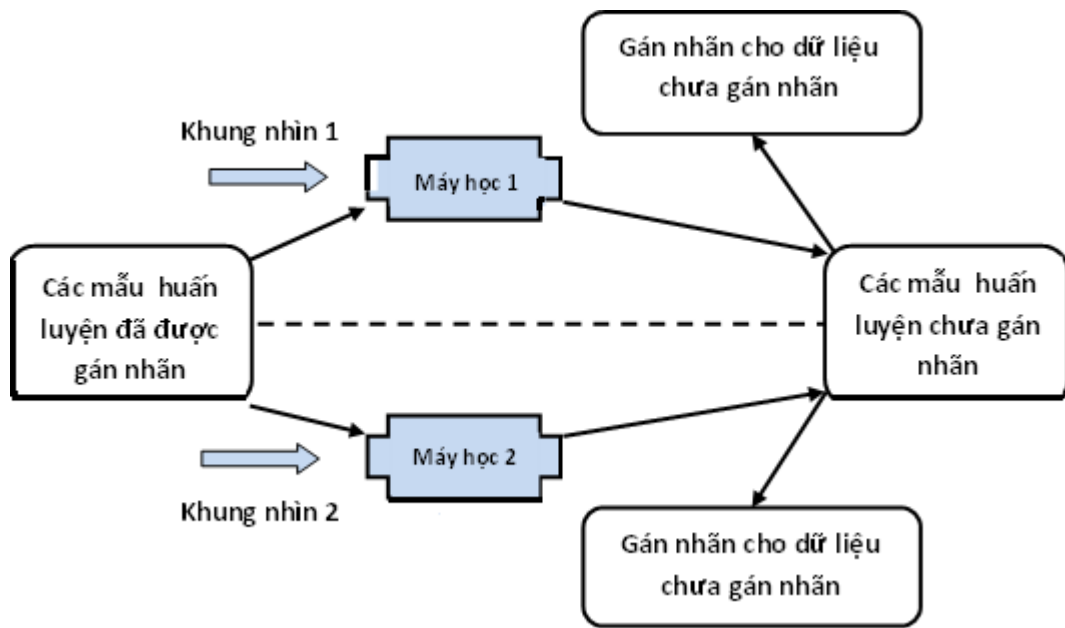
1.3.1.4. Thuật toán Co-training

a. Giới thiệu

Thuật toán Co-training dựa trên giả thiết rằng các đặc trưng có thể được phân chia thành hai tập con. Mỗi tập con phù hợp để huấn luyện một bộ phân lớp tốt. Hai tập con đó phải thoả mãn tính chất độc lập điều kiện khi cho trước lớp đó. Thuật toán được tiến hành như sau [13], [21], [53], [75]:

- Học hai bộ phân lớp riêng rẽ bằng dữ liệu đã gán nhãn trên hai tập thuộc tính con tương ứng.
- Mỗi bộ phân lớp sau đó lại phân lớp các dữ liệu chưa gán nhãn. Sau đó, chúng lựa chọn ra các dữ liệu chưa gán nhãn + nhãn dự đoán của chúng (các mẫu có độ tin cậy cao) để huấn luyện cho bộ phân lớp kia.
- Sau đó, mỗi bộ phân lớp được học lại với các mẫu huấn luyện được cho bởi bộ phân lớp kia và tiến trình lặp bắt đầu [97].

Sơ đồ Co-training đã được sử dụng trong rất nhiều lĩnh vực như phân tích thống kê và xác định cụm danh từ.



Hình 1.3 Sơ đồ biểu diễn trực quan thiết lập Co-training

Blum và Mitchell [2] đã công thức hoá hai giả thiết của mô hình Co-training và chứng minh tính đúng đắn của mô hình dựa trên thiết lập học giám sát theo mô hình PAC chuẩn. Cho trước một không gian các mẫu $X = X_1 \times X_2$, ở đây X_1 và X_2 tương ứng với hai khung nhìn khác nhau của cùng một mẫu. Mỗi mẫu x vì vậy có thể được biểu diễn bởi một cặp (x_1, x_2) . Chúng ta giả thiết rằng mỗi khung nhìn là phù hợp để phân lớp chính xác. Cụ thể, nếu D là một phân phối trên X , và C_1, C_2 là các lớp khái niệm được định nghĩa tương ứng trên X_1 và X_2 ; giả thiết rằng tất cả các nhãn trên các mẫu với xác suất lớn hơn không dưới phân phối D là trùng khớp với một hàm đích $f_1 \in C_1$ và cũng trùng khớp với hàm đích $f_2 \in C_2$. Nói cách khác, nếu f biểu diễn khái niệm đích kết hợp trên toàn bộ mẫu, thì với bất kỳ mẫu $x = x_1 \times x_2$ có nhãn 1, ta có $f(x) = f_1(x_1) = f_2(x_2) = 1$. Nghĩa là D gán xác suất bằng không mẫu (x_1, x_2) bất kỳ mà $f_1(x_1) \neq f_2(x_2)$.

• *Giả thiết thứ nhất*: Tính tương thích.

Với một phân phối D cho trước trên X , ta nói rằng hàm đích $f = (f_1, f_2) \in C_1 \times C_2$ là tương thích với D nếu thoả mãn điều kiện: D gán xác suất bằng không cho tập các mẫu (x_1, x_2) mà $f_1(x_1) \neq f_2(x_2)$. Nói cách khác, mức độ tương thích của một

hàm đích $f = (f_1, f_2)$ với một phân phối D có thể được định nghĩa bằng một số $0 \leq p \leq 1$:

$$P = 1 - P_{r_D}[(x_1, x_2): f_1(x_1) \neq f_2(x_2)] \quad (1.1)$$

• *Giả thiết thứ hai*: Độc lập điều kiện.

Ta nói rằng hàm đích f_1, f_2 và phân phối D thoả mãn giả thiết độc lập điều kiện nếu với bất kỳ một mẫu $(x_1, x_2) \in X$ với xác suất khác không thì,

$$\Pr_{(x_1, x_2) \in D} \left[x_1 = \hat{x}_1 \mid x_2 = \hat{x}_2 \right] = \Pr_{(x_1, x_2) \in D} \left[x_1 = \hat{x}_1 \mid f_2(x_2) = f_2(\hat{x}_2) \right] \quad (1.2)$$

và tương tự,

$$\Pr_{(x_1, x_2) \in D} \left[x_2 = \hat{x}_2 \mid x_1 = \hat{x}_1 \right] = \Pr_{(x_1, x_2) \in D} \left[x_2 = \hat{x}_2 \mid f_1(x_1) = f_1(\hat{x}_1) \right] \quad (1.3)$$

A.Blum & T. Mitchell đã chỉ ra rằng, cho trước một giả thiết độc lập điều kiện trên phân phối D , nếu lớp đích có thể học được từ nhiều phân lớp ngẫu nhiên theo mô hình PAC chuẩn, thì bất kỳ một bộ dự đoán yếu ban đầu nào cũng có thể được nâng lên một độ chính xác cao tùy ý mà chỉ sử dụng các mẫu chưa gán nhãn bằng thuật toán Co-training. A.Blum & T. Mitchell cũng đã chứng minh tính đúng đắn của sơ đồ Co-training bằng định lý sau:

Định lý (A.Blum & T. Mitchell)

Nếu C_2 có thể học được theo mô hình PAC với nhiều phân lớp và nếu giả thiết độc lập điều kiện thoả mãn, thì (C_1, C_2) có thể học được theo mô hình Co-training chỉ từ dữ liệu chưa gán nhãn, khi cho trước một bộ dự đoán yếu nhưng hữu ích ban đầu $h(x_1)$.

Blum và Mitchell đã tiến hành thực nghiệm Co-training trong phân lớp trang web trong thuật toán thể hiện rằng việc sử dụng dữ liệu chưa gán nhãn tạo ra một cải tiến quan trọng trong thực hành. Trong sơ đồ thiết lập trên, việc sử dụng U' sẽ tạo ra kết quả tốt hơn vì: Nó bắt buộc hai bộ phân lớp lựa chọn các mẫu có tính đại diện hơn cho phân phối D tạo ra tập U .

b. Thuật toán

❖ **Mục đích:** Mở rộng tập các mẫu gán nhãn ban đầu bằng cách sử dụng hai bộ phân lớp với hai khung nhìn của dữ liệu.

❖ **Giải thuật:**

Dữ liệu vào:

- L : Là tập các mẫu huấn luyện đã gán nhãn
- U : Là tập các mẫu chưa gán nhãn

Dữ liệu ra:

- Tạo một tập dữ liệu gán nhãn U' gồm u mẫu được chọn ngẫu nhiên từ U

For $i = 1$ to k do

- Sử dụng L huấn luyện bộ phân lớp h_1 trên phần x_1 của x
- Sử dụng L huấn luyện bộ phân lớp h_2 trên phần x_2 của x
- Cho h_1 gán nhãn p mẫu dương và n mẫu âm từ tập U'
- Cho h_2 gán nhãn p mẫu dương và n mẫu âm từ tập U'
- Thêm các mẫu tự gán nhãn này vào tập L
- Chọn ngẫu nhiên $2p + 2n$ mẫu từ tập U bổ sung vào tập U'

1.3.1.5. So sánh hai thuật toán

Kết quả đưa ra một số so sánh hai thiết lập Self-training và Co-training. Nói chung, sự khác nhau cơ bản giữa thuật toán Self-training và thuật toán Co-training là: Thuật toán Self-training chỉ sử dụng một khung nhìn dữ liệu, trong khi đó thuật toán Co-training sử dụng hai khung nhìn dữ liệu. Thuật toán Self-training không yêu cầu sự phân chia của các đặc trưng thành hai khung nhìn độc lập như thuật toán Co-training. Nó chỉ cần một bộ phân lớp với một khung nhìn của dữ liệu.

Thuật toán Co-training và thuật toán Self-training là hai thuật toán học bán giám sát có nhiệm vụ chính là mở rộng tập các mẫu đã gán nhãn ban đầu. Hiệu quả của thuật toán phụ thuộc vào chất lượng của các mẫu đã gán nhãn được thêm vào ở mỗi vòng lặp, được đo bởi hai tiêu chí:

- Độ chính xác của các mẫu được thêm vào đó.
- Thông tin hữu ích mà các mẫu mang lại cho bộ phân lớp.

Xem xét tiêu chí thứ nhất ta thấy, bộ phân lớp chứa càng nhiều thông tin thì độ tin cậy cho các dự đoán càng cao. Thuật toán Co-training sử dụng hai khung nhìn khác nhau của một mẫu dữ liệu với giả thiết là mỗi khung nhìn là đầy đủ (*sufficient*) để dự đoán nhãn cho các mẫu dữ liệu mới. Tuy nhiên, giả thiết này là không thực tế bởi vì nhiều khi tập tất cả các đặc trưng của một mẫu dữ liệu cũng chưa đủ để gán nhãn chúng một cách chính xác. Vì vậy, trong các ứng dụng thực, nếu xét theo tiêu chí này thì Self-training thường có độ tin cậy cao hơn.

Với tiêu chí thứ hai, ta biết rằng thông tin mà mỗi mẫu dữ liệu gán nhãn mới đem lại thường là các đặc trưng mới. Vì thuật toán Co-training huấn luyện trên hai khung nhìn khác nhau nên nó sẽ hữu ích hơn trong việc cung cấp các thông tin mới cho nhau.

Việc lựa chọn các mẫu gán nhãn mới có độ tin cậy cao là một vấn đề hết sức quan trọng, vì nếu tiêu chí thứ nhất không được thoả mãn, các mẫu bị gán nhãn sai thì thông tin mới do chúng đem lại chẳng những không giúp ích được mà thậm chí còn làm giảm hiệu quả của thuật toán.

1.3.2. Thuật toán học có giám sát SVM và bán giám sát SVM

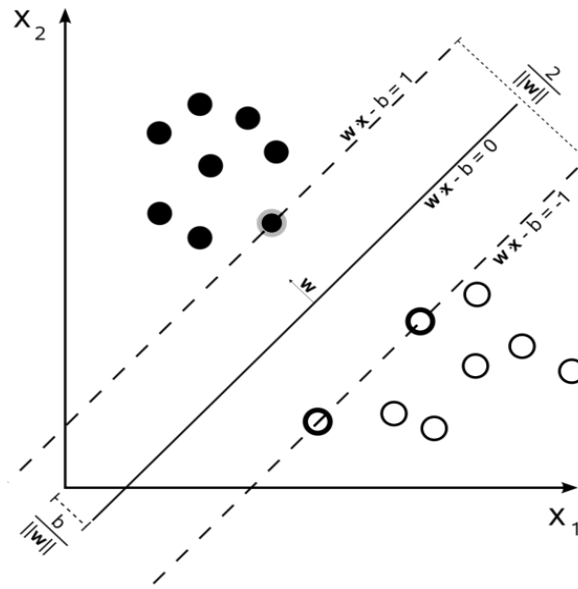
1.3.2.1. Giới thiệu

Trong lĩnh vực khai phá dữ liệu, các phương pháp phân loại văn bản đã dựa trên những phương pháp quyết định như quyết định Bayes, cây quyết định, K - láng giềng gần nhất, Những phương pháp này đã cho kết quả chấp nhận được và được sử dụng nhiều trong thực tế. Trong những năm gần đây, phương pháp phân lớp sử dụng tập phân lớp véc tơ hỗ trợ được quan tâm và sử dụng nhiều trong lĩnh vực nhận dạng và phân lớp. SVM là một họ các phương pháp dựa trên cơ sở các hàm nhân (kernel) để tối thiểu hoá rủi ro ước lượng. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân lớp khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...). Xét với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đối tốt và hiệu quả.

Thuật toán máy véc tơ hỗ trợ được Corters và Vapnik giới thiệu vào năm 1995 [33]. SVM rất hiệu quả để giải quyết các bài toán với dữ liệu có số chiều lớn như các véc tơ biểu diễn văn bản. Thuật toán SVM ban đầu chỉ được thiết kế để giải quyết bài toán phân lớp nhị phân, tức là số lớp hạn chế là hai lớp. Hiện nay, SVM được đánh giá là bộ phân lớp nhanh, chính xác nhất cho bài toán phân loại văn bản.

1.3.2.2. Thuật toán máy véc tơ hỗ trợ (SVM)

Đây là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau nhằm để phân loại và phân tích hồi quy [43], [44], [49], [85]. Mục đích của SVM là phân loại dữ liệu thành hai lớp khác nhau. Vì vậy, có thể nói SVM là một thuật toán phân loại nhị phân. Với một bộ các mẫu huấn luyện thuộc hai thể loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các mẫu khác vào hai thể loại đó. Thuật toán SVM chia hai lớp dữ liệu bằng một siêu mặt phẳng $d-1$ chiều khi số chiều của dữ liệu huấn luyện là d . **Hình 1.4** là ví dụ phân tách dữ liệu thuộc hai lớp sử dụng SVM. Trong đó, $w \cdot x - b = 0$ là siêu mặt phẳng thể hiện sự phân tách dữ liệu.



Hình 1.4 Siêu mặt tối ưu và biên

Cho trước tập huấn luyện được biểu diễn trong không gian véc tơ, trong đó mỗi văn bản là một điểm, phương pháp SVM dạng chuẩn tìm ra một siêu mặt phẳng

quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng với lớp (+) và lớp (-). Hiệu quả xác định siêu mặt phẳng này được quyết định bởi khoảng cách của điểm gần mặt phẳng nhất của mỗi lớp. Khoảng cách càng lớn thì mặt phẳng quyết định càng tốt đồng nghĩa với việc phân loại càng chính xác và ngược lại. Mục đích cuối cùng của phương pháp là tìm khoảng cách biên lớn nhất.

- **Mục đích:** Dùng để phân lớp dữ liệu mới thuộc lớp nào.
- **Dữ liệu vào:** Cho trước một tập dữ liệu huấn luyện đã gán nhãn thuộc lớp -1 hoặc +1.
- **Dữ liệu ra:** Dùng để tìm ra nhãn cho các dữ liệu mới.
- **Mô tả thuật toán:** Ta có một tập huấn luyện \mathcal{D} , một tập gồm n điểm có dạng.

$$\mathcal{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n$$

Với y_i mang giá trị +1 hoặc -1, ngầm định lớp chứa điểm x_i . Mỗi x_i là một véc tơ thực p - chiều. Ta cần tìm siêu mặt phẳng có lề lớn nhất chia tách các điểm x_i thuộc lĩnh vực quan tâm và được gán nhãn $y_i = 1$ và các điểm x_i thuộc lĩnh vực không quan tâm và được gán nhãn $y_i = -1$. Mỗi siêu mặt phẳng đều có thể được viết như là một tập các điểm x thỏa mãn phương trình:

$$w \cdot x - b = 0$$

Với (\cdot) kí hiệu cho tích vô hướng. Véc tơ trọng số w , nó vuông góc với siêu mặt phẳng. Tham số $\frac{b}{\|w\|}$ xác định độ lệch của siêu mặt phẳng từ nó đến véc tơ w .

Phương trình trên tương đương với phương trình sau:

$$C + w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n = 0 \quad (1.4)$$

Tương đương với công thức :

$$C + \sum_{i=1, \dots, n} w_i \cdot x_i = 0 \quad (1.5)$$

Với $w = w_1 + w_2 + \dots + w_n$ là bộ hệ số siêu mặt phẳng hay là véc tơ trọng số, C là

độ dịch, khi thay đổi w và b thì hướng và khoảng cách từ gốc tọa độ đến siêu mặt phẳng thay đổi.

Chúng ta cần chọn w và b để cực đại lề, hay khoảng cách giữa hai siêu mặt phẳng song song sao cho chúng càng xa càng tốt trong khi vẫn phân chia tốt dữ liệu. Các siêu mặt phẳng ấy được xác định bằng:

$$w \cdot x - b = 1$$

và

$$w \cdot x - b = -1$$

Để ý rằng nếu một dữ liệu huấn luyện có thể được chia tách một cách tuyến tính (bằng một đường thẳng), chúng ta có thể chọn hai siêu mặt phẳng của lề sao cho không có điểm nào ở giữa chúng và cố gắng cực đại khoảng cách giữa chúng. Bằng hình học, chúng ta tìm được khoảng cách giữa hai siêu mặt phẳng là $\frac{2}{\|w\|}$ và ta muốn cực tiểu giá trị $\|w\|$. Để tránh các điểm dữ liệu rơi vào bên trong lề, chúng ta thêm vào các điều kiện sau, với mỗi i ta có:

$$w \cdot x_i - b \geq 1 \text{ nếu } y_i = 1 \text{ (đối với lớp thứ nhất)}$$

hoặc

$$w \cdot x_i - b \leq -1 \text{ nếu } y_i = -1 \text{ (đối với lớp thứ hai)}$$

Có thể viết gọn lại như sau :

$$y_i (w \cdot x_i - b) \geq 1, \text{ với mọi } 1 \leq i \leq n$$

Ta có thể gom chúng lại trong một bài toán tối ưu:

Cực tiểu (theo w, b): $\|w\|$ với điều kiện (với mọi $i = 1, \dots, n$)

$$y_i (w \cdot x_i - b) \geq 1$$

Các vấn đề tối ưu hóa được trình bày trong phần trước là khó khăn để giải quyết bởi vì nó phụ thuộc vào $\|w\|$, chỉ tiêu w , trong đó bao gồm một căn bậc hai. May mắn thay, nó có thể làm thay đổi phương trình bằng cách thay thế $\|w\|$ với $\frac{1}{2} \|w\|^2$

(Yếu tố $\frac{1}{2}$ đang được sử dụng để thuận tiện trong toán học) mà không thay đổi các giải pháp (tối thiểu của bản gốc và phương trình sửa đổi có w và b). Đây là vấn đề tối ưu hóa một phương trình bậc hai. Rõ ràng hơn:

$$\text{Cực tiểu (trong } w, b) : \frac{1}{2} \|w\|^2 \text{ (với bất kỳ } i = 1, \dots, n)$$

$$y_i (w \cdot x_i - b) \geq 1$$

Đưa về phương trình Lagrange với số nhân α_i [7]:

$$\min_{w, b, \alpha} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i (w \cdot x_i - b) - 1] \right\} \quad (1.6)$$

Máy véc tơ hỗ trợ SVM là quá trình tìm ra các siêu mặt phẳng phụ thuộc vào tham số véc tơ trọng số w và độ dịch C . Mục tiêu của phương pháp SVM là ước lượng w và C để cực đại hoá lề giữa các lớp dữ liệu dương và âm. Các giá trị khác nhau của lề cho ta các họ siêu mặt phẳng khác nhau và lề càng lớn thì năng lực của máy học càng giảm. Như vậy, cực đại hoá lề thực chất là việc tìm một máy học có năng lực nhỏ nhất. Quá trình phân loại là tối ưu khi sai số phân loại là cực tiểu. Sau khi đã tìm được phương trình của siêu mặt phẳng bằng thuật toán SVM, áp dụng công thức này để tìm ra nhãn lớp cho các dữ liệu mới.

Nếu dữ liệu học không tách rời tuyến tính, thêm biến η_i và thay phương trình trên bằng phương trình:

$$\begin{aligned} \min_{w, b, \eta} C &= \sum_{i=1}^n \eta_i + \frac{1}{2} \|w\|^2 \\ \text{Với } y_i [w \cdot x_i - b] + \eta_i &\geq 1 \\ \eta_i &\geq 0, \quad i = 1, \dots, n \end{aligned} \quad (1.7)$$

Từ đó ta có phương trình tổng quát của siêu mặt phẳng tìm ra được bởi thuật toán SVM là:

$$f(x_1, x_2, \dots, x_n) = C + \sum w_i \cdot x_i \quad (1.8)$$

Với $i = 1, \dots, n$. Trong đó n là số dữ liệu huấn luyện.

1.3.3. SVM trong phân loại văn bản

Phân loại văn bản là một tiến trình đưa các văn bản chưa biết chủ đề vào các lớp văn bản đã biết (tương ứng với các chủ đề hay lĩnh vực khác nhau). Mỗi chủ đề được xác định bởi một số tài liệu mẫu của chủ đề đó. Để thực hiện quá trình phân loại, các phương pháp huấn luyện được sử dụng để xây dựng tập phân loại từ các tài liệu mẫu, sau đó dùng tập phân loại này để dự đoán loại của những tài liệu mới (chưa biết chủ đề).

Từ các thuật toán phân lớp hai lớp như SVM đến các thuật toán phân lớp đa lớp, đều có đặc điểm chung là yêu cầu văn bản phải được biểu diễn dưới dạng véc tơ đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu, trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp thì SVM là phương pháp sử dụng không gian véc tơ đặc trưng lớn nhất (hơn 10.000 chiều) trong khi đó các phương pháp khác có số chiều bé hơn nhiều (như Naïve Bayes là 2000, k-Nearest Neighbors là 2415, ...).

Trong các công trình nghiên cứu Joachims đã so sánh SVM với Naïve Bayes, k-Nearest Neighbour [41], [50], Joachims đã chứng minh rằng SVM xử lý rất tốt cùng với các đặc tính được đề cập trước đây của văn bản. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân loại tốt nhất khi so sánh với các phương pháp khác.

Theo nghiên cứu tại [94] thì trong các công trình nghiên cứu [90] đã chỉ ra rằng thuật toán SVM hoàn toàn được tiến hành tốt nhất so với các phương pháp phân loại văn bản khác. Tất cả các tài liệu nghiên cứu hiện nay cho thấy rằng SVM đưa ra kết quả chính xác nhất trong khía cạnh phân loại văn bản.

Những phân tích của các tác giả trên đây cho thấy SVM có nhiều điểm phù hợp cho việc ứng dụng phân loại văn bản. Và trên thực tế, các thí nghiệm phân loại văn bản tiếng Anh chỉ ra rằng SVM đạt độ chính xác phân loại cao và tỏ ra xuất sắc hơn so với các phương pháp phân loại văn bản khác.

Vấn đề căn bản của học bán giám sát là chúng ta có thể tận dụng dữ liệu chưa gán nhãn để cải tiến hiệu quả của độ chính xác trong khi phân loại, điều này được

đưa ra để so sánh với một tập phân loại được thiết kế mà không tính đến dữ liệu chưa gán nhãn.

Trong phần sau sẽ giới thiệu một phương thức cải tiến của SVM là bán giám sát S3VM (semi-supervised support vector machine) [41]. Bán giám sát S3VM được đưa ra nhằm nâng SVM lên một mức cao hơn, trong khi SVM là một thuật toán học có giám sát sử dụng dữ liệu đã gán nhãn, thì bán giám sát S3VM sử dụng cả dữ liệu đã gán nhãn kết hợp với dữ liệu chưa gán nhãn.

1.3.4. Bán giám sát SVM và phân loại trang Web

Máy véc tơ hỗ trợ SVM là một thuật toán học có giám sát sử dụng dữ liệu đã gán nhãn, thì S3VM sử dụng hỗn hợp dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn. Mục đích là để gán các lớp nhãn tới lớp chưa gán nhãn một cách tốt nhất, sau đó sử dụng hỗn hợp dữ liệu huấn luyện đã gán nhãn và dữ liệu chưa gán nhãn sau khi đã gán nhãn để phân lớp những dữ liệu mới. Nếu dữ liệu chưa gán nhãn rỗng thì phương pháp này trở thành phương pháp chuẩn SVM để phân lớp. Nếu dữ liệu đã gán nhãn rỗng, thì phương pháp này sẽ trở thành hình thể học không giám sát. Học bán giám sát xảy ra khi cả dữ liệu đã gán nhãn và chưa gán nhãn không rỗng.

Để hiểu một cách rõ ràng cụ thể về S3VM, thì chúng ta cần hiểu về SVM đã được trình bày ở trên, trong nghiên cứu này tìm hiểu về thuật toán S3VM là bài toán phân lớp nhị phân.

Cho trước một tập huấn luyện gồm những dữ liệu đã gán nhãn cùng với tập dữ liệu chưa gán nhãn bao gồm n dữ liệu. Mục đích là gán nhãn cho những dữ liệu chưa gán nhãn này.

Với hai lớp đã cho trước gồm lớp dương (lớp $+1$) và lớp âm (lớp -1). Mỗi dữ liệu được xem như một điểm trong không gian véc tơ. Mỗi điểm i thuộc tập dữ liệu huấn luyện có một sai số là η_i và mỗi điểm j thuộc dữ liệu chưa gán nhãn sẽ có hai sai số ξ_j (sai số phân lớp với giả sử rằng j thuộc lớp $+1$) và z_j (sai số phân lớp với giả sử rằng j thuộc lớp -1).

Thuật toán S3VM sẽ giải bài toán tối ưu sau (1.9) thay cho bài toán tối ưu (1.7) ở thuật toán SVM.

$$\min_{\mathbf{w}, b, \eta, \xi, z} C = \left[\sum_{i=1}^n \eta_i + \sum_{j=n+1}^{n+k} \min(\xi_j, z_j) \right] + \|\mathbf{W}\|$$

$$\begin{aligned} \text{Với} \quad & y_i(\mathbf{W} \cdot x_i - b) + \eta_i \geq 1 \quad \eta_i \geq 0 \quad i = 1 \dots n \\ & \mathbf{W} \cdot x_j - b + \xi_j \geq 1 \quad \xi_j \geq 0 \quad j = n+1 \dots n+k \\ & -(\mathbf{W} \cdot x_j - b) + z_j \geq 1 \quad z_j \geq 0 \end{aligned} \quad (1.9)$$

Sau khi đã tìm được ξ_i và z_j , chúng ta sẽ có được sai số nhỏ nhất của mỗi điểm j . Nếu $\xi_i < z_j$ thì điểm j thuộc lớp dương, ngược lại nếu $\xi_i > z_j$ thì điểm j thuộc lớp âm. Quá trình này diễn ra trên tất cả các điểm thuộc dữ liệu chưa gán nhãn, sau khi quá trình này đã hoàn thành, tất cả các điểm chưa gán nhãn sẽ được gán nhãn.

Tập dữ liệu chưa gán nhãn sau khi đã gán nhãn sẽ được đưa vào tập dữ liệu huấn luyện, tiếp theo đó sẽ sử dụng thuật toán SVM để học tạo ra SVM mới, SVM này chính là S3VM có một siêu phẳng mới. Sau đó áp dụng siêu phẳng này để phân lớp các mẫu dữ liệu mới được đưa vào.

1.3.5. Thuật toán phân loại văn bản

Có nhiều thuật toán phân loại văn bản thực hiện tốt như: thuật toán K - láng giềng gần nhất (kNN), cây quyết định hay Naïve Bayes, ... Ở đây, trình bày thuật toán Naïve Bayes được sử dụng trong thực nghiệm của nghiên cứu.

Thuật toán Naive Bayes

Thuật toán phân loại Naive Bayes thừa nhận một giả thiết là các đặc trưng là độc lập lẫn nhau. Thêm vào đó, bộ phân loại xác suất lựa chọn một vài dạng giả định cho phân phối của mỗi đặc trưng trong một lớp. Những mô hình xác suất phổ biến nhất là mô hình đa thức, mô hình độc lập nhị phân.

Qua tìm hiểu các mô hình phân loại Naive Bayes, chúng tôi quyết định sử dụng mô hình đa thức vì nó đã được chứng minh là tốt nhất so với các mô hình còn lại trong nhiều trường hợp của phân loại văn bản [50], [77]. Mô hình đa thức biểu diễn văn bản bằng tập các lần xuất hiện của các từ. Mô hình không quan tâm đến trật tự của từ mà chỉ quan tâm đến số lần xuất hiện của các từ trong một văn bản.

Nội dung mô hình học phân lớp đa thức Naïve Bayes được mô tả như sau. Giả thiết rằng văn bản được tạo ra bởi một mô hình trộn (mixture model) với tham số θ . Mô hình trộn bao gồm các thành phần trộn $c_j \in C = \{c_1, \dots, c_{|C|}\}$. Mỗi một văn bản d_i được tạo ra bằng cách:

- Lựa chọn một thành phần dựa theo các ưu tiên của nó, $P(c_j; \theta)$
- Sau đó, mô hình trộn tạo ra văn bản dựa trên các tham số của nó, với phân phối $P(d_i | c_j; \theta)$.

Chúng ta mô tả khả năng của một văn bản là tổng xác suất của tất cả các thành phần trộn.

$$P(d_i | \theta) = \sum_{j=1}^{|C|} P(c_j | \theta) P(d_i | c_j; \theta) \quad (1.10)$$

Mỗi văn bản có một nhãn lớp, giả sử rằng có sự tương ứng một-một giữa nhãn lớp và thành phần của mô hình trộn, vì vậy, ta sẽ sử dụng c_j vừa để biểu diễn thành phần trộn thứ j vừa biểu diễn phân lớp thứ j . Trong mô hình đa thức, ta giả thiết rằng:

- Độ dài của văn bản là độc lập với phân lớp của nó.
- Giả thiết Naive Bayes: Xác suất sự xuất hiện của từ trong một văn bản là độc lập với ngữ cảnh và vị trí của từ trong văn bản đó.

Vì vậy, mỗi văn bản d_i được tạo ra từ phân phối đa thức của các từ với nhiều lần thử nghiệm độc lập với độ dài của văn bản. Ta định nghĩa N_{it} là số lần xuất hiện của từ w_t trong văn bản d_i , thì xác suất của văn bản d_i khi biết trước phân lớp đơn giản là phân phối đa thức như:

$$P(d_i | c_j; \theta) = P(|d_i|) |d_i| \prod_{t=1}^{|V|} \frac{P(w_t | c_j; \theta)^{N_{it}}}{N_{it}!} \quad (1.11)$$

Dựa vào công thức, ta tính toán tối ưu Bayes cho những đánh giá này từ tập dữ liệu huấn luyện. Ở đây, ước lượng cho xác suất của từ w_t trong văn bản thuộc lớp c_j được tính theo:

$$P(w_t | c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_j | d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j | d_i)} \quad (1.12)$$

với $P(c_j | d_i) \in \{0, 1\}$ được xác định bởi nhãn lớp tương ứng của mỗi mẫu dữ liệu. Xác suất ưu tiên của mỗi lớp được tính đơn giản dựa trên mỗi lớp thay vì trên các từ như:

$$P(c_j) = \frac{\sum_{i=1}^{|D|} P(c_j | d_i)}{|D|} \quad (1.13)$$

Từ việc ước lượng các tham số trên dữ liệu huấn luyện theo các phương trình (1.11), (1.12), và (1.13) ta thực hiện phân loại các văn bản kiểm thử và lựa chọn phân loại với xác suất cao nhất theo qui tắc Bayes.

$$P(c_j | d_i) = \frac{P(c_j) P(d_i | c_j)}{P(d_i)} \quad (1.14)$$

Vì tính xác suất cho cùng một văn bản và do giả thiết Naive Bayes nên công thức (1.14) sẽ tương đương với công thức (1.15):

$$P(c_j | d_i) \propto P(c_j) P(d_i | c_j) = P(c_j) \prod_{k=1}^{|d_i|} P(W_{d_{i,k}} | c_j) \quad (1.15)$$

1.4. Bài toán phân loại văn bản

1.4.1. Văn bản

Văn bản được hiểu theo nghĩa rộng là một thực thể mang thông tin được ghi bằng ký hiệu ngôn ngữ của con người. Văn bản dùng để lưu trữ, ghi nhận và truyền đạt thông tin từ người này đến người khác. Có nhiều hình thức thể hiện văn bản. Thể hiện được dùng rộng rãi nhất là thể hiện trên giấy như hoàng phi, câu đối, tác phẩm văn học, khoa học kỹ thuật, công văn, khẩu hiệu, ... Ngoài ra còn có các thể hiện bằng âm thanh như băng ghi âm, đĩa nghe và thể hiện bằng bản vẽ, ... Hiện nay, với sự phát triển của khoa học máy tính, việc lưu trữ hay truyền tải thông tin còn có thể trên các tập tin như “.txt”, “.pdf”, hay “.doc”, “.docx”. Vì vậy những tập tin này cũng có thể được gọi là văn bản. Vì tất cả mọi thông tin, dữ liệu trên máy tính đều được lưu trữ dưới dạng hệ cơ số nhị phân, nên nghiên cứu này định nghĩa những văn bản thể hiện trên máy tính là “Văn bản số”. Cụ thể hơn, khi các văn bản số được viết bởi ngôn ngữ tiếng Việt thì gọi là “Văn bản số tiếng Việt”.

1.4.2. Biểu diễn văn bản bằng véc tơ đặc trưng

Như đã trình bày ở phần trước, bước đầu tiên trong qui trình phân lớp văn bản là

thao tác chuyển văn bản đang được mô tả dưới dạng chuỗi các từ thành một mô hình khác, sao cho phù hợp với các thuật toán phân lớp.

Thông thường người ta biểu diễn văn bản dưới dạng một véc tơ đặc trưng, cụ thể là véc tơ có trọng số. Ý tưởng của mô hình này là xem mỗi văn bản D_i được biểu diễn theo dạng $D_i = \vec{d}_i, i$, trong đó i là chỉ số dùng để nhận diện văn bản này và \vec{d}_i là véc tơ đặc trưng của văn bản D_i này, trong đó $\vec{d}_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ và n là số lượng đặc trưng của véc tơ văn bản, w_{ij} là trọng số của đặc trưng thứ j .

Trong quá trình chuyển thể văn bản sang dạng véc tơ đặc trưng, vấn đề mà chúng ta cần quan tâm là việc lựa chọn đặc trưng và số chiều cho không gian véc tơ, chọn bao nhiêu từ, chọn các từ nào, phương pháp chọn ra sao.

Việc lựa chọn phương pháp biểu diễn văn bản để áp dụng vào bài toán phân lớp tùy thuộc vào độ thích hợp, phù hợp, độ đo đánh giá mô hình phân lớp của phương pháp đó sử dụng so với bài toán mà chúng ta đang xem xét giải quyết. Ví dụ nếu văn bản là một trang Web thì sẽ có phương pháp để lựa chọn đặc trưng khác so với các loại văn bản khác.

Khi biểu diễn văn bản dưới dạng véc tơ, ta thấy chúng có các đặc điểm sau:

- Số chiều không gian đặc trưng thường rất lớn. Các văn bản càng dài, lượng thông tin nó đề cập đến nhiều vấn đề thì không gian đặc trưng càng lớn.
- Các đặc trưng độc lập khác nhau, sự kết hợp các đặc trưng này thường không có ý nghĩa trong phân loại.
- Các đặc trưng có tính rời rạc. Véc tơ đặc trưng \vec{d}_i có thể có nhiều thành phần mang giá trị 0, do đó có nhiều đặc trưng không xuất hiện trong văn bản \vec{d}_i (nếu chúng ta tiếp cận theo cách sử dụng giá trị nhị phân 0, 1 để biểu diễn cho việc có xuất hiện hay không một đặc trưng nào đó trong văn bản đang được biểu diễn thành véc tơ). Tuy nhiên, nếu đơn thuần cách tiếp cận sử dụng giá trị nhị phân 0, 1 này thì kết quả phân loại phần nào hạn chế là do có thể đặc trưng đó không có trong văn bản đang xét, nhưng trong văn bản đang xét lại có từ khóa khác với từ đặc trưng

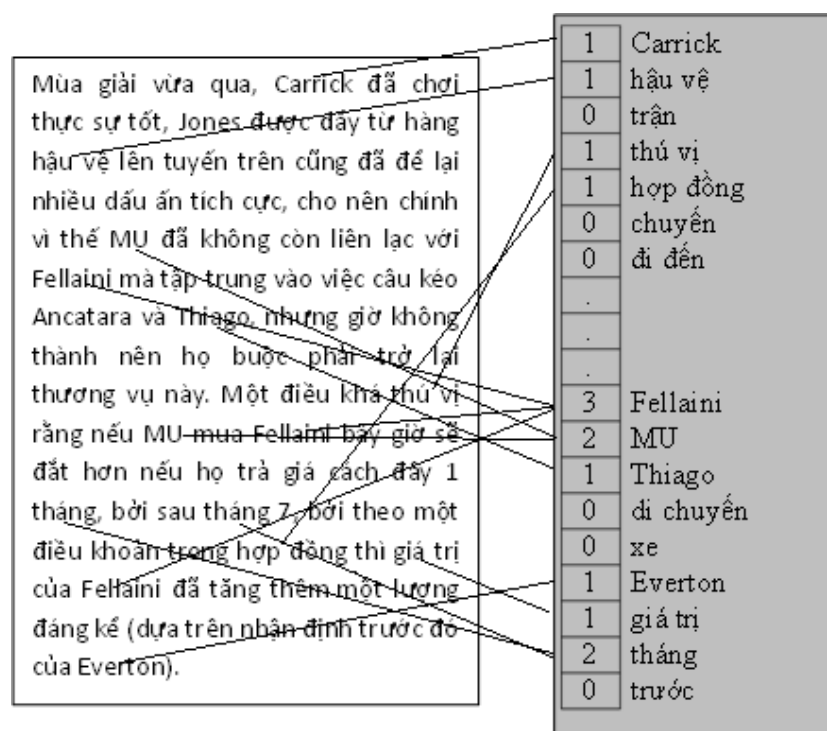
nhưng có ngữ nghĩa giống với từ đặc trưng này, do đó một cách tiếp cận khác là không sử dụng số nhị phân 0, 1 mà sử dụng giá trị số thực để phân nào giảm bớt sự rời rạc trong véc tơ văn bản.

Ví dụ: Từ điển (kịch bản, đội hình, bóng, cầu thủ)

$d1 = \text{"Đội hình hôm nay có các cầu thủ sau..."}$ $\rightarrow d1 = (0, 1, 0, 1)$

Hầu hết các văn bản có thể được phân chia một cách tuyến tính bằng các hàm tuyến tính. Như vậy, độ dài của véc tơ là số các từ khoá xuất hiện trong ít nhất một mẫu dữ liệu huấn luyện. Trước khi đánh trọng số cho các từ khoá cần tiến hành loại bỏ các từ dừng. Từ dừng là những từ thường xuất hiện nhưng không có ích trong việc đánh chỉ mục, nó không có ý nghĩa gì trong việc phân loại văn bản. Có thể nêu một số từ dừng trong tiếng Việt như “và”, “là”, “thì”, “như vậy”, ..., trong tiếng Anh như “and”, “or”, “the”, ... Thông thường từ dừng là các trạng từ, liên từ, giới từ.

Có thể lấy một ví dụ về việc biểu diễn văn bản dưới dạng véc tơ đặc trưng trọng số như sau:



Hình 1.5 Véc tơ đặc trưng biểu diễn văn bản mẫu

1.4.3. Phân loại văn bản

Phân loại văn bản là quá trình phân tích và gán một văn bản vào một hay nhiều lớp cho trước nhờ một mô hình phân loại. Mô hình phân loại này được xây dựng dựa trên một tập hợp các văn bản đã gán nhãn từ trước (đã xác định tên chủ đề trước) gọi là tập dữ liệu huấn luyện. Tập dữ liệu huấn luyện là tập các trang văn bản đã gán nhãn lớp tương ứng từng chủ đề. Quá trình xây dựng tập dữ liệu huấn luyện này thường được thực hiện bằng con người. Sau đó, mô hình được sử dụng để phân loại các trang văn bản chưa gán nhãn.

Bộ phân lớp có thể được xây dựng bằng tay dựa vào các kỹ thuật ứng dụng tri thức (thường là xây dựng một tập các tri thức) hoặc có thể được xây dựng một cách tự động bằng các kỹ thuật học máy thông qua một tập các dữ liệu huấn luyện được định nghĩa trước phân lớp tương ứng. Trong hướng tiếp cận học máy, ta chú ý đến các vấn đề sau:

- *Biểu diễn văn bản*: Một văn bản thông thường được biểu diễn bằng một véc tơ trọng số, độ dài của véc tơ là số các từ khóa xuất hiện trong ít nhất một mẫu dữ liệu huấn luyện. Biểu diễn trọng số có thể là nhị phân (từ khóa đó có hay không xuất hiện trong văn bản tương ứng) hoặc không nhị phân (từ khóa đó đóng góp tỷ trọng bao nhiêu cho ngữ nghĩa văn bản). Tồn tại một số phương pháp biểu diễn từ khóa điển hình như IDF, TF, TF-IDF, ...

- *Loại bỏ các từ dừng và lấy từ gốc*: Trước khi đánh trọng số cho các từ khóa cần tiến hành loại bỏ các từ dừng (*stop-word*). Từ điển Wikipedia định nghĩa: “Từ dừng là những từ xuất hiện thường xuyên nhưng lại không có ích trong đánh chỉ mục cũng như sử dụng trong các máy tìm kiếm hoặc các chỉ mục tìm kiếm khác”. Thông thường, các trạng từ, giới từ, liên từ là các từ dừng. Tuy nhiên, có thể liệt kê danh sách các từ dừng cho tiếng Việt mặc dù có thể là không đầy đủ. Việc lấy từ gốc và lưu lại các từ phát sinh từ mỗi từ gốc để nâng cao khả năng tìm kiếm được áp dụng cho các ngôn ngữ tự nhiên có chia từ.

- *Tiêu chuẩn đánh giá*: Phân loại văn bản được coi là không mang tính khách

quan theo nghĩa dù con người hay bộ phân loại tự động thực hiện việc phân loại thì đều có thể xảy ra sai sót. Tính đa nghĩa của ngôn ngữ tự nhiên, sự phức tạp của bài toán phân loại được coi là những nguyên nhân điển hình nhất của sai sót phân loại. Hiệu quả của bộ phân loại thường được đánh giá qua so sánh quyết định của bộ phân loại đó với quyết định của con người khi tiến hành trên một tập kiểm thử (test set) các văn bản đã gán nhãn lớp trước.

Phân loại văn bản là một giai đoạn quan trọng được sử dụng để hỗ trợ trong quá trình tìm kiếm thông tin, chiết lọc thông tin, lọc văn bản hoặc tự động dẫn đường cho các văn bản tới những chủ đề xác định trước, ...

Bài toán phân loại văn bản có rất nhiều ứng dụng trong thực tế, điển hình là các ứng dụng lọc trên Internet.

Bài toán phân loại văn bản có thể được phát biểu như sau:

Cho trước một tập văn bản $D = \{d_1, d_2, \dots, d_n\}$, d_i là văn bản thứ i và tập lớp $C = \{c_1, c_2, \dots, c_m\}$, c_j là lớp thứ j . Mục đích của bài toán là xác định và gán văn bản d_i thuộc về lớp c_j đã được định nghĩa. Mục tiêu của bài toán là đi tìm hàm f :

$$f : D \times C \rightarrow \{\text{True}, \text{False}\}$$

Trong đó: $f(d_i, c_j) = \text{True}$, nếu văn bản d_i thuộc lớp c_j

$$f(d_i, c_j) = \text{False}, \text{ nếu văn bản } d_i \text{ không thuộc lớp } c_j$$

Có nhiều bài toán phân loại văn bản như: phân lớp nhị phân (chỉ cần xác định một văn bản có thuộc một lớp cho trước hay không), phân lớp đa lớp (một văn bản thuộc một lớp nào đó trong danh sách các lớp cho trước), phân lớp đa trị (một văn bản có thể thuộc nhiều hơn một lớp trong danh sách các lớp cho trước).

Nếu có 3 lớp c_1, c_2, c_3 thì ta sẽ có các tổ hợp $(c_1, c_2), (c_1, c_3), (c_2, c_3)$

- Nếu có n lớp thì có $n * (n-1) / 2$ tổ hợp

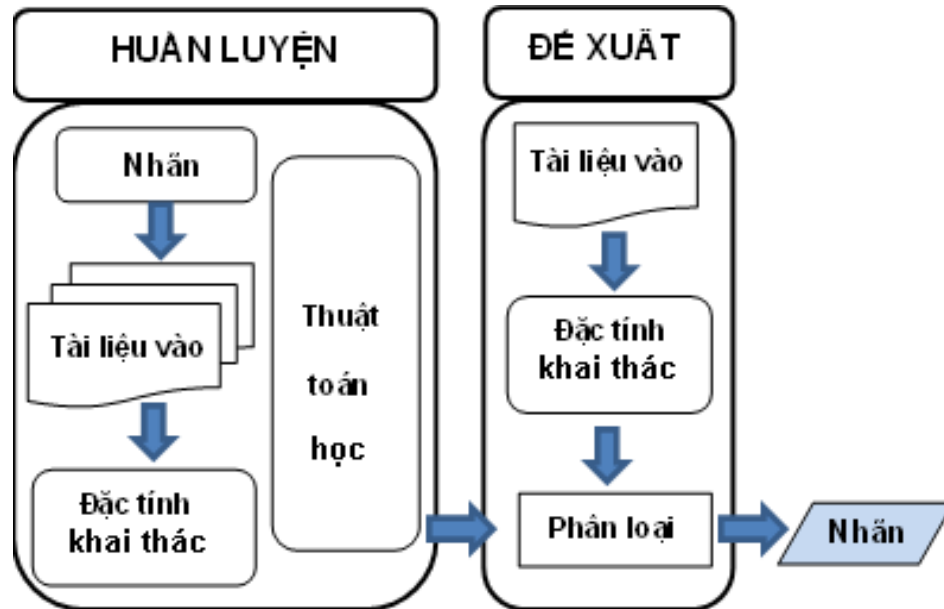
Ví dụ: Có văn bản d thông qua tổ hợp $(c_1, c_2) \rightarrow c_2$

d thông qua tổ hợp $(c_1, c_3) \rightarrow c_3$

d thông qua tổ hợp $(c_2, c_3) \rightarrow c_2$

$\rightarrow d$ thuộc lớp c_2 vì kết quả là c_2 xuất hiện nhiều nhất.

a. Mô hình tổng quát



Hình 1.6 Mô hình tổng quát của hệ thống phân loại văn bản

b. Các bước phân loại

Để tiến hành phân loại văn bản nói chung, chúng ta sẽ thực hiện các bước sau:

- **Bước 1:** Xây dựng bộ dữ liệu chủ quan dựa vào tài liệu văn bản đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lý và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng chủ đề.

- **Bước 2:** Dữ liệu cần phân loại được xử lý, rút ra đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và rút ra kết quả.

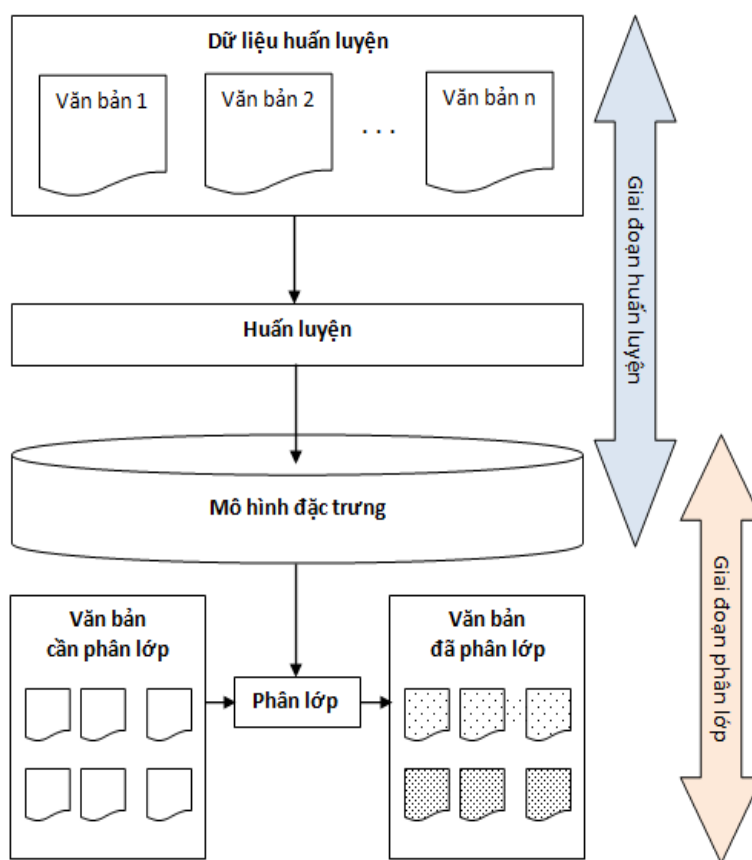
Dữ liệu đầu vào cho quá trình học máy hay dữ liệu đầu vào để phân loại đều là dạng văn bản đã qua công đoạn tiền xử lý. Công đoạn tiền xử lý này rất quan trọng và cần thiết, nó làm tối ưu hóa dữ liệu trong việc lưu trữ và xử lý. Các công đoạn trong quá trình tiền xử lý văn bản bao gồm: tách từ tiếng Việt, loại bỏ các từ dừng, từ tầm thường. Sau đó, rút trích đặc trưng và biểu diễn văn bản.

1.5. Đề xuất nghiên cứu

Qua kết quả nghiên cứu tổng quan ở trên, ta thấy SVM là một phương pháp phân loại văn bản được sử dụng phổ biến nhất hiện nay. SVM trở thành mô hình học máy

phổ biến nhất khi phát triển các ứng dụng thực tế nhằm mục đích phân loại văn bản. Đặc trưng cơ bản quyết định khả năng phân lớp là khả năng phân lớp những dữ liệu mới dựa vào những tri thức đã tích lũy được trong quá trình huấn luyện. Sau quá trình huấn luyện nếu hiệu suất tổng quát hoá của bộ phân lớp cao thì thuật toán huấn luyện được đánh giá là tốt. Hiệu suất tổng quát hoá phụ thuộc vào hai tham số là sai số huấn luyện và năng lực của học máy. Trong đó sai số huấn luyện là tỷ lệ lỗi phân lớp trên tập dữ liệu huấn luyện. Còn năng lực của học máy được xác định bằng kích thước Vapnik-Chervonenkis. Kích thước V-C là một khái niệm quan trọng đối với một họ hàm phân tích (hay là tập phân lớp). Đại lượng này được xác định bằng số điểm cực đại mà họ hàm có thể phân tích hoàn toàn trong không gian. Một tập phân lớp tốt là tập phân lớp có năng lực thấp nhất (có nghĩa là đơn giản nhất) và đảm bảo sai số huấn luyện nhỏ.

Mô hình tổng quát để phân loại văn bản có thể được mô tả lại như sau:



Hình 1.7 Mô hình phân lớp văn bản

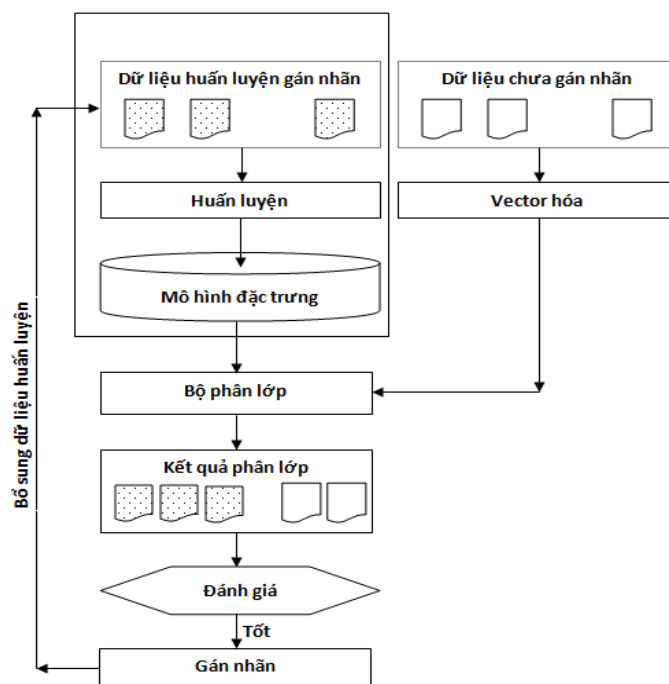
Đối với các phương pháp hiện tại, ta thường gặp phải 3 vấn đề như sau:

- Cần phải có một bộ dữ liệu huấn luyện tốt. Tuy nhiên, việc xây dựng các bộ dữ liệu huấn luyện đối với các ngôn ngữ ít phổ biến đang là một vấn đề lớn.
- Cần phải có một phương pháp, giải thuật hiệu quả để sử dụng trong bước huấn luyện và phân lớp. Phương pháp và giải thuật này phụ thuộc vào đặc điểm của ngôn ngữ, độ lớn của các dữ liệu.
- Khi biểu diễn văn bản bằng véc tơ đặc trưng thì số chiều của véc tơ là rất lớn nên đòi hỏi thời gian và chi phí tính toán rất lớn.

Nhằm góp phần giải quyết các vấn đề trên, tôi đề xuất 3 nội dung chính như sau:

- Xây dựng kho dữ liệu phục vụ phân loại văn bản tiếng Việt.
- Ứng dụng mô hình cự li trắc địa để cải tiến phương pháp và giải thuật phân lớp.
- Ứng dụng mô hình đồ thị Dendrogram để giảm số chiều của véc tơ.

Bên cạnh đó, tôi đề xuất mô hình cải tiến dựa trên học máy bán giám sát và mô hình không gian véc tơ như sau:



Hình 1.8 Mô hình đề xuất phân lớp văn bản sử dụng Self-training

Những giải pháp này sẽ được trình bày chi tiết trong các chương tiếp theo.

1.6. Tiểu kết chương

Chương này trình bày các kết quả nghiên cứu tổng quan về học máy, các ứng dụng, các dạng dữ liệu, các phương pháp học máy, tìm hiểu một số phương pháp, thuật toán học bán giám sát, thuật toán máy véc tơ hỗ trợ SVM áp dụng vào phân loại văn bản tiếng Việt và đề xuất giải pháp nhằm nâng cao chất lượng phân loại bằng mô hình phân loại văn bản, góp phần giải quyết các vấn đề nêu trên triển khai 3 nội dung:

- Xây dựng kho dữ liệu phục vụ phân loại văn bản tiếng Việt.
- Trình bày giải pháp sử dụng mô hình cự ly trắc địa trong phân loại văn bản.
- Đề xuất giải pháp gom cụm, rút gọn số chiều véc tơ phục vụ phân loại văn bản dựa trên đồ thị Dendrogram.

Từ mô hình trên đề xuất mô hình cải tiến dựa trên học bán giám sát và mô hình không gian véc tơ. Nội dung của chương là cơ sở quan trọng để triển khai các nội dung nghiên cứu đề xuất sẽ trình bày trong các chương sau.

Chương 2. XÂY DỰNG KHO DỮ LIỆU

Chương này giới thiệu các vấn đề cơ bản về kho dữ liệu như kiến trúc kho dữ liệu phục vụ cho phân loại văn bản tiếng Việt, phân tích các đặc tả dữ liệu, thiết kế kho dữ liệu và đưa ra giải pháp xây dựng kho dữ liệu để áp dụng phân loại văn bản tiếng Việt ở các chương sau.

2.1. Giới thiệu kho dữ liệu phân loại văn bản tiếng Việt

a. Giới thiệu

Ứng dụng kỹ thuật học bán giám sát vào phân loại văn bản tiếng Việt là lĩnh vực đang được các nhà nghiên cứu trong và ngoài nước quan tâm, để phục vụ phân loại văn bản, trước hết chúng ta phải có kho dữ liệu, cho đến nay vẫn chưa có kho dữ liệu văn bản tiếng Việt chuẩn để phục vụ cho phân loại văn bản tiếng Việt [84]. Các văn bản tiếng Việt được sử dụng trong những nghiên cứu trước đây của các nhà nghiên cứu Việt Nam đều được tạo bởi chính họ và chưa được kiểm chứng, do đó kết quả nghiên cứu mang nhiều tính chủ quan. Trong khi đó thế giới đã có nhiều kho dữ liệu được sử dụng rộng rãi với nhiều ngôn ngữ khác nhau, đặc biệt tiếng Anh như Reuters-21578, the RCV1 và 20 News Group [18], [19]. Việc xây dựng một kho dữ liệu lớn đây là một công việc mất rất nhiều thời gian, công sức. Các kho dữ liệu trên thế giới ra đời, đều trải qua một quá trình để từng bước hoàn thiện và tạo ra một kho dữ liệu đầy đủ. Chính vì lý do này việc xây dựng kho dữ liệu chuẩn là cần thiết.

Tuy nhiên, số lượng dữ liệu lớn không nói lên được nhiều điều, vì vấn đề quan trọng hơn đó chính là chất lượng của kho dữ liệu. Chất lượng của kho dữ liệu chính là độ phù hợp khi gán một văn bản mẫu vào một chủ đề. Đối với một văn bản mẫu có nội dung không rõ ràng thuộc chủ đề nào thì nó sẽ khó được sắp xếp ở vị trí nào trong các chủ đề liên quan, trong nghiên cứu này tôi sẽ sử dụng thuật toán Naïve Bayes để phân loại văn bản thông qua đó kiểm thử độ xác thực việc gán chủ đề lên văn bản mẫu [38], [50].

b. Mục đích của kho dữ liệu phục vụ phân loại văn bản tiếng Việt

Kho dữ liệu nhằm hỗ trợ để tổ chức thực hiện tốt, hiệu quả công việc phục vụ phân loại văn bản tiếng Việt, như có những quyết định hợp lý, nhanh một cách hiệu

quả và chính xác. Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau. Dữ liệu trong kho phải được xử lý để giảm thời gian và độ phức tạp khi phân loại văn bản. Xác định và làm sạch những dữ liệu thừa, không quan trọng của tài liệu giúp cho hệ thống phân loại văn bản tiếng Việt xác định độ tương tự giữa tài liệu cần phân loại và tập mẫu được hiệu quả hơn.

2.2. Tổng quan về kho dữ liệu

2.2.1. Khái niệm kho dữ liệu

Kho dữ liệu là tập hợp dữ liệu tương đối ổn định (ít hay thay đổi), được cập nhật theo thời gian và được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình tạo quyết định về mặt quản lý trong huấn luyện và kiểm thử cụ thể:

- Chứa số lượng lớn dữ liệu có liên quan trong quá khứ, thông tin luôn được cập nhật, truy xuất nhanh, không giới hạn kích thước.
- Được tối ưu hóa cho các thao tác đọc trong các yêu cầu truy vấn dữ liệu. Điều này đối lập với các cơ sở dữ liệu trong các hệ thống xử lý tác vụ được thiết kế để hỗ trợ cho tất cả các thao tác cập nhật, thay đổi, chỉnh sửa dữ liệu.
- Tải lên các dữ liệu mới hoặc dữ liệu được cập nhật định kỳ, rõ ràng và đồng nhất, dữ liệu được chuẩn hóa theo một chuẩn chung.

Kho dữ liệu gồm những đặc tính sau:

- **Hướng chủ đề:** nghĩa là dữ liệu sẽ cung cấp thông tin về một chủ đề cụ thể hơn. Kho dữ liệu theo hướng chủ đề nên nó sẽ cho phép phân tích thông tin được kết nối với một chủ đề cụ thể nào đó, để hỗ trợ trong việc phân tích dữ liệu.
- **Tích hợp:** là dữ liệu được thu thập trong kho dữ liệu có thể đến từ nhiều nguồn khác nhau, nhưng được kết hợp với nhau thành một thể thống nhất.
- **Tính ổn định:** có nghĩa là sẽ không có việc cập nhật dữ liệu được lưu trữ trong kho dữ liệu mà thay vào đó là các thông tin được tổ chức để hiển thị các thay đổi của dữ liệu đó. Dữ liệu trong kho được sử dụng cho việc phân tích nên các thao tác cập nhật hay xóa có thể làm ảnh hưởng tới việc phân tích này. Vì vậy, dữ liệu trong kho không bao giờ được cập nhật và xóa bỏ. Khi nào một thuộc tính cụ thể hoặc mục dữ liệu được cập nhật tại nguồn thì phiên bản mới của nó được lưu trữ trong kho dữ liệu để vô hiệu hóa phiên bản dữ liệu cũ.

- **Có tính lịch sử:** các thông tin trong kho dữ liệu được cập nhật tập trung theo thời gian và lưu trữ lâu dài, toàn bộ lịch sử dữ liệu được lưu vết.

- **Gắn thời gian:** kho dữ liệu lưu trữ dữ liệu từ quá khứ cũng như hiện tại, mỗi tập tin chứa một yếu tố thời gian như một phần của khóa chính để bảo đảm tính duy nhất của mỗi tập tin và cung cấp một đặc trưng về thời gian cho dữ liệu. Toàn bộ dữ liệu trong kho được tạo ra và gắn với một giá trị thời gian nhất định.

Kho dữ liệu phục vụ phân loại văn bản là một tập hợp các văn bản được tạo ra, gồm tập dữ liệu huấn luyện (training) và tập dữ liệu kiểm thử (testing).

- Tập dữ liệu huấn luyện: chứa các văn bản đã được gán vào các chủ đề cho trước, dùng để huấn luyện cho giải thuật “máy học” cách nhận biết chủ đề của các văn bản, máy học bằng cách tập hợp các từ trong tập văn bản này vào cơ sở tri thức. Do đó tập dữ liệu cho giai đoạn này cực kỳ quan trọng trong việc phân loại văn bản với cơ sở tri thức đó.

- Tập dữ liệu kiểm thử: dùng để đánh giá tính khả thi và độ chính xác của giải thuật phân loại sau khi xây dựng thành công chương trình.

2.2.2. Đặc điểm của kho dữ liệu

Trước tiên kho dữ liệu là cơ sở dữ liệu lớn, kho dữ liệu thường chỉ đọc, kho dữ liệu hướng về tính ổn định, thông tin có thể lấy từ nhiều nguồn khác nhau, thông tin đưa vào sẽ được làm sạch và đưa vào cấu trúc của dữ liệu đó chính là cơ sở dữ liệu rất lớn. Kho dữ liệu rất lớn có thể khai thác thông tin dễ dàng thì bản thân kho dữ liệu phải được chuyển hóa, phân ra thành những chủ đề do đó những chủ đề chuyên môn hóa đó tạo thành một cơ sở dữ liệu chuyên biệt đó là dữ liệu chủ đề. Mọi quản trị cơ sở dữ liệu hỗ trợ cho việc truy vấn thông tin trong dữ liệu chủ đề rồi đưa ra quyết định, nhận định những thông tin trong dữ liệu chủ đề đó là OLAP (On line Analytical Processing) là bộ phân tích trực tuyến. Để đảm bảo độ chính xác cao trong kết quả phân loại cuối cùng thì không chỉ cần một thuật toán tốt, đáng tin cậy mà cần phải có một kho dữ liệu tốt. Điều kiện đủ của một kho dữ liệu tốt là: nguồn gốc, tính đầy đủ, tính hiệu quả.

a. Nguồn gốc: Một vấn đề luôn luôn được đặt ra khi xây dựng một kho dữ liệu, đó là dữ liệu sẽ được lấy ở đâu? Nguồn gốc của một kho dữ liệu chính là nơi mà

người xây dựng kho lấy về, từ đó các dữ liệu thô được tinh chỉnh thành các dữ liệu dùng trong kho. Do đó, nếu nguồn gốc của dữ liệu đáng tin cậy, cơ sở dữ liệu lớn thì kho dữ liệu có các văn bản đầy đủ và khá chính xác.

b. Tính đầy đủ: Một kho dữ liệu tốt nếu như nó cung cấp đủ các thành phần mà người dùng cần. Tức là kho dữ liệu phải có văn bản học thì đúng là văn bản học, văn bản để kiểm thử thì đúng là văn bản kiểm thử. Các dữ liệu trong kho phải có nhiệm vụ và vị trí rõ ràng, không có sự mập mờ cũng như dư thừa hay thiếu sót.

c. Tính hiệu quả: được đánh giá trên hai mặt: tốc độ và sự chính xác. Với hai kho dữ liệu như nhau, nếu ta cùng kiểm nghiệm một thuật toán thì việc sử dụng kho dữ liệu với thời gian nhanh hơn sẽ giúp ta tiết kiệm thời gian. Nhưng nếu chỉ nhanh thì không đủ, điều ta cần là phải chính xác.

Tính hiệu quả của một kho dữ liệu sẽ được tăng lên qua một thời gian dài sử dụng và liên tục chỉnh sửa, cập nhật. Để thu được một kho dữ liệu hoàn chỉnh và có hiệu quả cao cần có một thời gian dài, xây dựng và phát triển. Và chính các kết quả thực nghiệm kho dữ liệu sẽ khẳng định nó có hiệu quả hay không, cụ thể hơn là có dùng được hay không.

2.2.3. Mục đích của kho dữ liệu

Đáp ứng mọi yêu cầu thông tin cho người sử dụng. Thông tin phải trực quan và dễ hiểu với người dùng. Hỗ trợ đưa ra những quyết định nhanh và hợp lý.

- Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau.
- Thông tin trong kho dữ liệu phải đảm bảo tính nhất quán.
- Thích nghi với sự thay đổi và có tính bảo mật cao.

Để đạt được những mục tiêu trên cần thực hiện các công việc sau:

- Truy cập dễ dàng: thông tin lưu trữ trong kho dữ liệu phải trực quan và dễ hiểu với người dùng, dữ liệu được trình bày thông qua các tên gọi quen thuộc và gắn gũi với nhiệm vụ người dùng. Nâng cao chất lượng dữ liệu bằng phương pháp làm sạch, dữ liệu được truy xuất dễ dàng, hệ thống dữ liệu một cách nhất quán, thích nghi và thay đổi linh hoạt. Tốc độ truy cập nhanh, do phải xử lý số lượng tập tin lớn cùng một lúc nên đây là một trong những yêu cầu phải có của một kho dữ liệu.

- Tính nhất quán: Dữ liệu trong kho thường đến từ nhiều nguồn khác nhau. Do vậy trước khi được đưa vào kho cần phải đảm bảo về chất lượng giúp cho việc đồng nhất dữ liệu trở nên dễ dàng. Tổng hợp và kết nối nguồn dữ liệu đồng thời đồng bộ hóa các nguồn dữ liệu với kho dữ liệu. Quản lý các siêu dữ liệu, đồng nhất các hệ cơ sở dữ liệu, dữ liệu phải kiểm soát việc truy cập một cách hiệu quả.

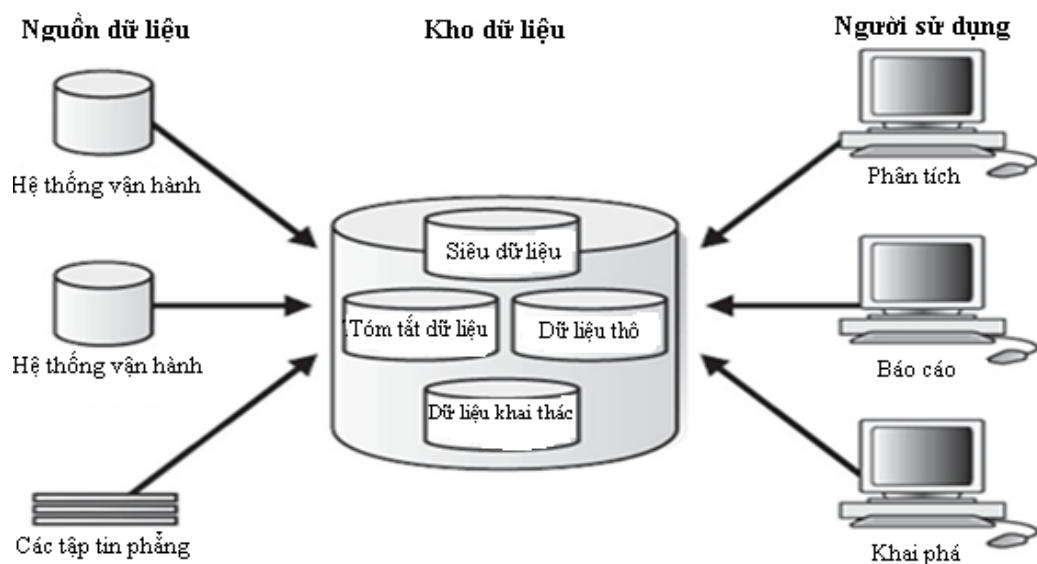
- Thích nghi với sự thay đổi: Dữ liệu cần phải được thiết kế để xử lý những thay đổi có thể xảy ra. Vì thay đổi là điều không thể tránh khỏi cho bất cứ ứng dụng nào, nói vậy có nhiều khi có thay đổi mới dữ liệu cũ vẫn phải đảm bảo tính đúng đắn. Tích hợp, tóm tắt và tổ chức dữ liệu theo từng chủ đề.

- Kho dữ liệu phải chính xác để hỗ trợ quá trình ra quyết định. Đây là mục tiêu quan trọng của yêu cầu xây dựng kho dữ liệu, những giá trị muốn đưa vào thông tin để từ đó đưa ra những chiến lược góp phần đem lại kết quả xử lý tốt nhất.

- Tính bảo mật: Dữ liệu trong kho đến từ nhiều nguồn khác nhau. Vì vậy việc bảo mật thông tin là một điều vô cùng quan trọng.

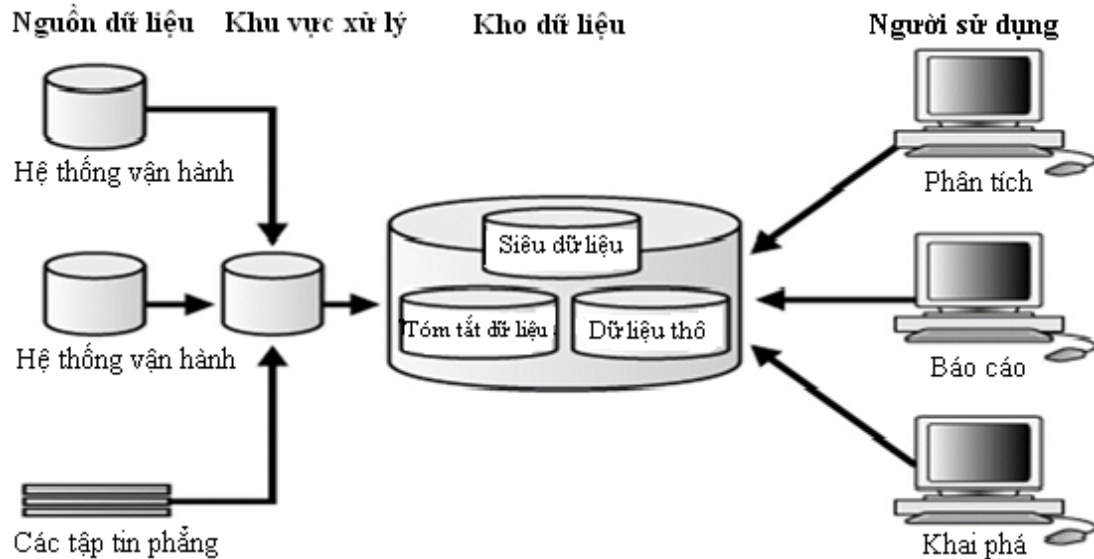
2.2.4. Kiến trúc kho dữ liệu

a. Kiến trúc DWH cơ bản: đây là kiến trúc đơn giản cho kho dữ liệu (Data warehouse (DWH))



Hình 2.1 Kiến trúc DWH cơ bản

b. Kiến trúc DWH với khu vực xử lý (Staging Area): thêm thành phần làm sạch và xử lý dữ liệu trước khi đưa vào kho.



Hình 2.2 Kiến trúc DWH với khu vực xử lý

Các thành phần của kho dữ liệu:

Nguồn dữ liệu (Data Sources): rất nhiều nguồn khác nhau và có cấu trúc dữ liệu khác nhau.

- Nguồn dữ liệu bên trong: là nguồn dữ liệu chính để xây dựng kho dữ liệu, chứa các dữ liệu chi tiết hiện tại của hệ thống tác nghiệp.
- Dữ liệu từ hệ thống phân tích: là dữ liệu được tổng hợp từ dữ liệu nguồn đã cũ và tổ chức lại theo nhiều phương pháp khác nhau.
- Dữ liệu từ bên ngoài: là các dữ liệu từ các nguồn ngoài hệ thống tác nghiệp như từ Internet. Có thể do các tổ chức khác thu thập và tạo ra, nó được sử dụng cho các yêu cầu phân tích dữ liệu.
- Dữ liệu từ các hệ thống nguồn thường hỗn tạp và chứa nhiều cấu trúc khác nhau ví dụ: các cơ sở dữ liệu, từ các tập tin Excel, các tập tin thô, hay dạng XML, ... Vì thế trước khi đưa vào kho dữ liệu cần phải chuyển đổi và tích hợp dữ liệu.

Khu vực xử lý (Staging Area): Ở khu vực này dữ liệu được sử dụng các kỹ thuật làm sạch và chuyển đổi để đảm bảo tính nhất quán dữ liệu trước khi đưa vào kho dữ liệu đích.

- Tiền xử lý: định dạng tài liệu và chuyển đổi tài liệu sang định dạng có thể chấp nhận được trong kho dữ liệu. Chứng thực và chuẩn hóa ngôn ngữ.
- Xử lý chính: dữ liệu đưa về hình thức chuẩn phù hợp cho việc tìm kiếm và khai phá dữ liệu hiệu quả. Quá trình này bao gồm: phân cụm, trích chọn và tổng hợp đặc trưng, lập chỉ mục.

Siêu dữ liệu (Metadata): là thành phần cơ bản để xây dựng và quản lý kho dữ liệu. Siêu dữ liệu không phải là dữ liệu phân tích, nó chỉ mô tả thông tin của tài liệu. Miêu tả dữ liệu trong quá trình xây dựng, quản lý và hoạt động của kho.

Kho dữ liệu (Data Warehouse): là cơ sở dữ liệu được tổ chức lại theo mô hình hình sao hay mô hình bông tuyết. Mô hình được phi chuẩn hóa, chấp nhận sự dư thừa dữ liệu trong lưu trữ dữ liệu chính vì thế mô hình dữ liệu đơn giản hơn nên việc truy vấn dễ dàng hơn và tốc độ xử lý cũng nhanh hơn mô hình dữ liệu chưa được chuẩn hóa. Ngoài ra kho dữ liệu còn chứa các siêu dữ liệu.

Kho dữ liệu chủ đề (Data Marts): Kho dữ liệu chủ đề có đặc điểm giống với kho dữ liệu nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chủ đề. Các kho dữ liệu chủ đề là một tập con của kho dữ liệu hoặc được xây dựng độc lập từ đó tích hợp lại thành kho dữ liệu.

2.3. Phân tích yêu cầu

Phân tích và định rõ yêu cầu là bước kỹ thuật đầu tiên trong tiến trình kỹ nghệ phần mềm. Tại bước này các phát biểu chung về phạm vi phần mềm được làm mịn thành một bản đặc tả cụ thể để trở thành nền tảng cho mọi hoạt động kỹ nghệ phần mềm sau đó. Việc phân tích phải tập trung vào các miền thông tin, chức năng và hành vi của vấn đề. Việc làm bản mẫu thường giúp chỉ ra cách tiếp cận khác để từ đó có thể làm mịn thêm yêu cầu. Đặc tả cần được xét duyệt để đảm bảo rằng người phát triển và khách hàng có cùng nhận biết về hệ thống cần phát triển.

2.3.1. Xây dựng kho

Quản trị hệ thống thông tin đáp ứng được những yêu cầu ở mức độ cao nghĩa là thông tin mang tính phân tích và có khả năng hỗ trợ quyết định. Tuy nhiên việc xây dựng một hệ thống như vậy vấp phải một số hạn chế về mặt kỹ thuật, đặc biệt là khi

kích thước cũng như độ phức tạp của môi trường thông tin tăng lên. Những hệ thống thông tin xây dựng theo phương pháp truyền thống không làm hài lòng người sử dụng và các nhà quản lý hệ thống thông tin. Những mục tiêu này không thể đạt được bởi dữ liệu ngày càng tăng, lưu trữ phân tán ở nhiều dạng không tương thích với nhau. Nhiều hệ cơ sở dữ liệu đã được xây dựng không tương thích với nhau, quản trị dữ liệu phức tạp. Giải pháp cho tất cả các vấn đề nêu trên chính là việc xây dựng một kho dữ liệu.

Những yêu cầu đặt ra khi xây dựng kho dữ liệu:

- Kho dữ liệu được xây dựng trực tuyến phục vụ yêu cầu mọi lúc mọi nơi.
- Kho dữ liệu được xây dựng dựa theo hai nguồn cung cấp: nguồn dữ liệu ban đầu trong quá trình xây dựng kho dữ liệu và nguồn dữ liệu do người quản trị tải lên trực tiếp vào kho khi kho dữ liệu hoàn thành.
- Dữ liệu của kho được sưu tập từ các bài viết trên website theo các chủ đề đã được xác định như: bóng đá, giáo dục, pháp luật, quốc tế, xã hội, ... nguồn dữ liệu đó được tổng hợp từ 4 website điện tử khác nhau được đọc nhiều nhất: vnexpress, tuoitre, dantri, vietnamnet.
- Dữ liệu trước khi đưa vào kho cần được mô tả thông tin liên quan như: chủ đề, ngày phát hành, phong chữ, kích thước, nguồn gốc văn bản, tác giả, văn bản (tiêu đề, nội dung tóm tắt (nếu có), câu đầu tiên của văn bản, nội dung).
- Số lượng bài được tải về lưu trữ kho tài liệu hiển thị danh sách tài liệu theo chủ đề, số lượng, dung lượng, định dạng dữ liệu chủ yếu dưới dạng .TXT và chưa được xử lý cụ thể:

Bảng 2.1 Dữ liệu thô tải về

| STT | Loại tài liệu | Số lượng bài đã tải về | Tổng dung lượng |
|-----|---------------|------------------------|-----------------|
| 1 | Bóng đá | 1512 | 363411 KB |
| 2 | Giáo dục | 1231 | 335561 KB |
| 3 | Pháp luật | 1194 | 175410 KB |
| 4 | Quốc tế | 1208 | 255815 KB |
| 5 | Xã hội | 1152 | 232633 KB |

2.3.2. Khai thác kho

Khai thác kho dữ liệu là một quá trình trích xuất thông tin có mối quan hệ hoặc có mối tương quan nhất định từ một kho dữ liệu lớn nhằm mục đích dự đoán các xu thế, các hành vi trong tương lai hoặc tìm kiếm những tập thông tin hữu ích mà bình thường không thể nhận diện được. Trên thực tế, khai thác kho dữ liệu chỉ là một bước thiết yếu trong quá trình khai thác tri thức trong cơ sở dữ liệu, quá trình này bao gồm các bước sau:

- **Bước 1:** Làm sạch dữ liệu là loại bỏ nhiễu hoặc các dữ liệu không thích hợp.
- **Bước 2:** Tập hợp dữ liệu là tích hợp dữ liệu từ nhiều nguồn khác nhau: Cơ sở dữ liệu, kho dữ liệu, file văn bản, ...
- **Bước 3:** Chọn dữ liệu ở bước này, những dữ liệu liên quan trực tiếp đến nhiệm vụ sẽ được thu thập từ các nguồn dữ liệu ban đầu.
- **Bước 4:** Chuyển đổi dữ liệu là dữ liệu sẽ được chuyển đổi về dạng phù hợp cho việc khai thác dữ liệu bằng cách thực hiện các thao tác nhóm hoặc tập hợp.
- **Bước 5:** Khai thác dữ liệu đây là giai đoạn thiết yếu, trong đó các phương pháp thông minh sẽ được áp dụng để trích xuất ra các mẫu dữ liệu.
- **Bước 6:** Đánh giá mẫu là đánh giá sự hữu ích của các mẫu biểu diễn tri thức dựa vào một số phép đo.
- **Bước 7:** Mô tả tri thức là sử dụng các kỹ thuật trình diễn và trực quan hóa dữ liệu để biểu diễn tri thức khai thác được cho người sử dụng, quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là quá trình lặp đi lặp lại các bước.

Dữ liệu của kho được khai thác bởi hai đối tượng chính:

- **Người sử dụng:** Thông qua môi trường web, người sử dụng có thể xem danh sách tài liệu có trong kho theo nhiều tiêu chí và tải dữ liệu về.
- **Hệ thống phân loại văn bản tiếng Việt:** có thể sử dụng dữ liệu trong kho để phân tích, so khớp nhằm đưa ra kết quả đánh giá cho một tài liệu cần phân loại. Các tài liệu phục vụ cho cả 2 giai đoạn là huấn luyện và kiểm thử.

2.3.3. Cập nhật kho

Việc cập nhật dữ liệu giúp kho dữ liệu có nguồn dữ liệu phong phú với nhiều chủ đề khác nhau và đảm bảo dữ liệu có độ bao phủ lớn phục vụ cho phân loại văn bản tiếng Việt. Đối với kho dữ liệu phục vụ phân loại tiếng Việt, để nâng cao mức độ chính xác của tài liệu trong việc chọn chủ đề trước khi tải lên sẽ qua bước phân loại học máy kiểm thử, kết quả phân loại nếu phù hợp với nhận xét khách quan ban đầu của người quản trị sẽ thực hiện tải lên.

Việc cập nhật kho dữ liệu thông qua 4 bước sau:

- **Bước 1:** Người quản trị chọn đường dẫn đến tập tài liệu và thực hiện phân loại kiểm thử tài liệu.
- **Bước 2:** Hệ thống sẽ xử lý và đưa ra kết quả kiểm tra (bao nhiêu % so với chủ đề gốc do người quản trị gán), người quản trị xem kết quả.
- **Bước 3:** Người quản trị chọn upload (nếu kết quả hiển thị từ 90% trở lên), hệ thống xử lý và sao lưu tập tin gốc vào kho.
- **Bước 4:** Sử dụng công cụ chuyển đổi dữ liệu phục vụ cho hệ thống phân loại văn bản tiếng Việt lưu vào trong kho.

2.4. Phân tích và đặc tả dữ liệu

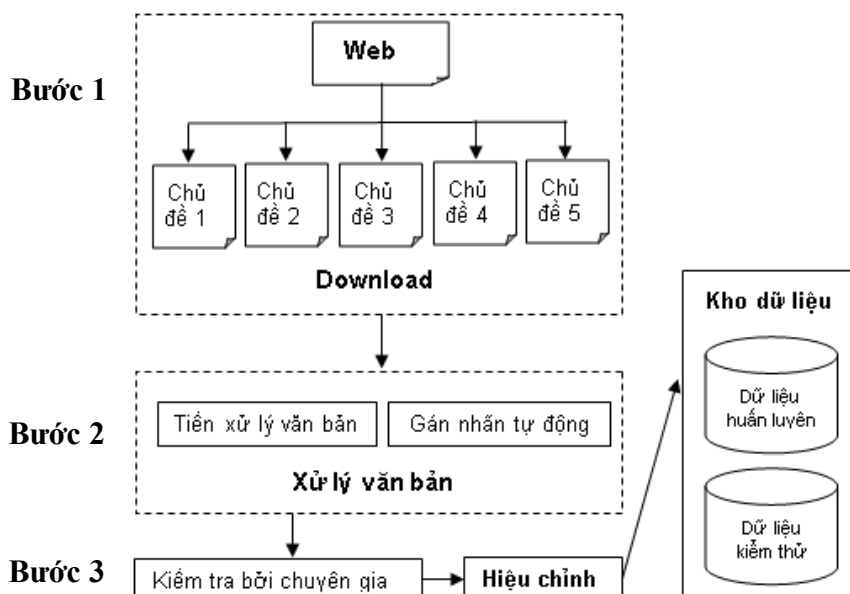
Phân tích và định rõ yêu cầu đặc tả dữ liệu là bước kỹ thuật đầu tiên trong quá trình xây dựng kho dữ liệu, làm mịn thành một bản đặc tả dữ liệu cụ thể để trở thành nền tảng cho mọi hoạt động xây dựng kho dữ liệu. Để hiểu rõ đặc tả dữ liệu, người ta tạo ra mô hình, phân hoạch vấn đề và tạo ra những biểu diễn mô tả cho bản chất của dữ liệu rồi sau đó đi vào các chi tiết. Trong nhiều trường hợp, không thể nào đặc tả được đầy đủ mọi vấn đề tại giai đoạn đầu. Việc làm bản mẫu thường giúp chỉ ra cách tiếp cận khác để từ đó có thể làm mịn thêm dữ liệu. Kết quả của việc phân tích là tạo ra bản đặc tả các dữ liệu. Đặc tả cần được xét duyệt để đảm bảo rằng người phát triển và sử dụng có cùng nhận biết về hệ thống cần phát triển.

Tài liệu được sưu tập từ các bài viết trên các website điện tử theo các chủ đề Bóng đá, giáo dục, pháp luật, quốc tế, xã hội.

Nguồn tài liệu tổng hợp từ 4 website điện tử tiếng Việt phổ biến như: vnexpress, vietnamnet, dantri, tuoitre. Định dạng của tài liệu chủ yếu dưới dạng .txt.

2.5. Giải pháp xây dựng kho

2.5.1. Đề xuất mô hình tổng quát



Hình 2.3 Mô hình đề xuất tổng quát kho dữ liệu

2.5.2. Quá trình xây dựng kho dữ liệu

Dựa trên chuyên gia (con người) để xây dựng kho dữ liệu, quá trình xây dựng thực hiện qua 4 bước.

- Bước 1: Sưu tập dữ liệu

Nguồn dữ liệu được tổng hợp download từ các trang website điện tử được đọc nhiều nhất: vnexpress, vietnamnet, dantri, tuoitre, bao gồm nội dung các bài viết thuộc năm chủ đề bởi chuyên gia (người) như: Bóng đá, giáo dục, pháp luật, quốc tế, xã hội, ... nên nội dung của các bài viết có thể được coi là chuẩn.

Viết module tải tự động các bài viết, đồng thời loại bỏ các thông tin không cần thiết, các hình ảnh, ...) xuất về máy dạng *.txt theo từng chủ đề.

Đọc lần lượt từng bài báo để xem xét lại và chọn ra các bài có nội dung có phù hợp với chủ đề hay không.

- Bước 2: Xử lý văn bản (Tiền xử lý văn bản, gán nhãn tự động, ...)

Chuẩn hóa dữ liệu đầu vào phù hợp lưu trữ trong kho như:

- Loại bỏ các hình ảnh, những dữ liệu thừa, ...

- Chuyển đổi phong: Ở bước này việc xử lý chủ yếu bằng phương pháp thủ công.
- Chuyển đổi cấu trúc, thể thức trình bày văn bản của tài liệu về dạng chuẩn, tài liệu phải ở định dạng *.txt. Như đã trình bày ở bước 1 sẽ viết module để xử lý.

- Bước 3: Upload tập dữ liệu vào kho

Để nâng cao độ chính xác của tập dữ liệu trước khi đưa vào kho, tôi đề xuất cách upload tài liệu như sau:

- Kiểm tra lại dữ liệu bởi chuyên gia và hiệu chỉnh.
- Chọn số lượng (theo chủ quan) tập mẫu ban đầu để huấn luyện, số tập mẫu ban đầu này “được coi” là chỉ liên quan đến một chủ đề. Sau khi gán nhãn chủ đề xong, sẽ sử dụng chương trình phân loại văn bản ta xây dựng để phân loại các văn bản còn lại và đưa nó vào chủ đề thích hợp trong kho. Cụ thể gán cho các văn bản ban đầu chủ đề gốc (theo chủ quan) của người quản trị, máy duyệt từng văn bản mẫu, kết quả mà chương trình đưa ra đạt so với chủ đề gốc thì văn bản mẫu đó sẽ được người quản trị upload vào kho. Cứ như vậy, số lượng tập mẫu trong kho sẽ tăng dần lên.

- Bước 4: Đánh giá chất lượng kho

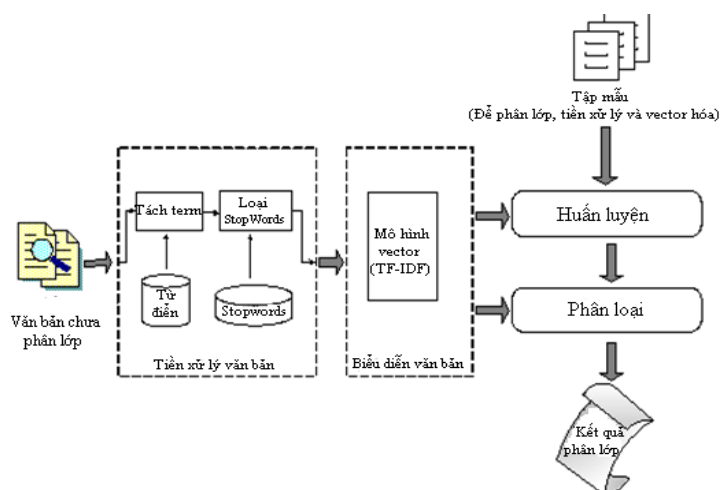
Sau khi hoàn thành kho dữ liệu, ta sẽ tiến hành thử nghiệm phân loại nhiều lần trên chính tập mẫu trong kho để cải thiện và đánh giá chất lượng kho.

Chất lượng kho được đánh giá tốt khi mức độ chính xác của kết quả phân loại tài liệu dựa trên nguồn mẫu kho càng cao.

2.5.3. Quy trình của chương trình phân loại văn bản

Việc xây dựng chương trình phân loại văn bản chính xác có ý nghĩa quan trọng trong bước kiểm tra tính đúng đắn (chất lượng của dữ liệu: có thuộc đúng chủ đề ta muốn gán) trước khi upload vào kho. Quy trình chương trình phân loại văn bản gồm các bước:

- Bước 1: Tiền xử lý dữ liệu (Tách term, loại bỏ từ dừng Stopword)
- Bước 2: Biểu diễn văn bản (trích chọn đặc trưng)
- Bước 3: Huấn luyện tập mẫu
- Bước 4: Phân loại văn bản



Hình 2.4 Quy trình phân loại văn bản

a. Tiền xử lý dữ liệu

Để có thể có được kết quả huấn luyện và phân loại tốt các văn bản (phân loại chính xác), chúng ta cần có một hệ từ vựng chuẩn. Việc tách các từ và gán nhãn từ loại cho tiếng Việt các cụm từ này trên thực tế là rất khó khăn và đòi hỏi sử dụng đến nhiều thuật toán khác nhau do vậy việc xây dựng một modul như vậy là không khả thi. Thay vào đó chúng ta có thể tích hợp các hệ thống nghiên cứu khác vào hệ thống giúp cho việc xây dựng được tốt hơn.

Trong bước này ta sử dụng lại các công cụ đã có sẵn cần tách được các danh từ để đưa vào cơ sở dữ liệu, chúng ta nên sử dụng tách từ tiếng Việt và gán từ loại tiếng Việt nằm trong hệ phân loại tiếng Việt để tích hợp vào hệ thống.

Để tiết kiệm không gian lưu trữ và gia tăng tốc độ tìm kiếm, các công cụ tìm kiếm sẽ không ghi nhận lại những từ quá phổ biến, quá chung chung như: chỉ, cho, cái, có, cứ, do, gì, là, mà, nếu, thì, vì, ... những từ này gọi là stopwords.

Tách từ tiếng Việt

- Dựa vào mô hình Maximum Entropy là phương pháp học máy và mô hình Maximum Entropy Markov Model (MEM) [64] với giải thuật tối ưu BLMVM có hỗ trợ giá trị thực [76], tận dụng thông tin tri thức từ nhiều nguồn khác nhau làm tăng độ chính xác của bộ tách từ, cách tiếp cận chương trình vnTokenizer [46] được sử dụng để tách từ, kết hợp nhiều đặc trưng hữu ích từ các mô hình khác gồm: Mô hình

phân đoạn từ dựa vào từ điển, mô hình nhận dạng tên thực thể và mô hình n-gram, cách trích chọn đặc trưng hữu ích từ các mô hình dựa vào từ điển và mô hình nhận dạng thực thể huấn luyện và kiểm thử mô hình, sử dụng Corpus tiếng Việt về tách từ trong đó mô hình n-gram được huấn luyện sử dụng treebank tiếng Việt (70,000 câu đã được tách từ)

- Độ chính xác trên 97%.

Gán nhãn từ loại tiếng Việt

Áp dụng các mô hình học máy Maximum Entropy và CRFs [58], dùng mô hình ngôn ngữ cho trước một số cách tách từ của toàn bộ câu, một mô hình ngôn ngữ có thể đánh giá được cách nào có khả năng cao hơn. Đây là cách tiếp cận sử dụng nguồn của vnTokenizer, JvnTagger [46]. Thực hiện gán nhãn từ loại POS (part – of – speech tagging) sử dụng phương pháp học máy MEM đã được sử dụng thành công cho tách từ. Thực hiện POS để phân lớp với các lớp chính là nhãn từ loại. Quá trình gán nhãn quan tâm tới kiến trúc theo kiểu đường ống (pipeline), nghĩa là việc gán nhãn từ loại được thực hiện sau khi đã có thông tin về từ vựng. Kiến trúc tổng thể gán nhãn POS tiếng Việt. Trong đó có hai pha chính là pha huấn luyện mô hình và pha giải mã.

- Pha huấn luyện mô hình: Đầu vào là văn bản đã được tách từ đưa qua bộ trích chọn đặc trưng rồi đưa vào mô hình MEM để huấn luyện.
- Pha giải mã: Văn bản đầu vào sẽ đưa qua pha giải mã kết quả sẽ cho ra chuỗi thể tốt nhất ứng với mỗi câu đầu vào.
- Được huấn luyện sử dụng dữ liệu treebank tiếng Việt (20,000 câu đã gán nhãn từ loại).
- Độ chính xác trên 93%.

Loại bỏ Stopword

Nội dung của văn bản thể hiện ở trong tập hợp các `term` mà nó chứa, bước đầu tiên của tiền xử lý văn bản là tách `term`. Chức năng này sẽ tìm ra một tập hợp các `term` xuất hiện trong văn bản và sau đó chỉ giữ lại những `term` có ý nghĩa, đây là những `term` còn lại sau khi loại bỏ `stopword`.

Stopword là những từ cùng một lúc xuất hiện nhiều ở nhiều văn bản, như liên từ, thực từ, hư từ, ...ngoài ra cũng có một số động từ, tính từ, phó từ. Ví dụ : “nếu”, “thì”, “nếu không thì”, “ hầu như là“, ... những stopwords này chúng ta đã có một danh sách các từ. Việc tạo ra một danh sách stopwords càng nhiều với độ chính xác cao sẽ giúp cho việc tách term từ văn bản chính xác hơn, các term được tách ra sẽ có ý nghĩa lớn hơn trong việc phân loại văn bản sau này.

b. Biểu diễn văn bản

Hiện nay, để giải quyết hầu hết những vấn đề liên quan đến văn bản chúng ta dùng các mô hình biểu diễn. Vì vậy, các mô hình biểu diễn không ngừng phát triển, hàm chứa được nhiều hơn những suy nghĩ mà con người muốn diễn đạt, đồng thời nâng cao hiệu quả sử dụng.

Mô hình biểu diễn văn bản truyền thống được sử dụng phổ biến nhất như: Mô hình không gian véc tơ và mô hình túi từ. Mô hình không gian véc tơ biểu diễn văn bản như một véc tơ đặc trưng của các thuật ngữ (từ) xuất hiện trong toàn bộ tập văn bản. Trọng số các véc tơ đặc trưng thường được tính qua độ đo TF-IDF [42]. Mô hình này nắm bắt được các thông tin cấu trúc quan trọng như vị trí xuất hiện của từ, vùng lân cận của từ, trật tự xuất hiện của các từ trong văn bản. Mô hình biểu diễn văn bản được đề xuất và được đánh giá cao vì tận dụng được các thông tin quan trọng về cấu trúc mà không gian véc tơ và mô hình túi từ đã sử dụng. Khi ứng dụng vào từng loại bài toán khác nhau, các thành phần thích hợp nhất trong văn bản trở thành mối quan hệ hiệu quả nhất như: trật tự xuất hiện, tần số đồng hiện, vị trí xuất hiện, độ tương đồng và có thể biểu diễn câu, từ, hay câu kết hợp từ. Kết thúc quá trình tiền xử lý văn bản, chương trình sẽ thực hiện biểu diễn văn bản theo mô hình không gian véc tơ. Nhiệm vụ đầu tiên trong việc xử lý phân loại văn bản là chọn được một mô hình biểu diễn văn bản thích hợp. Một văn bản ở dạng thô (dạng chuỗi) cần được chuyển sang một mô hình khác để tạo thuận lợi cho việc biểu diễn và tính toán. Tùy thuộc vào từng thuật toán phân loại khác nhau mà chúng ta có mô hình biểu diễn riêng. Một trong những mô hình đơn giản và thường được sử dụng trong nhiệm vụ này là mô hình không gian véc tơ.

Mô hình không gian véc tơ

Mô hình không gian véc tơ là một trong những mô hình được sử dụng rộng rãi nhất cho việc tìm kiếm (truy hồi) thông tin. Nguyên nhân chính là bởi vì sự đơn giản của nó.

Mỗi văn bản sẽ được biểu diễn dưới dạng véc tơ $d_i = (w_{i1}, w_{i2}, \dots, w_{im})$ trong đó w_{ij} là giá trị (trọng số) của từ T_j trong văn bản d_i , m là số lượng đặc trưng (số chiều không gian véc tơ). Mỗi một đặc trưng tương ứng với một từ xuất hiện trong tập huấn luyện sau khi loại bỏ các từ dừng ra khỏi các văn bản.

Ví dụ: Biểu diễn một tập hợp gồm ba văn bản D_1, D_2, D_3 với số chiều $m = 3$ nghĩa là trong từ điển có 3 từ là T_1, T_2, T_3 .

- Văn bản D_1 chứa 3 từ T_1 , 4 từ T_2 và 4 từ $T_3 \Rightarrow D_1 = 3T_1 + 4T_2 + 4T_3$

- Văn bản D_2 chứa 4 từ T_1 , 6 từ T_2 và 1 từ $T_3 \Rightarrow D_2 = 4T_1 + 6T_2 + 1T_3$

- Văn bản D_3 chứa 0 từ T_1 , 0 từ T_2 và 3 từ $T_3 \Rightarrow D_3 = 0T_1 + 0T_2 + 3T_3$

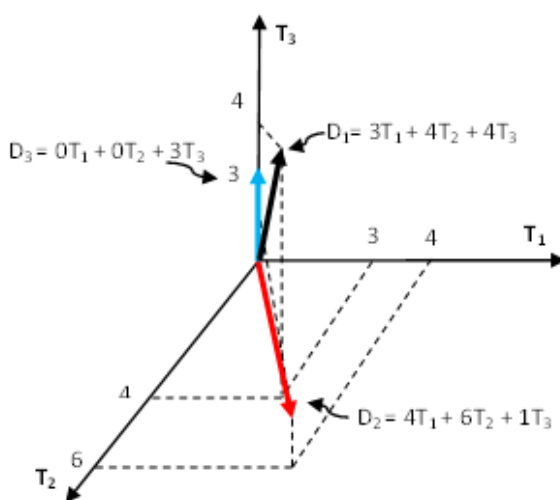
Ta có cách viết dưới dạng véc tơ như sau:

- Văn bản D_1 : $D_1(3, 4, 4)$

- Văn bản D_2 : $D_2(4, 6, 1)$

- Văn bản D_3 : $D_3(0, \text{[redacted]})$

Do đó các véc tơ D_1, D_2, D_3 sẽ được biểu diễn trên mô hình không gian véc tơ 3 chiều như hình vẽ.



Hình 2.5 Mô hình không gian véc tơ 3 chiều

Các đặc trưng của mỗi văn bản khi biểu diễn dưới dạng véc tơ thường:

- Có số chiều không gian đặc trưng lớn.
- Các đặc trưng độc lập nhau.
- Các đặc trưng có tính rời rạc.

Biểu diễn các đặc trưng của véc tơ bằng số thực và sử dụng một cấu trúc để chỉ lưu trữ những đặc trưng có giá trị khác 0. Đây là một cải tiến giúp giảm kích thước lưu trữ, tiết kiệm bộ nhớ. Có nhiều phương pháp để xác định giá trị của trọng số w_{ij} cho mỗi đặc trưng (mỗi từ) trong mỗi văn bản.

2.5.4. Sử dụng thuật toán Naïve Bayes để phân loại văn bản

Bước 1: Huấn luyện [9].

- Từ tập huấn luyện, ta rút trích tập từ vựng (các đặc trưng)
- Tính xác suất $P(C_i)$ và $P(x_k|C_i)$

❖ Đầu vào:

- Các véc tơ đặc trưng của văn bản trong tập huấn luyện (Ma trận $m \times n$, với m là số véc tơ đặc trưng trong tập huấn luyện, n là số đặc trưng của véc tơ).
- Tập nhãn/lớp cho từng véc tơ đặc trưng của tập huấn luyện.

❖ Đầu ra:

- Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.
- Công thức tính:

$$P(C_i) = \frac{|docs_i|}{|total\#documents|} \quad (2.1)$$

$docs_i$: số tài liệu của tập huấn luyện thuộc lớp C_i

$total\#documents$: số tài liệu có trong tập huấn luyện.

$$P(x_k/C_i) = \frac{|n_k|}{|Text_i|} \quad (2.2)$$

hoặc

$$P(x_k/C_i) = \frac{n_k+1}{n+|Text_i|} \quad (2.3)$$

(làm mịn với luật Laplace)

Trong đó:

- n : tổng số từ đôi một khác nhau của lớp C_i .

- n_k : tổng số từ x_k trong tập từ vựng trong lớp C_i .
- $|\text{Text}_i|$: tổng số từ vựng (không phân biệt đôi một) trong lớp C_i .

Bước 2: Phân lớp

❖ Đầu vào:

- Véc tơ đặc trưng của văn bản cần phân lớp.
- Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

❖ Đầu ra:

- Nhãn/lớp của văn bản cần phân loại.
- Công thức tính xác suất thuộc phân lớp i khi biết trước mẫu X

$$P^{new} = \max_{c_j=1 \rightarrow |C|} (P(C_i) \prod_{k \in positions} (P(x_k/C_i) * |x_k|^{new})) \quad (2.4)$$

positions: tập từ vựng trong bộ huấn luyện.

Dựa vào véc tơ đặc trưng của văn bản cần phân lớp, áp dụng công thức trên tính xác suất thuộc từng phân lớp cho văn bản, và chọn ra lớp có xác suất cao nhất.

Xét ví dụ: ta có tập tài liệu để huấn luyện sau khi đã véc tơ hoá (sử dụng phương pháp đơn giản đếm số lần xuất hiện) và rút trích đặc trưng như sau:

Bộ từ vựng (đặc trưng): Tự tin, Sáng tạo, Khéo léo, Nhiệt tình

Bảng 2.2 Dữ liệu huấn luyện

| Văn bản | Tự tin | Sáng tạo | Khéo léo | Nhiệt tình | Lớp |
|-----------|--------|----------|----------|------------|---------|
| Văn bản 1 | 44 | 28 | 8 | 58 | Bóng đá |
| Văn bản 2 | 12 | 31 | 40 | 4 | Xã hội |
| Văn bản 3 | 14 | 26 | 24 | 6 | Xã hội |
| Văn bản 4 | 35 | 42 | 10 | 47 | Bóng đá |
| Văn bản 5 | 29 | 34 | 11 | 64 | Bóng đá |
| Văn bản 6 | 10 | 24 | 32 | 3 | Xã hội |

Bước huấn luyện:

- Tính xác suất các lớp C_i trong tập huấn luyện:

$$P(C_1 = \text{"Xã hội"}) = 3/6 = 0.5$$

$$P(C_2 = \text{"Bóng đá"}) = 3/6 = 0.5$$

- Tính xác suất $P(x_k|C_i)$

$$\text{Lớp } C_1 = \text{"Xã hội"}: \text{Tổng} = 226$$

$$P(\text{Tự tin} | \text{Xã hội}) = (12+14+10)/226 = 36/226$$

$$\begin{aligned}
P(\text{Sáng tạo}|\text{Xã hội}) &= (31+26+24)/226 = 81/226 \\
P(\text{Khéo léo}|\text{Xã hội}) &= (40+24+32)/226 = 96/226 \\
P(\text{Nhiệt tình}|\text{Xã hội}) &= (4+6+3)/226 = 13/226 \\
\text{Lớp } C_2 = \text{"Bóng đá"}: \text{Tổng} &= 410 \\
P(\text{Tự tin}|\text{Bóng đá}) &= (44+35+29)/410 = 108/410 \\
P(\text{Sáng tạo}|\text{Bóng đá}) &= (28+42+34)/410 = 104/410 \\
P(\text{Khéo léo}|\text{Bóng đá}) &= (8+10+11)/410 = 29/410 \\
P(\text{Nhiệt tình}|\text{Bóng đá}) &= (58+47+64)/410 = 169/410
\end{aligned}$$

Bước phân lớp: cho văn bản có véc tơ đặc trưng sau:

Docnew = (21, 42, 16, 52) . Xác định lớp cho văn bản mới ?

Tính các xác suất :

Xác suất Docnew thuộc Bóng đá:

$$P(\text{Bóng đá}) * p(\text{Tự tin}|\text{Bóng đá}) * 21 * p(\text{Sáng tạo}|\text{Bóng đá}) * 42 * p(\text{Khéo léo}|\text{Bóng đá}) * 16 * p(\text{Nhiệt tình}|\text{Bóng đá}) * 52 = \mathbf{0.5861}$$

Xác suất Docnew thuộc Xã hội:

$$P(\text{Xã hội}) * p(\text{Tự tin}|\text{Xã hội}) * 21 * p(\text{sáng tạo}|\text{Xã hội}) * 42 * p(\text{khéo léo}|\text{Xã hội}) * 16 * p(\text{nhiệt tình}|\text{Xã hội}) * 52 = \mathbf{0.2313}$$

Kết quả: Văn bản Docnew thuộc về lớp Bóng đá do $\max(P^{\text{new}}) = \mathbf{0.5861}$

2.5.5. Định dạng đầu ra của dữ liệu trong kho

Kho dữ liệu văn bản tiếng Việt đã được xây dựng có được các văn bản mẫu dưới dạng .txt. Định dạng này không thể hiện được nhiều thông tin khác như tên bài báo, nội dung chính của bài báo, phân lớp nó thuộc về từng chủ đề, tác giả của bài báo, ngày viết, kích thước, ... Bởi vậy cần phải tạo ra một định dạng chuẩn khác của các tập mẫu là định dạng XML, định dạng này thể hiện được các thông tin đó nhằm phục vụ cho mục đích phân loại văn bản tiếng Việt.

Định dạng văn bản mẫu

<BKTEXTS>

<METADATA>

<TOPIC></TOPIC> // Các phân lớp mà văn bản thuộc về chủ đề

<DATE></DATE> // Ngày tháng phát hành văn bản

<VNFONT></VNFONT> // Phong chữ sử dụng của văn bản

```

    <SIZE></SIZE> // Kích thước văn bản
<SOURCE>// Nguồn gốc của văn bản
    <DATELINE></DATELINE>
    <ORGS></ORGS>
    <COUNTRIES></COUNTRIES>
</SOURCE>
    <AUTHOR>// Tác giả văn bản
    <FULLNAME></FULLNAME>
    <ORGS></ORGS>
    <COUNTRIES></COUNTRIES>
</AUTHOR>
</METADATA>
<TEXT>
    <TITLE></TITLE> // Tiêu đề văn bản
    <SUMMARY></SUMMARY> // Nội dung tóm tắt văn bản nếu có
    <HEADLINE></HEADLINE> // Câu đầu tiên của văn bản, sử dụng
    khi văn bản không có tiêu đề
    <BODY></BODY> // Nội dung văn bản
</TEXT>
</BKTEXTS>

```

2.6. Kết quả kho dữ liệu thử nghiệm và đánh giá

2.6.1. Kết quả kho dữ liệu thử nghiệm

Sau các bước thu thập dữ liệu tải về, các dữ liệu này sẽ được tác giả duyệt nội dung phù hợp với từng chủ đề tương ứng trước khi đưa văn bản vào chương trình phân loại kiểm nghiệm theo từng chủ đề để cập nhật vào kho. Số lượng văn bản sau khi đã được duyệt lưu trữ vào kho tới thời điểm thực hiện thử nghiệm là 5027 bài viết, cụ thể như sau:

Bảng 2.3 Kết quả kho dữ liệu thử nghiệm

| Chủ đề | Số lượng bài viết |
|-----------|-------------------|
| Bóng đá | 1023 |
| Giáo dục | 1014 |
| Pháp luật | 987 |
| Quốc tế | 1009 |

| | |
|--------|-----|
| Xã hội | 994 |
|--------|-----|

2.6.2. Đánh giá kho dữ liệu

Kho dữ liệu phục vụ phân loại văn bản tiếng Việt được xây dựng từ các trang web điện tử có nguồn đảm bảo tin cậy, được sử dụng rộng rãi ở Việt Nam. Đồng thời, dữ liệu được kiểm chứng dựa trên hai nguồn dữ liệu trước khi đưa vào kho, đó là dữ liệu thu thập thủ công và dữ liệu nguồn có trước.

2.7. Tiểu kết chương

Trong chương này, đã nghiên cứu lý thuyết về kho dữ liệu, hệ thống phân loại văn bản tiếng Việt. Phân tích và tìm hiểu các qui trình như công cụ, kỹ thuật xây dựng kho dữ liệu. Nêu được giải pháp xây dựng kho dữ liệu và cập nhật dữ liệu vào kho, sử dụng thuật toán phân loại văn bản Naïve Bayes, so khớp sự tương đồng giữa thu thập thủ công dữ liệu và dữ liệu nguồn lưu trữ trước, kiểm chứng dữ liệu từ hai nguồn đó (chấp nhận được) mới cập nhật vào kho. Kết quả đã xây dựng được một kho dữ liệu với số lượng 5027 văn bản mẫu với 5 chủ đề. Hệ thống kho dữ liệu đã xây dựng đáp ứng được yêu cầu phi chức năng như khả năng lưu trữ dữ liệu lớn, hệ thống ổn định, giao diện đơn giản, dễ sử dụng.

Mục tiêu xây dựng kho dữ liệu là phục vụ cho phân loại văn bản tiếng Việt dựa trên mô hình cự ly đường trắc địa và áp dụng để rút gọn số chiều của véc tơ dựa trên đồ thị Dendrogram ở các chương sau.

Chương 3. PHÂN LOẠI VĂN BẢN DỰA TRÊN MÔ HÌNH CỤ LY TRẮC ĐỊA

Trong chương này, trình bày kết quả nghiên cứu mô hình cụ ly đường trắc địa để áp dụng vào học máy phục vụ phân loại văn bản tiếng Việt. Kết hợp sử dụng một số mẫu dữ liệu đã gán nhãn với số lượng lớn dữ liệu chưa gán nhãn. Vì vậy, nghiên cứu mô hình cụ ly đường trắc địa cho phép nâng cao hiệu quả phân loại văn bản tiếng Việt so với các mô hình phân loại trước đây. Kết quả đạt được trong nghiên cứu đề xuất một giải pháp mới trong phân loại văn bản dựa trên mô hình cụ ly trắc địa và lý thuyết đồ thị là đúng và chưa được công bố trong bất kỳ công trình nghiên cứu khoa học nào khác.

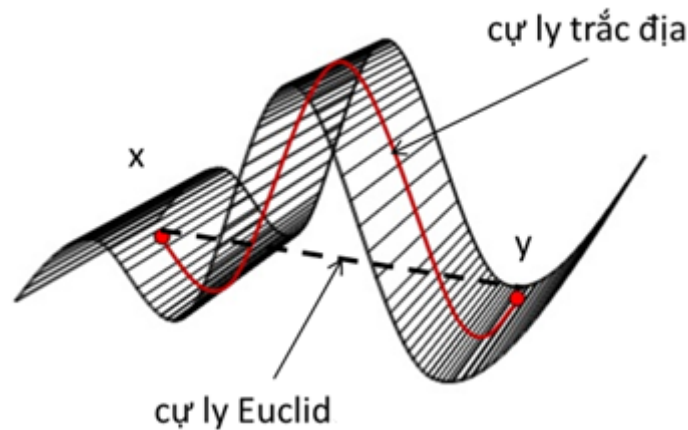
3.1. Mô hình cụ ly trắc địa trên máy véc tơ hỗ trợ

3.1.1. Mô hình cụ ly trắc địa

Để xây dựng mô hình phân loại văn bản tiếng Việt, việc gán nhãn cho các tập tin là điều cần thiết. Số lượng các tập tin được gán nhãn càng nhiều thì việc phân loại càng chính xác. Tuy nhiên, công việc gán nhãn này tiêu tốn rất nhiều chi phí hay thời gian để các chuyên gia trong từng lĩnh vực thực hiện. Vậy bài toán đặt ra là làm sao có thể giảm chi phí gán nhãn mà vẫn nâng cao được hiệu quả phân loại. Vì vậy, trong nghiên cứu này đề xuất phương pháp phân loại văn bản dựa trên máy véc tơ hỗ trợ và mô hình cụ ly đường trắc địa nhằm tăng hiệu suất phân loại cho các văn bản tiếng Việt số.

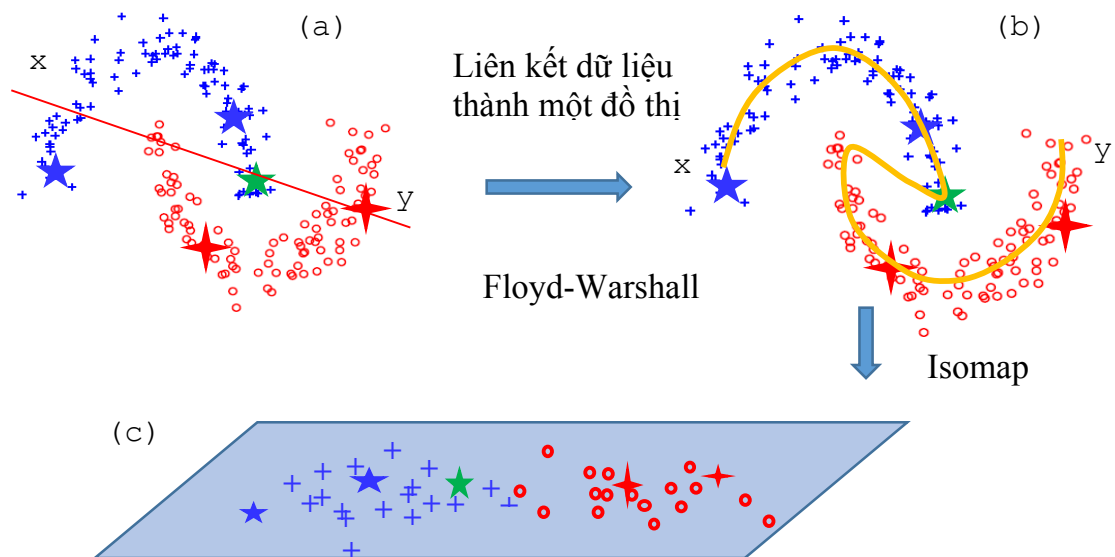
Mô hình cụ ly đường trắc địa là mô hình sử dụng cụ ly đo đặc dựa trên dữ liệu thực đã được cho trước, sử dụng hệ tương quan ngắn nhất (trong phân loại văn bản là mức độ gần nhau giữa các văn bản) để tính khoảng cách giữa hai véc tơ [4], [66] [92]. Khoảng cách này được gọi là cụ ly trắc địa và khác với khoảng cách Euclid. Xây dựng mô hình đường trắc địa hợp lý, nghĩa là xây dựng được mối tương quan giữa các văn bản hợp lý thì việc phân loại văn bản tự động sẽ dễ dàng hơn. Các mô hình phân loại trước đây thường sử dụng cụ ly Euclid để tính khoảng cách dữ liệu.

Vì vậy, trên những không gian bị uốn, việc sử dụng cự ly Euclid sẽ khó có thể xây dựng được một mô hình phân loại tốt. Hình vẽ dưới đây phân biệt cự ly Euclid và cự ly trắc địa của điểm x và điểm y trên không gian 3 chiều [28].



Hình 3.1 Cự ly Euclid và cự ly trắc địa

Cự ly Euclid chính là cự ly được đo bởi khoảng cách nối từ hai điểm x và y tính theo đường chim bay. Ý tưởng mô hình đề xuất.



Hình 3.2 Mô hình đề xuất

Trong không gian 3 chiều véc tơ dữ liệu các lớp nằm độc lập với nhau và các không gian dữ liệu này được ngăn cách bởi một khoảng cách. Các mô hình phân loại trước đây thường sử dụng cự ly Euclid để tính khoảng cách dữ liệu giữa hai

điểm x và y theo đường thẳng sẽ không chính xác. Vì vậy, trên những không gian bị uốn (*hình a*), việc sử dụng tính khoảng cách Euclid sẽ khó có thể xây dựng được mô hình phân loại tốt. Dẫn đến việc phân loại không thành công. Để khắc phục mô hình trên trong nghiên cứu này đề xuất mô hình tính khoảng cách bằng cự ly trắc địa, ta có thể chuyển văn bản đó thành một véc tơ đặc trưng liên kết dữ liệu thành một đồ thị (*hình b*) và kết nối dữ liệu với đồ thị đã xây dựng tại đỉnh (véc tơ) có cự ly Euclid gần nhất. Dựa trên phương pháp Isometric Feature Mapping (Isomap) ta có thể dễ dàng tìm thấy (*hình c*) và phân loại được các không gian này bằng cách sử dụng cự ly trắc địa.

Trong khi đó, cự ly trắc địa chính là khoảng cách thực tế men theo địa hình bị uốn cong giữa x và y . Khi so sánh đến hai khoảng cách này, ta thường liên tưởng đến việc tính độ dài đường đi khi qua một ngọn núi trong thực tế. Chúng ta không thể tính chi phí của đường đi bằng cách tính khoảng cách đường chim bay vì trên thực tế chúng ta phải đi theo sườn núi, lên đỉnh núi rồi lại xuống bằng cách theo sườn núi phía bên kia để tới đích. Vì vậy, đo khoảng cách thực tế thì mới có thể tính toán chính xác chi phí của một quãng đường. Các dữ liệu trong phân loại văn bản cũng vậy, văn bản thường được phân bố trên các không gian bị uốn cong. Vì vậy, việc sử dụng cự ly trắc địa sẽ có thể làm tăng hiệu quả nhận dạng hơn.

Thuật toán Floyd-Warshall để duyệt đồ thị

Khi nhắc đến các thuật toán duyệt đồ thị, có thể chúng ta đã biết (và đã từng thực hiện) Depth-First Search, Breadth-First Search, hoặc Dijkstra. Ý nghĩa của từng thuật toán, đứng ở khía cạnh bài toán tìm đường đi ngắn nhất. Thuật toán DFS dùng để giải các bài toán mà chúng ta muốn tìm được lời giải (không nhất thiết phải là quãng đường ngắn nhất), hoặc ta muốn thăm tất cả các đỉnh của đồ thị. Thuật toán duyệt theo chiều sâu BFS cũng để duyệt các đỉnh của đồ thị, nhưng có một tính chất quan trọng là: nếu tất cả các cạnh *không có trọng số*, lần đầu tiên một đỉnh được thăm, ta có ngay đường đi ngắn nhất đến đỉnh đó. Bây giờ đến thuật toán Dijkstra, đây là thuật toán nổi tiếng dùng để tìm đường đi ngắn nhất từ một đỉnh cho trước đến các đỉnh còn lại, trong một đồ thị có các cạnh *có trọng số không âm*.

Như vậy, thuật toán Dijkstra đã tiến hơn một bước so với thuật toán BFS.

Trong nghiên cứu này, để đi tìm đường đi ngắn nhất sử dụng một thuật toán ít biết đến hơn để duyệt đồ thị, đó là thuật toán Floyd-Warshall.

Nếu như thuật toán Dijkstra giải quyết bài toán tìm đường đi ngắn nhất từ *một đỉnh cho trước* đến mọi đỉnh khác trong đồ thị, thì thuật toán Floyd-Warshall sẽ tìm đường đi ngắn nhất *giữa mọi đỉnh* sau một lần chạy thuật toán. Một tính chất nữa là thuật toán Floyd-Warshall có thể chạy trên đồ thị có các cạnh có trọng số *có thể âm*, tức là không bị giới hạn như thuật toán Dijkstra. Tuy nhiên, lưu ý là trong đồ thị không được có vòng nào có tổng các cạnh là âm, nếu có vòng như vậy ta không thể tìm được đường đi ngắn nhất (mỗi lần đi qua vòng này độ dài quãng đường lại giảm, nên ta có thể đi vô hạn lần). Thuật toán Floyd-Warshall so sánh tất cả các đường đi có thể giữa từng cặp đỉnh. Nó là một dạng của quy hoạch động.

Ta có thể dễ dàng sử dụng thuật toán Floyd-Warshall để tìm cự ly trắc địa cho tất cả các cặp véc tơ như sau: Thuật toán này cho phép chúng ta tìm đường đi ngắn nhất giữa mọi cặp đỉnh. Nếu đỉnh k nằm trên đường đi ngắn nhất từ đỉnh i tới đỉnh j thì đoạn đường từ i tới k và từ k tới j phải là đường đi ngắn nhất từ i tới k và từ k tới j tương ứng. Do đó ta sử dụng ma trận D để lưu độ dài đường đi ngắn nhất giữa mọi cặp đỉnh.

Procedure Floyd-Warshall

Input:

N : Là tập các văn bản (n đỉnh)

Output:

D : Ma trận chứa khoảng cách ngắn nhất giữa mọi cặp đỉnh của đồ thị

Begin //Xây dựng đồ thị

for i from 1 to n do

for j from 1 to n do

$D[i, j] = \|x_k - x_l\|^2$

```

        if (D[i,j]> $\varepsilon$ ) then D[i,j]= $\infty$ 
        endif
    endfor
endfor

//Thuật toán Floyd- Warshall tìm đường đi ngắn nhất giữa
mọi cặp đỉnh
for k from 1 to n do
    for i from 1 to n do
        for j from 1 to n do
            D[i,j]= min(D[i,j], D[i,k]+D[k,j])
        endfor
    endfor
endfor

end

```

3.1.2. Kỹ thuật phân cụm đa dạng sử dụng cự ly trắc địa

Là kỹ thuật phân cụm khi cấu trúc của từng lớp tương đối phức tạp. Kỹ thuật phân cụm đa dạng (Manifold Clusterings based on Geodesic Distance) [4], [7], [31] [66] là một kỹ thuật phân loại dữ liệu không có nhãn nằm trên các không gian bị uốn một cách phi tuyến. Trong trường hợp dữ liệu của các cụm nằm trên từng không gian và các không gian này được ngăn cách bởi một khoảng cách, thì dựa trên phương pháp isometric feature mapping (Isomap) ta có thể dễ dàng tìm thấy và phân loại được các không gian này bằng cách sử dụng cự ly trắc địa. Sau đây là sơ lược về phương pháp phân cụm sử dụng cự ly trắc địa [1], [28], [36], [62].

Mục đích chính của phương pháp này là sử dụng cự ly trắc địa, nên ban đầu ta tính các cự ly trắc địa D_{kl} của tất các cặp đỉnh tương ứng với các cặp dữ liệu trong tập huấn luyện thứ k và l . Cự ly trắc địa này được tính toán bằng cách tính cự ly nhỏ nhất giữa các cặp đỉnh trên đồ thị láng giềng (neighborhood graph) của tập

huấn luyện. Đồ thị láng giềng này được xây dựng bằng cách kết nối các cặp đỉnh k và l có cự ly nhỏ hơn ε (ε -Isomap) hay k là một trong K đỉnh láng giềng gần nhất của l (K -Isomap). Cự ly trắc địa D_{kl} được xác định như sau:

$$D_{kl} = \min\{d_{G;kl}, d_{G;ki} + d_{G;il}\} \quad (3.1)$$

Với $d_{G;kl} = \|\mathbf{x}_k - \mathbf{x}_l\|^2$ nếu đỉnh k và l có kết nối trong đồ thị láng giềng G , ngược lại ta gán $d_{G;kl} = \infty$.

Tiếp theo, phương pháp phân cụm sử dụng cự ly trắc địa tối ưu bài toán sau:

$$\min_{u,c} \sum_{i=1}^N \sum_{j=1}^C u_{ij} d(i,j) \quad (3.2)$$

$$\text{Với:} \quad \sum_{j=1}^C u_{ij} = 1, \quad u_{ij} \in \{0, 1\}$$

Với cự ly của từng dữ liệu đến véc tơ \mathbf{c}_j được tính bởi:

$$\begin{aligned} d(i,j) &= \|\varphi(\mathbf{x}_i) - \mathbf{c}_j\|^2 \\ &= K_{ii} - \frac{2}{\sum_l^N u_{lj}} \sum_{l=1}^N u_{lj} K_{il} + \frac{2}{(\sum_l^N u_{lj})^2} \sum_{k=1}^N \sum_{l=1}^N u_{ki} u_{lj} K_{kl} \end{aligned} \quad (3.3)$$

Trong đó $K_{kl} = \varphi(\mathbf{x}_k)^* \varphi(\mathbf{x}_l)$ là hàm nhân của bộ phân loại, là tích vô hướng của hai véc tơ được ánh xạ lên một không gian mới nhằm phân loại dễ dàng hơn. Không gian này đã được [36] đề xuất trong các nghiên cứu trước đây. Tùy thuộc việc lựa chọn hàm nhân cho bộ phân loại văn bản, ta có thể xây dựng được nhiều bộ phân cụm khác nhau. Chẳng hạn, nhân đa thức $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T (\mathbf{x}_j + 1)^p$ dẫn đến bộ phân cụm đa thức, nhân Gaussian $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ dẫn đến bộ phân cụm RBF (Radial Basis Functions)[54] và nhân sigmoid $k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i \mathbf{x}_j + c)$ dẫn tới mạng nơ ron sigmoid hai lớp.

3.1.3. Phương pháp tính toán cự ly trắc địa

Trước khi xây dựng mô hình cự ly đường trắc địa cho văn bản tiếng Việt số, mỗi văn bản cần phải thông qua các bước tách từ và trích chọn đặc tính để được biểu diễn bằng một véc tơ. Véc tơ này thể hiện các đặc tính của văn bản mà ta cần phân loại và gọi là véc tơ đặc trưng.

Trong các nghiên cứu trước đây, có rất nhiều phương pháp tách từ và trích chọn đặc tính. Tuy nhiên, trong nghiên cứu này không chú trọng và đi sâu vào nghiên cứu và phát triển các phương pháp tách từ hay trích chọn đặc tính mà chỉ sử dụng như một phương pháp sẵn có. Nghiên cứu này sử dụng phương pháp thống kê unigram cho việc tách từ, và mô hình không gian véc tơ cho việc trích chọn đặc tính. Cụ thể là, khi cho một tập huấn luyện các văn bản tiếng Việt số được lưu trữ dưới dạng tập tin .doc hay .pdf ta chuyển đổi các văn bản này thành một dạng tập tin chuẩn là .txt. Sau đó tách các từ đơn tiếng Việt ra và sắp xếp chúng lại thành một danh sách các từ đơn. Từ đơn nào có tần số xuất hiện ít chúng tôi đó là những từ đó không phổ biến và có khả năng là những từ có thể gây nhiễu khi xây dựng mô hình phân loại. Vì vậy, những từ đơn tần số xuất hiện thấp ta loại bỏ ra khỏi danh sách các từ đơn đã sắp xếp trước đó. Từ danh sách này, ta xây dựng một mô hình véc tơ sao cho ứng với một từ trong danh sách sẽ có một phần tử trong véc tơ. Nghiên cứu này sử dụng hai phương pháp chính để xây dựng véc tơ cho văn bản là phương pháp sử dụng tần số xuất hiện và phương pháp TF-IDF.

Vấn đề được đặt ra ở đây là làm thế nào để tính được cự ly trắc địa trong một bài toán chỉ cho trước các véc tơ đặc trưng trong dữ liệu, mà hoàn toàn không biết không gian đã bị uốn cong như thế nào. Cự ly trắc địa này được tính toán bằng cách tính cự ly nhỏ nhất giữa các cặp đỉnh trên đồ thị láng giềng của tập huấn luyện.

Từ tập huấn luyện bao gồm các véc tơ đặc trưng đã được trích chọn ở bước tách từ và trích chọn đặc tính, ta xây dựng đồ thị kết nối giữa các véc tơ tương ứng với mỗi điểm gần nhau trên đồ thị lại với nhau. Hai điểm được kết nối với nhau thể hiện sự giống nhau về mặt tần số xuất hiện các từ trong hai văn bản. Nghĩa là, hai văn bản có véc tơ kết nối với nhau trên đồ thị, thì khả năng có cùng một chủ đề nào đó là cao. Sau đó, sử dụng đồ thị đã xây dựng tìm tất cả các khoảng cách trắc địa giữa 2 đỉnh bất kỳ sử dụng thuật toán Warshall-Floyd và gọi nó là D_{kl} . Trong đó k, l là số thứ tự của văn bản trong tập huấn luyện.

Đối với văn bản cần phân loại nhưng không thuộc tập huấn luyện, ta có thể

chuyển văn bản đó thành một véc tơ đặc trưng và kết nối với đồ thị đã xây dựng tại đỉnh (véc tơ) có cự ly Euclid gần nhất. Sau đó ta tính cự ly trắc địa của văn bản x tới các văn bản x_k trong tập huấn luyện như công thức sau:

$$D_k(x) = D_{kl} + \|x - x_l\|^2 \quad (3.4)$$

Với l là số thứ tự của văn bản trong tập huấn luyện gần với văn bản cần phân loại nhất:

$$l = \arg \min_u (\|x - x_u\|^2) \quad (3.5)$$

3.1.4. Hàm nhân trong máy véc tơ hỗ trợ sử dụng cự ly trắc địa

Máy véc tơ hỗ trợ (SVM) ban đầu là một thuật toán phân lớp tuyến tính, nhờ áp dụng các hàm nhân, thuật toán có thể tìm ra các siêu phẳng trong không gian phi tuyến đặc trưng biến đổi.

Mở rộng tích vô hướng thông qua hàm ánh xạ cho biến trong không gian H lớn hơn và thậm chí có thể vô hạn chiều, theo đó đẳng thức vẫn được giữ đúng. Trong mỗi đẳng thức, khi chúng ta có tích vô hướng thì chúng ta cũng tính được tích vô hướng thông qua phép biến đổi các véc tơ và nó được gọi là hàm nhân. Hàm nhân được sử dụng để xác định nhiều quan hệ đầu vào không tuyến tính.

Đối với hàm nhân tuyến tính ta có thể xác định được nhiều hàm bậc hai hoặc hàm mũ. Trong những năm gần đây, nhiều nghiên cứu đã đi sâu vào nghiên cứu các hàm nhân khác nhau cho sự phân lớp SVM và cho nhiều thống kê thử nghiệm khác.

Đối với véc tơ hỗ trợ, có rất nhiều hàm nhân có thể kể tên như sau [59]:

- Hàm Polynomial (homogeneous): $k(x_k, x_l) = (x_k \cdot x_l)^d$ (3.6)

- Hàm Polynomial (inhomogeneous): $k(x_k, x_l) = (x_k \cdot x_l + 1)^d$ (3.7)

- Hàm Hyperbolic tangent: $k(x_k, x_l) = \tanh(\beta x_k \cdot x_l + c)$
với $\beta > 0$ và $c < 0$. (3.8)

+ Hàm Gaussian: $k(x_k, x_l) = \exp(-\gamma \|x_k - x_l\|^2)$ với $\gamma > 0$ (3.9)

Trong nghiên cứu này đề xuất hàm nhân của máy véc tơ hỗ trợ sử dụng cự ly trắc địa kết hợp với hàm Gauss như sau:

$$k(x_k, x_l) = \exp(-\gamma D_{kl}) \quad (3.10)$$

$$k(x_k, x_l) = \exp(-\gamma D_k(x)) \quad (3.11)$$

Trong đó, $k(x_k, x_l)$ là tích vô hướng của 2 véc tơ của văn bản k và văn bản l trên không gian phân loại của máy véc tơ hỗ trợ $k(x_k, x_l)$ là tích vô hướng của véc tơ chứa văn bản bất kỳ và véc tơ của văn bản k trên không gian phân loại của máy véc tơ hỗ trợ.

Để tính toán trong SVM thì ma trận nhân thỏa mãn điều kiện: Tất cả các giá trị riêng phải không âm. Bằng thực nghiệm hàm nhân trong mô hình đề xuất thỏa mãn điều kiện đó.

3.2. Phương pháp phân loại văn bản dựa trên mô hình cự ly trắc địa

Phương pháp này đề xuất sử dụng cự ly đường trắc địa nhằm để tính hàm nhân trong ma trận nhân của máy véc tơ hỗ trợ. Đây là phương pháp sử dụng tất cả các dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn của tập huấn luyện để tính cự ly trắc địa [52], nên có thể nói phương pháp đề xuất là một trong những phương pháp học bán giám sát [68].

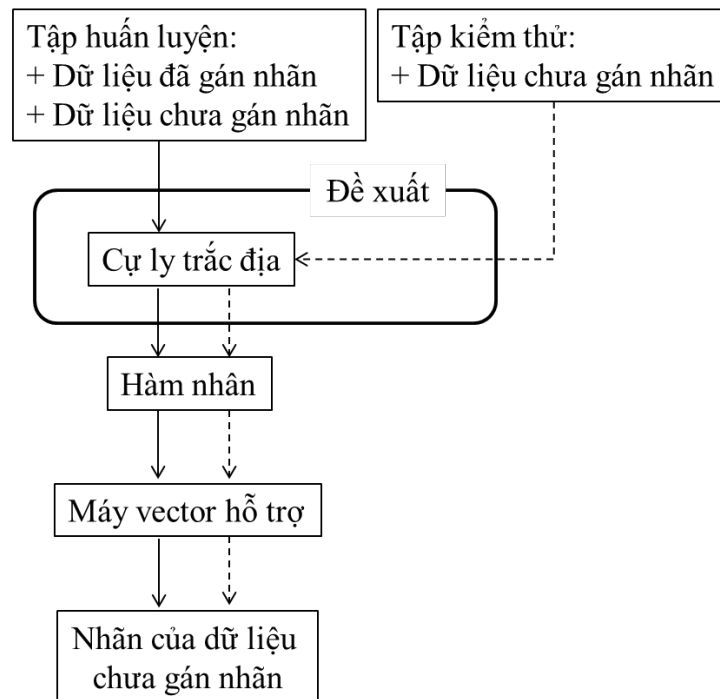
Phương pháp đề xuất này chủ yếu dựa trên mô hình véc tơ hỗ trợ. Điểm khác biệt của phương pháp đề xuất là sử dụng cự ly trắc địa để tính hàm nhân cho máy véc tơ hỗ trợ thay vì sử dụng cự ly Euclid. Mô hình đề xuất gồm có 3 phần chính:

Phần thứ nhất: tính cự ly trắc địa cho dữ liệu bao gồm cả dữ liệu của tập huấn luyện và dữ liệu của tập kiểm thử. Phần này bao gồm cả các phương pháp trích chọn đặc tính của văn bản tiếng Việt.

Phần thứ hai: sử dụng cự ly trắc địa để tính hàm nhân trong ma trận nhân của máy véc tơ hỗ trợ.

Phần thứ ba: sử dụng máy véc tơ hỗ trợ để phân loại (gán nhãn) cho văn bản số tiếng Việt.

Mô hình đề xuất như sau:



Hình 3.3 Mô hình đề xuất phân loại văn bản dựa trên cụm ly trích địa

3.3. Thực nghiệm phân loại văn bản dựa trên mô hình cụm ly trích địa

3.3.1. Phát triển chương trình ứng dụng

Trong phần này, tập hợp tất cả các nghiên cứu đã trình bày để xây dựng chương trình phân loại văn bản và sử dụng một số thư viện hỗ trợ trong việc xử lý số liệu, đó là các thư viện Accord.Controls.dll, Accord.dll, Accord.MachineLearning.dll, Accord.Math.dll, Accord.Statistics.dll, AForge.dll, AForge.Math.dll, xd2txlib.dll và ZedGraph.dll. Ngoài ra, chương trình không đi sâu vào giới thiệu các chức năng của công cụ hỗ trợ mà trực tiếp đi vào việc thực nghiệm bộ phân loại văn bản dựa trên mô hình trích địa đã đề xuất.

3.3.2. Chuẩn bị dữ liệu

Trong phân loại văn bản, khối lượng dữ liệu lớn không là yếu tố quyết định mà yếu tố được quan tâm nhất là chất lượng của kho dữ liệu. Chất lượng của kho dữ liệu chính là độ phù hợp khi gán một văn bản mẫu vào một chủ đề.

Để phục vụ quá trình nghiên cứu phân loại văn bản tiếng Việt, trước hết đã tiến hành xây dựng một kho dữ liệu để sử dụng thống nhất cho cả giai đoạn huấn luyện

và kiểm thử để đánh giá. Kho dữ liệu này gồm tập hợp các văn bản được tạo ra theo từng chủ đề để đã xác định trước nhằm phục vụ phân loại văn bản tiếng Việt, gồm các tập dữ liệu huấn luyện (Training) và các tập dữ liệu kiểm thử (Testing).

Tập dữ liệu huấn luyện chứa các văn bản đã được gán chủ đề trước, dùng để phục vụ cho giai đoạn huấn luyện. Tập dữ liệu kiểm thử chứa các văn bản chưa dán nhãn và dùng làm dữ liệu thử nghiệm độ chính xác của phần mềm phân loại.

Kho dữ liệu này bao gồm các bài viết trên Website của các báo điện tử được nhiều người quan tâm nhất đó là (Vnexpress, Vietnamnet, Dantri và Tuoitre) theo các chủ đề Bóng đá, giáo dục, pháp luật, quốc tế, xã hội. Tất cả đều đã được xử lý và lưu trữ dưới dạng .TXT với tổng dung lượng khoảng 144 MB.

Bảng 3.1 Thống kê số tập tin trong kho dữ liệu

| STT | Loại tài liệu | Huấn luyện | | Kiểm thử | Tổng |
|-----|---------------|------------|--------------|----------|-------------|
| | | Gán nhãn | Chưa có nhãn | | |
| 1 | Bóng đá | 10 | 613 | 400 | 1023 |
| 2 | Giáo dục | 10 | 604 | 400 | 1014 |
| 3 | Pháp luật | 10 | 577 | 400 | 987 |
| 4 | Quốc tế | 10 | 599 | 400 | 1009 |
| 5 | Xã hội | 10 | 584 | 400 | 994 |

Khi áp dụng trong thực tế số lượng văn bản chưa gán nhãn rất nhiều (vài triệu văn bản) còn số lượng văn bản được gán nhãn rất ít (vài trăm, vài nghìn văn bản) cho nên tỷ lệ việc phân như trên là hợp lý.

Trong quá trình thực nghiệm, nghiên cứu này chia ngẫu nhiên từng thể loại văn bản thành hai phần, phần dữ liệu đã gán nhãn và phần dữ liệu chưa gán nhãn. Phần dữ liệu đã gán nhãn bao gồm 10 văn bản đã biết trước thể loại (đã được gán nhãn) và phần thứ hai là các văn bản chưa biết trước thể loại (chưa được gán nhãn). Trong phần thứ 2, nghiên cứu lại tiếp tục chia làm hai phần. Đầu tiên là chọn ngẫu nhiên 400 văn bản cho việc kiểm thử, các văn bản còn lại được sử dụng trong việc xây dựng mô hình cụ li đường trắc địa kết hợp với 10 văn bản đã được gán nhãn ở phần một. Bảng 3.1 là bảng thống kê số tập tin trong kho dữ liệu cho từng thể loại.

- **Ví dụ:** Thẻ loại Bóng đá có 1023 văn bản, trong đó có 10 văn bản được gán nhãn và 1013 văn bản chưa được gán nhãn. Nghiên cứu tiếp tục chia số văn bản chưa gán nhãn làm hai phần phục vụ trong việc huấn luyện và kiểm thử. Trong đó số 613 văn bản chưa được gán nhãn sử dụng trong xây dựng mô hình cụ li đường trắc địa để huấn luyện. Phần còn lại, 400 văn bản được sử dụng để kiểm thử. Khi đánh giá việc phân loại văn bản, nghiên cứu này sử dụng cả 1013 văn bản cho việc tính tỷ lệ phân loại đúng, bởi vì thực tế 1013 văn bản này đều chưa gán nhãn trước khi thực nghiệm. Tương tự thẻ loại Giáo dục có 1014 văn bản, trong đó có 10 văn bản đã được gán nhãn và 1004 văn bản chưa được gán nhãn, ...

3.3.3. Triển khai chương trình

Để chương trình có thể chạy tốt thì cần có một số yêu cầu tối thiểu về môi trường cài đặt như sau:

- Microsoft Windows XP
- Nền .Net Framework 4.0
- Ram 256MB
- Tốc độ CPU lớn hơn 1.72 GHz

Việc cài đặt rất đơn giản, chỉ cần copy tất cả các tập tin trong thư mục có chứa tập tin thực hiện “ChuongTrinhPhanLoaiVanBan.exe”qua máy của người dùng rồi khởi động chương trình.

Trong nghiên cứu này đã xây dựng một phần mềm thử nghiệm để phân loại văn bản gồm có 2 chức năng chính đó là chức năng huấn luyện và chức năng phân loại văn bản dựa trên mô hình phân loại đã huấn luyện.

- **Chức năng huấn luyện (Learning):** chuyển tự động tất cả các các tập tin văn bản ở dạng khác nhau thành dạng chuẩn là tập tin .txt, tập tin này chỉ gồm các chuỗi ký tự tiếng Việt và không có thêm các thông tin phụ khác như hình ảnh, kiểu chữ hay kích cỡ chữ... cho phép người sử dụng thêm tập tin văn bản vào chương trình, gán nhãn cho các tập tin văn bản cần huấn luyện.

✓ Cho phép người sử dụng lựa chọn các phương pháp trích chọn đặc tính gồm: Đặc tính theo tần số xuất hiện không khử nhiễu, đặc tính theo tần số xuất hiện có

khử nhiễu, đặc tính IT-IDF, tạo và lưu trữ tập huấn luyện và tự động học để tạo ra mô hình phân loại.

✓ Cho phép người sử dụng lựa chọn các phương pháp học máy gồm: KDA sử dụng hàm nhân là hàm Gaussian, KDA sử dụng hàm nhân là hàm Gaussian với cự ly trắc địa, SVM sử dụng hàm nhân là hàm Gaussian, SVM sử dụng hàm nhân là hàm Gaussian với cự ly trắc địa, huấn luyện và lưu mô hình đã huấn luyện thành một project.

Các thao tác sử dụng mô hình phân loại văn bản: Sau khi khởi động chương trình phân loại văn bản, giao diện chính xuất hiện, chọn “Model Learning” chương trình sẽ xuất hiện ra cửa sổ giáo diện “Learning” gồm các chức năng:

Add File Names: Cho phép thêm các tập tin văn bản tiếng Việt để bổ sung các văn bản dùng để huấn luyện.

Add Label: Để gán nhãn cho các văn bản, ta có thể thêm các nhãn bằng cách nhập tên vào ô bên phải nút “Add Label” và nhấn nút này. Ví dụ: Ta nhập “Giao Duc” và nhấn nút “Add Label” thì nhãn “Giao Doc with Label” để gán nhãn cho các tập tin văn bản tiếng Việt đã chọn. Tương tự theo các thao tác trên ta có thể thêm nhiều văn bản và nhiều loại nhãn khác nhau vào tập huấn luyện. Lưu tập huấn luyện nhấn nút “Save”, khi cần sử dụng lại nhấn nút “Load”.

Muốn xây dựng mô hình phân loại, ta chọn mục “Feature Extraction” cho việc chọn phương pháp trích chọn đặc tính và chọn mục “Machine Learning Model” cho việc chọn phương pháp phân loại của học máy. Sau đó, điền các thông tin cần thiết ở bên góc phải dưới như tên “Project”, địa chỉ lưu trữ mô hình phân loại hay một số thông tin khi thực hiện mô hình. Cuối cùng nhấn nút “Learn and Save” để chương trình tự động xây dựng và lưu trữ mô hình phân loại với các phương pháp đã chọn.

- **Chức năng phân loại văn bản** (Document Classification): cho phép người dùng nhập một mô hình phân loại đã xây dựng và chỉ định tập tin cần phân loại. Kết quả là phần mềm sẽ tự động gán nhãn cho tập tin này.

Sau khi khởi động chương trình chọn nhấn vào nút “Document Classification”

chương trình sẽ hiện ra cửa sổ giao diện” DocClassification”, để nhập một mô hình phân loại đã xây dựng vào chương trình ta nhấn nút “Load Model”. Cửa sổ “Browse For Folder” hiện ra và cho phép người dùng chỉ định đường dẫn của thư mục có chứa mô hình phân loại đã xây dựng.

Người sử dụng có thể chọn một hay nhiều tập tin văn bản cùng một lúc để phân loại bằng cách nhấn vào nút “Open files” trong mục “Document Classification”.

3.3.4. Kết quả thực nghiệm

Tiến hành thực hiện phương pháp phân loại văn bản dựa trên học bán giám sát SVM không hỗ trợ mô hình cự ly trắc địa và phân loại văn bản dựa trên học bán giám sát SVM có hỗ trợ mô hình cự ly trắc địa, sau đó so sánh và đánh giá kết quả của hai phương pháp bằng cách gán nhãn cho hơn 1000 bài báo đúng với từng loại chủ đề đã xây dựng trong kho dữ liệu như đã trình bày ở trên và phân loại tất cả các bài báo theo từng chủ đề đã thu thập, cột dọc thể hiện nhãn thực tế, cột ngang thể hiện nhãn có được từ kết quả phân loại.

Nghiên cứu này đã tiến hành thử nghiệm 100 lần ngẫu nhiên, với mỗi lần thử nghiệm việc chia dữ liệu gán nhãn (10 văn bản cho từng thể loại) và các văn bản còn lại không gán nhãn là ngẫu nhiên, từ đó chọn ngẫu nhiên kết quả 5 lần thực nghiệm được trình bày ở các bảng sau.

a. Lần thử nghiệm thứ nhất

Bảng 3.2 Kết quả phân loại lần 1 sử dụng SVM

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 887 | 0 | 58 | 78 | 0 | 86.7% |
| Giáo dục | 0 | 516 | 225 | 159 | 114 | 51.0% |
| Pháp Luật | 24 | 0 | 864 | 62 | 37 | 87.5% |
| Quốc Tế | 0 | 65 | 16 | 895 | 34 | 88.7% |
| Xã hội | 0 | 108 | 277 | 253 | 356 | 35.8% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 69.9% |

Bảng 3.3 Kết quả phân loại lần 1 sử dụng SVM với mô hình cự ly trắc địa

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 769 | 105 | 34 | 115 | 0 | 75.2% |
| Giáo dục | 0 | 821 | 104 | 89 | 0 | 81.0% |
| Pháp Luật | 25 | 41 | 864 | 47 | 10 | 87.5% |
| Quốc Tế | 17 | 23 | 21 | 932 | 16 | 92.4% |
| Xã hội | 73 | 67 | 172 | 326 | 356 | 35.8% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 74.4% |

Trong thử nghiệm lần thứ nhất, kết quả dễ dàng thấy rằng tỷ lệ phân loại thành công của SVM trong mục bóng đá, pháp luật, xã hội cao hơn khi sử dụng SVM có kết hợp với mô hình trắc địa (phương pháp đề xuất). Tuy nhiên, các mục khác lại có tỷ lệ phân loại thành công thấp hơn. Kết quả trung bình của tỷ lệ phân loại thành công của tất các mục là 69.9% khi sử dụng SVM và 74.4% khi sử dụng phương pháp đề xuất.

b. Lần thử nghiệm thứ 2

Bảng 3.4 Kết quả phân loại lần 2 sử dụng SVM

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 868 | 63 | 34 | 0 | 58 | 84.8% |
| Giáo dục | 0 | 888 | 43 | 0 | 83 | 87.6% |
| Pháp Luật | 0 | 35 | 878 | 6 | 68 | 89.0% |
| Quốc Tế | 0 | 18 | 122 | 826 | 43 | 81.9% |
| Xã hội | 45 | 29 | 502 | 29 | 389 | 39.1% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 76.5% |

Bảng 3.5 Kết quả phân loại lần 2 sử dụng SVM với mô hình cự ly trắc địa

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 808 | 0 | 0 | 184 | 31 | 79.0% |
| Giáo dục | 0 | 676 | 0 | 279 | 59 | 66.7% |
| Pháp Luật | 0 | 0 | 593 | 276 | 118 | 60.1% |
| Quốc Tế | 15 | 0 | 0 | 899 | 95 | 89.1% |
| Xã hội | 0 | 0 | 54 | 378 | 562 | 56.5% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 70.3% |

Trong thử nghiệm lần 2, ta có kết quả ngược lại với kết quả lần thứ nhất. Kết quả trung bình của tỷ lệ phân loại thành công của tất các mục là 76.5% khi sử dụng SVM và 70.3% khi sử dụng phương pháp đề xuất.

c. Lần thử nghiệm thứ 3

Bảng 3.6 Kết quả phân loại lần 3 sử dụng SVM

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 721 | 0 | 7 | 295 | 0 | 70.5% |
| Giáo dục | 0 | 763 | 0 | 234 | 17 | 75.2% |
| Pháp Luật | 0 | 22 | 674 | 291 | 0 | 68.3% |
| Quốc Tế | 0 | 19 | 0 | 990 | 0 | 98.1% |
| Xã hội | 0 | 51 | 83 | 557 | 303 | 30.5% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 68.5% |

Bảng 3.7 Kết quả phân loại lần 3 sử dụng SVM với mô hình cự ly trắc địa

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 750 | 0 | 126 | 147 | 0 | 73.3% |
| Giáo dục | 0 | 879 | 117 | 18 | 0 | 86.7% |
| Pháp Luật | 0 | 81 | 840 | 41 | 23 | 85.1% |
| Quốc Tế | 0 | 33 | 242 | 720 | 14 | 71.4% |
| Xã hội | 0 | 74 | 261 | 208 | 451 | 45.3% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 72.4% |

Trong thử nghiệm lần 3, ta có kết quả trung bình của tỷ lệ phân loại thành công của tất các mục là 68.5% khi sử dụng SVM và 72.4% khi sử dụng phương pháp đề xuất. Phương pháp đề xuất cho kết quả tốt hơn khi chỉ sử dụng SVM.

d. Lần thử nghiệm thứ 4

Bảng 3.8 Kết quả phân loại lần 4 sử dụng SVM

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 759 | 25 | 22 | 217 | 0 | 74.2% |
| Giáo dục | 14 | 737 | 71 | 179 | 13 | 72.7% |
| Pháp Luật | 0 | 48 | 689 | 181 | 69 | 69.8% |
| Quốc Tế | 21 | 54 | 68 | 808 | 58 | 80.1% |
| Xã hội | 3 | 83 | 177 | 158 | 573 | 57.6% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 70.9% |

Bảng 3.9 Kết quả phân loại lần 4 sử dụng SVM với mô hình cự ly trắc địa

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 834 | 25 | 28 | 136 | 0 | 81.5% |
| Giáo dục | 14 | 778 | 31 | 179 | 12 | 76.7% |
| Pháp Luật | 0 | 50 | 689 | 178 | 70 | 69.8% |
| Quốc Tế | 21 | 52 | 54 | 824 | 58 | 81.7% |
| Xã hội | 3 | 83 | 209 | 156 | 543 | 54.6% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 72.9% |

Trong thử nghiệm lần 4, ta có kết quả trung bình của tỷ lệ phân loại thành công của tất các mục là 70.9% khi sử dụng SVM và 72.9% khi sử dụng phương pháp đề xuất. Phương pháp đề xuất cho kết quả tốt hơn khi chỉ sử dụng SVM.

e. Lần thử nghiệm thứ 5

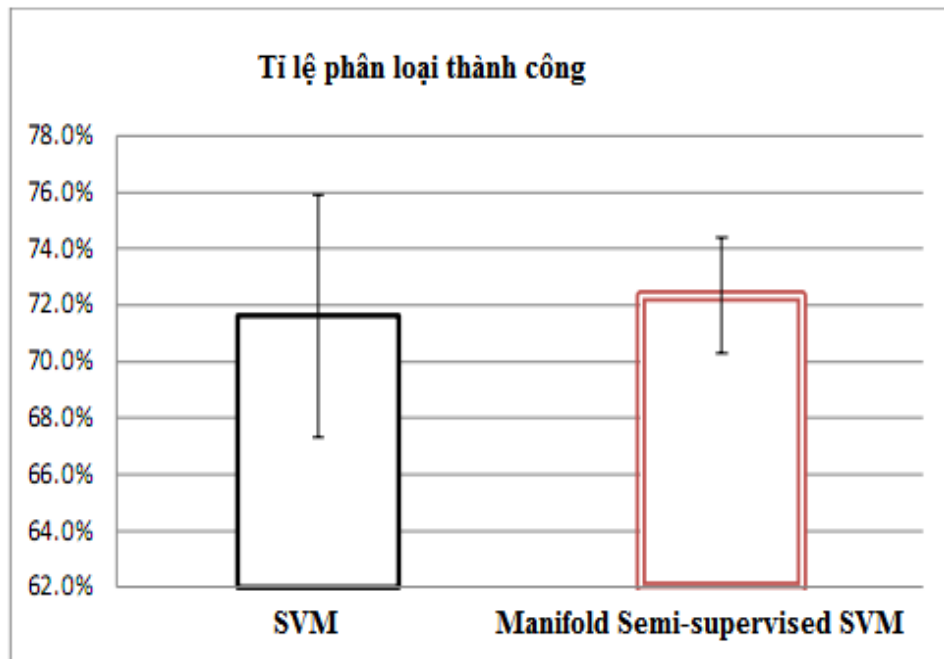
Bảng 3.10 Kết quả phân loại lần 5 sử dụng SVM

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 776 | 34 | 19 | 194 | 0 | 75.9% |
| Giáo dục | 14 | 725 | 75 | 179 | 21 | 71.5% |
| Pháp Luật | 0 | 46 | 692 | 184 | 65 | 70.1% |
| Quốc Tế | 12 | 41 | 54 | 805 | 97 | 79.8% |
| Xã hội | 11 | 83 | 241 | 156 | 503 | 50.6% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 69.6% |

Bảng 3.11 Kết quả phân loại lần 5 sử dụng SVM với mô hình cự ly trắc địa

| Nhãn thực tế | Nhãn có được từ kết quả phân loại | | | | | |
|---------------------------------------|-----------------------------------|------------|------------|------------|------------|-----------------|
| | Bóng Đá | Giáo dục | Pháp Luật | Quốc Tế | Xã hội | Tỷ lệ phân loại |
| Bóng Đá | 736 | 26 | 43 | 218 | 0 | 71.9% |
| Giáo dục | 0 | 799 | 121 | 42 | 52 | 78.8% |
| Pháp Luật | 17 | 55 | 795 | 98 | 42 | 80.5% |
| Quốc Tế | 0 | 27 | 134 | 792 | 56 | 78.5% |
| Xã hội | 49 | 51 | 168 | 153 | 573 | 57.6% |
| Tỷ lệ phân loại thành công trung bình | | | | | | 73.5% |

Trong thử nghiệm lần 5, ta có kết quả trung bình của tỷ lệ phân loại thành công của tất các mục là 69.6% khi sử dụng SVM và 73.5% khi sử dụng phương pháp đề xuất. Phương pháp đề xuất cho kết quả tốt hơn khi chỉ sử dụng SVM.



Hình 3.4 Giá trị trung bình và độ lệch chuẩn của tỷ lệ phân loại

Hình trên biểu diễn giá trị trung bình và căn phương sai của tỷ lệ phân loại thành công sử dụng SVM và phương pháp đề xuất.

Đánh giá phương pháp đề xuất bằng thực nghiệm với chương trình đã xây dựng. Tôi đã so sánh phương pháp đề xuất với SVM thuần túy trên cùng một bộ dữ liệu. Ta có thể thấy rằng mặc dù giá trị trung bình tỷ lệ phân loại của hai phương pháp không chênh lệch nhiều về kết quả, tuy nhiên độ lệch chuẩn của phương pháp đề xuất ($\pm 2\%$) nhỏ hơn nhiều so với SVM ($\pm 4\%$). Điều đó cho thấy phương pháp đề xuất ổn định hơn so với sử dụng SVM thuần túy.

3.4. Tiểu kết chương

Kết quả nghiên cứu, qua thực nghiệm nghiên cứu ứng dụng kỹ thuật phân loại văn bản tiếng Việt dựa trên đề xuất giải pháp sử dụng mô hình cự ly đường trắc địa kết hợp với máy véc tơ hỗ trợ. Mô hình cự ly đường trắc địa sử dụng hệ tương quan ngắn nhất (mức độ gần nhau giữa các văn bản) để tính khoảng cách giữa hai véc tơ. Cự ly trắc địa này khác với khoảng cách Euclid và giúp cho việc phân loại văn bản tự động sẽ chính xác hơn và cho phép phân thành nhiều loại thay vì chỉ phân ra hai loại (dựa trên phân lớp nhị phân).

Về cơ bản, chương trình phân loại văn bản đã thực hiện hoàn thành được các chức năng đã đặt ra là giúp người sử dụng xây dựng mô hình phân loại cho các loại văn bản tiếng Việt. Tự động phân loại các văn bản mới dựa trên mô hình đã xây dựng. Tuy nhiên việc thu thập dữ liệu ban đầu chỉ mới ở mức thử nghiệm.

Trong thời gian tới, sẽ tiếp tục nghiên cứu bổ sung một số tính năng mới và hoàn thiện chương trình để nâng cao hiệu quả phân loại, đồng thời xây dựng kho dữ liệu đủ lớn nhằm mục đích phân loại văn bản một cách chính xác hơn.

Tuy nhiên, trong giải pháp đề xuất sử dụng mô hình cự ly đường trắc địa kết hợp với máy véc tơ hỗ trợ (SVM) để phân loại văn bản tiếng Việt, còn tồn tại một số hạn chế như tỷ lệ cho việc nhận dạng và phân loại sai do ảnh hưởng của các từ đồng nghĩa, các từ đồng âm nhưng khác nghĩa, các từ không có hoặc các từ ít có nghĩa, tốc độ xử lý tính toán còn chậm, do số chiều véc tơ đặc trưng trong văn bản rất lớn. Từ đó đề xuất giải pháp tiếp theo là gom cụm các từ đồng nghĩa, gần nghĩa, rút gọn số chiều véc tơ bằng phương pháp phân cụm dựa trên đồ thị.

Chương 4. RÚT GỌN SỐ CHIỀU VEC TƠ DỰA TRÊN ĐỒ THỊ DENDROGRAM

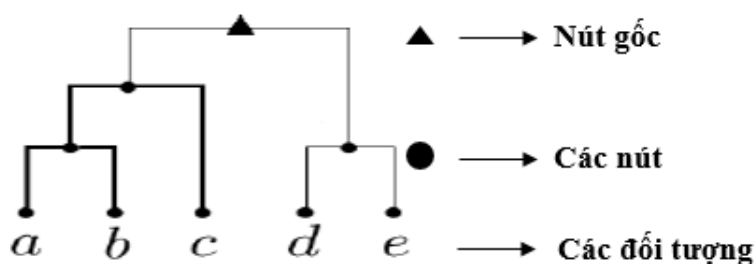
Nội dung chương này trình bày về giải pháp đề xuất để rút gọn số chiều vec tơ biểu diễn văn bản tiếng Việt dựa trên đồ thị Dendrogram và tập văn bản lấy từ Wikipedia. Việc rút gọn số chiều vec tơ sẽ được áp dụng vào quá trình phân loại văn bản tiếng Việt thông qua các thử nghiệm. Kết quả đạt được trong nghiên cứu đề xuất một giải pháp mới để rút gọn số chiều của vec tơ biểu diễn văn bản dựa trên đồ thị Dendrogram là đúng và chưa được công bố trong bất kỳ công trình nghiên cứu khoa học nào khác.

4.1. Giới thiệu

4.1.1. Định nghĩa đồ thị Dendrogram

Định nghĩa : Một Dendrogram là một đồ thị nhị phân bắt nguồn từ một gốc với các loại đỉnh như sau [86]:

- Các đỉnh của mức 1, gọi là các đối tượng.
- Các đỉnh của mức 2, được gọi là các nút.
- Chỉ có một đỉnh của mức 2 được gọi là nút gốc.



Hình 4.1. Đồ thị Dendrogram

4.1.2. Giải pháp đề xuất

Chúng ta có nhiều cách để biểu diễn một văn bản nhưng cách thông dụng nhất đó là biểu diễn thông qua vec tơ. Khi biểu diễn bằng vec tơ, một trong những vấn đề gặp phải đó là số chiều vec tơ quá lớn dẫn đến việc xử lý chậm. Cách thông dụng

nhất để giải quyết vấn đề này đó là tìm cách rút gọn số chiều véc tơ bằng cách loại bỏ bớt những phần tử không quan trọng trong véc tơ (ví dụ: tần suất xuất hiện quá bé, các từ không quan trọng về mặt ngữ nghĩa, ...). Tuy nhiên, đối với tiếng Việt chưa có giải pháp nào hiệu quả để rút gọn số chiều véc tơ phục vụ phân loại văn bản.

Giải pháp của luận án đề xuất dựa trên ý tưởng là chỉ giữ lại những từ đại diện cho một nhóm các từ mà tần suất xuất hiện cùng nhau lớn. Ví dụ, khi nói về bóng đá, tập các từ như *sút*, *phạt góc*, *ném biên*, *việt vị*, ... thường xuất hiện cùng nhau trong một văn bản khá cao nên trong véc tơ biểu diễn ta chỉ cần giữ lại một phần tử đại diện cho nhóm từ này và bỏ đi các phần tử tương ứng với các từ khác để giảm số phần tử của véc tơ (giảm số chiều của véc tơ). Câu hỏi đặt ra là làm thế nào để biết những từ nào thường xuất hiện cùng nhau? Giải pháp ở đây là thống kê tần suất xuất hiện cùng nhau của các từ trên một tập dữ liệu lớn đó là các tài liệu trên Wikipedia. Khi đã tìm ra tần suất xuất hiện cùng nhau trên các văn bản, ta sẽ gom chúng lại thành các cụm và biểu diễn chúng thông qua đồ thị Dendrogram. Đồ thị Dendrogram là công cụ biểu diễn dạng cây, phân thành nhiều tầng và mỗi đồ thị con theo tầng là những từ có tần suất xuất hiện cùng nhau lớn. Dựa trên đồ thị biểu diễn này, hệ thống phân loại (hoặc người sử dụng hệ thống) sẽ quyết định tỉ lệ rút gọn số chiều của véc tơ sao cho đáp ứng được yêu cầu phân loại chính xác và đảm bảo thời gian tính toán mong muốn [40], [65].

Quá trình xây dựng đồ thị Dendrogram được thực hiện như sau:

1) **Bước 1:** Tiền xử lý dữ liệu Wikipedia. Tải tất cả các tài liệu trên Wikipedia và tiền xử lý để có được một tập gồm n tài liệu $D = \{d_1, d_2, \dots, d_n\}$.

2) **Bước 2:** Xây dựng ma trận P thể hiện tần số xuất hiện chung của các cặp từ trên cùng một trang Wikipedia. Ma trận P ở đây là một ma trận vuông m (với m là số mục từ có trong từ điển tiếng Việt, trường hợp thử nghiệm dùng từ điển Hồ Ngọc Đức có 39.000 mục từ), giá trị của mỗi phần tử của ma trận này chính là tần suất xuất hiện cùng nhau của mỗi cặp từ trên các văn bản lấy từ Wikipedia.

$$\begin{array}{c}
a_1 \quad a_2 \quad a_3 \quad \dots \quad a_m \\
\begin{array}{c} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_m \end{array} \left[\begin{array}{ccccc} p_{1,1} & p_{1,2} & p_{1,3} & \dots & p_{1,m} \\ p_{2,1} & p_{2,2} & p_{2,3} & \dots & p_{2,m} \\ p_{3,1} & p_{3,2} & p_{3,3} & \dots & p_{3,m} \\ \dots & \dots & \dots & \dots & \dots \\ p_{m,1} & p_{m,2} & p_{m,3} & \dots & p_{m,m} \end{array} \right]
\end{array}$$

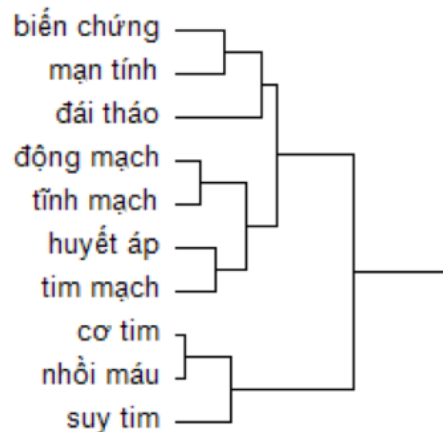
Trong đó, a_i là mục từ trong từ điển và $p_{i,j}$ là tần suất mà cặp từ a_i và a_j cùng xuất hiện trên các tập các tài liệu Wikipedia. Khi thực hiện, ta sẽ quét qua tất cả các tài liệu Wikipedia và thay đổi giá trị $p_{i,j}$ nếu hai từ a_i và a_j cùng xuất hiện trong văn bản.

3) **Bước 3:** Xây dựng ma trận xác suất W dựa trên ma trận tần suất P , các phần tử của ma trận mới sẽ thể hiện xác suất xuất hiện cùng nhau của các cặp từ thay cho tần suất xuất hiện. Phép biến đổi này cho phép tăng vai trò của các từ có tần suất xuất hiện không lớn nhưng thường đi cùng nhau. Các phần tử ma trận W được tính theo công thức:

$$w[i, j] = \frac{p_{i,j}}{p_{i,i} + p_{j,j} - p_{i,j}}$$

4) **Bước 4:** Xây dựng đồ thị Dendrogram dựa trên xác suất xuất hiện cùng nhau của các cặp từ. Cách xây dựng đồ thị là tạo các đồ thị con và sau đó nhóm chúng lại với nhau dựa trên xác suất xuất hiện cùng nhau của các từ.

Ví dụ: sau khi tính được xác suất xuất hiện cùng nhau giữa các từ, ta tạo được một đồ thị Dendrogram như sau:



5) Bước 5: Rút gọn số chiều véc tơ bằng cách xoá bớt các phần tử cùng cây con và áp dụng véc tơ rút gọn vào bài toán phân loại. Ví dụ: trong véc tơ biểu diễn văn bản khi phân loại, trong cụm các từ *biến chứng*, *mạn tính*, *đái tháo* ta xoá bớt các phần tử tương ứng với các từ *biến chứng* và *mạn tính*, chỉ giữ lại từ *đái tháo*; cụm các từ *động mạch*, *tĩnh mạch*, *huyết áp*, *tim mạch* ta chỉ giữ lại từ *tim mạch*; cụm các từ *cơ tim*, *nhồi máu*, *suy tim* ta chỉ giữ lại *suy tim*. Như vậy, thay vì véc tơ biểu diễn các từ trên gồm 10 phần tử thì được rút gọn lại còn 3 phần tử.

Các bước xây dựng đồ thị Dendrogram ở **Bước 4** được trình bày một cách tổng quát như sau:

- Bước 1.** Đặt tất cả các dữ liệu thành từng nhóm riêng lẻ (dựa trên ma trận xác suất). Ban đầu mỗi phần tử ma trận được xem như một nhóm.
- Bước 2.** Từ ma trận khoảng cách các nhóm, gom hai nhóm có khoảng cách gần nhất thành một nhóm.
- Bước 3.** Nếu số lượng nhóm là một thì kết thúc. Ngược lại thì thực hiện Bước 4.
- Bước 4.** Tính khoảng cách nhóm vừa được tạo ra ở Bước 2 với các nhóm còn lại và cập nhật ma trận khoảng cách.
- Bước 5.** Quay lại Bước 2.

Có rất nhiều phương pháp tính để tính khoảng cách giữa hai nhóm. Dựa theo tính chất của từng dữ liệu, ta có thể sử dụng các phương pháp tính khoảng cách sau:

- Phương pháp láng giềng gần nhất (Nearest neighbor method): Khoảng cách giữa hai nhóm được tính bởi khoảng cách nhỏ nhất trong tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

- Phương pháp xa (Furthest neighbor method): Khoảng cách giữa hai nhóm được tính bởi khoảng cách lớn nhất trong tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

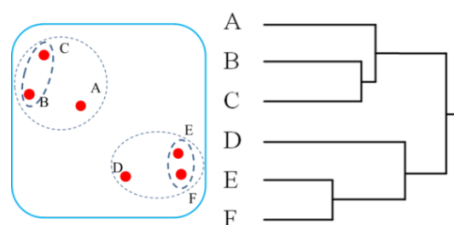
- Phương pháp trung bình nhóm (Group average method): Khoảng cách giữa hai nhóm được tính bởi khoảng cách trung bình của tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

- Phương pháp trọng tâm (Centroid method): Khoảng cách giữa hai nhóm được tính bởi khoảng cách trọng tâm của hai nhóm.

- Phương pháp Wards (Wards method): Khoảng cách giữa hai nhóm được tính bởi tổng bình phương khoảng cách của tất cả các cặp dữ liệu thuộc hai nhóm khác nhau.

Khoảng cách ở đây có thể được tính bằng nhiều cách khác nhau. Nếu các dữ liệu được thể hiện bằng các véc tơ hay các điểm trong không gian Euclid thì ta có thể sử dụng khoảng cách Euclid hay khoảng cách Minkowski để tính. Tuy nhiên tùy theo tính chất của bài toán hay dữ liệu mà chúng ta có thể định nghĩa khoảng cách bằng các phương pháp khác như sử dụng khoảng cách Manhattan, khoảng cách Mahalanobis, xác suất, hệ số tương quan, ... Đối với văn bản thì ta còn có thể tính khoảng cách dựa theo hệ số tương quan về từ, về cấu trúc câu, về ngữ nghĩa của hai văn bản.

Trong luận án này, sử dụng xác suất xuất hiện cùng nhau trên một văn bản để tính khoảng cách giữa hai từ trong tiếng Việt.



Hình 4.2 Ví dụ về đồ thị Dendrogram

Hình trên là một ví dụ cách xây dựng đồ thị dendrogram dựa trên phương pháp láng giềng gần nhất với khoảng cách Euclid. Phần bên trái là các véc tơ “A”, ”B”,

“C”, “F”, “E”, “D”, trong không gian hai chiều. Ta thấy “E” và “F” có khoảng cách nhỏ nhất nên được gom thành một nhóm gồm hai phần tử. Tương tự ta cũng có “B” và “C” cũng được gom thành một nhóm. Từ các nhóm nhỏ, ta lại có được các nhóm lớn hơn nhờ việc gom các nhóm nhỏ lại với nhau. Ta được các nhóm “A, B, C” và nhóm “D, E, F”. Kết quả là cuối cùng tất cả các đối tượng được gom lại thành một nhóm.

4.2. Xây dựng đồ thị Dendrogram từ dữ liệu Wikipedia

4.2.1. Thuật toán xử lý Wikipedia

Wikipedia là một bách khoa toàn thư mở với nhiều ngôn ngữ thể hiện dưới một trang web trên Internet. Tôi đã lấy tất cả các dữ liệu tiếng Việt của Wikipedia với dung lượng 1.3 GB định dạng *.7z. Sau khi giải nén thu được tập tin *.xml và có kích thước 91.8 GB.

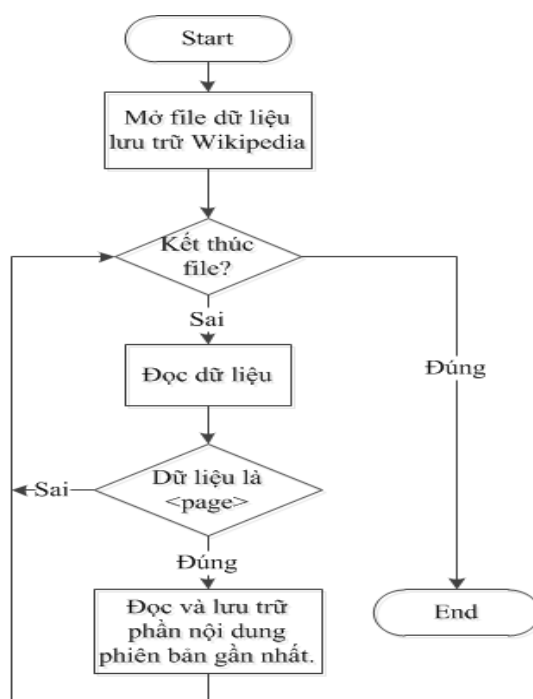
Định dạng dữ liệu Wikipedia được rút gọn như sau:

```
<page>
<title></title>
<id></id>
<revision>
<comment></comment>
<timestamp></timestamp>
<text></text>
...
</revision>
<revision>
<comment></comment>
<timestamp></timestamp>
<text></text>
...
</revision>
...
</page>
```

Mỗi trang Wikipedia được lưu trong thẻ `<page></page>`, trong đó bao gồm nhiều trường khác nhau như `<title></title>`, `<id></id>`, `<comment></comment>`, ... Trong số đó có nhiều thẻ `<revision></revision>` tương ứng với nhiều phiên bản của một trang Wikipedia. Tương ứng trong mỗi thẻ `<revision>` có nhiều trường khác nhau, trong đó chỉ chú ý đến hai trường là thời gian `<timestamp></timestamp>` và nội dung `<text></text>`.

Vì dữ liệu lưu trữ quá lớn, nên cần xử lý để rút gọn bớt tập tin dữ liệu. Trong nghiên cứu này, tôi chỉ tập trung vào phần nội dung của mỗi `<page>`, do đó cần loại bỏ tất cả các trường còn lại. Thêm nữa, mỗi `<page>` lại chứa nhiều phiên bản, nên chỉ cần lấy phiên bản gần nhất để lưu trữ rồi xử lý.

Với đầu vào là một file HTML được tải về trên Wikipedia và đầu ra là một file `.txt` chứa nội dung là bài viết mới nhất, ta có sơ đồ thuật toán như sau:



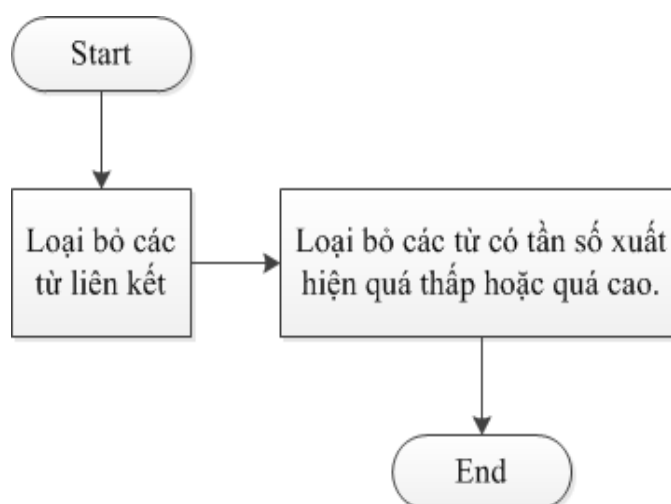
Hình 4.3 Lưu đồ thuật toán xử lý tập tin dữ liệu Wikipedia

4.2.2. Thuật toán xử lý từ điển

Tập tin từ điển được lấy từ bộ từ điển online của Hồ Ngọc Đức gồm 39000 từ tiếng Việt.

Để không gây nhiễu trong quá trình tính toán, nhiều từ đã được lược bỏ. Đầu tiên lược bỏ các từ liên kết từ như “là”, “và”, “hoặc”, ... từ điển còn lại 34520 từ. Thông qua việc phân tích tần số xuất hiện trên Wikipedia, các từ có tần số quá thấp hoặc quá cao sẽ được loại bỏ vì khả năng gom thành cụm của các từ có tần số xuất hiện thấp là rất thấp. Các từ có tần số xuất hiện quá cao là các từ khóa của Wikipedia như “tham khảo”, “chú thích”, ..., các từ này có mặt hầu như mọi trang Wikipedia. Qua quá trình này, từ điển tiếp tục được rút gọn còn 14015 từ.

Thuật toán:



Hình 4.4 Sơ đồ thuật toán xử lý từ điển

4.2.3. Thuật toán tính toán ma trận P tần số xuất hiện chung

Từ đầu vào là các tập tin .txt tương ứng với các văn bản đã được xử lý của Wikipedia và danh sách các từ rút gọn từ từ điển ta sẽ có đầu ra là một ma trận tần suất xuất hiện cùng nhau của từng cặp từ (xem ma trận tần suất P đã trình bày ở phần giải pháp đề xuất).

Ví dụ : Có 3 câu sau.

- Cầu thủ sút bóng vào khung thành.
- Cầu thủ bị phạt thẻ đỏ sẽ bị đuổi ra khỏi sân.
- Cầu thủ sút bóng ra ngoài.

Như vậy ta sẽ có tần số xuất hiện chung của các cặp từ trong một câu như sau:

$$P_{\text{cầu thủ, bóng}} = 2$$

$$P_{\text{cầu thủ, thẻ đỏ}} = 1$$

Tính toán ma trận P tần số xuất hiện chung trên một trang Wikipedia của các cặp từ qua hai bước sau:

Bước 1. Tính toán ma trận xuất hiện của các từ trên một trang Wikipedia.

Bước 2. Tính ma trận P .

Thông qua việc tính toán ma trận xuất hiện của các từ, ta cũng có ma trận tần số xuất hiện của từng từ, từ đó giúp cho quá trình loại bỏ các từ trong từ điển.

4.2.4. Thuật toán xây dựng đồ thị Dendrogram

Để xây dựng đồ thị Dendrogram dựa vào ma trận P đã tính toán, ta có thể thực hiện như sau:

Bước 1 : Khởi tạo ma trận w thể hiện xác suất xuất hiện cùng nhau của các cặp từ thứ i và j trên cùng một trang Wikipedia.

$$w[i, j] = \frac{P_{ij}}{P_{ii} + P_{jj} - P_{ij}}$$

Bước 2 : Xây dựng đồ thị dendrogram bằng cách lặp đi lặp lại việc dưới đây đến khi tất cả các từ đã được đánh dấu:

+ Tìm phần tử lớn nhất trong w thể hiện tần số xuất hiện cao nhất của cặp từ x và y .

+ Cập nhật lại ma trận w với mọi i

$$w[x, i] = \min(w[x, i], w[y, i])$$

$$w[i, y] = \min(w[i, x], w[i, y])$$

Với tần số xuất hiện chung P_{ij} là tổng số trang Wikipedia xuất hiện cả hai từ thứ i và j trong từ điển. Ta có, $P_{ii} + P_{jj} + P_{ij}$ là tổng số trang có ít nhất một trong hai từ thứ i và j . Suy ra $w[i, j]$ là xác suất xuất hiện cùng nhau trong tập chứa tất cả các trang có ít nhất một trong hai từ thứ i và j .

4.2.5. Triển khai phân cụm

a. Xử lý Wikipedia

Vì file dữ liệu lưu trữ Wikipedia quá lớn, nên để đọc được tập tin này tôi sử dụng lớp *XmlTextReader* để đọc. Ưu điểm của *XmlTextReader* là lớp này đọc trực tiếp từ stream từng nút một, vết bộ nhớ nhỏ nên thích hợp với việc đọc file Wikipedia 91.8 GBytes.

Vấn đề được đặt ra ở đây là không thể biết trước cũng như ước chừng được độ dài của mỗi phần nội dung của mỗi phiên bản và các phiên bản lại lưu trữ không theo thứ tự thời gian cho nên việc lưu trữ trong chương trình là hoàn toàn không thể. Do đó, nội dung của mỗi <page> được lưu trữ ở file tạm trước khi được lưu trữ vào file rút gọn.

Sau khi rút gọn, file rút gọn chứa 1.184.476 trang Wikipedia tiếng Việt. File này tiếp tục được rút gọn bằng:

- Chuyển tất cả các kí tự thành kí tự thường.
- Xóa tất cả dòng trống.
- Xóa bỏ dãy các kí tự nằm liên tiếp nhau.

Kết quả cuối cùng là kích thước file rút gọn còn 3.2 GBytes.

b. Từ điển

Từ điển sau khi được lấy về thì chỉ lấy phần từ, không lấy phần nghĩa và các nội dung khác, đồng thời các từ giống nhau cũng được loại bỏ. Sau đó tất cả các từ trong từ điển đều được chuyển thành các kí tự thường. Để thuận tiện cho việc tìm kiếm xử lý, từ điển được sắp xếp theo thứ tự từ điển. Rồi sau đó mới tiến hành loại bỏ các từ theo thuật toán trình bày ở trên

c. Tính toán ma trận tần số xuất hiện chung

Do dữ liệu lớn nên việc tính ma trận xuất hiện các từ trên một trang Wikipedia rất tốn thời gian. Do đó, 1.184.476 trang Wikipedia được chia nhỏ thành 12 file với mỗi file chứa 100.000 trang Wikipedia. Chương trình tính toán ma trận này được viết đa luồng nên sẽ tăng thời gian tính toán.

Chương trình tính toán ma trận xuất hiện cũng có chức năng resume để tránh trường hợp mất điện hay sự cố đột ngột.

d. Tổ chức dữ liệu trong chương trình

Dữ liệu trong chương trình bao gồm:

- File dữ liệu Wikipedia : *.xml - 91.8 Gbytes
- File từ điển : *.index
- File kết quả phân cụm từ : *.csv và *.index
- File dữ liệu văn bản : *.doc

e. Thư viện hỗ trợ

- Accord.NET

Accord.NET là một nền tảng cho khoa học máy tính trong .NET. Nền tảng này được xây dựng dựa trên nền tảng AForge.NET phổ biến, tập trung vào việc cung cấp các phương pháp thống kê, học máy, nhận dạng và xử lý âm thanh. Nền tảng cung cấp một số lượng lớn các phân bố xác suất, bộ kiểm tra giả thuyết, hàm hạt nhân và hỗ trợ hầu hết các kỹ thuật đo lường.

Accord.NET còn cung cấp một danh sách đầy đủ các ví dụ và các ứng dụng mẫu.

Phiên bản: Accord.NET framework 2.8.1

- xd2txlib.dll

Thư viện này giúp trích xuất văn bản từ một loạt các tài liệu dưới dạng PDF, WORD, EXCEL thành dạng text.

Phiên bản: xdoc2txt 2.01

f. Công cụ sử dụng

- Microsoft Visual Studio 2010
- Dot net Bar for Windows Forms
- Dev – C++ 4.9.9.2
- Windows 7

4.2.6. Thử nghiệm

Để thử nghiệm giải pháp đề xuất ở trên, luận án đã triển khai thực hiện như sau.

4.2.6.1. Cấu trúc hệ thống

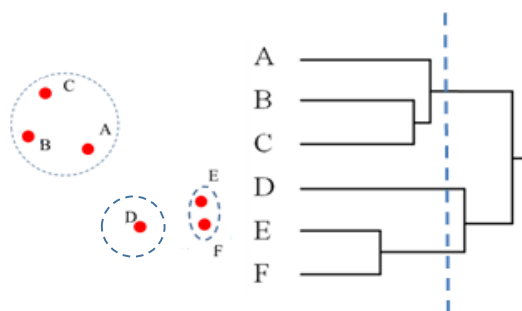
Hệ thống bao gồm các chương trình sau:

- Chương trình tiền xử lý dữ liệu Wikipedia trước khi tiến hành tính toán.
- Chương trình xây dựng ma trận P thể hiện tần số xuất hiện chung của các cặp từ trên cùng một trang Wikipedia.
- Chương trình xây dựng đồ thị Dendrogram từ ma trận P tần số xuất hiện chung.
- Chương trình chính được xây dựng để thực hiện các chức năng sau: phân cụm các từ (hiển thị kết quả qua xây dựng đồ thị Dendrogram và tiến hành phân cụm các từ), xây dựng mô hình phân loại và tiến hành phân loại văn bản tiếng Việt.

4.2.6.2. Các chức năng

a. Chức năng phân cụm: Thông qua việc biểu diễn sự tương quan giữa các từ bằng đồ thị Dendrogram, người dùng sẽ dễ dàng điều chỉnh việc phân cụm theo tỷ lệ mong muốn. Người dùng điều chỉnh việc phân cụm bằng cách cắt tại điểm của đồ thị Dendrogram theo chiều cao của đồ thị.

Ví dụ ở hình 4.5 thể hiện vết cắt trên đồ thị Dendrogram đã chia các điểm thành 3 nhóm phân biệt: “A,B,C”, “D” và “E,F”. Tương tự, chúng ta chỉ cần di chuyển vị trí cắt sẽ nhận được các kết quả phân cụm khác nhau.



Hình 4.5 Ví dụ cho việc cắt đồ thị Dendrogram, kết quả nhận được 3 cụm

b. Chức năng xây dựng mô hình phân loại: Người dùng có thể chọn mô hình phân loại trước khi phân loại văn bản trong 2 mô hình chính: SVM và SVM đề xuất. Với SVM đề xuất là mô hình sử dụng SVM có sử dụng kết quả phân cụm từ tiếng Việt.

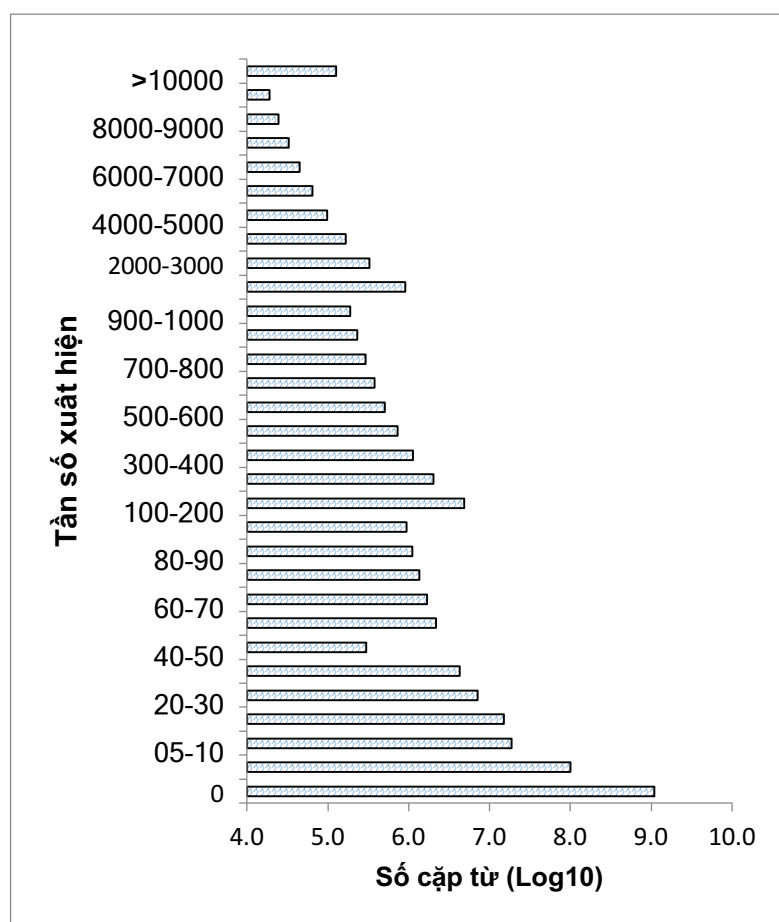
Kết quả phân cụm từ tiếng Việt được áp dụng vào việc rút gọn véc tơ thuộc tính. Như vậy mô hình SVM đề xuất ở đây chính là sử dụng SVM với véc tơ thuộc tính

đã được rút gọn số chiều, tỷ lệ rút gọn tương ứng với tỷ lệ nhận được khi cắt tại điểm tương ứng trên đồ thị Dendrogram.

c. Chức năng phân loại: Chức năng này giúp người dùng phân loại tự động văn bản tiếng Việt từ mô hình phân loại đã được xây dựng từ trước.

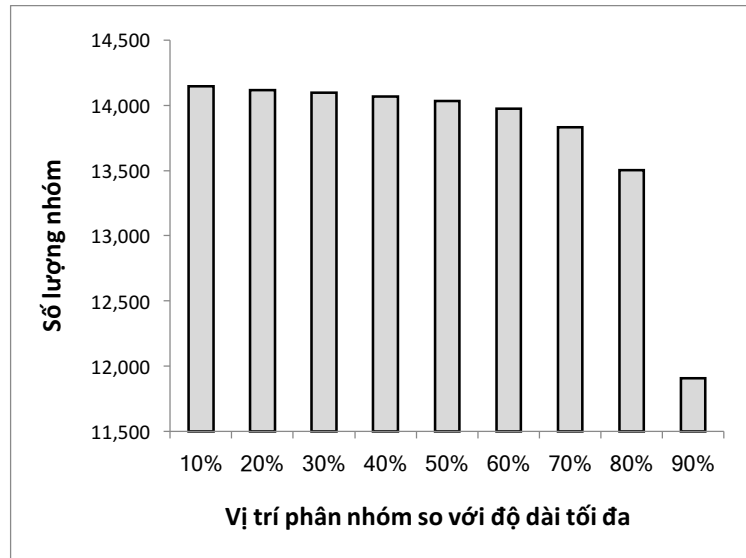
4.2.6.3. Kết quả thực nghiệm

Tiến hành phân cụm với bộ từ điển cho được các kết quả sau:



Hình 4.6 Số lượng cặp từ theo tần số xuất hiện chung

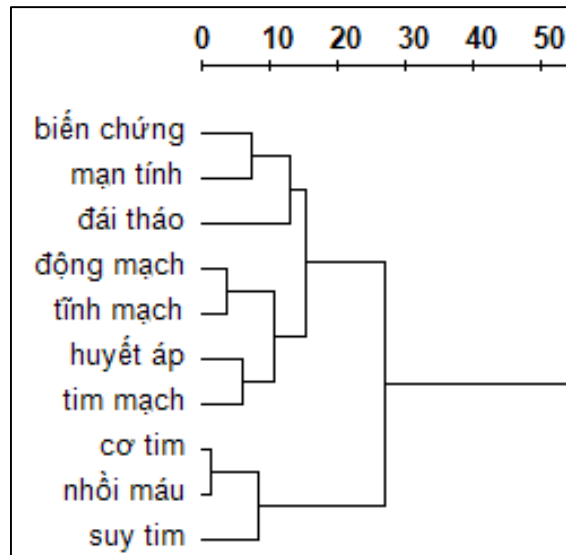
Hình trên biểu diễn số lượng cặp từ theo tần số xuất hiện chung. Dễ dàng thấy rằng số cặp từ không xuất hiện chung trên một trang bất kỳ có số lượng lớn nhất (1.1×10^9 cặp từ). Số lượng cặp từ tỉ lệ nghịch với tần số xuất hiện chung.



Hình 4.7 Số lượng nhóm phụ thuộc phân cụm trên đồ thị Dendrogram

Hình trên biểu diễn kết quả của việc số lượng nhóm phụ thuộc vào vị trí phân cụm trên phương pháp phân tích nhóm dựa trên đồ thị Dendrogram.

Tại vị trí cắt là 20% so với độ dài tối đa, nghiên cứu đã tìm được các nhóm từ có liên quan hoặc gần nghĩa thể hiện như sau:

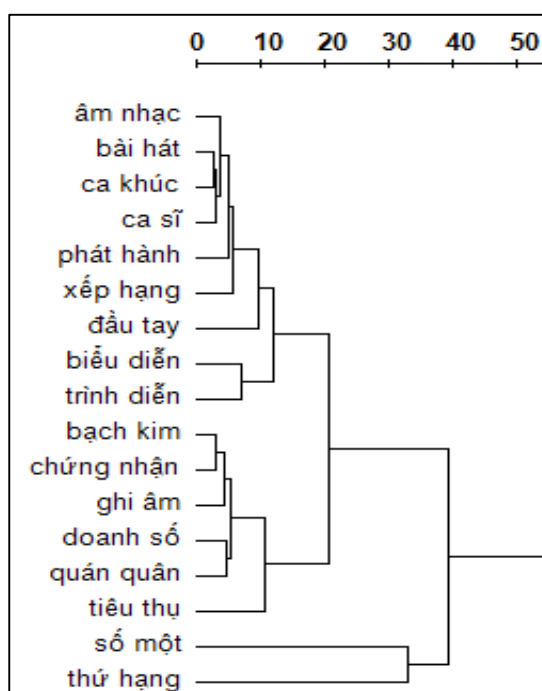


Hình 4.8 Kết quả phân cụm với Dendrogram

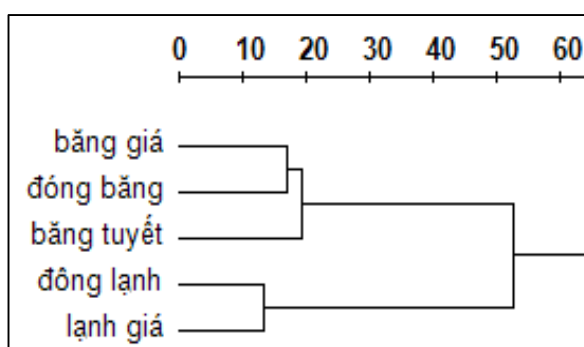
Chiều dài của trục hoành thể hiện mối quan hệ W_{ij} của cụm từ i và j . W_{ij} càng lớn thì khoảng cách các cụm từ càng nhỏ, hình trên biểu diễn khoảng cách các cụm

từ theo cách tính: $d_{ij} = 1 / W_{ij}$

Theo hình trên ta có khoảng cách của 2 từ “nhồi máu” và “cơ tim” rất thấp, có thể thấy được 2 từ này thường xuyên đi chung với nhau theo cụm từ “nhồi máu cơ tim”. Từ “suy tim” có quan hệ gần với “nhồi máu | cơ tim” và nhóm từ còn lại có quan hệ xa hơn so với “nhồi máu | cơ tim | suy tim”. Tuy nhiên, các từ này được gom đúng thành một nhóm chứng tỏ phương pháp đề xuất đã phân cụm thành công các cụm từ có liên quan chặt chẽ với nhau.



Hình 4.9 Một ví dụ khác thể hiện những từ liên quan đến âm nhạc

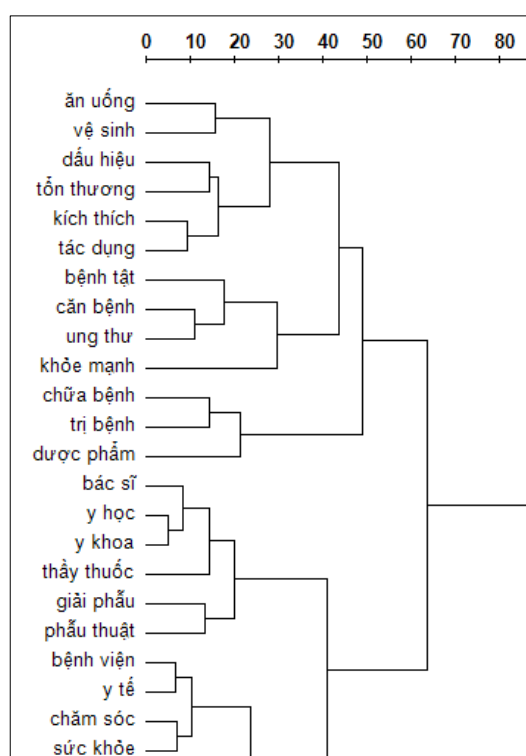


Hình 4.10 Một ví dụ đồ thị Dendrogram cho các từ

Hình trên là một số ví dụ đồ thị Dendrogram cho các từ “băng giá”, “đóng băng”,

“băng tuyết”, “đông lạnh”, “lạnh giá” đó là kết quả phân cụm đúng sử dụng phương pháp đề xuất. Ta dễ dàng nhận thấy được rằng các nhóm từ được phân cụm thành các chủ đề [31].

Trong kết quả thực nghiệm, luận án đã tiến hành chọn ngẫu nhiên 1000 nhóm từ và tiến hành đếm thủ công số lượng nhóm đồng nghĩa đúng. Kết quả thu được là có 56% nhóm bao gồm hai từ đồng nghĩa. Ngoài ra còn phát hiện một số cụm từ bao gồm cả danh từ, động từ và tính từ cho một chủ đề.



Hình 4.11 Ví dụ đồ thị Dendrogram cho các từ thuộc chủ đề y học

Tuy nhiên, vẫn còn có một số từ không mang cùng một ý nghĩa nhưng có chung một nhóm từ như, “sự tích” và “tích tụ” (do cùng là chuỗi con của “sự tích tụ”). Những từ này thông thường cùng là chuỗi con của một từ có nghĩa, dẫn tới việc hay xuất hiện cùng nhau nên kết quả phân cụm chưa được chính xác. Ngoài ra, trong tiếng Việt còn có rất nhiều từ, cụm từ không có trong từ điển mà tác giả đã sử dụng như “cà chớn”, “cà cháo”. Hơn nữa, nghiên cứu này chỉ giới hạn trên các trang Wikipedia nên chưa thể phát hiện hết tất cả các từ, cụm từ liên quan với nhau trong

tiếng Việt.

Không có ngưỡng chính xác mà chỉ bằng thực nghiệm mới tìm ra ngưỡng tốt nhất, thực ra trong thực nghiệm chúng ta chưa xác định ngưỡng nào là tốt nhất mà phải tùy thuộc vào dữ liệu.

4.3. Áp dụng véc tơ rút gọn vào phân loại văn bản

4.3.1. Dữ liệu đầu vào

Gồm tập dữ liệu đã phân loại và kho dữ liệu được xây dựng ở chương trước (trong nội dung xây dựng dữ liệu để thử nghiệm mô hình đường trắc địa).

4.3.2. Kết quả thực nghiệm

a. Mô hình huấn luyện

Việc phân cụm tiếng Việt dẫn đến việc giảm số chiều không gian véc tơ thuộc tính của văn bản, từ đó kéo theo giảm dung lượng lưu trữ không gian véc tơ mẫu. Tôi đã tiến hành huấn luyện mô hình phân loại dựa trên 5 tập mẫu đã gán nhãn. Với mỗi chủ đề 1000 văn bản.

Lần thứ 1: 15 mẫu cho mỗi loại nhãn.

Lần thứ 2: 20 mẫu cho mỗi loại nhãn.

Lần thứ 3: 40 mẫu cho mỗi loại nhãn.

Lần thứ 4: 80 mẫu cho mỗi loại nhãn.

Lần thứ 5: 120 mẫu cho mỗi loại nhãn.

Bảng 4.1 Dữ liệu huấn luyện, kiểm thử

| STT | Loại tài liệu | Huấn luyện | | | | | Kiểm thử |
|-----|---------------|------------|-------|-------|-------|-------|----------|
| | | Lần 1 | Lần 2 | Lần 3 | Lần 4 | Lần 5 | |
| 1 | Bóng đá | 15 | 20 | 40 | 80 | 120 | 400 |
| 2 | Giáo dục | 15 | 20 | 40 | 80 | 120 | 400 |
| 3 | Pháp luật | 15 | 20 | 40 | 80 | 120 | 400 |
| 4 | Quốc tế | 15 | 20 | 40 | 80 | 120 | 400 |
| 5 | Xã hội | 15 | 20 | 40 | 80 | 120 | 400 |

Tạo véc tơ thuộc tính dựa trên véc tơ đã rút gọn số chiều. Tiến hành phân loại văn bản tiếng Việt trên máy véc tơ hỗ trợ (SVM).

Đối với các phương pháp giảm số chiều véc tơ như PCA (phân tích thành phần chính), MDA (Phân tích đa biệt thức), LDA (phân tích biệt thức tuyến tính) là một trong những giải pháp thường hay được sử dụng trong nhiều bài toán khác nhau. Tuy nhiên để thực hiện tính toán trong PCA hay LDA việc tính véc tơ riêng là mấu chốt của thuật toán. Khi số chiều là n thì ma trận trong quá trình tính toán là $n \times n$ phần tử và độ phức tạp là $O(N^3)$. Trên thực tế mỗi văn bản trong nghiên cứu này sử dụng véc tơ 14015 từ để biểu diễn (số từ này lấy từ quá trình rút gọn từ thuật toán xử lý từ điển), nên độ phức tạp của thuật toán khoảng chừng là $O(n)$. Vì vậy một máy tính bình thường hay kể cả siêu máy tính cùng tính toán rất khó khăn. Từ đó trong nghiên cứu này đề xuất phương pháp tính toán như đã trình bày hợp lý hơn.

Trong nghiên cứu này, đã nghiên cứu tiến hành xây dựng bộ dữ liệu kiểm thử bao gồm k dữ liệu huấn luyện và 1 dữ liệu kiểm thử bằng cách chọn ngẫu nhiên từ kho dữ liệu đã xây dựng. Chúng tôi thay đổi k lần lượt như mô hình huấn luyện trên.

- + *Lần thứ 1*: 15 mẫu cho mỗi loại nhãn.
- + *Lần thứ 2*: 20 mẫu cho mỗi loại nhãn.
- + *Lần thứ 3*: 40 mẫu cho mỗi loại nhãn.
- + *Lần thứ 4*: 80 mẫu cho mỗi loại nhãn.
- + *Lần thứ 5*: 120 mẫu cho mỗi loại nhãn.

Ví dụ: Trong thử nghiệm lần 1, nghiên cứu chọn $k = 15$, có nghĩa là ứng với từng thể loại, nghiên cứu chọn ngẫu nhiên 15 văn bản để huấn luyện và dữ liệu kiểm thử là ($1 = 400$) cho từng thể loại. Các bước thực hiện lần 1 như sau:

- + Lựa chọn bộ thực nghiệm huấn luyện và kiểm thử như trên.
- + Xây dựng véc tơ ban đầu cho từng văn bản (số chiều mỗi véc tơ là 14015)
- + Rút gọn véc tơ bằng cách sử dụng kết quả phân cụm Dendrogram với tỷ lệ rút gọn là: 0%; 10%; ... ; 90%, ứng với chiều dài véc tơ là 14015, 12613, ... , 1401.
- + Trong thử nghiệm nghiên cứu đã sử dụng SVM với hàm Gaussian là:

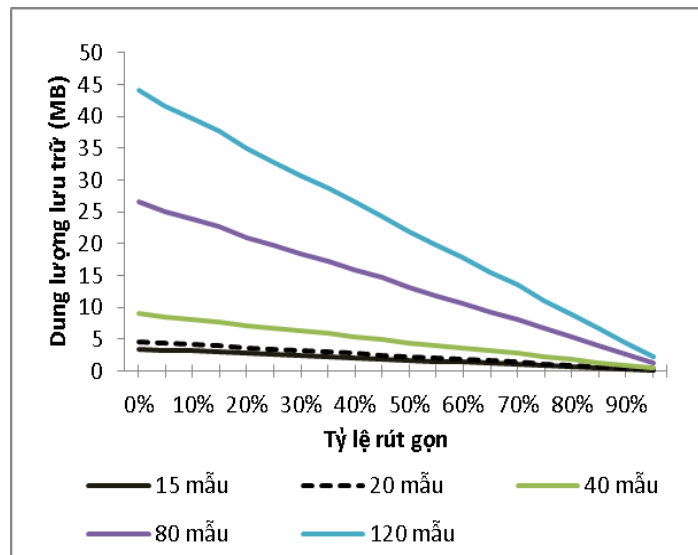
$k(\mathbf{x}_i, \mathbf{x}_j) = e^{(-1/2 ||\mathbf{x}_i - \mathbf{x}_j||^2)}$, với $\mathbf{x}_i, \mathbf{x}_j$ là véc tơ đã rút gọn biểu diễn văn bản i và j .

+ Đánh giá kết quả dựa trên 400 văn bản của từng thể loại chưa được gán nhãn.

Tương tự cho các thử nghiệm lần 2, 3, 4, 5.

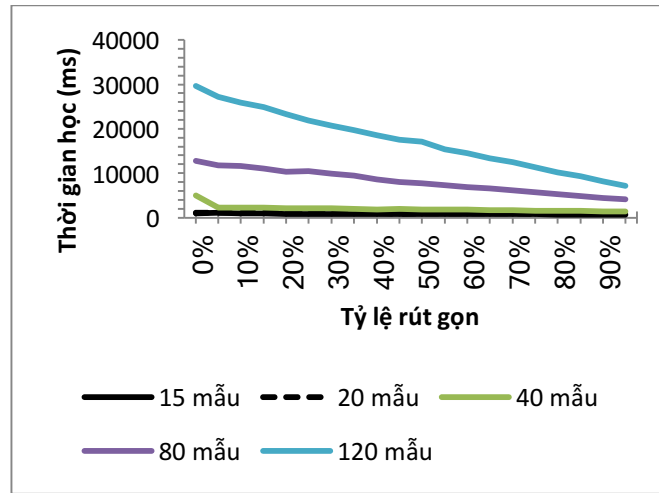
Kết quả phân cụm từ tiếng Việt được áp dụng vào việc rút gọn véc tơ thuộc tính. Như vậy mô hình SVM đề xuất ở đây chính là sử dụng SVM với véc tơ thuộc tính đã được rút gọn số chiều, tỷ lệ rút gọn tương ứng với tỷ lệ nhận được khi cắt đồ thị Dendrogram.

Hình 4.12 cho thấy dung lượng lưu trữ véc tơ phụ thuộc vào số lượng từ được phân cụm, việc lưu các véc tơ hỗ trợ trong SVM dưới dạng file ở từng mô hình phân loại với tỷ lệ khác nhau, qua đó biểu diễn sự phụ thuộc giữa dung lượng lưu trữ không gian véc tơ mẫu và tỷ lệ (%) rút gọn số từ dựa trên phương pháp phân cụm đã đề xuất. Tỷ lệ 0% tương ứng với việc không rút gọn véc tơ thuộc tính (phương pháp trước đó). Ta thấy rằng tỷ lệ rút gọn càng tăng thì dung lượng lưu trữ càng giảm.



Hình 4.12 Dung lượng lưu trữ véc tơ phụ thuộc vào số lượng từ

Tôi lưu các véc tơ hỗ trợ trong SVM dưới dạng file ở từng mô hình phân loại với các tỷ lệ phân cụm từ khác nhau. Hình trên cho thấy khi rút gọn thuộc tính véc tơ đem lại độ phức tạp tính toán giảm

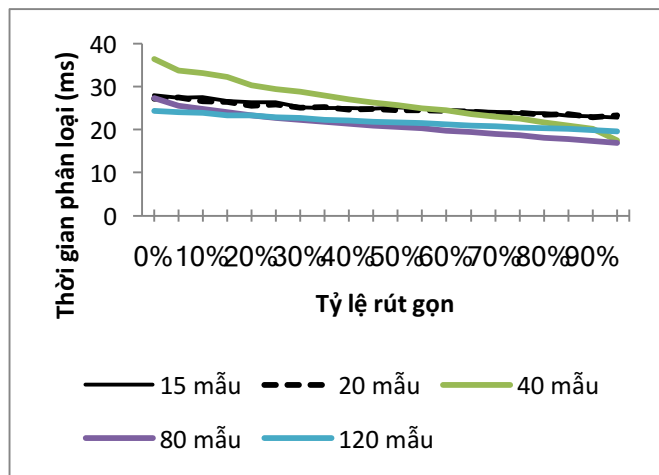


Hình 4.13 Đồ thị thể hiện thời gian gán nhãn của 5 lần huấn luyện

Hình 4.13 đồ thị thể hiện thời gian gán nhãn của năm lần huấn luyện khác nhau. Trong thời gian học phụ thuộc vào số lượng văn bản đã gán nhãn, số lượng văn bản gán nhãn lớn sẽ có thời gian học cao, biểu diễn thời gian huấn luyện tập mẫu phụ thuộc vào tỷ lệ rút gọn. Ta thấy việc rút gọn từ điển cho phép tăng tốc độ huấn luyện trong mô hình SVM.

Thời gian học phụ thuộc vào số lượng văn bản đã gán nhãn. Số lượng gán nhãn lớn sẽ có thời gian học cao, cho thấy khi rút gọn thuộc tính véc tơ đem lại độ phức tạp tính toán giảm

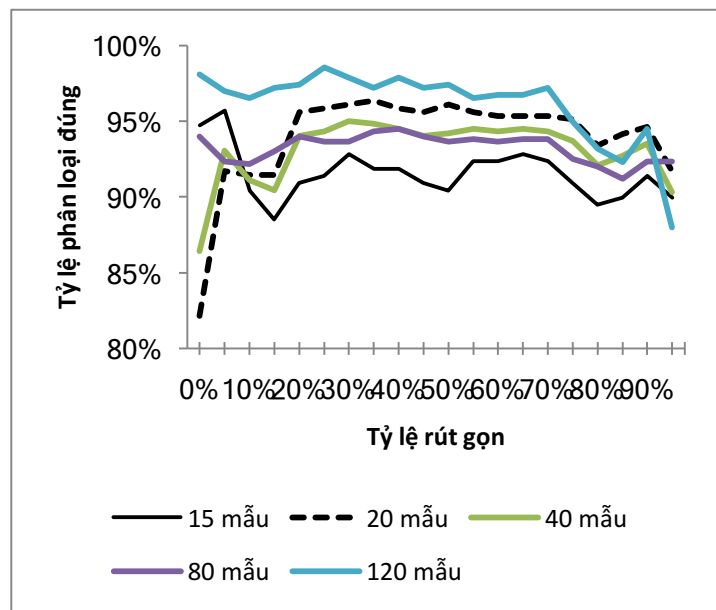
b. Phân loại văn bản



Hình 4.14 Thời gian phân loại văn bản trung bình của 5 lần huấn luyện

Hình trên biểu diễn thời gian phân loại trung bình một văn bản phụ thuộc vào số lượng từ được phân cụm cho thấy tỷ lệ rút gọn. Việc phân cụm cũng góp phần giảm thời gian phân loại văn bản.

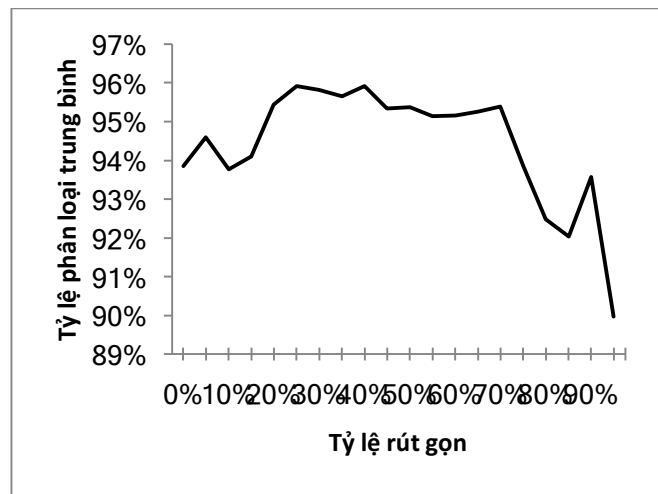
c. Độ chính xác Phân loại văn bản



Hình 4.15 Đồ thị thể hiện độ phân loại của 5 lần HL theo tỷ lệ phân cụm

Từ hình trên ta thấy rằng khi không phân cụm từ, véc tơ thuộc tính có số chiều bằng số lượng từ, kết quả phân loại không ổn định ở 5 lần huấn luyện khác nhau. Dù lần huấn luyện thứ 5 nhiều nhãn hơn lần huấn luyện thứ nhất nhưng lại có kết quả phân loại thấp hơn. Cho thấy sự không ổn định và mang tính ngẫu nhiên cao. Còn khi ta gom cụm từ ở một tỷ lệ nhất định ta thấy được sự ổn định trong phương pháp đề xuất. Và dù ít nhãn nhưng kết quả phân loại vẫn cao.

d. Độ chính xác phân loại văn bản trung bình



Hình 4.16 Đồ thị thể hiện sự thay đổi của kết quả theo tỷ lệ phân loại

Dựa vào hình trên ta thấy việc rút gọn từ điển cho phép cải thiện việc phân loại đúng nếu ta chọn đúng tỷ lệ rút gọn từ điển (từ 30% --> 70%) so với không gian véc tơ ban đầu thì tỷ lệ phân loại đúng vẫn bản cao hơn so với khi chưa phân cụm và rút gọn từ.

4.4. Tiểu kết chương

Kết quả đạt được qua các phương pháp đề xuất nhằm nâng cao chất lượng phân loại văn bản tiếng Việt tự động. Phương pháp thứ nhất sử dụng từ điển bách khoa toàn thư Wikipedia và đồ thị Dendrogram trong việc rút gọn số chiều véc tơ biểu diễn văn bản tiếng Việt. Phương pháp thứ hai là áp dụng véc tơ đã rút gọn để phân loại văn bản. Thực nghiệm cho thấy việc áp dụng không gian véc tơ được rút gọn dựa trên đồ thị Dendrogram và thư viện Wikipedia giúp tiết kiệm dung lượng lưu trữ và thời gian phân loại văn bản tiếng Việt mà vẫn đảm bảo tỷ lệ phân loại đúng, tỷ lệ phân loại văn bản cao hơn so với khi chưa phân cụm.

Hạn chế của phương pháp đề xuất này là chỉ mới thử nghiệm xác suất xuất hiện chung của các cặp từ trong một trang Wikipedia để phân nhóm từ dẫn tới có khả năng sai lệch về mặt ngữ nghĩa, nếu như trang Wikipedia đầy có quá nhiều thông tin. Chẳng hạn như một trang bao gồm nhiều thông tin về Bóng đá, Giáo dục, Pháp luật, ... Trong nghiên cứu tiếp theo sẽ khắc phục những hạn chế nêu trên.

KẾT LUẬN

Kết quả đạt được

Luận án này đã trình bày các kết quả nghiên cứu về phân loại văn bản tiếng Việt kết hợp giữa kỹ thuật học máy bán giám sát và dựa trên máy véc tơ hỗ trợ (SVM).

Kết quả đạt được là:

- Đã xây dựng kho dữ liệu phục vụ cho các thực nghiệm khi phân loại văn bản tiếng Việt.
- Đề xuất và thử nghiệm giải pháp phân loại văn bản dựa trên cự ly trắc địa.
- Đề xuất và thử nghiệm giải pháp rút gọn số chiều véc tơ khi biểu diễn văn bản tiếng Việt để tăng tốc độ xử lý nhưng vẫn đảm bảo độ chính xác khi phân loại văn bản.

Dựa trên kết quả thử nghiệm, luận án đã so sánh phương pháp đề xuất dựa trên mô hình cự ly trắc địa với mô hình SVM thuần túy trên cùng một bộ dữ liệu. Tỷ lệ phân loại trung bình của hai phương pháp không chênh lệch nhiều về kết quả, tuy nhiên căn phương sai của phương pháp đề xuất ($\pm 2\%$) nhỏ hơn nhiều so với SVM ($\pm 4\%$). Điều đó cho thấy phương pháp đề xuất ổn định hơn so với sử dụng SVM thuần túy.

Thực nghiệm cũng đã cho thấy việc áp dụng không gian véc tơ được rút gọn bằng Dendrogram và Wikipedia giúp giảm đáng kể dung lượng lưu trữ và thời gian phân loại văn bản tiếng Việt mà vẫn đảm bảo tỷ lệ phân loại đúng. Ở mức rút gọn 30%-70% so với không gian véc tơ ban đầu, tỷ lệ phân loại đúng văn bản cao hơn so với khi chưa phân cụm.

Giới hạn của luận án

Về cơ bản, chương trình phân loại văn bản đã thực hiện hoàn thành được các chức năng đã đặt ra là giúp người sử dụng xây dựng mô hình phân loại cho các loại văn bản tiếng Việt. Tự động phân loại các văn bản mới dựa trên mô hình đã xây dựng. Tuy nhiên việc thu thập dữ liệu ban đầu chỉ mới ở mức thử nghiệm.

Điểm hạn chế của luận án, đó là chưa sử dụng WORDNET hoặc xây dựng đồ thị

đồng hiện để xem xét mối tương quan ngữ nghĩa giữa các từ trước khi xây dựng véc tơ đặc trưng cho cụm văn bản. Chính điều này có thể làm giảm khả năng tối ưu khi gom cụm thông qua giải thuật gom cụm.

Rút gọn số chiều véc tơ văn bản chỉ mới thử nghiệm xác suất xuất hiện chung của các cặp từ trong một trang Wikipedia để phân nhóm từ dẫn tới có khả năng sai lệch về mặt ngữ nghĩa nếu như trang Wikipedia đây có quá nhiều thông tin. Chẳng hạn như một trang bao gồm cả thông tin về Bóng đá, Giáo dục, Pháp luật, Quốc tế, Xã hội, ...

- Chỉ mới thực nghiệm trên máy véc tơ hỗ trợ (VSM).
- Chưa so sánh các thuật toán Dendrogram khác nhau.

Trong thời gian tới, tôi sẽ bổ sung một số tính năng mới và hoàn thiện chương trình để nâng cao hiệu quả, đồng thời xây dựng kho dữ liệu đủ lớn nhằm mục đích phân loại văn bản một cách chính xác hơn.

Đề xuất hướng nghiên cứu tiếp theo

Tóm tắt văn bản là một hướng nghiên cứu đang được quan tâm của các nhà khoa học hiện nay, đặc biệt trong vấn đề ngôn ngữ tiếng Việt còn nhiều vấn đề cần được quan tâm nghiên cứu. Chính vì thế, hướng nghiên cứu tóm tắt văn bản vẫn đang là một hướng nghiên cứu mở. Trong giới hạn nghiên cứu của luận án, tôi xin đề xuất hướng nghiên cứu trong tương lai của đề tài này là:

- Tiếp tục nghiên cứu WORDNET trợ giúp tra cứu ngữ nghĩa tiếng Anh, từ đó xây dựng WORDNET cho tra cứu tiếng Việt. Hoặc sử dụng đồ thị đồng hiện để tối ưu khả năng tương tác khi tạo véc tơ đặc trưng cho cụm văn bản.
- Để nâng cao tính hiệu quả của mô hình học bán giám sát có kết hợp tóm tắt nội dung văn bản, tôi sẽ tiếp tục nghiên cứu các phương pháp xử lý tách từ tiếng Việt, nhằm tăng độ chính xác của phương pháp trích rút ý chính nội dung văn bản, đồng thời tiến hành thực nghiệm nhiều tỷ lệ nén nội dung khác nhau để tìm ra tỷ lệ nén nội dung có độ chính xác cao hơn, nhằm cải thiện thêm độ chính xác của kết quả phân lớp văn bản dựa vào mô hình đề xuất.

- Thử nghiệm với tần số xuất hiện chung trong một đoạn văn, một câu.
- Thử nghiệm với bộ dữ liệu khác Wikipedia, ví dụ các bài báo trên các trang báo mạng Việt Nam.
- Thử nghiệm với các phương pháp học máy khác và so sánh các thuật toán Dendrogram khác nhau.

CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

1. Vo Duy Thanh, Vo Trung Hung, Pham Minh Tuan, Doan Van Ban, “Text classification based on semi-supervised learning”, Proceeding of the SoCPaR 2013, IEEE catalog number CFP1395H-ART, ISBN 978-1-4799-3400-3/13/\$31.00, pp. 238-242, 2013.
2. Vo Duy Thanh, Vo Trung Hung, Phạm Minh Tuan and Ho Khắc Hưng, “Text Classification Based On Manifold Semi-Supervised Support Vector Machine”, Proceeding of the ISDA 2014, 14th International Conference on Intelligent Systems Design and Applications, Okinawa, Japan 27-29, November 2014, IEEJ catalog, ISSN: 2150-7996, pp. 13-19.
3. Pham Minh Tuan, Nguyen Thi Le Quyen, Vo Duy Thanh, Vo Trung Hung, “Vietnamese Documents Classification Based on Dendrogram and Wikipedia”, Proceedings of Asian Conference on Information Systems 2014, ACIS 2014, December 1-3, 2014, Nha Trang, Viet Nam, © 2014 by ACIS 2014, ISBN: 978-4-88686-089-7, pp. 247-253.
4. Vo Duy Thanh, Vo Trung Hung, Ho Khắc Hưng, Tran Quoc Huy, “Text Classification Based On SVM And Text Summarization”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 4, Issue 02, February-2015, pp. 181-186.
5. Võ Trung Hùng, Nguyễn Thị Ngọc Anh, Hồ Phan Hiếu, Nguyễn Ngọc Huyền Trân, Võ Duy Thanh, “So sánh văn bản dựa trên mô hình véc tơ”, Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng, ISSN: 1859-1531, số 3(112)-2017, quyển 1, Trang: 105-109.

TÀI LIỆU THAM KHẢO

- [1] Asgharbeygi. N and A. Maleki. (2008), “Geodesic K-means Clustering“. Proc. ICPR08: pp. 1-4.
- [2] A. Blum and T. Mitchell. (1998), “Combining labeled and unlabeled data with Co-training”. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98), pp. 92-100.
- [3] A. P. Dempster. Et al. (1997), “Maximum likelihood from incomplete data via the EM algorithm”. Journal of the Royal Statistical Society, Series B, 39(1): pp. 1-38.
- [4] A. Murari, P. Boutot, “Clustering Based on the Geodesic Distance on Gaussian Manifolds for the Automated Classification of Disruptions”, EFDA–JET–PR(12)27.
- [5] Balcan. M. F and Blum. A. (2006), “An augmented pac model for semi-supervised learning.“ In O. Chapelle, B. Schölkopf and A. Zien (Eds.), Semi-supervised learning. MIT Press, pp. 61-89.
- [6] Bengio, Y. et al. (2007). “Greedy layer-wise training of deep networks”. Advances in Neural Information Processing Systems, NIPS 19.
- [7] Belkin, M. et al. (2006). “Manifold regularization: a geometric framework for learning from Labeled and Unlabeled Examples”. Journal of Machine Learning Research, 7, 2399–2434.
- [8] Bennett. K. P. (1998), “Semi-Supervised Support Vector Machines.“ Department of Mathematical Sciences Rensselaer Polytechnic Institute Troy, pp. 368-374.
- [9] Bùi Khánh Linh. et al. (2016),”Phân loại văn bản tiếng Việt dựa trên mô hình chủ đề”, Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX – “Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)”; Cần Thơ, ngày 4-5/8/2016, DOI: 10.15625/vap.2016.00065.
- [10] C. C. Kemp. et al. (2003),“Semi-Supervised learning with trees.“ Advances in Neural Information Processing System 16, NIPS 2003: pp. 257-264.

- [11] Carlson. A. (2010), *Coupled Semi-Supervised Learning*. Machine Learning Department School of Computer Science, Carnegie Mellon University Pittsburgh, PA 15213, CMU-ML-10-104, May.
- [12] Chapelle. et al. (2006), *Semi-Supervised Learning*, The MIT Press Cambridge, Massachusetts Institute of Technology, ISBN 978-0-262-03358-9, London – England.
- [13] C. Liu and P. C. Yuen. (2011). “A boosted co-training algorithm for human action recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 9, pp. 1203–1213.
- [14] Collins. M and Singer. Y. (1999), “Unsupervised models for named entity classification“, *EMNLP/VLC-99*, pp. 100-110.
- [15] Cozman. F. G and Cohen. I. (2002), “Unlabeled data can degrade classification performance of generative classifiers.“, *Int’ l Florida Artificial Intell. Society Conf*, pp. 327-331.
- [16] C. Rosenberg. et al. (2005). “Semisupervised self-training of object detection models,” in *Proceedings of the 7th IEEE Workshop on Applications of Computer Vision (WACV ’05)*, IEEE, January 2005, pp. 29–36.
- [17] Craven. M., et al. (1998). “Learning to extract symbolic knowledge from the World Wide Web”. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pp. 509-516.
- [18] David D. Lewis. Et al. (2004), “RCV1: A New Benchmark Collection for Text Categorization Research”, *Journal of Machine Learning Research* (5), pp 361-397.
- [19] Dennis Ramdass & Shreyes Seshasai (2009), “Document Classification for Newspaper Articles”, 6.863 Final Project Spring 2009, pp. 1-12.
- [20] Diederik. et al. (2014), “Semi-supervised Learning with Deep Generative Models“, *NIPS Neural Information Processing Systems*, Montreal, Canada, 8-11th December.

- [21] Didaci, Luca. et al. (2012). “Analysis of Co-training Algorithm with Very Small Training Sets”. Lecture Notes in Computer Science. Springer Berlin Heidelberg. ISBN: 9783642341656, pp. 719–726.
- [22] Dinh Dien. et al. (2001), “Vietnamese Word Segmentation“, Proceedings of the NLPRS 2001, Tokyo, Japan, 27-30 November, pp. 749-756.
- [23] Đinh Thị Phương Thu. et al. (2005), “Phương án xây dựng tập mẫu cho bài toán phân lớp văn bản tiếng Việt, nguyên lý, giải thuật, thử nghiệm và đánh giá kết quả”. Tạp chí khoa học công nghệ.
- [24] Đỗ Phúc và Trần Thế Lâm. (2004), “Phân loại văn bản tiếng Việt dựa trên tập thô“, Hội thảo Quốc gia về CNTT, Đà Nẵng, pp. 125-131.
- [25] Đỗ Phúc. (2006), “Nghiên cứu ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản tiếng Việt có xem xét ngữ nghĩa”, Tạp chí phát triển KH & CN, tập 9, số 2, pp. 23-32.
- [26] Đỗ Phúc. et al. (2008), “Gom cụm đồ thị và ứng dụng vào việc rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận“, Tạp chí phát triển khoa học công nghệ, Tập 11, số 05, pp. 21-32.
- [27] Fazakis, Nikos. et al. (2015). "Self-Trained LMT for Semi-supervised Learning". Computational Intelligence and Neuroscience. 2016: 1–13. doi: 10.1155/2016/3057481.
- [28] Feil. B and Abonyi. J. (2007), “Geodesic Distance Based Fuzzy Clustering“, Lecture Notes in Computer Science, Soft Computing in Industrial Applications 39, pp. 50-59.
- [29] Giang Nguyễn Linh và Nguyễn Mạnh Hiển. (2006), “Phân loại văn bản tiếng Việt với bộ phân loại véc tơ hỗ trợ SVM“, Tạp chí CNTT&TT.
- [30] Glenn Fung and O. L. Mangasarian. (2001), “Semi-supervised Support Vector Machines for Unlabeled Data Classification“, Optimization Methods and Software, pp. 1-14.(26)
- [31] Goh. A. (2011), *Riemannian manifold clustering and dimensionality reduction for vision-based analysis*. Machine Learning for Vision-Based Motion Analysis: Theory and Techniques, Springer-Verlag: pp. 27-53.

- [32] Hamel. L. (2008), *Knowledge Discovery With Support vector machines*. University of Rhode Island, ISBN 978-0-470-37192-3.
- [33] Houda benbrahim. (2011), “Fuzzy Semi-supervised Support Vector Machines“, Machine Learning and Data Mining in Pattern Recognition, of the series Lecture Notes in Computer Science, 7th International Conference, MLDM, New York, USA, Vol 6871, pp. 127-139.
- [34] Hồ Thị Ngọc. (2012), “Nghiên cứu ứng dụng học bán giám sát”, Luận văn thạc sĩ, Đại học Đà Nẵng.
- [35] Hung Nguyen. et al. (2005), “Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese“, Proceedings of 4th IEEE International Conference on Computer Science - Research, Innovation and Visio.
- [36] J. Kim. et al. (2007), “Soft Geodesic Kernel K-means“, Proc. ICASSP20072, pp. 429-432.
- [37] Jafar Tanha. et al. (2015), “Semi-supervised Self-training for decision tree classifiers“, International Journal of Machine Learning and Cybernetics, pp. 1–16.
- [38] Jason D.M Rennie (2001), *Improving Multi-class Text Classification with Naive Bayes*, Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Master of Science.
- [39] Jason Weston. et al. (2008). “Deep learning via semi-supervised embedding“, Proceeding ICML '08 Proceedings of the 25th international conference on Machine learning, Helsinki, Finland, July 05-09, 2008, ACM New York, NY, USA ©2008, ISBN: 978-1-60558-205-4 doi>10.1145/1390156.1390303, pp. 1168-1175.
- [40] Jin Chen. et al. (2009), “Constructing Overview + Detail Dendrogram – Matrix Views“, IEEE Trans Vis Comput Graph. Nov-Dec, pp. 889-896.
- [41] Joachims, T. (1999), “Transductive inference for text classification using

- support véc to machines“, Proc. 16th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 200–209.
- [42] Joachims, T. (1997), “A probabilistic analysis of the Rocchio algorithm with TF-IDF for text categorization”, ICML 97 Proceedings of the Fourteenth International Conference on Machine Learning, pp 143-151.
 - [43] Joachims, T. (1998), “Text Categorization with Support Vector Machines: Learning with Many Relevant Features“, In European Conference on Machine Learning (ECML), pp. 137-142.
 - [44] Kristin. et al. (1998), “Semi-supervised Support Vector Machines“, Advances in neural information processing systems, pp. 368-374.
 - [45] Lang, K. (1995). “Newsweeder: Learning to filter netnews”. In Machine Learning: Proceeding of th Twelfth International Conference (ICML-95), pp. 331-339.
 - [46] Le, Hong Phuong. et al, (2008). “A Hybrid Approach to Word Segmentation of Vietnamese Texts”. 2nd International Conference on Language and Automata Theory and Applications - LATA 2008, Mar, Tarragona, Spain. Springer Berlin / Heidelberg, 5196, pp. 240-249.
 - [47] Lewis, D. D., & Gale, W. A. (1994). “A sequential algorithm for training text classifiers”. In SIGIR '94: Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3-12.
 - [48] Lewis, D. D, & Knowles, K. A. (1997). “Threading electronic mail: A preliminary study”. Information Processing and Management, 33 (2), pp. 209-217.
 - [49] Li Cunhe and Wu Chenggang. (2010), “A new semi-supervised support vector machine learning algorithm based on active learning“, Future Computer and Communication (ICFCC), 2nd International Conference on Vol: 3, pp. 638-641.
 - [50] McCallum. A and Nigam. K. (1998), “A comparison of event models for naïve

- bayes text classification“, AAAI-98 Workshop on “Learning for Text Categorization”, Press, pp 335-343.
- [51] Min Song. et al. (2011), “Combining active learning and semi-supervised learning techniques to extract protein interaction sentences“, BMC Bioinformatics, December, pp. 1471-1480.
- [52] Mitchells. T. (2006), *The discipline of machine learning*, Technical Report CMU-ML- 06-108, Carnegie Mellon University, pp. 1-7.
- [53] M.-L. Zhang and Z.-H. Zhou. (2011), “CoTrade: confident co-training with data editing,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 41, no. 6, pp. 1612–1626.
- [54] M. Iggane. (2012). “Self-training using a k-nearest neighbor as a base classifier reinforced by support vector machines”. International Journal of Computer Applications, vol. 56, no. 6, pp. 43–46.
- [55] Mohamed Farouk Abdel Hady. et al. (2010), “Semi-supervised learning for tree-structured ensembles of RBF networks with Co-training“, Neural Networks, The 18th International Conference on Artificial Neural Networks, ICANN, Vol 23, Issue 4, May, pp. 497–509.
- [56] Nikos, Fazakis. et al. (2016), ”Self-trained LMT for semisupervised learning”, Journal Computational Intelligence and Neuroscience Volume 2016, January 2016 Article No. 10, Hindawi Publishing Corp. New York, NY, United States doi>10.1155/2016/3057481. pp. 1-13.
- [57] Neil D. Lawrence and Michael I. Jordan. (2004), “Semi-supervised Learning via Gaussian Processes“, Neural Information Processing Systems 17, pp. 753-760.
- [58] Nguyen, Cam Tu. et al, (2006). “Vietnamese word segmentation with CRFs and SVMs: An investigation”. In 20th Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 215-222.
- [59] Nguyễn Ngọc Bình. (2004), “Dùng lý thuyết tập thô và các kỹ thuật khác để phân loại, phân cụm văn bản tiếng Việt“, Kỷ yếu hội thảo ICT.rda’04. Hà nội.

- [60] Nigam. K. (2001), *Using unlabeled data to improve text classification*. Technical Report CMU-CS-01-126. Carnegie Mellon University. Doctoral Dissertation.
- [61] Nigam. K. et al. (2000), *Text classification from labeled and unlabeled documents using EM*. Machine Learning, pp. 103–134.
- [62] Pham. M. T and K. Tachibana. (2013), “An Algorithm for Fuzzy Clustering Based on Conformal Geometric Algebra“, Knowledge and Systems Engineering Advances in Intelligent Systems and Computing 245, pp. 83-94.
- [63] Pazzani, M. J. et al. (1996). “Syskill & Webert: Identifying interesting Web sites”. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp. 54-59.
- [64] Ratnaparkhi A. (1997), “A Simple Introduction to Maximum Entropy Model For Natural Language Processing”. In Technical Report 97-08 Institute for Reseach In Cognitive Science University of Pensylvania, pp. 1-11.
- [65] Renaud Blanch. et al. (2015). “Dendrogramix: a Hybrid Tree-Matrix Visualization Technique to Support Interactive Exploration of Dendrograms”, To appear in IEEE Transactions on Visualization and Computer Graphics (Proceedings of PacificVis 2015) pp. 31-38.
- [66] R. Souvenir and R. Pless. (2005), “Manifold clustering“, IEEE International Conference on Computer Vision I: pp. 648–653.
- [67] Sahami, M. et al. (1998). “A Baysian approach to _ltering junk e-mail”. In AAAI-98 Workshop on Learning for Text Categorization. Tech. rep. WS-98-05, AAAI Press. <http://robotics.stanford.edu/users/sahami/papers.html>.
- [68] S. Poria. et al. (2012), “Fuzzy clustering for semi-supervised learning - Case study: Construction of an emotion lexicon“, Proceedings of MICA, pp. 73-86.
- [69] Seege. M. (2001), *Learning with labeled and unlabeled data*. Technical Report. University of Edinburgh.
- [70] Shavlik, J., & Eliassi-Rad, T. (1998). “Intelligent agents for web-based tasks: An advice-taking approach”. In AAAI-98 Workshop on Learning for Text

- Categorization. Tech. rep. WS-98-05, AAAI Press.
<http://www.cs.wisc.edu/~shavlik/mlrg/publications.html>.
- [71] Shifei Ding. et al. (2015), “An overview on semi-supervised support vector machine“, in *Neural Computing and Applications*, pp. 1-10.
 - [72] Sidorov Grigori and Velasquez Francisco. et al. (2009), “Syntactic n-Grams as Machine Learning Features for Natural Language Processing“, *Expert Systems with Applications* 41 (3), pp. 853–860.
 - [73] Sidorov Grigori. et al. (2012), “Syntactic Dependency-based n-grams as Classification Features“, *LNAI 7630*, pp. 1–11.
 - [74] Stamatis Karlos. Et al. (2016). “A Semisupervised Cascade Classification Algorithm”, *Applied Computational Intelligence and Soft Computing*, Volume 2016, Article ID 5919717, 14 pages, <http://dx.doi.org/10.1155/2016/5919717>.
 - [75] S. Sun and F. Jin. (2011). “Robust co-training”. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 7, pp. 1113–1126.
 - [76] Steven J. Benson and Jone J. Moré, (2001). “A Limited Memory Variable Metric Method In Subspace for Bound-constrained Optimization Problem”. In *Preprint ANL/MCS, P909-0901*.
 - [77] Susana Eyheramendy, et al. (2003), “On the Naive Bayes Model for Text Classification”. In *Proceedings of the ninth international workshop on Artificial Intelligence & Statistics*, eds, C.M. Bishop and B.J. Frey.
 - [78] Thắng Huỳnh Quyết và Đinh Thị Thu Phương. (2005), “Tiếp cận phương pháp học không giám sát trong học có giám sát với bài toán phân lớp văn bản tiếng Việt và đề xuất cải tiến công thức tính độ liên quan giữa hai văn bản trong mô hình véc tơ“, *Kỷ yếu Hội thảo ICT.rda’04*, pp. 251-261.
 - [79] Tongguang Ni. et al. (2015), “Locality Preserving Semi-Supervised Support Vector Machine“, *Journal of information Science and Engineering* 31, pp. 2009-2024.
 - [80] Trần Cao Đệ và Phạm Nguyên Khang. (2012), “Phân loại với máy học vector

- hỗ trợ và cây quyết định“, Tạp chí khoa học Trường Đại học Cần Thơ, 21a, pp. 52-63.
- [81] Trần Mai Vũ. et al. (2008), “Độ tương đồng ngữ nghĩa giữa hai câu và áp dụng vào bài toán sử dụng tóm tắt đa văn bản để đánh giá chất lượng phân cụm dữ liệu trên máy tìm kiếm VNSSEN“, Hội thảo CN Thông tin Truyền thông lần thứ nhất (ICTFIT08) ĐHKHTN, ĐHQG TP HCM, pp. 94-102.
- [82] Trần Ngọc Phúc. et al. (2013), ”Phân loại nội dung tài liệu web tiếng Việt, Tạp chí Khoa học và Công nghệ 51 (6), pp. 669-680.
- [83] Triguero Isaac. et al. (2013), *Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study*. Knowledge and Information Systems 42 (2): pp. 245–284.
- [84] Van Nguyen. et al. (2014), “Kernel-based semi-supervised learning for novelty detection“, International Joint Conference on Neural Networks (IJCNN), Conference Location, July. pp: 4129 - 4136.
- [85] Vipin Kumar. et al. (2010), “Optimizing F-Measure with Support Vector Machines“, Proceedings of the 16 International, Florida, Artificial Intelligence Research Society Conference, pp. 356-360.
- [86] Võ Trung Hùng. (2017), *Một số phương pháp và mô hình áp dụng trong xử lý ngôn ngữ tự nhiên*. ISBN 987-604-80-2014-7. NXB Thông tin và Truyền thông.
- [87] Vu Cong Duy Hoang. et al. (2007), “A Comparative Study on Vietnamese Text Classification Methods“, Research, Innovation and Vision for the Future, IEEE International Conference on, pp. 267-273.
- [88] Xiaojin Zhu. (2008), *Semi-Supervised Learning Literature Survey*. Computer Sciences TR 1530, University of Wisconsin, Last modified on July.
- [89] Y. Wang and S. Chen. (2013), “Safety-aware semi-supervised classification“, IEEE Transaction on Neural Network and Learning System, Vol. 24, pp. 1763-1772.
- [90] Yu, H. et al. (2003). “Text classification from positive and unlabeled

- documents”. In O. Frieder, J. et al. (Eds.), CIKM 2003: Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management. pp. 232-239.
- [91] Yitan Li. Et al. (2015). “Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective”. Proceedings of the Twenty-Fourth international joint conference on Artificial Intelligence (IJCAI 2015). pp 3650-3656.
- [92] Yun Jin. et al. (2011), “A Semi-Supervised Learning Algorithm Based on Modified Self-training SVM“, in Journal of Computers 6, pp.1438-1443.
- [93] Yves Grandvalet and Yoshua Bengio. (2005), “Semi-supervised Learning by Entropy Minimization“, Advances in neural information processing systems 17, pp. 1-8.
- [94] Z. H. Zhou. et al. (2007), “Semi-supervised learning with very few labeled training examples“, in Proceedings of the 22nd Conference on Artificial Intelligence and the 19th Innovative Applications of Artificial Intelligence Conference (AAAI '07), pp. 675-680.
- [95] Zhu. et al. (2009), *introduction to semi-supervised learning*. Morgan & Claypool. ISBN 9781598295481.
- [96] Zhou, D., Huang, J., & Scholkopf, B. (2005). “Learning from labeled and unlabeled data on a directed graph”. ICML05, 22nd International Conference on Machine Learning. Bonn, Germany.
- [97] Zhou, Z.-H., & Li, M. (2005). “Semi-supervised regression with co-training”. International Joint Conference on Artificial Intelligence (IJCAI).
- [98] Zhu, X. (2005). “Semi-supervised learning with graphs”. Doctoral dissertation, Carnegie Mellon University (mã số CMU-LTI-05-192).
- [99] Piyush Rai. (2011). *Semi-supervised learning*, CS5350/6350: Machine Learning, November 8- 2011.