

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

LÊNG HOÀNG LÂM

**PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT VÀ
ỨNG DỤNG VÀO CÁC CƠ QUAN NHÀ NƯỚC TỈNH BẮC KẠN**

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 0101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS. ĐOÀN VĂN BAN

Thái Nguyên - 2017
LỜI CAM ĐOAN

Tôi xin cam đoan đây là sản phẩm nghiên cứu, tìm hiểu của cá nhân tôi. Các số liệu, kết quả trình bày trong luận văn là trung thực. Những nội dung trình bày trong luận văn hoặc là của bản thân, hoặc là được tổng hợp từ những nguồn tài liệu có nguồn gốc rõ ràng và được trích dẫn hợp pháp, đầy đủ.

Tôi xin hoàn toàn chịu trách nhiệm cho lời cam đoan của mình.

Thái Nguyên, tháng 4 năm 2017
HỌC VIÊN

Lèng Hoàng Lâm

LỜI CẢM ƠN

Trân trọng cảm ơn các thầy giáo, cô giáo trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên; các giảng viên đến từ Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Trường Đại học Quốc gia Hà Nội... đã tạo điều kiện tốt nhất cho học viên trong quá trình học tập và làm luận văn. Đặc biệt, xin được bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới thầy giáo, PGS.TS. Đoàn Văn Ban, người đã định hướng và luôn tận tình chỉ bảo, hướng dẫn em trong việc nghiên cứu, thực hiện luận văn này.

Trong suốt quá trình học tập và thực hiện đề tài, học viên luôn nhận được sự ủng hộ, động viên của gia đình, đồng nghiệp, đặc biệt là sự quan tâm tạo điều kiện của Ban lãnh đạo Trung tâm Công nghệ thông tin và Truyền thông tỉnh Bắc Kạn - nơi học viên đang công tác. Xin trân trọng cảm ơn!

Thái Nguyên, tháng 4 năm 2017

HỌC VIÊN

Lèng Hoàng Lâm

MỤC LỤC

LỜI CAM ĐOAN	i
LỜI CẢM ƠN	ii
MỤC LỤC.....	iii
DANH MỤC CÁC TỪ VIẾT TẮT	v
DANH MỤC CÁC HÌNH.....	vi
DANH MỤC CÁC BẢNG.....	vii
MỞ ĐẦU.....	1
CHƯƠNG I. TỔNG QUAN VỀ PHÂN LOẠI VĂN BẢN TIẾNG VIỆT	3
1.1. Khai phá dữ liệu.....	4
1.2. Khai phá dữ liệu văn bản	7
1.3. Phân loại văn bản	11
1.3.1. Giới thiệu bài toán phân loại văn bản.....	11
1.3.2. Quy trình phân loại văn bản.....	12
1.3.3. Phân loại văn bản tiếng Việt.....	13
1.4. Đặc trưng của văn bản tiếng Việt	14
1.4.1. Các đơn vị của tiếng Việt	14
1.4.2. Ngữ pháp của tiếng Việt.....	17
1.4.3. Từ tiếng Việt.....	18
1.4.4. Câu tiếng Việt	20
1.4.5. Các đặc điểm chính tả và văn bản tiếng Việt	23
1.5. Công tác quản lý văn bản tại các cơ quan tỉnh Bắc Kạn	23
1.6. Kết luận chương 1	25
CHƯƠNG II. CÁC KỸ THUẬT TRONG PHÂN LOẠI VĂN BẢN TIẾNG VIỆT	25
2.1. Tách từ trong văn bản	26
2.1.1. Phương pháp khớp tối đa.....	27
2.1.2. Mô hình tách từ bằng WFST và mạng Neural.....	28
2.1.3. Phương pháp học dựa vào sự biến đổi trạng thái	29
2.1.4. Loại bỏ từ dừng.....	31
2.2. Trọng số của từ trong văn bản	31
2.2.1. Phương pháp Boolean.....	32
2.2.2. Phương pháp dựa trên tần số	32

2.3. Các mô hình biểu diễn văn bản.....	33
2.3.1. Mô hình Boolean	33
2.3.2. Mô hình xác suất.....	33
2.3.3. Mô hình không gian vector.....	34
2.4. Độ tương đồng văn bản.....	36
2.5. Thuật toán phân loại văn bản	39
2.5.1. Thuật toán Support Vector Machine (SVM)	39
2.5.2. Thuật toán K-Nearest Neighbor (kNN)	43
2.5.3. Thuật toán Naïve Bayes (NB)	44
2.6. Phân loại văn bản tiếng Việt	47
2.6.1. Trích chọn đặc trưng văn bản	47
2.6.2. Sử dụng thuật toán SVM để phân loại văn bản	50
2.7. Kết luận chương 2	53
CHƯƠNG III. ÁP DỤNG THUẬT TOÁN SUPPORT VECTOR MACHINE PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT.....	54
3.1. Ứng dụng SVM vào bài toán phân loại văn bản hành chính tiếng Việt tại các cơ quan nhà nước tỉnh Bắc Kạn.....	54
3.2. Áp dụng phân loại văn bản	56
3.3. Xây dựng chương trình thử nghiệm ứng dụng phân loại văn bản áp dụng vào máy tìm kiếm văn bản hành chính tiếng Việt	57
3.3.1. Mô tả bài toán	57
3.3.2. Quá trình tiền xử lý văn bản	59
3.3.3. Vector hóa và trích chọn đặc trưng văn bản	60
3.3.4. Đánh giá bộ phân lớp.....	60
3.3.5. Chương trình thực nghiệm.....	62
3.3.6. Kết quả thực nghiệm.....	62
3.4. Kết luận chương 3	63
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	64
TÀI LIỆU THAM KHẢO.....	65

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Giải thích
CSDL	Cơ sở dữ liệu
KDD	Knowledge Discovery from Data
IDF	Inverse Document Frequency
kNN	K-Nearest Neighbor
NB	Naïve Bayers
SVM	Support Vector Machine
S ³ VM	Semi-Supervised Support Vector Machine
TBL	Transformation - based Learning
TF	Term Frequency
WFST	Weighted Finite - State Transducer

DANH MỤC CÁC HÌNH

Hình 1.1. Các bước trong quá trình phát hiện tri thức từ CSDL (KDD)	5
Hình 1.2. Quy trình phân loại văn bản.....	13
Hình 2.1. Biểu diễn văn bản theo mô hình xác suất	34
Hình 2.2. Minh họa hình học thuật toán SVM.....	40
Hình 2.3. Chi tiết giai đoạn huấn luyện	50
Hình 2.4. Mô hình SVM	51
Hình 3.1. Chi tiết giai đoạn huấn luyện	58
Hình 3.2. Chi tiết giai đoạn phân lớp	59

DANH MỤC CÁC BẢNG

Bảng 3.1. Bộ dữ liệu thử nghiệm	62
Bảng 3.2. Kết quả phân lớp bộ dữ liệu kiểm tra	63
Bảng 3.3. Đánh giá hiệu suất phân lớp	63

MỞ ĐẦU

1. Đặt vấn đề

Trong thời đại bùng nổ Công nghệ thông tin hiện nay, phương thức sử dụng văn bản giấy truyền thống đã dần được số hóa, chuyển sang dạng các văn bản điện tử lưu trữ trên máy tính và được chia sẻ, truyền tải trên mạng. Với rất nhiều tính năng ưu việt của tài liệu số như: Lưu trữ gọn nhẹ, linh hoạt; thời gian lưu trữ lâu dài; dễ hiệu chỉnh và đặc biệt tiện dụng trong trao đổi, chia sẻ nên ngày nay, số lượng văn bản điện tử được sử dụng trong các cơ quan nhà nước tăng lên rất nhanh chóng. Do đó, một vấn đề đặt ra là làm thế nào để có thể tìm kiếm và khai thác thông tin từ nguồn dữ liệu phong phú này. Các kỹ thuật để giải quyết vấn đề này được gọi là “Text Mining” hay Khai phá dữ liệu văn bản.

Khai phá dữ liệu văn bản đề cập đến tiến trình trích lọc các mẫu hình thông tin hay tri thức đáng quan tâm hoặc có giá trị từ các tài liệu văn bản. Trong đó, phân loại văn bản là một bài toán cơ bản nhất của lĩnh vực khai phá dữ liệu văn bản. Phân loại văn bản là công việc phân tích nội dung của văn bản và sau đó ra quyết định (hay dự đoán) văn bản thuộc nhóm nào trong các nhóm văn bản đã cho trước. Văn bản được phân loại có thể thuộc một nhóm, nhiều nhóm, hoặc không thuộc nhóm văn bản mà ta đã định nghĩa trước. Phân loại văn bản có thể thực hiện bằng nhiều cách như sử dụng tiếp cận lý thuyết tập thô, cách tiếp cận theo luật kết hợp hoặc dựa trên cách tiếp cận máy học. Đây là một lĩnh vực mang tính khoa học cao, ứng dụng được rất nhiều trong các bài toán thực tế hiện nay như tìm kiếm thông tin, lọc văn bản, tổng hợp tin tức tự động, thư viện điện tử,... Do vậy, học viên quyết định chọn đề tài “*Phân loại văn bản hành chính tiếng Việt và ứng dụng vào các cơ quan nhà nước tỉnh Bắc Kạn*” để nghiên cứu, thực hiện luận văn tốt nghiệp của mình.

Mục tiêu của đề tài luận văn là khảo sát, tìm hiểu một số phương pháp

phân loại văn bản thường được sử dụng hiện nay, trên cơ sở đó đề xuất lựa chọn một phương án phân loại văn bản tiếng Việt tự động và ứng dụng thử nghiệm phân loại cho một đối tượng cụ thể là văn bản hành chính tiếng Việt.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu bao gồm: Các thuật toán phân loại văn bản và các vấn đề liên quan đến bài toán phân loại văn bản tiếng Việt.

Phạm vi nghiên cứu của luận văn tập trung vào một số thuật toán phân loại văn bản thông dụng; các đặc trưng của văn bản tiếng Việt; các kỹ thuật liên quan trong xử lý phân loại văn bản và ứng dụng thuật toán học bán giám sát trong phân loại văn bản tiếng Việt.

3. Hướng nghiên cứu của đề tài

Nghiên cứu lý thuyết cơ bản về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản với một số thuật toán phân loại văn bản thông dụng như Naïve Bayes, K-Nearest Neighbor, Support Vector Machine.

Nghiên cứu về các đặc trưng của văn bản tiếng Việt và các kỹ thuật liên quan trong xử lý phân loại văn bản tiếng Việt như tách từ, biểu diễn văn bản, đánh trọng số của từ, tính độ tương đồng văn bản.

Từ kết quả thu được tiến hành cài đặt ứng dụng trong bài toán phân loại văn bản hành chính tiếng Việt.

4. Những nội dung chính

Nội dung chính của luận văn được trình bày trong 3 chương với tổ chức cấu trúc như sau:

Chương 1. Tổng quan về phân loại văn bản tiếng Việt.

Chương này trình bày khái quát về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản tiếng Việt; đồng thời làm rõ các đặc trưng của văn bản tiếng Việt và giới thiệu sơ bộ về công tác quản lý văn bản tại các cơ quan thuộc tỉnh Bắc Kạn.

Chương 2: Các kỹ thuật trong phân loại văn bản tiếng Việt.

Chương này trình bày về bài toán phân loại văn bản tiếng Việt với các thuật toán phân loại và các kỹ thuật cơ bản trong việc xử lý văn bản tiếng Việt để phân loại; sử dụng thuật toán SVM vào bài toán phân loại văn bản.

Chương 3: Áp dụng thuật toán Support Vector Machine phân loại văn bản hành chính tiếng Việt.

Chương này trình bày về một phương thức cải tiến của SVM là thuật toán bán giám sát SVM và sử dụng bán giám sát SVM vào bài toán phân loại văn bản tiếng Việt; tiến hành cài đặt thử nghiệm thuật toán.

5. Phương pháp nghiên cứu

Nghiên cứu cơ sở lý thuyết về phân loại văn bản, cơ sở lý thuyết về các thuật toán phân loại, cơ sở lý thuyết về xử lý văn bản tiếng Việt và thực nghiệm, tập trung vào việc xây dựng kho dữ liệu huấn luyện và xây dựng chương trình thử nghiệm để đánh giá kết quả phân loại văn bản.

6. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học: Đề tài nghiên cứu các vấn đề liên quan đến bài toán phân loại văn bản tiếng Việt và một số thuật toán thường được sử dụng trong phân loại văn bản. Ứng dụng thuật toán học bán giám sát SVM vào bài toán phân loại văn bản tiếng Việt.

Ý nghĩa thực tiễn: Luận văn đề xuất sử dụng thuật toán SVM trong bài toán phân loại văn bản tiếng Việt. Đây là thuật toán phân loại hiệu quả có độ chính xác cao, thích hợp áp dụng giải quyết các bài toán thực tế như tìm kiếm thông tin, phân loại văn bản, phân loại trang web,... Ứng dụng thử nghiệm được xây dựng có thể tiếp tục phát triển để áp dụng thực tiễn vào bài toán phân loại và tìm kiếm văn bản hành chính tiếng Việt với độ chính xác cao.

CHƯƠNG I. TỔNG QUAN VỀ PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

1.1. Khai phá dữ liệu

Khai phá dữ liệu là một quá trình khám phá ra các mẫu và tri thức thú vị từ một lượng lớn dữ liệu. Các nguồn dữ liệu có thể bao gồm các CSDL, kho dữ liệu, Web, các kho thông tin khác hoặc dữ liệu được truyền trực tiếp vào hệ thống. Đây là một bước quan trọng trong quá trình phát hiện tri thức trong CSDL [6].

Phát hiện tri thức trong CSDL (Knowledge Discovery from Data - KDD) là một quá trình không tầm thường nhận ra những mẫu có giá trị, mới, hữu ích tiềm năng và hiểu được trong dữ liệu [1]. Quá trình KDD gồm một số bước sau:

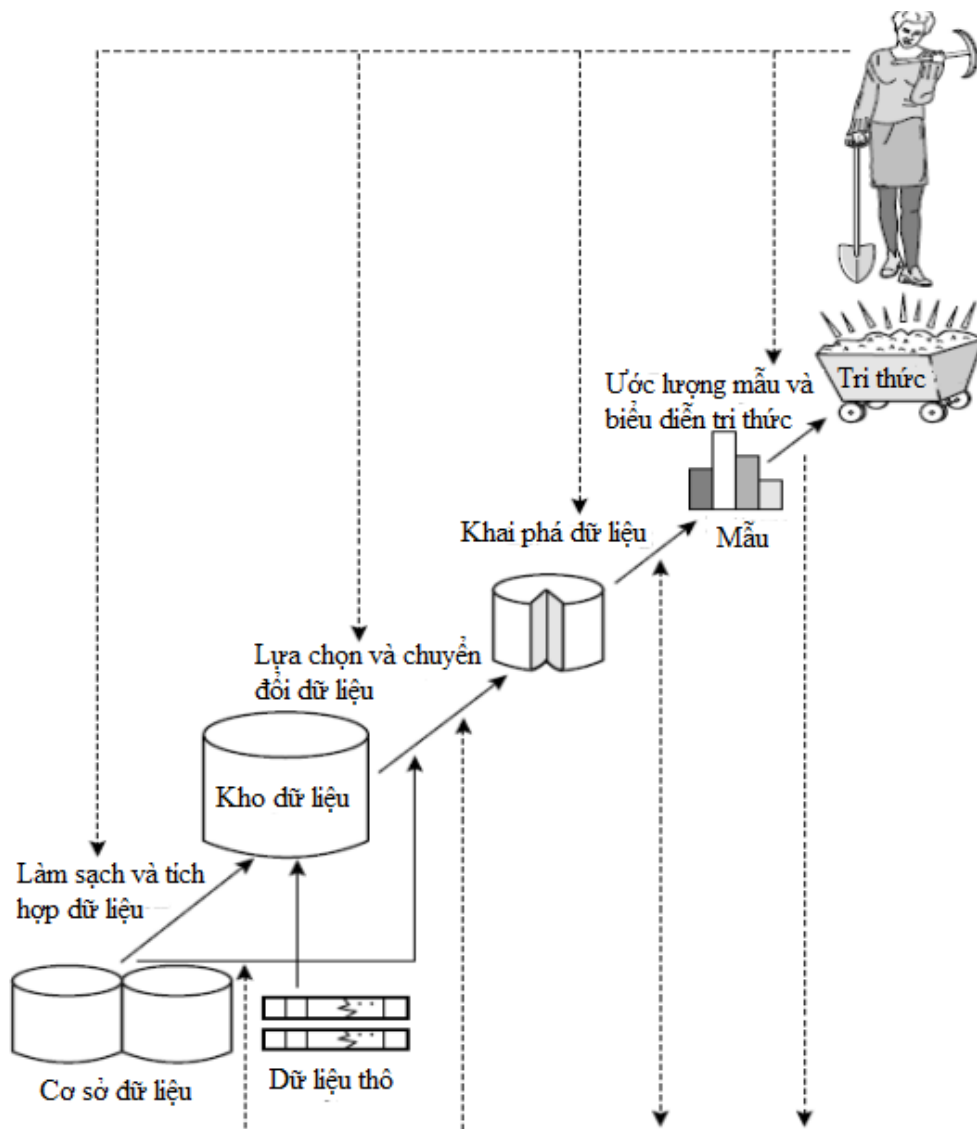
(1) Làm sạch và tích hợp dữ liệu (Cleaning and Integration): Loại bỏ nhiễu và các dữ liệu không cần thiết; tích hợp các nguồn dữ liệu lại với nhau.

(2) Lựa chọn, chuyển đổi dữ liệu (Selection and Transformation): Lựa chọn các dữ liệu có liên quan đến quá trình phân tích từ CSDL, chuyển đổi sang các dạng phù hợp cho quá trình xử lý.

(3) Khai phá dữ liệu (Data Mining): Là một trong những bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để trích chọn ra các mẫu dữ liệu.

(4) Ước lượng mẫu và biểu diễn tri thức (Evaluation and Presentation): Quá trình đánh giá kết quả thông qua một độ đo nào đó và biểu diễn các kết quả một cách trực quan cho người dùng.

Khai phá dữ liệu là giai đoạn chủ yếu của quá trình KDD, được thực hiện sau các quá trình thu thập và tinh lọc dữ liệu, có nghĩa là chỉ tìm các mẫu tri thức (pattern) có ý nghĩa trên tập dữ liệu có hy vọng chứ không phải là trên toàn bộ CSDL như các phương pháp thống kê trước đây.



Hình 1.1. Các bước trong quá trình phát hiện tri thức từ CSDL (KDD) [6]

Một số bài toán khai phá dữ liệu điển hình:


🚦 **Bài toán phân lớp (Classification/Categorization):** Phân lớp thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo, nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho dự báo tiếp theo. Một số phương pháp điển hình là cây quyết định, mạng neuron. Nội dung của phân lớp chính là một hàm ánh xạ các dữ liệu vào một trong một số lớp đã biết.

Ví dụ, phân lớp một văn bản vào trong một số lớp văn bản đã biết.


🚦 **Bài toán phân cụm (Clustering):** Phân cụm thực hiện nhóm dữ liệu

thành các “cụm” (có thể coi là các lớp mới) để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng. Phân cụm là bài toán mô tả hướng tới việc nhận biết một tập hữu hạn các cụm hoặc các lớp để mô tả dữ liệu. Các cụm (lớp) có thể tách rời nhau và toàn phần (tạo nên một phân hoạch cho tập dữ liệu), hoặc được trình bày đẹp hơn như phân lớp có thứ bậc hoặc có thể chồng lên nhau (giao nhau).

Ví dụ, phát hiện các nhóm người tiêu dùng trong CSDL tiếp thị, hoặc nhận biết các loại quang phổ trong tập phép đo không gian hồng ngoại.

 *Bài toán hồi quy (Regression)*: Hồi quy là một bài toán điển hình trong phân tích thống kê và dự báo, trong đó tiến hành việc dự đoán các giá trị của một hoặc một số biến phụ thuộc vào giá trị của một tập hợp các biến độc lập. Trong khai phá dữ liệu, bài toán hồi quy được quy về việc học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một số biến khác.

Ví dụ, bài toán dự báo nhu cầu người tiêu dùng đối với một sản phẩm mới được coi như một hàm của quảng cáo tiêu dùng.

 *Bài toán mô tả khái niệm (Concept Description)*: Nội dung của bài toán mô tả khái niệm là tìm ra các đặc trưng và tính chất của khái niệm (dùng để “mô tả” khái niệm đó). Điển hình nhất trong lớp bài toán này là các bài toán như tổng quát hóa, tóm tắt, phát hiện các đặc trưng dữ liệu ràng buộc.

Ví dụ, bài toán tóm tắt văn bản trong khai phá văn bản (Text Mining).

Ứng dụng của khai phá dữ liệu:

Khai phá dữ liệu tuy là một hướng tiếp cận mới nhưng thu hút được sự quan tâm của rất nhiều nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của nó. Chúng ta có thể liệt kê ra đây vài ứng dụng điển hình như:

- Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support);
- Điều trị y học (medical treatment);

- Khai phá văn bản và web (text mining & web mining);
- Nhận dạng (pattern recognition);
- ...

1.2. Khai phá dữ liệu văn bản

Khai phá dữ liệu văn bản (text mining) hay phát hiện tri thức từ các CSDL văn bản (textual databases) là quá trình trích chọn ra các mẫu hình thông tin (pattern) hay các tri thức (knowledge) mới, có giá trị và tác động được đang tiềm ẩn trong các văn bản để sử dụng các tri thức này vào việc tổ chức thông tin tốt hơn nhằm hỗ trợ con người [1].

Khai phá dữ liệu văn bản có thể được coi là việc mở rộng kỹ thuật khai phá dữ liệu truyền thống.

Thông tin được lưu trữ dưới dạng nguyên sơ nhất chính là văn bản (dữ liệu phi cấu trúc). Thậm chí ta có thể thấy rằng dữ liệu tồn tại dưới dạng văn bản còn có khối lượng lớn hơn rất nhiều so với các dữ liệu có cấu trúc khác. Thực tế, những nghiên cứu gần đây đã cho thấy rằng có đến 80% thông tin của một tổ chức nằm dưới dạng văn bản. Đó có thể là các công văn giấy tờ, các biểu mẫu điều tra, các yêu cầu khiếu nại, các thư tín điện tử (email), thông tin trên các website... Khi các nghiên cứu về CSDL ra đời vào những năm 60, người ta tưởng rằng có thể lưu mọi loại thông tin dưới dạng dữ liệu có cấu trúc. Nhưng thực tế sau hơn 50 năm phát triển, người ta vẫn dùng các hệ thống lưu trữ ở dạng văn bản và thậm chí còn có xu hướng dùng thường xuyên hơn. Từ đó người ta có thể tin rằng các sản phẩm khai phá dữ liệu văn bản có thể có giá trị thương mại cao hơn rất nhiều lần so với các sản phẩm khai phá dữ liệu truyền thống khác. Tuy nhiên, ta cũng có thể thấy ngay rằng các kỹ thuật khai phá dữ liệu văn bản phức tạp hơn nhiều so với các kỹ thuật khai phá dữ liệu truyền thống bởi vì phải thực hiện trên dữ liệu văn bản vốn đã ở dạng phi cấu trúc và có tính mờ (fuzzy).

Một ví dụ cho bài toán khai phá dữ liệu văn bản, khi nói đến các thiết bị văn phòng, ta có các thông tin sau:

- “Máy in là thiết bị ngoại vi đi kèm với máy tính cá nhân”
- “Máy tính cá nhân thường được sử dụng tại các văn phòng”

Sau khi phân tích các thông tin quan trọng này, hệ thống cần phải đưa ra các suy luận cụ thể:

- “Khi trang bị máy tính cá nhân cho các nhân viên văn phòng phải trang bị kèm theo máy in”.

Rõ ràng ở đây có sự phân tích suy luận ở mức độ cao. Để đạt được như vậy cần phải có những công trình nghiên cứu về trí tuệ nhân tạo tiên tiến hơn.

Bài toán khai phá dữ liệu văn bản là một bài toán nghiên cứu đa lĩnh vực, bao gồm nhiều kỹ thuật cũng như các hướng nghiên cứu khác nhau: Thu thập thông tin (information retrieval), phân tích văn bản (text analysis), chiết xuất thông tin (information extraction), phân loại văn bản (categorization), học máy (machine learning),... và bản thân các kỹ thuật khai phá dữ liệu.

Trong khuôn khổ đề tài này học viên tập trung đề cập đến một bài toán cụ thể, đó là bài toán *phân loại dữ liệu văn bản* (text categorization).

Quá trình khai phá văn bản:

Quá trình khai phá văn bản là cụ thể hóa quá trình khai phá dữ liệu nói chung đối với dữ liệu văn bản. Với giả thiết đã xác định được: (1) bài toán khai phá văn bản và (2) miền dữ liệu văn bản thuộc miền ứng dụng, quá trình khai phá văn bản thường bao gồm bốn bước chính [1]:

1- Bước tiền xử lý, bao gồm hai giai đoạn:

+ Thu thập dữ liệu văn bản thuộc miền ứng dụng. Có hai điều cần được lưu ý ở giai đoạn này. *Thứ nhất*, chỉ cần thu thập dữ liệu văn bản thuộc miền ứng dụng mà không phải là tập tất cả các văn bản có thể có của thế giới thực. Ví dụ, trong bài toán khai phá văn bản thuộc lĩnh vực công nghệ thông tin thì

chỉ cần quan tâm thu thập các văn bản về công nghệ thông tin. *Thứ hai*, yêu cầu cốt lõi của giai đoạn này là tập dữ liệu văn bản thu thập được phải đại diện được cho toàn bộ dữ liệu văn bản thuộc miền ứng dụng, nhưng không phải là toàn bộ dữ liệu văn bản thuộc miền ứng dụng.

+ Biểu diễn dữ liệu văn bản thu thập được sang khuôn dạng phù hợp với bài toán khai phá văn bản. Ở giai đoạn này, hệ thống sẽ chuyển văn bản từ dạng phi cấu trúc về dạng có cấu trúc. Ví dụ, với nội dung: “*Luận văn này khó lắm*”, hệ thống sẽ cố gắng phân tích thành *Luận văn|này|khó|lắm*. Các từ được lưu riêng rẽ một cách có cấu trúc để tiện cho việc xử lý.

2- Lựa chọn tập dữ liệu đầu vào cho thuật toán khai phá dữ liệu. Trong hầu hết trường hợp, tập dữ liệu thuộc miền ứng dụng đã thu thập được là rất lớn, vì vậy nhiều trường hợp vượt quá khả năng xử lý (về không gian, thời gian) đối với các thuật toán khai phá dữ liệu. Do đó, cần chọn ra từ tập dữ liệu thu thập được một tập con để thực hiện bài toán khai phá dữ liệu. Tập con này được xác định bằng cách loại bỏ các thông tin dư thừa, giữ lại các yếu tố đảm bảo tính đại diện của tập dữ liệu thu thập được. Bước này phụ thuộc nhiều vào ngôn ngữ đang được phân tích và kỹ thuật sẽ được dùng để phân tích ở bước tiếp theo. Ví dụ, nếu kỹ thuật phân tích văn bản chỉ dựa vào xác suất xuất hiện từ khoá, khi đó ta có thể loại bỏ các từ phụ như: *Nếu, thì, thế nhưng...*

3- Thực hiện thuật toán khai phá dữ liệu đối với tập dữ liệu đã được lựa chọn để tìm ra các mẫu, các tri thức. Ví dụ, đối với bài toán phân lớp văn bản, mẫu (tri thức) được tích hợp thành bộ phân lớp kết quả và bộ phân lớp này sẽ được sử dụng vào việc phân lớp đối với các văn bản mới.

4- Thực hiện việc khai thác sử dụng các mẫu, các tri thức nhận được từ quá trình khai phá văn bản vào thực tiễn hoạt động.

Có rất nhiều kỹ thuật, phương pháp được sử dụng cho khai phá văn bản. Các bước tiền xử lý là các kỹ thuật rất phức tạp nhằm phân tích một phân lớp

đặc biệt thành các thuộc tính đặc biệt, sau đó tiến hành áp dụng các phương pháp khai phá dữ liệu kinh điển tức là phân tích thống kê và phân tích các liên kết. Các bước còn lại sẽ khai phá cả văn bản đầy đủ từ tập các văn bản, ví dụ như phân lớp văn bản.

Các kỹ thuật chính của khai phá văn bản có thể được phân ra thành các nhiệm vụ mà chúng thực hiện khi xử lý khai phá văn bản: Loại thông tin mà chúng có thể trích ra và loại phân tích được thực hiện bởi chúng.

Các loại thông tin được trích ra có thể là:

- **Các nhãn:** Giả sử, được liên kết với mỗi văn bản là tập các nhãn, các thao tác khai phá tri thức được thực hiện trên các nhãn của mỗi văn bản. Nói chung, có thể giả sử rằng các nhãn tương ứng với các từ khoá, mỗi một từ khoá có quan hệ với một chủ đề cụ thể nào đó.

- **Các từ:** Ở đây giả sử rằng một văn bản được gán nhãn với từng từ xuất hiện trong văn bản đó.

- **Các thuật ngữ:** Với mỗi văn bản tìm thấy các chuỗi từ, mỗi chuỗi từ thuộc về một lĩnh vực nào đó, và việc khai phá văn bản được thực hiện trên các khái niệm gán nhãn cho mỗi văn bản. Thường thì các thuật ngữ được tách ra ít và có xu hướng tập trung vào các thông tin quan trọng của văn bản.

Các loại kết hợp:

- **Kết hợp thông thường:** Một số thuật toán trước đây giả sử rằng dữ liệu nguyên mẫu được tạo lập chủ dẫn để trợ giúp cho các kỹ thuật xử lý ngôn ngữ tự nhiên. Các cấu trúc có chủ dẫn trên thực tế có thể được sử dụng như một cơ sở cho việc xử lý khai phá tri thức.

- **Các phân cấp thuật ngữ:** Ở đây mỗi văn bản được đánh với các thuật ngữ lấy ra từ một phân cấp các thuật ngữ. Sau đó, một hệ thống sẽ phân tích sự phân bố nội dung của các thuật ngữ hậu duệ của từng thuật ngữ liên quan đến các hậu duệ khác do các phân bố liên kết và các phép đo khác nhằm khai thác

các quan hệ mới giữa chúng. Loại liên kết này có thể cũng được sử dụng để lọc và tổng hợp chủ đề của các tin tức.

- **Khai phá văn bản đầy đủ:** Không giống như loại liên kết thông thường thực hiện thao tác “mù quáng” trên các chú dẫn của văn bản, kỹ thuật này sử dụng lợi thế của nội dung nguyên mẫu của các văn bản. Kỹ thuật này được gọi là “trích văn bản nguyên mẫu”.

1.3. Phân loại văn bản

1.3.1. Giới thiệu bài toán phân loại văn bản

Bài toán phân loại văn bản (Text categorization) giải quyết việc gán tên các chủ đề (tên lớp/nhãn lớp) đã được xác định cho trước vào các văn bản dựa trên nội dung của nó. Phân loại văn bản được sử dụng để hỗ trợ trong quá trình tìm kiếm thông tin (information retrieval), chiết lọc thông tin (information extraction) hoặc lọc văn bản... [1],[12]. Đây là một tác vụ liên quan đến việc ra quyết định xử lý. Với mỗi xử lý phân loại, khi đưa ra một văn bản, một quyết định được đưa ra nó có thuộc một lớp nào hay không. Nếu nó thuộc một phân lớp nào đó thì phải chỉ ra phân lớp mà nó thuộc vào. Ví dụ, đưa ra một chủ đề về *công nghệ thông tin*, cần phải đưa ra quyết định rằng chủ đề đó thuộc các phân lớp *phần cứng*, *phần mềm*, *hệ thống thông tin* hay bất cứ một khái niệm nào khác thuộc về lĩnh vực công nghệ thông tin.

Nói cách khác, phân loại văn bản là tiến trình đưa các văn bản chưa biết chủ đề vào các lớp văn bản đã biết chủ đề. Các chủ đề này được xác định bởi một tập các tài liệu mẫu. Để thực hiện quá trình phân loại văn bản, một giải thuật máy học được sử dụng để xây dựng bộ phân loại từ tập huấn luyện bao gồm nhiều văn bản, sau đó dùng bộ phân loại này để dự đoán lớp của những tài liệu mới. Đây là một trong những bài toán cơ bản nhất của lĩnh vực khai phá dữ liệu văn bản.

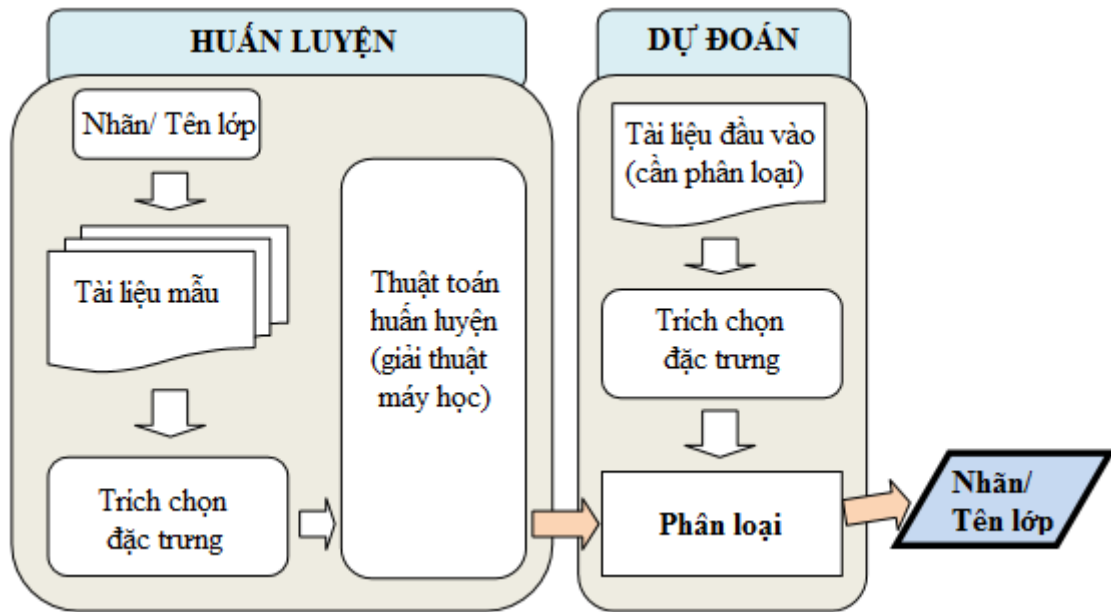
Đặc điểm nổi bật của bài toán phân loại văn bản là sự đa dạng của chủ đề văn bản và tính đa chủ đề của văn bản. Tính đa chủ đề của văn bản làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, nếu do con người thực hiện, và dễ bị nhập nhằng khi phân loại tự động. Ví dụ, một tài liệu về *Văn hóa* có thể xếp vào *Kinh tế* nếu như viết về kinh phí đầu tư cho du lịch và tác động của đầu tư này đến *kinh tế - xã hội*. Về bản chất, một văn bản là một tập hợp từ ngữ có liên quan với nhau tạo nên nội dung ngữ nghĩa của văn bản. Từ ngữ của văn bản là đa dạng do tính đa dạng của ngôn ngữ (đồng nghĩa, đa nghĩa, từ vay mượn nước ngoài,...) và số lượng từ cần xét là lớn. Cần lưu ý rằng, một văn bản có thể có số lượng từ không nhiều, nhưng số lượng từ cần xét rất nhiều vì phải bao hàm tất cả các từ của ngôn ngữ đang xét.

Việc tự động phân loại văn bản vào một chủ đề nào đó giúp cho việc sắp xếp, lưu trữ và truy vấn tài liệu dễ dàng hơn về sau. Một trong những ứng dụng quan trọng nhất của phân loại văn bản tự động là ứng dụng trong các hệ thống tìm kiếm văn bản. Từ một tập con văn bản đã phân lớp sẵn, tất cả các văn bản trong miền tìm kiếm sẽ được gán chỉ số lớp tương ứng. Trong truy vấn của mình, người dùng có thể xác định chủ đề hoặc lớp văn bản mà mình mong muốn tìm kiếm để hệ thống cung cấp đúng yêu cầu của mình.

Trong phân lớp văn bản, sự tương ứng giữa một văn bản với một lớp thông qua việc gán giá trị đúng sai (True - văn bản thuộc lớp, hay False - văn bản không thuộc lớp) hoặc thông qua một độ phụ thuộc (đo độ phụ thuộc của văn bản vào lớp). Trong trường hợp có nhiều lớp thì phân loại đúng sai sẽ là việc xem một văn bản có thuộc vào một lớp duy nhất nào đó hay không.

1.3.2. Quy trình phân loại văn bản

Qua tìm hiểu, học viên nghiên cứu áp dụng quy trình phân loại văn bản chung cho hầu hết các phương pháp phân loại như sơ đồ sau:



Hình 1.2. Quy trình phân loại văn bản [7]

Để tiến hành phân loại văn bản nói chung, ta thực hiện qua hai bước:

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.

Các phần xử lý của từng quá trình sẽ được trình bày chi tiết trong các chương tiếp theo của luận văn.

1.3.3. Phân loại văn bản tiếng Việt

Bài toán phân loại văn bản tiếng Việt được đưa ra nhằm giải quyết việc xây dựng một hệ thống có thể phân loại được văn bản tiếng Việt. Hay nói khác đi, khi đưa ra một văn bản tiếng Việt, hệ thống cần chỉ ra rằng văn bản đó là loại văn bản thuộc chủ đề nào (kinh tế, chính trị, giáo dục, thể thao,...).

Để giải quyết được bài toán phân loại văn bản tiếng Việt, cần phải dựa vào những kết quả nghiên cứu về văn bản nói chung, về dữ liệu văn bản và các kỹ thuật xử lý đã được phát triển trên thế giới. Tuy nhiên, các văn bản tiếng

Việt lại có những đặc trưng riêng của nó. Ta có thể dễ dàng nhận thấy sự khác biệt về mặt kí pháp, cú pháp và ngữ pháp của tiếng Việt trong văn bản so với các ngôn ngữ phổ biến trên thế giới như tiếng Anh, tiếng Pháp. Do vậy, chúng ta cần phải tìm hiểu về những đặc trưng riêng của các văn bản tiếng Việt, trên cơ sở đó lựa chọn các kỹ thuật xử lý phù hợp áp dụng cho bài toán phân loại văn bản tiếng Việt.

1.4. Đặc trưng của văn bản tiếng Việt

Tiếng Việt là một *ngôn ngữ đơn lập* [2], đặc điểm này bao quát toàn bộ đặc trưng tiếng Việt về mặt ngữ âm, ngữ nghĩa và ngữ pháp. Do đó, chúng ta phải tiến hành nghiên cứu đặc điểm này của tiếng Việt để có thể có được hướng nghiên cứu phù hợp cho bài toán xử lý phân loại văn bản tiếng Việt.

1.4.1. Các đơn vị của tiếng Việt

a. Tiếng và đặc điểm của tiếng

Trong tiếng Việt, cũng như trong các văn bản tiếng Việt, *tiếng* là một thành phần khá quan trọng. Trong ký pháp, mỗi tiếng đứng độc lập, và ta có thể phát hiện được ngay các *tiếng* trong tiếng nói cũng như trong văn bản [2].

Tiếng và giá trị ngữ âm:

Ngữ âm chính là mặt âm của ngôn ngữ. Trên thực tế, các ứng dụng liên quan đến tiếng Việt như dịch thuật, lưu trữ người ta vẫn ghi lại âm thành dạng văn bản, sau đó mới tiến hành các thao tác xử lý. Mỗi tiếng chính là một âm tiết và được ghi lại thành một cụm trong văn bản.

Tiếng và giá trị ngữ nghĩa:

Nếu xét về mặt ngữ nghĩa thì *tiếng* là đơn vị nhỏ nhất có thể có nghĩa [2]. Thực ra ta có thể thấy rằng đơn vị ngữ âm thấp nhất là âm vị thì hoàn toàn không có nghĩa (ví dụ như các chữ cái đứng riêng rẽ). Tuy nhiên cũng có những tiếng có nghĩa (ví dụ như *a*, *ừ*).

Theo [2], ta có thể phân biệt các tiếng như sau:

- Các tiếng tự nó có nghĩa (ví dụ như *chuông, bút, gió*) có thể được dùng để gọi tên sự vật, hiện tượng, có thể được dùng như một từ.
- Các tiếng có nghĩa nhưng không dùng để gọi tên sự vật, hiện tượng mà chỉ được dùng với tư cách là bộ phận để cấu thành nên từ có nghĩa ở bậc cao hơn. Ví dụ: Ta không thể nói *tôi thực* mà chỉ có thể nói *tôi ăn*, nhưng có những từ như *thực phẩm*.
- Các tiếng bản thân không hề có nghĩa mà chỉ dùng để kết hợp tạo thành nghĩa cho đơn vị trực tiếp cao hơn, đó là từ. Ví dụ: Các tiếng *lãng, dăng* tự nó không có nghĩa nhưng có thể tạo thành từ có nghĩa là *lãng dăng*.

Tiếng và giá trị ngữ pháp:

Khía cạnh ngữ pháp bao gồm những quy tắc cấu tạo từ, cấu tạo câu. Và ta có thể thấy rằng tiếng là *đơn vị ngữ pháp dùng để cấu tạo từ* [2].

Về việc dùng tiếng để cấu tạo từ, ta có hai trường hợp như sau:

- Từ một tiếng: Đây là trường hợp một tiếng dùng để làm một từ, ví dụ như *cây, đá*. Các tiếng (đóng vai trò là từ) là một bộ phận cấu thành nên câu.
- Từ nhiều tiếng: Là một khối hai hay nhiều hơn các tiếng kết hợp với nhau, gắn bó tương đối chặt chẽ.

Việc nghiên cứu cấu trúc từ (nhiều tiếng hay một tiếng) rất quan trọng trong quá trình nghiên cứu và cài đặt ứng dụng phân tích cú pháp tiếng Việt.

b. Từ và các đặc điểm của từ

Từ là đơn vị nhỏ nhất để đặt câu:

Như trên vừa trình bày, ta thấy từ có thể gồm có một tiếng nhưng cũng có thể gồm hai hay nhiều tiếng, tuy nhiên từ là *đơn vị nhỏ nhất để đặt câu* [2].

Có một lưu ý là để đặt câu, tức là để viết, để nói, để suy nghĩ thì chúng ta dùng *từ* chứ không phải là dùng *tiếng*. Đây là một lưu ý rất quan trọng, vì trong thực tế thành phần riêng rẽ có thể phát hiện trong một câu (ở dạng nói hay viết) là một *tiếng*, nhưng để có thể hiểu ý nghĩa của câu ta phải dùng *từ*.

Do đó bất kì một nghiên cứu về tiếng Việt trên máy tính nào **cũng phải quan tâm đến việc ghép các tiếng thành từ.**

✚ Từ có nghĩa hoàn chỉnh và cấu tạo ổn định:

Ta có thể nhận ra điều này ở các từ tiếng Việt một tiếng, còn đối với những từ nhiều tiếng thì đó là những đặc điểm xác định lẫn nhau, cấu tạo ổn định dẫn đến nghĩa hoàn chỉnh và ngược lại. Ví dụ như từ hai tiếng *cây cối* có cấu tạo ổn định và nghĩa hoàn chỉnh, nhưng cụm không phải là từ như *cây* và *cối* không có cấu tạo ổn định và nghĩa hoàn chỉnh.

Đối với những từ nhiều tiếng, tính hoàn chỉnh về nghĩa và ổn định về cấu tạo được hình thành theo mối quan hệ giữa các tiếng cấu thành nên từ. Đó là mối quan hệ phối hợp, có thể theo ngữ âm (các từ láy âm), hoặc về nghĩa (ví dụ như nghĩa của hai từ *xe* và *đạp* trong từ *xe đạp*).

c. Câu và các đặc điểm của câu

Trong ngữ pháp tiếng Việt, từ và câu là những đơn vị ngữ pháp rất quan trọng. Đối với con người, từ được coi như sẵn có trong kho từ vựng được tích lũy trong quá trình sống. Còn để có thể hiểu, giao tiếp thì con người phải dùng đến câu. Trong ngôn ngữ, câu là đơn vị ở bậc cao hơn cả. Nói gì, viết gì cũng phải thành câu.

✚ Câu có ý nghĩa hoàn chỉnh:

Tính hoàn chỉnh về nghĩa của câu là tính hoàn chỉnh của cả một quá trình tư duy, quá trình thông báo diễn ra trong một hoàn cảnh nhất định [2].

Trong một câu bao giờ cũng có hai thành phần, một thành phần nêu sự vật hiện tượng và một thành phần giải thích của sự vật hiện tượng đó.

✚ Câu có cấu tạo đa dạng:

Câu có dạng đơn giản như là *câu đơn*, và còn có những cấu trúc phức tạp hơn gọi là *câu ghép*. Xét về mặt ngữ nghĩa, câu đơn có nhiều dạng khác nhau, biểu lộ những ý nghĩa, trạng thái, nội dung cần thông báo khác nhau.

Tính chất đa dạng không trái ngược với tính chất chặt chẽ của câu về mặt ngữ pháp. Nói chung, cấu tạo ngữ pháp có thay đổi thì nghĩa cũng có thay đổi và ngược lại [2].

1.4.2. Ngữ pháp của tiếng Việt

a. Trong phạm vi cấu tạo từ

Trong phạm vi cấu tạo từ, phương tiện ngữ pháp chủ yếu là *sự kết hợp* các tiếng. *Trật tự* sắp xếp các tiếng rất quan trọng trong cấu tạo từ. Kết hợp hai phương tiện này, có hai phương thức cấu tạo từ chủ yếu là *láy* và *ghép*.

Láy là việc sắp đặt các tiếng thành đôi, kề cận nhau, có sự phối hợp về ngữ âm tạo nên nghĩa.

Ghép là việc sắp đặt các tiếng thành nhóm, kề cận nhau, có sự phối hợp về ngữ nghĩa tạo nên nghĩa của từ ghép.

b. Trong phạm vi cấu tạo câu (phạm vi cú pháp)

Ta có các phương tiện *trật tự*, *hư từ* và *ngữ điệu*.

Trật tự sắp đặt các từ là phương tiện chính để biểu thị quan hệ ngữ pháp - tức là quan hệ cú pháp - giữa các từ trong một câu [2]. Trong tiếng Việt, khi trật tự các yếu tố cấu thành thay đổi thì nghĩa của câu cũng thay đổi theo. Ví dụ ta có các hoán vị các tiếng của một tổ hợp như sau:

Sai đâu sửa đấy. | Sửa đâu sai đấy. | Sửa đấy sai đâu. | Đấy sai sửa đâu.

Trật tự theo hướng thuận biểu hiện ở chỗ yếu tố chính trước, yếu tố phụ sau, yếu tố được xác định trước, yếu tố xác định sau, yếu tố dùng để khai triển đứng liền sau từ, nếu có cách ly cũng đứng không quá xa.

Hư từ là những từ dùng để biểu thị một số những quan hệ cú pháp nhất định. Ví dụ trong câu *Anh, chị đã đi chơi rồi*, có thể dùng từ *và* để nối hai từ *anh, chị* để làm rõ hơn mối quan hệ liên hợp, bình đẳng giữa hai từ. Như thế từ *và* là một hư từ. Tuy nhiên ta có thể thấy rằng hư từ có những sắc thái về nghĩa [2], ví dụ như ta thấy “*Anh với chị*” khác với “*Anh và chị*”.

Điều đó đặt ra một khó khăn khi tiến hành nghiên cứu tiếng Việt trên máy tính, đó là việc xác định ngữ nghĩa của một câu, vì trong văn bản tiếng Việt thì cách viết hay biểu thị trên khá phổ biến.

1.4.3. Từ tiếng Việt

a. Từ đơn - từ ghép

Như đã trình bày ở trên, từ trong tiếng Việt có thể có một tiếng hay gồm nhiều tiếng. Những từ nhiều tiếng lại có thể được ghép bởi những tiếng hay từ khác có nghĩa. Ví dụ hai từ *đất*, *nước* ghép với nhau thành một từ có ý nghĩa trừu tượng hơn là *đất nước*. Những từ này được gọi là các từ ghép.

Do tồn tại cả hai dạng từ đơn (một tiếng) và từ ghép, ta phải nghiên cứu để có thể đề xuất phương án hữu hiệu cho bài toán nhận dạng từ trong câu.

Khi xem xét từ ghép, chúng ta có thể thấy có hai loại như sau:

- Từ ghép song song: Mỗi tiếng thường là một tiếng có nghĩa, có thể dùng làm từ một tiếng, gắn bó với nhau theo quan hệ song song và nói chung có thể đổi chỗ cho nhau. Trong sự phối hợp về ngữ nghĩa thì thường có sự biến đổi nghĩa riêng thành một nghĩa hình tượng, như ví dụ của từ *đất nước* đã nêu trên, hoặc trong các từ *quần áo*, *giày dép*....

- Từ ghép chính phụ: Mỗi tiếng có thể là một tiếng có nghĩa, nhưng thông thường có một tiếng chính có thể được dùng làm từ còn tiếng kia không có chức năng ngữ pháp đó, ví dụ như *nhà thương*, *bánh mì*.

Ta cũng có thể thấy trong tiếng Việt tồn tại một số các từ ghép có nhiều tiếng hơn, phát triển từ loại từ ghép chính phụ, qua đó có thể chia thành các phần chính, phần phụ, thuận tiện hơn trong việc phân tích từ.

Chính sự tồn tại của từ ghép (ghép bởi các tiếng có nghĩa) mà có sự nhập nhằng về nghĩa của một câu. Ta có thể lấy ví dụ sau:

Ông già đi nhanh quá

Ở đây, hai tiếng *già* và *đi* đều là các tiếng có nghĩa, do đó câu trên có thể hiểu theo nhiều cách như sau:

Ông / già đi nhanh quá | Ông già / đi nhanh quá

Giải quyết được vấn đề này rất phức tạp, tuy nhiên chắc chắn muốn phân tích câu hoặc xử lý văn bản tiếng Việt thì bài toán đầu tiên được đặt ra là *làm thế nào để tách các từ trong câu*.

b. Từ loại

Có thể phân loại các từ theo cách thức cấu tạo như đã xét ở phần trên, cũng có thể phân loại theo các chữ cái đầu như khi ta làm từ điển. Tuy nhiên có một cách phân loại đặc biệt quan trọng về mặt cấu tạo câu, đó là xác định *từ loại* cho mỗi từ tiếng Việt.

Theo [2], tiếng Việt có thể có những từ loại sau:

- Danh từ (Việt Nam, Hà Nội, Nguyễn Ái Quốc)
- Động từ (đi, đứng, soạn bài)
- Tính từ (xấu, đẹp, trắng, đỏ)
- Phó từ (sẽ, đã, rồi, rất)
- Liên từ (của, thì)
- Đại từ (tôi, nó, anh, em, hắn)
- Trợ từ (nhỉ, hả, nhé)
- Cảm từ (ái chà, chao ôi, vâng, dạ)
- Số từ (một, hai)
- Loại từ (con, cái)
- Giới từ (cùng, với, bằng, để)
- Trạng từ (hôm qua)

Tuy nhiên việc phân loại trên chỉ có ý nghĩa tương đối. Nếu xem xét kỹ hơn nữa về mặt cú pháp, trong mỗi loại từ lại còn có thể chia nhỏ hơn được

nữa, ví dụ *rất* là *phó từ đứng trước* vì trong một câu nó chỉ đứng trước các tính từ để nhấn mạnh hiệu quả biểu đạt của tính từ.

c. Dừng từ cấu tạo ngữ

Ngữ là đơn vị ngữ pháp bậc trung gian giữa từ và câu [2].

Việc tìm hiểu cấu tạo cũng như các loại ngữ là cần thiết để tìm hiểu cấu tạo của câu. Qua cấu tạo của ngữ, có thể nhận rõ thêm đặc điểm ngữ pháp của từ loại và các tiểu loại.

Theo [2], ta có một số nhận xét như sau:

- Ngữ là một cấu tạo theo quan hệ cú pháp chính phụ.
- *Kết từ* cũng được dùng để biểu hiện quan hệ chính phụ giữa chính tố với một số loại phụ tố sau. Ví dụ "*luận văn của tôi*".
- Khi phụ tố sau do thực từ đảm nhiệm thì nói chung phụ tố ấy có thể là một ngữ. Ví dụ: "*một người / học sinh / rất thông minh*".

Ta có thể xét một số ngữ loại như sau [2]:

- Danh ngữ: Ngữ có danh từ làm trung tâm.
- Động ngữ: Ngữ có động từ làm trung tâm.
- Tính ngữ: Ngữ có tính từ làm trung tâm.
- Giới ngữ: Ngữ bắt đầu bằng giới từ.

Cũng như đã phân tích trong phần từ loại, để có thể xây dựng được một hệ thống luật cú pháp tốt, ta cần phải phân chia các ngữ loại một cách chặt chẽ hơn. Ví dụ: Ta có danh ngữ "*cái cầu*", nếu thêm một *số từ* nữa, ví dụ từ "*một*" thì danh ngữ mới "*một cái cầu*" phải là một *danh ngữ kết thúc trái* vì rõ ràng ta không thể mở rộng về phía trái danh ngữ này nữa.

1.4.4. Câu tiếng Việt

Câu là đơn vị dừng từ, hay đúng hơn là dừng ngữ mà cấu tạo nên trong quá trình tư duy, thông báo; nó có nghĩa hoàn chỉnh, có cấu tạo ngữ pháp và có tính chất độc lập [2].

Xét về cấu trúc câu, tiếng Việt có hai loại câu là *câu đơn* và *câu ghép*.

a. Câu đơn

Câu đơn là loại câu cơ sở của tiếng Việt, bao gồm một nòng cốt đơn hay một kết cấu chủ vị. Về mặt ngữ nghĩa, câu đơn mang nghĩa tự thân, còn câu ghép mang nghĩa kết hợp. Câu đơn có thể là câu khẳng định, câu phủ định, câu nghi vấn, câu tường thuật, câu cầu khiến, câu biểu cảm. Ví dụ:

- *Tôi chưa làm xong việc này.*
- *Em đi làm chưa?*
- *Bông hoa mới đẹp làm sao!*

Nòng cốt đơn của một câu đơn là một kết cấu chủ vị. Ngoài ra, câu đơn còn có các thành phần ngoài nòng cốt [2]:

- Thành phần than gọi. Ví dụ: "**Em ơi**, chúng mình đi nào".
- Thành phần chuyển tiếp. Ví dụ: "*Còn hẵn*, **trái lại**, không làm gì cả".
- Thành phần chú thích. Ví dụ: "*Nó*, **em tôi**, rất thông minh".
- Thành phần tình huống. Ví dụ: "**Ở nhà**, tôi ít tuổi nhất".
- Thành phần khởi ý. Ví dụ: "**Thuốc**, anh ấy không hút".

Để biểu diễn một câu đơn, người ta thường dùng mô hình suy diễn câu đơn như sau: Px - Cx - Vx - Bx.

Trong đó, P: Thành phần phụ; C: Chủ ngữ; V: Vị ngữ; B: Bỏ ngữ, định ngữ; x: Thành phần có thể khai triển tiếp.

Cách biểu diễn này rõ ràng rất thuận tiện trong việc xây dựng bộ luật cú pháp và tiến hành phân tích cú pháp cho một câu đầu vào.

b. Câu ghép

Về mặt ngữ pháp, câu ghép bao gồm bộ phận chủ yếu là một nòng cốt ghép, được tạo nên bởi ít nhất hai vế và mỗi vế thường bao gồm một nòng cốt đơn. Ví dụ:

- *Mây tan, mưa tạnh.*

Tuy rằng câu đơn chỉ có một nòng cốt đơn nhưng không phải bao giờ câu đơn cũng ngắn hơn câu ghép. Người ta có thể chia câu ghép thành hai loại: Câu ghép song song và câu ghép qua lại [2].

✚ Câu ghép song song:

Là loại câu ghép có thể có hai vế hay nhiều hơn, tuy nhiên sự liên kết giữa các vế là lỏng lẻo, có thể tách thành các câu đơn mà vẫn bảo toàn nghĩa. Trong một số trường hợp các vế có quan hệ, sử dụng các kết từ, tuy nhiên ý nghĩa độc lập của các vế vẫn tương đối rõ ràng. Ví dụ:

- *Khán giả hò reo, cờ phát rực trời, cuộc đấu diễn ra quyết liệt.*
- *Nó vẫy tôi và tôi tiến lại phía nó.*

✚ Câu ghép qua lại:

Là loại câu có hai vế và vế này là điều kiện tồn tại của vế kia. Có cả hai vế thì câu mới có ý nghĩa trọn vẹn. Nối giữa hai vế là các liên từ, thông thường người ta dùng cả cặp liên từ. Câu ghép có thể biểu diễn như sau:

$$xN1 + yN2$$

Một trong các liên từ có thể được loại bỏ. Ta có một số ví dụ như sau:

- *(Bởi) vì N1 (cho) nên/ mà N2.*
- *Nếu N1 thì N2.*
- *Không những N1 mà còn N2.*
- ...

*(Vi) phở ngon **nên** cửa hàng của anh ấy luôn đông khách.*

*Tôi làm **thì** nó được nghỉ.*

✚ Các thành phần câu:

- Chủ ngữ: Thành phần chủ yếu của câu.
- Vị ngữ: Thành phần chính, bổ sung, giải thích ý nghĩa cho chủ ngữ.
- Trạng ngữ: Thành phần thứ yếu, bổ sung ý nghĩa cho câu, chỉ nơi chốn, thời gian, không gian.

- Bỏ ngữ: Thành phần phụ thuộc, bổ sung nghĩa cho động từ làm vị ngữ.
- Định ngữ: Thành phần phụ thuộc, bổ sung ý nghĩa cho vị ngữ.

1.4.5. Các đặc điểm chính tả và văn bản tiếng Việt

Việc nghiên cứu các đặc điểm chính tả có ý nghĩa rất quan trọng trong khâu tiền xử lý dữ liệu, tạo nguồn dữ liệu đầu vào cho những pha sau như phân tích cú pháp hay đánh trọng số cho các từ, lập chỉ mục.

Một số vấn đề về chính tả tiếng Việt mà ta cần quan tâm như sau:

- Các từ đồng âm thường sử dụng tùy tiện như “*Mĩ*”/“*Mỹ*”, “*kĩ*”/“*kỹ*”...
- Từ địa phương: Người ta vẫn hay sử dụng một số từ địa phương thay cho các từ phổ thông trong văn bản. Ví dụ “*cây kiếng*” thay cho “*cây cảnh*”.
- Vị trí dấu: Theo quy định đánh dấu tiếng Việt, dấu được đặt trên nguyên âm có ưu tiên cao nhất. Tuy nhiên khi soạn văn bản, do có nhiều bộ gõ tiếng Việt khác nhau nên nhiều khi dấu được đặt không theo chuẩn. Ví dụ hai chữ: “*hỏa*” hay “*hoả*”.
- Cách viết hoa: Theo quy định, đầu câu và đầu tên riêng phải viết hoa. Tuy nhiên vẫn tồn tại một số cách viết như sau: “*Công ty Xi măng Bỉm sơn*”.
- Phiên âm tiếng nước ngoài: Các cách viết như ví dụ sau vẫn mặc định được chấp nhận trong văn bản tiếng Việt: “*Singapore*”/ “*Xinh-ga-po*”...
- Từ gạch nối: Do cách viết dấu gạch nối tùy tiện nên không thể phân biệt giữa nối tên riêng hay chú thích.

Những vấn đề vừa nêu trên thực sự gây ra nhiều trong dữ liệu đầu vào, đòi hỏi phải có một hệ thống tiền xử lý tốt, đảm bảo cho việc phân tích cú pháp được thực hiện có hiệu quả.

1.5. Công tác quản lý văn bản tại các cơ quan tỉnh Bắc Kạn

Văn bản quản lý nhà nước là một trong những phương tiện quan trọng và chủ yếu để tiến hành tổ chức mọi hoạt động của các cơ quan nhà nước nói chung. Đồng thời, đây là một nguồn tư liệu xác thực và có giá trị cần thiết cho

việc nghiên cứu ở trên tất cả các lĩnh vực như: Chính trị, kinh tế, văn hóa - xã hội, giáo dục, lịch sử, khoa học - công nghệ,...

Thực hiện tốt công tác quản lý văn bản và giải quyết văn bản tại các cơ quan nhà nước là một trong những yếu tố có tính chất quyết định đến hiệu quả công việc và có ảnh hưởng rất tích cực đến chất lượng hoạt động chung của cơ quan, đơn vị. Đặc biệt, trong công cuộc đổi mới hiện nay, khi Đảng và Nhà nước ta đã và đang thực hiện chủ trương từng bước cải cách nền hành chính quốc gia, thì công tác quản lý và giải quyết văn bản tại các cơ quan nhà nước có ý nghĩa và tầm quan trọng hơn bao giờ hết.

Thời gian qua, cùng với tất cả các địa phương trong cả nước, tỉnh Bắc Kạn đã có nhiều quan tâm trong việc đầu tư ứng dụng công nghệ thông tin phục vụ công tác quản lý văn bản và điều hành tại các cơ quan thuộc Tỉnh. Bắt đầu từ các module phần mềm riêng lẻ, triển khai phân tán, độc lập tại từng cơ quan những năm 2003-2005, đến năm 2011 đã triển khai trên diện rộng hệ thống phần mềm quản lý văn bản và hồ sơ công việc (Phần mềm) theo mô hình tập trung tại 28 cơ quan nhà nước cấp sở/huyện. Hệ thống phần mềm được quản lý, cài đặt tập trung tại Trung tâm tích hợp dữ liệu của tỉnh, các cơ quan truy cập từ xa qua mạng Internet để khai thác, sử dụng.

Từ đầu năm 2017 đến nay, hệ thống phần mềm đã được nâng cấp và đang được triển khai nhân rộng tại tất cả các cơ quan, đơn vị thuộc Tỉnh. Đến tháng 12 năm 2017, hệ thống phần mềm sẽ được triển khai tới tất cả các cơ quan khối Đảng, đoàn thể và các cơ quan chính quyền 3 cấp tỉnh, huyện, xã trên địa bàn toàn tỉnh, đảm bảo hoàn thành chỉ tiêu UBND tỉnh đề ra trong Kế hoạch hành động thực hiện Nghị quyết số 36a/NQ-CP ngày 14/10/2015 của Chính phủ về Chính phủ điện tử (phần đầu đến 01/01/2018: Đảm bảo trên 80% văn bản hành chính ở cấp tỉnh, 60% ở cấp huyện/thành phố được trao đổi dưới dạng điện tử).


Hệ thống phần mềm được triển khai đã tin học hóa toàn bộ các quy trình tiếp nhận, xử lý văn bản đến; dự thảo, phát hành văn bản đi tại nội bộ từng cơ quan; kết nối liên thông phục vụ việc trao đổi văn bản điện tử trên môi trường mạng giữa các đơn vị từ cấp tỉnh đến cấp xã; đồng thời hình thành hệ thống cơ sở dữ liệu văn bản chung của toàn tỉnh. Toàn bộ cơ sở dữ liệu văn bản (đi, đến) của các cơ quan, đơn vị được lưu trữ tập trung tại Trung tâm tích hợp dữ liệu của Tỉnh, với số lượng văn bản điện tử tăng lên nhanh chóng từng ngày. Tuy nhiên, hệ thống phần mềm hiện nay mới chỉ quản lý, phân loại các văn bản được số hóa theo góc độ quản lý nhà nước, với các cách phân loại văn bản như: Phân loại theo hiệu lực pháp lý, theo hình thức văn bản, thẩm quyền ban hành văn bản,...; chưa có chức năng phân loại theo chủ đề, lĩnh vực của văn bản. Vấn đề đặt ra hiện nay cần phải có một công cụ hỗ trợ việc phân loại các văn bản trong kho dữ liệu đồ sộ này một cách tự động theo từng lĩnh vực như kinh tế, chính trị, giáo dục, thể thao,... để phục vụ nhu cầu tra cứu của người sử dụng.

1.6. Kết luận chương 1

Nội dung của chương 1 đã trình bày một số kiến thức cơ bản về khai phá dữ liệu và khai phá dữ liệu văn bản nói riêng, thông qua đó ta có thể nắm bắt được các kiến thức nền tảng như: Khái niệm thế nào là khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản; làm thế nào để giải quyết được bài toán phân loại văn bản; nắm được các bước cơ bản của bài toán phân loại văn bản. Nội dung của chương cũng đã phân tích, làm rõ các đặc trưng của văn bản tiếng Việt và giới thiệu sơ bộ về công tác quản lý văn bản tại các cơ quan thuộc tỉnh Bắc Kạn. Đây chính là những kiến thức cơ sở để định hướng, xác định mục tiêu cụ thể của đề tài và lựa chọn, tìm hiểu các thuật toán về phân loại văn bản tiếng Việt và các kỹ thuật liên quan trong các chương tiếp theo.


CHƯƠNG II. CÁC KỸ THUẬT TRONG PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

2.1. Tách từ trong văn bản

 Khó khăn trong tách từ tiếng Việt:

Như đã phân tích ở phần trên, tiếng Việt là ngôn ngữ đơn lập [2]. Khác với các ngôn ngữ Châu Âu, mỗi từ là một nhóm các ký tự có nghĩa được cách nhau bởi một khoảng trắng, việc xác định từ chỉ đơn giản dựa vào khoảng trắng để tách từ. Ví dụ, câu “*I am a doctor*” sẽ được tách thành 4 từ: *I, am, a, doctor*. Với tiếng Việt, nếu dựa vào khoảng trắng để tách ta chỉ thu được các tiếng. Từ có thể được ghép từ một hay nhiều tiếng, phải có ý nghĩa hoàn chỉnh và có cấu tạo ổn định. Câu “*Tôi là một bác sỹ*” được tách thành 4 từ: *Tôi, là, một, bác sỹ*; trong đó, từ “*bác sỹ*” được hình thành từ hai tiếng “*bác*” và “*sỹ*”.

Hiện nay có rất nhiều phương pháp được sử dụng để tách từ tiếng Việt. Tuy nhiên, do sự phức tạp của ngữ pháp tiếng Việt nên chưa có phương pháp nào đạt được chính xác 100%. Và việc lựa chọn phương pháp nào là tốt nhất cũng đang là vấn đề tranh cãi.

 Các khó khăn khác liên quan đến từ trong tiếng Việt:

Tiếng Việt có các từ đồng nghĩa nhưng khác âm. Các công cụ tìm kiếm hiện nay còn nhiều hạn chế trong việc hỗ trợ xác định các từ đồng nghĩa. Vì vậy, kết quả trả về sẽ không đầy đủ.


Ngược lại, có những từ đồng âm khác nghĩa. Các hệ thống tìm kiếm trả về các văn bản có chứa các từ được tách trong câu truy vấn mà không xác định chúng có thực sự liên quan không. Vì vậy, kết quả không chính xác.


Một số từ xuất hiện rất nhiều nhưng không có ý nghĩa trong văn bản. Các từ như: “*Và/với/...*” có tần số xuất hiện rất lớn trong các văn bản. Nếu tìm cách trả về các văn bản có chứa những từ này thì kết quả sẽ vô nghĩa.

Có nhiều phương pháp để tách từ trong tiếng Việt. Trong khuôn khổ nội dung luận văn này, học viên sẽ trình bày một số phương pháp tách từ phổ biến đang được sử dụng hiện nay.

2.1.1. Phương pháp khớp tối đa

Tư tưởng của phương pháp khớp tối đa (Maximum Matching) là duyệt một câu từ trái qua phải và chọn từ có nhiều tiếng nhất mà có mặt trong từ điển tiếng Việt [1]. Thuật toán có 2 dạng sau:

 *Dạng đơn giản*: Giả sử có một chuỗi các tiếng trong câu là t_1, t_2, \dots, t_N . Thuật toán kiểm tra xem t_1 có mặt trong từ điển hay không, sau đó kiểm tra tiếp t_1-t_2 có trong từ điển hay không. Tiếp tục như vậy cho đến khi tìm được từ có nhiều tiếng nhất có mặt trong từ điển và đánh dấu từ đó. Sau đó tiếp tục quá trình trên với tất cả các tiếng còn lại trong câu và trong toàn bộ văn bản. Dạng này khá đơn giản, nhưng nó gặp phải rất nhiều nhập nhằng trong tiếng Việt. Ví dụ, nó bị gặp phải lỗi khi phân đoạn từ câu sau: “*học sinh | học sinh | học*”, câu đúng phải là “*học sinh | học | sinh học*”.

 *Dạng phức tạp*: Dạng này có thể tránh được một số nhập nhằng gặp phải trong dạng đơn giản. Đầu tiên thuật toán kiểm tra xem t_1 có mặt trong từ điển không, sau đó kiểm tra tiếp t_1-t_2 có mặt trong từ điển không. Nếu t_1-t_2 đều có mặt trong từ điển, thì thuật toán thực hiện chiến thuật *chọn 3-từ tốt nhất*, cụ thể như sau:

- Độ dài trung bình của 3 từ là lớn nhất. Ví dụ, chuỗi “*cơ quan tài chính*” được phân đoạn đúng thành “*cơ quan | tài chính*”, tránh được việc phân đoạn sai thành “*cơ | quan tài | chính*” vì cách phân đúng phải có độ dài trung bình lớn nhất.

- Sự chênh lệch độ dài của 3 từ là ít nhất. Ví dụ, chuỗi “*công nghiệp hoá chất phát triển*” được phân đoạn đúng thành “*công nghiệp | hoá chất | phát triển*”, thay vì phân đoạn sai thành “*công nghiệp hoá | chất | phát triển*”. Cả 2 cách phân đoạn này đều có độ dài trung bình bằng nhau, nhưng cách phân đoạn đúng có sự chênh lệch độ dài 3 từ ít hơn.

Phương pháp này thực hiện tách từ đơn giản, nhanh và chỉ cần dựa vào từ điển để thực hiện. Tuy nhiên, hạn chế của phương pháp này cũng chính là từ điển, bởi độ chính xác khi thực hiện tách từ phụ thuộc hoàn toàn vào tính đủ, tính chính xác của từ điển.

2.1.2. Mô hình tách từ bằng WFST và mạng Neural

Phương pháp WFST (Weighted Finite - State Transducer) còn gọi là phương pháp chuyển dịch trạng thái hữu hạn có trọng số. Ý tưởng chính của phương pháp này áp dụng cho phân đoạn từ tiếng Việt là các từ được gán trọng số bằng xác suất xuất hiện của từ đó trong dữ liệu. Sau đó duyệt qua các câu, cách duyệt có trọng số lớn nhất được chọn là cách dùng để phân đoạn từ [1]. Phương pháp WFST đã được áp dụng trong công trình [8] đã được công bố của tác giả Đinh Điền năm 2001. Trong đó, tác giả đã sử dụng WFST kèm với mạng Neural để xây dựng hệ thống tách từ gồm hai tầng: Tầng WFST để tách từ; tầng mạng Neural dùng để khử nhập nhằng về ngữ nghĩa (nếu có).

Tầng WFST: Gồm có ba bước [1]

Bước 1: Xây dựng từ điển trọng số.

Từ điển trọng số D được xây dựng như là một đồ thị biến đổi trạng thái hữu hạn có trọng số. Giả sử:

- + H là tập các tiếng trong tiếng Việt (hay còn gọi là các từ chính tả).
- + P là tập các loại từ trong tiếng Việt.
- + Mỗi cung của D có thể là:
 - Từ một phần tử của H tới một phần tử của H ;
 - Từ phần tử ε (xâu rỗng) đến một phần tử của P .

Mỗi từ trong D được biểu diễn bởi một chuỗi các cung bắt đầu bởi một cung tương ứng với một phần tử của H , kết thúc bởi một cung có trọng số tương ứng với một phần tử của $\varepsilon \times P$. Trọng số biểu diễn một chi phí ước lượng (estimated cost) cho bởi công thức:

$$C = -\log\left(\frac{f}{N}\right)$$

Trong đó, f là tần số xuất hiện của từ; N là kích thước tập mẫu.

Đối với các trường hợp từ mới chưa gặp, tác giả áp dụng xác suất có điều kiện Goog-Turning (Baayen) để tính toán trọng số.

Bước 2: Xây dựng các khả năng tách từ.

Bước này thống kê tất cả các khả năng phân đoạn của một câu. Giả sử câu có n tiếng, thì có tới 2^{n-1} cách phân đoạn khác nhau. Để giảm sự bùng nổ các cách phân đoạn, thuật toán loại bỏ ngay những nhánh phân đoạn mà chứa từ không xuất hiện trong từ điển.

Bước 3: Lựa chọn khả năng tách tối ưu.

Sau khi liệt kê tất cả các khả năng phân đoạn từ, thuật toán chọn cách tách từ tốt nhất, đó là cách tách từ có trọng số bé nhất.

 **Tầng mạng Neural:**

Tầng này được sử dụng để khử nhập nhằng khi tách từ bằng cách kết hợp so sánh với từ điển.

Phương pháp này có độ chính xác khá cao (>98% đối với tách từ trong lĩnh vực khoa học - kỹ thuật; >94% đối với tiểu thuyết văn học [8]), bằng việc kết hợp mạng Neural với từ điển để khử các nhập nhằng có thể có khi tách ra nhiều từ từ một câu. Khi đó tầng mạng Neural sẽ loại bỏ đi các từ không phù hợp bằng cách kết hợp với từ điển. Tuy nhiên, việc xây dựng tập ngữ liệu học đầy đủ đáp ứng yêu cầu là rất công phu, tốn kém về thời gian và công sức.

2.1.3. Phương pháp học dựa vào sự biến đổi trạng thái

Học trên sự biến đổi trạng thái (TBL - Transformation-Based Learning) là một phương pháp học “hướng lỗi” (error-driven) dựa trên tập luật đã được sắp xếp. TBL được Eric Brill (1995) phát triển cho bài toán gán nhãn từ loại (Part-Of-Speech tagging) [9].

Mô hình học TBL bao gồm 3 thành phần quan trọng: Ngữ liệu đã gán nhãn, heuristic cơ sở để dự đoán giá trị khởi đầu cho các thể hiện, và một tập các khung luật được sử dụng để quyết định không gian cho các luật chuyển đổi. Các bước để tạo mô hình học cụ thể như sau:

- + Gỡ nhãn của ngữ liệu huấn luyện.
- + Áp dụng heuristic cơ sở để tạo các giả thuyết ban đầu cho ngữ liệu vừa được gỡ nhãn.
- + Phát sinh các luật có thể sửa ít nhất một lỗi dựa trên khung luật. Các luật này sau đó được kiểm tra dựa trên điểm của chúng khi áp dụng cho tập huấn luyện. Luật có điểm cao nhất (điểm được tính bằng hiệu số giữa thay đổi đúng và thay đổi sai) sẽ được chọn. Luật được chọn sẽ đem áp dụng lại cho ngữ liệu học.
- + Toàn bộ quá trình học như trên sẽ được lặp lại trên ngữ liệu sau khi chuyển đổi (ngữ liệu được áp dụng luật được lựa chọn ở bước trước). Quá trình này sẽ dừng nếu điểm của luật nhỏ hơn một ngưỡng T nào đó. Kết quả là chúng ta có một dãy các luật chuyển đổi.

Sau khi có được dãy các luật chuyển đổi, chúng ta có thể áp dụng dãy luật này cho một văn bản mới bằng cách áp dụng heuristic cơ bản, rồi lần lượt áp dụng từng luật trong dãy luật được rút trích trong quá trình học để xác định các tham số (các xác suất) cần thiết cho mô hình nhận diện từ.

Đặc điểm của phương pháp này là khả năng tự rút ra quy luật của ngôn ngữ. Nó có những ưu điểm của cách tiếp cận dựa trên luật, nhưng khắc phục được hạn chế của việc xây dựng các luật một cách thủ công bởi các chuyên gia. Các luật được thử nghiệm tại chỗ để đánh giá độ chính xác và hiệu quả của luật (dựa trên ngữ liệu huấn luyện).

Sử dụng TBL có khả năng khử được một số nhập nhằng như: “*The singer sang a lot of a??as*” thì hệ thống có thể xác định được “*a??as*” là “*arias*” (dân

ca) thay vì “*areas*” (*khu vực*). Tuy nhiên, để xây dựng được tập ngữ liệu đầy đủ, chính xác phục vụ cho máy “học” đòi hỏi tốn kém rất nhiều thời gian và công sức, cài đặt phức tạp và máy tính phải trải qua một thời gian huấn luyện khá lâu để có thể rút ra được các luật.

2.1.4. Loại bỏ từ dừng

Như đã nêu ở trên, trong văn bản thường có một số từ xuất hiện rất nhiều nhưng không có ý nghĩa, không có ích trong việc phân biệt nội dung của các văn bản, đó là các *từ dừng* (stop-words). Cần phải thực hiện lọc và loại bỏ chúng trong các hệ thống xử lý, phân loại hay tìm kiếm văn bản.

Ví dụ: “và”, “hoặc”, “với”, “nhưng”, “cũng”, “là”, “mỗi”, “bởi”...

2.2. Trọng số của từ trong văn bản

Dữ liệu văn bản là một trong những dạng dữ liệu truyền thống và quan trọng nhất được con người sử dụng để lưu trữ thông tin. Một trong những vấn đề khó khăn nhất của máy tính là làm thế nào để biểu diễn văn bản phản ánh đúng nội dung. Công việc này còn gọi là đánh chỉ số văn bản. Trước đây, quá trình này được làm thủ công với sự giúp đỡ của người dùng hoặc các chuyên gia trong một số lĩnh vực chuyên ngành. Tuy nhiên, với số lượng lớn các văn bản ngày càng tăng, thì việc đánh chỉ số thủ công là không khả thi, do vậy, việc đánh chỉ số một cách tự động là cần thiết [1]. Trong đó, việc lựa chọn các từ để đánh chỉ số, hay còn gọi là lựa chọn đặc trưng, là một công việc rất quan trọng nhưng lại không dễ dàng. Đây cũng là một chủ đề nghiên cứu trong học máy và trong khai phá dữ liệu nói chung.

Giả sử từ một văn bản d thuộc miền ứng dụng, sử dụng các phương pháp lựa chọn từ ta nhận được một tập từ vựng T , được dùng để biểu diễn văn bản d . Đặc trưng cho độ quan trọng của từ thuộc tập T trong một văn bản bất kỳ là một giá trị số được gán cho từ đó trong văn bản d đã cho. Công việc tính trọng

số của từ còn được gọi là đánh trọng số các từ trong văn bản. Bài toán đánh trọng số được phát biểu như sau:

Input: Cho một từ $t_i \in T$ và một văn bản d_j thuộc miền ứng dụng.

Output: Giá trị w_{ij} là trọng số (độ quan trọng) của từ t_i trong văn bản d_j .

Ta xét một số phương pháp đánh trọng số từ điển hình như sau [1]:


2.2.1. Phương pháp Boolean

Giả sử có một tập gồm m văn bản $D = \{d_1, d_2, \dots, d_m\}$. Tập từ vựng T gồm n từ khoá $T = \{t_1, t_2, \dots, t_n\}$. Gọi $W = (w_{ij})$ là ma trận trọng số, với w_{ij} là trọng số của từ khoá t_i trong văn bản d_j . Phương pháp Boolean là phương pháp đánh trọng số đơn giản nhất, giá trị trọng số w_{ij} được tính như sau:

$$w_{ij} = \begin{cases} 1 & t_i \in d_j \\ 0 & t_i \notin d_j \end{cases}$$

2.2.2. Phương pháp dựa trên tần số

Phương pháp dựa trên tần số xác định giá trị các số trong ma trận $W=(w_{ij})$, dựa vào tần số xuất hiện của các từ khoá trong văn bản và tần số xuất hiện của văn bản trong tập D gồm m văn bản đang được xem xét. Dưới đây là hai phương pháp đánh trọng số dựa trên tần số phổ biến.

 *Phương pháp dựa trên tần số từ khoá (TF - Term Frequency):*

Giá trị của một từ khoá được tính dựa trên số lần xuất hiện của từ khoá trong văn bản. Gọi tf_{ij} là số lần xuất hiện của từ khoá t_i trong văn bản d_j , khi đó có thể chọn cách tính w_{ij} theo một trong ba công thức dưới đây:

$$w_{ij} = \sqrt{tf_{ij}} \quad \text{hoặc} \quad w_{ij} = 1 + \log(tf_{ij}) \quad \text{hoặc} \quad w_{ij} = tf_{ij}$$

Phương pháp này được lý giải từ lập luận rằng, trong một văn bản thì một từ xuất hiện nhiều thường quan trọng hơn một từ xuất hiện ít.

 *Phương pháp TFIDF:*

Giá trị của ma trận trọng số theo phương pháp này được tính như sau:

$$w_{ij} = \begin{cases} \left[1 + \log(tf_{ij})\right] \log\left(\frac{m}{df_i}\right) & \text{nếu } tf_{ij} \geq 1 \\ 0 & \text{nếu } tf_{ij} = 0 \end{cases} \quad (2.1)$$

Với df_i là số lượng văn bản trong tập văn bản m mà trong đó từ t_i có xuất hiện. Khi đó, n từ t_i có giá trị trọng số lớn nhất sẽ được chọn làm n đặc trưng của văn bản.

2.3. Các mô hình biểu diễn văn bản

Bước đầu tiên của mọi phương pháp phân loại văn bản là chuyển việc mô tả văn bản dùng chuỗi kí tự thành một dạng mô tả khác, phù hợp với các thuật toán học theo mẫu và phân lớp. Bởi vậy, ở phần này luận văn tập trung tìm hiểu một số phương pháp biểu diễn văn bản phổ biến hiện nay [1].

2.3.1. Mô hình Boolean

Giả sử có một tập gồm m văn bản $D = \{d_1, d_2, \dots, d_m\}$. Mỗi văn bản gồm n từ khoá $T = \{t_1, t_2, \dots, t_n\}$. Gọi $W = (w_{ij})$ là ma trận trọng số, trong đó w_{ij} là trọng số của từ khoá t_i trong văn bản d_j .

Mô hình Boolean là mô hình đơn giản nhất, với trọng số các từ trong văn bản là 0 hoặc 1. Mỗi văn bản được biểu diễn dưới dạng tập hợp như sau:

$d_i = \{t_{ij}\}$, trong đó t_{ij} là từ t_i có trọng số w_{ij} trong văn bản d_j là 1.

Ví dụ: Giả sử có một văn bản đơn giản d gồm các từ:

$d = \text{"Hello world ! Hello Vietnam !"}$

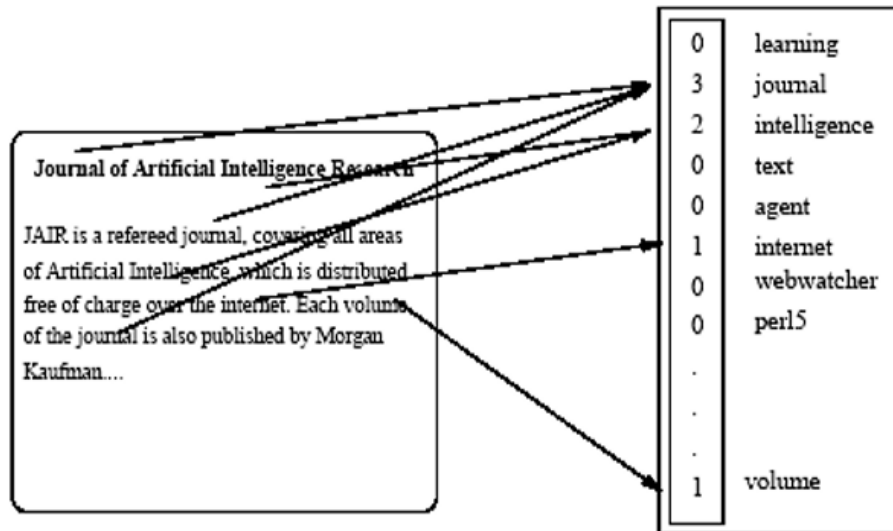
Khi đó văn bản d được biểu diễn như sau:

$d = \{\text{"Hello"}, \text{"world"}, \text{"Vietnam"}, \text{"!"}\}.$

2.3.2. Mô hình xác suất

Mô hình xác suất là mô hình toán học làm việc với các biến ngẫu nhiên và phân bố xác suất của nó. Theo thuật ngữ toán học, một mô hình xác suất có thể được coi như một cặp (Y, P) , trong đó Y là tập các quan sát (biến ngẫu

nhien) và P là tập các phân bố xác suất trên Y . Khi đó, sử dụng suy diễn xác suất sẽ cho ta kết luận về các phần tử của tập Y . Các phương pháp suy diễn có thể là các phương pháp hồi quy hoặc suy diễn Bayes.



Hình 2.1. Biểu diễn văn bản theo mô hình xác suất [1]

Văn bản trong mô hình xác suất được coi như một quan sát trong tập Y , trong đó các từ trong văn bản được giả thiết là độc lập, không phụ thuộc vào vị trí cũng như ngữ pháp trong văn bản. Khi đó văn bản sẽ gồm các từ mà nó chứa trong đó, chính vì vậy mà phương pháp này được gọi là biểu diễn *túi - các - từ* (bag - of - word). Để đơn giản, người ta còn gọi là *mô hình biểu diễn theo túi - các - từ*. Mô hình này được sử dụng nhiều trong phân lớp văn bản khi áp dụng suy diễn Bayes trong bài toán phân lớp.

Ví dụ, văn bản d đưa tin về Tạp chí Nghiên cứu về trí tuệ nhân tạo được biểu diễn theo mô hình *túi - các - từ* một cách đơn giản theo bảng với trọng số là tần số từ có trong văn bản (Hình 2.1).

2.3.3. Mô hình không gian vector

Mô hình không gian vector là một trong những mô hình toán học được sử dụng rộng rãi nhất trong biểu diễn văn bản bởi tính chất dễ hiểu của nó. Mô hình này được đề xuất bởi Salton và cộng sự năm 1975 khi giải quyết bài toán truy vấn thông tin. Theo cách biểu diễn này, mỗi văn bản được biểu diễn trong

một không gian nhiều chiều, trong đó mỗi chiều tương ứng với một từ trong văn bản. Một từ với độ quan trọng của nó được xác định bằng một phương pháp đánh chỉ số trong văn bản và giá trị trọng số được chuẩn hoá trong đoạn $[0, 1]$ [1].

Tổng quát hơn, một văn bản d trong không gian vector, ký hiệu là \mathbf{v}_d sẽ được biểu diễn như sau trong một không gian vector gồm N chiều, trong đó N là số lượng từ có trong tập văn bản: $\mathbf{v}_d = [w_{1,d}, w_{2,d}, \dots, w_{N,d}]^T$. Khi đó độ giống nhau giữa hai văn bản sẽ được tính bằng độ đo cosine giữa hai vector:

$$\cos \theta = \frac{(v_1 * v_2)}{\|v_1\| * \|v_2\|} \quad (2.2)$$

Mô hình không gian vector là mô hình toán học hết sức quan trọng trong biểu diễn văn bản, đặc biệt là trong lĩnh vực truy vấn thông tin. Với mô hình này, các văn bản được sắp xếp theo mức độ liên quan đến nội dung yêu cầu. Việc lưu trữ dữ liệu và tìm kiếm đơn giản hơn khi sử dụng mô hình logic.

- ***Các đặc trưng của văn bản khi biểu diễn dưới dạng vector***

- Không gian đặc trưng thường lớn. Các văn bản càng dài, lượng thông tin trong nó đề cập đến nhiều vấn đề thì không gian đặc trưng càng lớn.
- Các đặc trưng độc lập nhau. Sự kết hợp các đặc trưng này thường không có ý nghĩa trong phân lớp.
- Các đặc trưng rời rạc: Vector đặc trưng d_i có thể có nhiều thành phần mang giá trị 0 do có nhiều đặc trưng không xuất hiện trong văn bản d_i (nếu 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 vector). 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 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 phần nào giảm bớt sự rời rạc trong vector văn bản.

2.4. Độ tương đồng văn bản

Độ tương đồng là một đại lượng dùng để so sánh hai hay nhiều đối tượng với nhau, phản ánh cường độ của mối quan hệ giữa các đối tượng với nhau. Ví dụ: Xét 2 câu “*Tôi là nam*” và “*Tôi là nữ*”, ta có thể nhận thấy hai câu trên có sự tương đồng khá cao [3].

Phát biểu bài toán: Xét 2 văn bản d_i và d_j . Mục tiêu của bài toán là tìm ra một giá trị của hàm $S(d_i, d_j)$, với $S \in (0,1)$, thể hiện độ tương đồng giữa hai văn bản d_i và d_j . Hàm $S(d_i, d_j)$ được gọi là độ đo sự tương đồng giữa 2 văn bản d_i và d_j . Giá trị của hàm $S(d_i, d_j)$ càng cao thì sự giống nhau về nghĩa của hai văn bản càng nhiều.

Ví dụ, trong mô hình không gian vector, sử dụng độ đo Cosine để tính độ tương đồng giữa hai văn bản, mỗi văn bản được biểu diễn bởi một vector.

Độ tương đồng ngữ nghĩa là một giá trị tin cậy phản ánh mối quan hệ ngữ nghĩa giữa các câu, các văn bản. Đó là một giá trị tỷ lệ dựa trên sự giống nhau về nội dung ý nghĩa của các tập văn bản hoặc các câu, các thuật ngữ trong một danh sách các thuật ngữ. Thực tế, khó có được một giá trị có độ chính xác cao bởi ngữ nghĩa chỉ được hiểu đầy đủ trong một ngữ cảnh cụ thể.

Bài toán độ tương đồng ngữ nghĩa được sử dụng phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và có nhiều kết quả khả quan. Một số ứng dụng quan trọng của bài toán này trong thực tế đó là: Tìm kiếm thông tin; phân lớp văn bản; tóm tắt văn bản; đánh giá tính chặt chẽ của văn bản,...

Trong phạm vi đề tài này, luận văn tập trung tìm hiểu một số phương pháp tính độ tương đồng văn bản dựa trên vector biểu diễn, đó là: Độ đo Cosine; độ đo khoảng cách Euclide; độ đo khoảng cách Manhattan.

a. Tính độ tương đồng sử dụng độ đo Cosine [3]

Trong phương pháp này, các văn bản được biểu diễn theo mô hình không gian vector. Mỗi thành phần trong vector chỉ đến một từ tương ứng trong danh sách mục từ chính. Danh sách mục từ chính thu được từ quá trình tiền xử lý văn bản đầu vào, với các bước tiền xử lý gồm: Tách câu, tách từ, gán nhãn từ loại, loại bỏ những câu không hợp lệ (không phải là câu thực sự) và biểu diễn câu trên không gian vector.

Không gian vector (hay số chiều của vector) có kích thước bằng số mục từ trong danh sách mục từ chính. Giá trị mỗi phần tử là độ quan trọng của mục từ trong câu. Độ quan trọng của từ được tính theo công thức sau:

$$w_{ij} = \frac{tf_{ij}}{\sqrt{\sum_j tf_{ij}^2}} \quad \text{Với } tf_{ij} \text{ là tần số xuất hiện của mục từ } i \text{ trong câu } j.$$

Với không gian biểu diễn tài liệu được chọn là không gian vector và trọng số TF. Giả sử vector biểu diễn cho hai văn bản lần lượt có dạng:

$D_i = \langle w_1^i, \dots, w_t^i \rangle$, với w_t^i là trọng số của từ thứ t trong không gian i .

$D_j = \langle w_1^j, \dots, w_t^j \rangle$, với w_t^j là trọng số của từ thứ t trong không gian j .

Độ đo tương đồng được tính là Cosine của góc giữa 2 vector biểu diễn cho hai văn bản là D_i và D_j . Độ tương tự giữa chúng được tính theo công thức:

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^t w_k^i w_k^j}{\sqrt{\sum_{k=1}^t (w_k^i)^2 * \sum_{k=1}^t (w_k^j)^2}} \quad (2.3)$$

b. Tính độ tương đồng dựa vào độ đo khoảng cách Euclide [3]

Sử dụng khoảng cách Euclide là một phương pháp phổ biến để xác định mức độ tương đồng giữa các vector đặc trưng của hai văn bản.

Cho hai vector \vec{v}_a và \vec{v}_b là các vector đặc trưng của hai văn bản trong không gian Euclide n chiều: $\vec{v}_a = (w_{a1}, w_{a2}, \dots, w_{an})$; $\vec{v}_b = (w_{b1}, w_{b2}, \dots, w_{bn})$. Khoảng cách Euclide được định nghĩa như sau:

$$euc_dist(\vec{v}_a, \vec{v}_b) = \sqrt{\sum_{i=1}^n (w_{ai} - w_{bi})^2}$$

$$\frac{euc_dist(\vec{v}_a, \vec{v}_b)}{n} \text{ nằm trong khoảng } 0 \text{ và } 1.$$

Mức độ tương đồng giữa hai vector này được xác định bằng công thức:

$$euc_sim(\vec{v}_a, \vec{v}_b) = 1 - \frac{euc_dist(\vec{v}_a, \vec{v}_b)}{n} = 1 - \frac{1}{n} \sqrt{\sum_{i=1}^n (w_{ai} - w_{bi})^2} \quad (2.4)$$

c. Tính độ tương đồng dựa vào độ đo khoảng cách Manhattan [3]

Khoảng cách Manhattan là một phương pháp khác dùng để xác định mức độ tương đồng giữa các vector đặc trưng của hai văn bản.


Cho hai vector \vec{v}_a và \vec{v}_b : $\vec{v}_a = (w_{a1}, w_{a2}, \dots, w_{an})$; $\vec{v}_b = (w_{b1}, w_{b2}, \dots, w_{bn})$. Khoảng cách Manhattan được định nghĩa như sau:

$$man_dist(\vec{v}_a, \vec{v}_b) = \sum_{i=1}^n |w_{ai} - w_{bi}|$$

$$\frac{man_dist(\vec{v}_a, \vec{v}_b)}{n} \text{ nằm trong khoảng } 0 \text{ và } 1.$$

Mức độ tương đồng giữa hai vector này được xác định bằng công thức:

$$man_sim(\vec{v}_a, \vec{v}_b) = 1 - \frac{man_dist(\vec{v}_a, \vec{v}_b)}{n} = 1 - \frac{1}{n} \sum_{i=1}^n |w_{ai} - w_{bi}| \quad (2.5)$$

 **Nhận xét:** Các phương pháp nêu trên cho kết quả tốt như nhau trong việc xác định mức độ tương đồng giữa các vector, nên tùy vào mục tiêu mà chọn phương pháp nào là phù hợp.

2.5. Thuật toán phân loại văn bản

2.5.1. Thuật toán Support Vector Machine (SVM)

Thuật toán máy vector hỗ trợ (Support Vector Machine – SVM) dựa trên phương pháp tiếp cận thống kê được Vapnik đề xuất năm 1995. Thuật toán SVM rất hiệu quả để giải quyết các bài toán với dữ liệu có số chiều lớn, như các vector biểu diễn văn bản, và được coi là một trong những thuật toán khai phá dữ liệu điển hình nhất [1],[12],[13].

Xét bài toán phân loại đơn giản nhất - phân loại hai phân lớp văn bản với một tập huấn luyện cho trước được biểu diễn trong không gian vector:

$$\{(x_i, y_i) | x_i \in \mathbb{R}^m; i = 1, 2, \dots, n\}$$

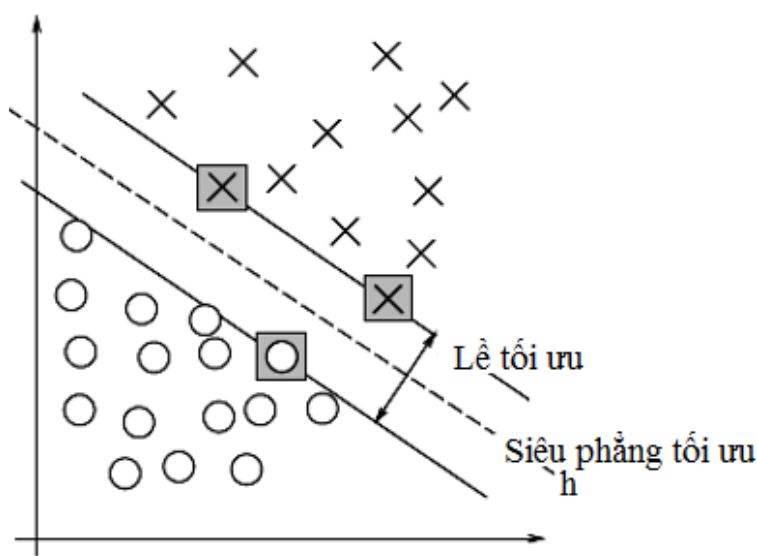
Trong đó, mẫu là các vector đối tượng được phân lớp thành các mẫu dương và mẫu âm:

- Các mẫu dương (+) là các mẫu x_i thuộc lĩnh vực quan tâm và được gán nhãn $y_i = 1$.
- Các mẫu âm (-) là các mẫu x_i không thuộc lĩnh vực quan tâm và được gán nhãn $y_i = -1$.

Ý tưởng của thuật toán là tìm ra một siêu mặt phẳng h quyết định tốt nhất để chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng, tạm gọi là lớp dương (+) và lớp âm (-). Chất lượng của siêu mặt phẳng này được quyết định bởi một khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng. Biên càng lớn thì kết quả phân chia các điểm thành hai lớp càng tốt, nghĩa là sẽ đạt được kết quả phân loại tốt. Tập các vector hỗ trợ hình thành siêu phẳng tốt nhất được gọi là các vector hỗ trợ.

Thực chất phương pháp này là một bài toán tối ưu, mục tiêu là tìm ra một không gian H và siêu mặt phẳng quyết định h trên H sao cho sai số phân lớp là thấp nhất, kết quả phân loại sẽ là tốt nhất.

Trong trường hợp này, tập phân lớp SVM là *mặt siêu phẳng phân tách* các mẫu dương khỏi các mẫu âm với *độ chênh lệch cực đại*, trong đó *độ chênh lệch* – còn gọi là **Lề** (margin) xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất. Mặt siêu phẳng này được gọi là *mặt siêu phẳng lề tối ưu*. Hình 2.2 mô tả một minh họa hình học cho thuật toán SVM bao gồm cả chỉ dẫn các vector hỗ trợ.



Các điểm mang dấu cộng hoặc hình tròn biểu diễn các mẫu (dương hoặc âm). Các đường thẳng biểu diễn các siêu phẳng quyết định, trong đó đường rời nét (nằm giữa hai đường song song) là siêu phẳng tốt nhất vì khoảng cách của nó tới các tập huấn luyện là nhỏ nhất. Các hộp vuông nhỏ chứa các dấu cộng và hình tròn biểu diễn các vector hỗ trợ.

Hình 2.2. Minh họa hình học thuật toán SVM [1]

Trong không gian đối tượng, mỗi siêu phẳng đều có thể được viết dưới dạng một tập hợp các điểm thỏa mãn: $w x + b = 0$.

Với w là một vector pháp tuyến của siêu phẳng hay còn gọi vector trọng số, b là độ dịch. Khi thay đổi w và b thì hướng và khoảng cách từ gốc tọa độ đến mặt siêu phẳng thay đổi.

Bộ phân lớp SVM được xác định như sau:

Bộ phân lớp SVM phụ thuộc vào vector trọng số w và độ dịch b . Mục tiêu của phương pháp SVM là ước lượng w và b sao cho cực đại hóa lề giữa các lớp dữ liệu dương và âm.

Vậy ta cần chọn w và b để cực đại hóa lề sao cho khoảng cách giữa hai siêu phẳng song song ở xa nhau nhất có thể trong khi vẫn phân chia được dữ liệu. Các siêu phẳng được xác định bằng các công thức sau:

$$wx + b = 1 \text{ và } wx + b = -1 \quad (2.6)$$

Nếu dữ liệu huấn luyện có thể được chia tách một cách tuyến tính, ta có thể chọn hai siêu phẳng của lề sao cho không có điểm nào nằm giữa chúng, sau đó tăng khoảng cách giữa chúng đến tối đa có thể. Bằng phương pháp hình học ta tính được khoảng cách giữa hai siêu phẳng là $\frac{2}{\|w\|}$. Do vậy, để cực đại hóa khoảng cách giữa hai siêu phẳng lề ta phải cực tiểu hóa giá trị $\|w\|$.

Để không có điểm dữ liệu nào nằm trong lề, ta có các điều kiện sau:

$w x_i + b \geq 1$ đối với x_i thuộc lớp thứ nhất (lớp dương).

$w x_i + b \leq -1$ đối với x_i thuộc lớp thứ hai (lớp âm).

Ta có thể viết gọn lại như sau: $y_i (w x_i + b) \geq 1$ với $1 \leq i \leq n$.

Vậy ta có bài toán tối ưu hóa sau:

$$\begin{cases} \text{Cực tiểu hóa } \|w\| \text{ theo } w \text{ và } b \\ y_i (w x_i + b) \geq 1, i = 1, \dots, n. \end{cases}$$

Thay $\|w\|$ bằng hàm mục tiêu $\frac{1}{2} \|w\|^2$, ta có bài toán quy hoạch toàn phương với lời giải có cùng kết quả w và b với bài toán tối ưu ban đầu:

$$\begin{cases} \text{Cực tiểu hóa } \frac{1}{2} \|w\|^2 \text{ theo } w \text{ và } b \\ y_i (w x_i + b) \geq 1, \quad i = 1, \dots, n. \end{cases}$$

Tiến hành thêm các nhân tử Lagrange α , ta sẽ nhận được bài toán sau:

$$\min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i (w x_i + b) - 1] \right\} \quad (2.7)$$

Giải bài toán này bằng các kỹ thuật giải bài toán quy hoạch toàn phương. Theo điều kiện Karush-Kuhn-Tucker, lời giải của bài toán có thể được viết dưới dạng tổ hợp tuyến tính của các vector huấn luyện:

$$w = \sum_{i=1}^n \alpha_i y_i x_i$$

Xác định b như sau:

Từ phương trình: $y_i(wx_i + b) = 1 \Leftrightarrow wx_i + b = 1/y_i \Leftrightarrow b = wx_i - y_i$

Trong thực tế, người ta hay sử dụng cách tính giá trị trung bình từ tất cả vector huấn luyện để tính b:

$$b = \frac{1}{N_{TB}} \sum_{i=1}^{N_{TB}} (wx_i + y_i)$$

Sau khi tìm được vector trọng số w và độ dịch b, ta xây dựng được phương trình tổng quát của siêu phẳng tìm được bởi thuật toán SVM như sau:

$$f(x) = wx + b$$

Với: $f(x) \geq 0$ thì x thuộc lớp thứ nhất; $f(x) < 0$ thì x thuộc lớp thứ hai.

Sau khi đã tìm được phương trình của siêu phẳng bằng thuật toán SVM, ta áp dụng công thức này để tìm ra nhãn lớp cho các dữ liệu mới.

Theo phương diện toán học, bài toán được phát biểu như sau:

Đầu vào: Cho tập huấn luyện: $D = \{(x_i, y_i) | x_i \in \mathbb{R}^m; i = 1, 2, \dots, n\}$

với $y_i \in \{-1, 1\}$ xác định dữ liệu (mẫu) dương hay âm.

Đầu ra: Phương trình của siêu phẳng phân chia tập huấn luyện thành hai miền rời nhau với khoảng cách của siêu phẳng tới tập huấn luyện D là lớn nhất (siêu phẳng tối ưu).

Thực hiện thuật toán [4],[13]:

$$\text{Giải bài toán tối ưu: } \min_{w,b} \max_{a \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(wx_i + b) - 1] \right\}$$

$$\text{Tính: } w = \sum_{i=1}^n \alpha_i y_i x_i \quad (2.8)$$

$$\text{Tính: } b = wx_i - y_i \text{ hoặc } b = \frac{1}{N_{TB}} \sum_{i=1}^{N_{TB}} (wx_i + y_i) \quad (2.9)$$

$$\text{Xây dựng hàm: } f(x) = wx + b \quad (2.10)$$

là phương trình của siêu phẳng cần tìm.

2.5.2. Thuật toán K-Nearest Neighbor (kNN)

K-Nearest Neighbor (kNN) là thuật toán truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trong nhiều năm qua, được đánh giá là một trong những phương pháp tốt nhất được sử dụng từ những thời kỳ đầu trong nghiên cứu về phân loại văn bản [5].

Ý tưởng: Khi cần phân loại một văn bản mới, thuật toán sẽ xác định khoảng cách (có thể áp dụng khoảng cách Euclide, Cosine, Manhattan, ...) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất, gọi là k “*láng giềng gần nhất*” (Nearest Neighbor), sau đó dùng các khoảng cách này đánh trọng số cho tất cả các chủ đề. Khi đó, trọng số của một chủ đề chính là tổng tất cả các khoảng cách ở trên của các văn bản trong k láng giềng gần nhất có cùng chủ đề, chủ đề nào không xuất hiện trong k láng giềng gần nhất sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo giá trị trọng số giảm dần, chủ đề có trọng số cao sẽ được chọn là chủ đề của văn bản cần phân loại.

Theo phương diện toán học, bài toán được phát biểu như sau:

Đầu vào: Tập huấn luyện gồm các văn bản (mẫu) \vec{d}_i đã được phân lớp vào các chủ đề c_j cho trước và văn bản mới cần phân loại \vec{x} .

Đầu ra: Chủ đề của văn bản \vec{x} .

Thuật toán:

Bước 1. Xác định giá trị tham số k (số láng giềng gần nhất).

Bước 2. Tính khoảng cách giữa văn bản cần phân loại \vec{x} với tất cả các mẫu \vec{d}_i trong tập huấn luyện. Ví dụ, sử dụng độ đo Cosine để tính:

$$\text{sim}(\vec{x}, \vec{d}_i) = \cos(\vec{x}, \vec{d}_i) = \frac{\vec{x} * \vec{d}_i}{\|\vec{x}\| * \|\vec{d}_i\|}$$

Bước 3. Sắp xếp các mẫu trong tập huấn luyện theo thứ tự khoảng cách với \vec{x} tăng dần rồi chọn ra k láng giềng gần nhất (kNN).

Bước 4. Đánh trọng số cho tất cả các chủ đề.

Trọng số của chủ đề c_j đối với văn bản \vec{x} được tính như sau:

$$W(\vec{x}, c_j) = \sum_{\vec{d}_i \in \{kNN\}} \text{sim}(\vec{x}, \vec{d}_i) * y(\vec{d}_i, c_j) - b_j \quad (2.11)$$

Trong đó:

- $y(\vec{d}_i, c_j) \in \{0, 1\}$, với:
 - + $y = 0$: Văn bản \vec{d}_i không thuộc về chủ đề c_j
 - + $y = 1$: Văn bản \vec{d}_i thuộc về chủ đề c_j
- b_j là ngưỡng phân loại của chủ đề c_j được tự động học sử dụng một tập văn bản hợp lệ được chọn ra từ tập huấn luyện.

Bước 5. Sắp xếp các chủ đề theo giá trị trọng số tăng dần.

Bước 6. Gán \vec{x} vào chủ đề (lớp) có trọng số cao nhất.

Để chọn được tham số k tốt nhất cho thao tác phân loại, thuật toán cần được chạy thử nghiệm trên nhiều giá trị k khác nhau, giá trị k càng lớn thì thuật toán càng ổn định và sai sót càng thấp.

2.5.3. Thuật toán Naïve Bayes (NB)

Naïve Bayes là phương pháp phân loại dựa vào xác suất, được coi là một trong những thuật toán phân lớp điển hình nhất trong học máy và khai phá dữ liệu, đặc biệt được sử dụng rộng rãi trong phân lớp văn bản. Trong học máy,

Naïve Bayes thường được coi như thuật toán học máy chuẩn để so sánh với các thuật toán khác [1].

Ý tưởng cơ bản của cách tiếp cận này là sử dụng xác suất có điều kiện giữa từ hoặc cụm từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Như thế NB không tận dụng được sự phụ thuộc của nhiều từ vào một chủ đề cụ thể. Chính giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác với độ phức tạp theo số mũ vì nó không sử dụng cách kết hợp các từ để đưa ra phán đoán chủ đề.

Mục đích chính của thuật toán là làm sao tính được xác suất $P(C_j, d')$, xác suất để văn bản d' nằm trong lớp C_j . Theo luật Bayes, văn bản d' sẽ được gán vào lớp C_j nào có xác suất $P(C_j, d')$ cao nhất. $P(C_j, d')$ được xác định theo công thức [10]:

$$H_{BAYES}(d') = \arg \max_{C_j \in C} \left(\frac{P(C_j) * \prod_{i=1}^{|d'|} P(w_i | C_j)}{\sum_{C' \in C} P(C') * \prod_{i=1}^{|d'|} P(w_i | C')} \right) \quad (2.12)$$

$$= \arg \max_{C_j \in C} \left(\frac{P(C_j) * \prod_{w \in F} P(w | C_j)^{TF(w, d')}}{\sum_{C' \in C} P(C') * \prod_{w \in F} P(w | C')^{TF(w, d')}} \right)$$

Trong đó:

- + $TF(w, d')$ là số lần xuất hiện của từ w trong văn bản d'
- + $|d'|$ là số lượng các từ trong văn bản d'
- + w là một từ trong không gian đặc trưng F với số chiều là $|F|$

+ $P(C_j)$ được tính dựa trên tỷ lệ phần trăm của số văn bản mỗi lớp tương ứng trong tập dữ liệu huấn luyện:

$$P(C_j) = \frac{\|C_j\|}{\|C\|} = \frac{\|C_j\|}{\sum_{C' \in C} \|C'\|}$$

+ $P(w_i | C_j)$ được tính sử dụng phép ước lượng Laplace:

$$P(w_i | C_j) = \frac{1 + TF(w_i, C_j)}{|F| + \sum_{w' \in |F|} TF(w', C_j)}$$

Biểu diễn thuật toán:

Đầu vào: Tập huấn luyện D đã được vector hóa dưới dạng:

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

Tập xác định các nhãn lớp: $C = (c_1, c_2, \dots, c_m)$

Các thuộc tính độc lập điều kiện đôi một với nhau.

Văn bản mới cần phân loại X_{new} .

Đầu ra: Nhãn lớp của văn bản X_{new} .

Thuật toán:

Theo định lý Bayes:

$$P(C_i | X) = \frac{P(X | C_i) * P(C_i)}{P(X)} \quad (2.13)$$

Theo tính chất độc lập điều kiện:

$$P(X | C_i) = \prod_{k=1}^n P(x_k | C_i) \quad (2.14)$$

Trong đó:

+ $P(C_i | X)$: Xác suất thuộc phân lớp i khi biết trước mẫu X .

+ $P(C_i)$: Xác suất là phân lớp i .

+ $P(x_k | C_i)$: Xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân lớp i .

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1. Huấn luyện Naïve Bayes (dựa vào tập huấn luyện D); tính xác suất $P(C_i)$ và $P(x_k|C_i)$.

Bước 2. X_{new} được gán vào lớp có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} \left(P(C_i) = \prod_{k=1}^n P(x_k|C_i) \right) \quad (2.15)$$

Nhìn chung, Naïve Bayes là một công cụ rất hiệu quả trong một số trường hợp. Tuy nhiên, kết quả có thể rất xấu nếu dữ liệu huấn luyện nghèo nàn và các tham số dự đoán (như không gian đặc trưng) có chất lượng kém. Đây là một thuật toán phân loại tuyến tính thích hợp trong phân loại văn bản nhiều chủ đề. NB có ưu điểm cài đặt đơn giản, tốc độ thực hiện thuật toán nhanh, dễ cập nhật dữ liệu huấn luyện mới và có tính độc lập cao với tập huấn luyện, có thể sử dụng kết hợp nhiều tập huấn luyện khác nhau.

2.6. Phân loại văn bản tiếng Việt

2.6.1. Trích chọn đặc trưng văn bản

Các phương pháp rút trích thông tin cổ điển coi mỗi một văn bản như là tập các từ khóa và gọi tập các từ khóa này là tập các term. Một phần tử trong tập term đơn giản là một từ, mà ngữ nghĩa của từ này giúp tạo nên nội dung của văn bản. Vì vậy, tập term được sử dụng để tạo các chỉ mục và tóm lược nội dung của văn bản [5].

Giả sử cho một tập term của một văn bản nào đó, có thể nhận thấy rằng không phải tất cả các từ trong tập term này đều có mức độ quan trọng như nhau trong việc mô tả nội dung văn bản. Ví dụ, xét một tập gồm một trăm ngàn văn bản, giả sử có một từ A nào đó xuất hiện trong một trăm ngàn văn bản này thì có thể khẳng định rằng từ A này không quan trọng và ta sẽ không quan tâm đến nó, bởi chắc chắn nó sẽ không cho ta biết được về nội dung của các văn bản này. Vì vậy từ A sẽ bị loại ra khỏi tập các term, khi chúng ta xây dựng tập term cho văn bản để miêu tả nội dung ngữ nghĩa của các văn bản này. Kết quả này

có được thông qua thao tác xác định trọng số cho mỗi một từ trong tập term của một văn bản.

Đặt k_i là từ thứ i trong tập term, d_j là văn bản j , và $w_{ij} \geq 0$ là trọng số của từ k_i trong văn bản d_j . Giá trị của trọng số này rất quan trọng trong việc miêu tả nội dung của văn bản.

Đặt t là số lượng các từ trong tập term của hệ thống. $K = \{k_1, k_2, k_3, \dots, k_t\}$ là tập tất cả các từ trong tập term, trong đó k_i là từ thứ i trong tập term. Trọng số $w_{ij} > 0$ là trọng số của từ k_i trong văn bản d_j . Với mỗi một từ, nếu nó không xuất hiện trong văn bản thì $w_{ij} = 0$. Do đó, văn bản d_j thì được biểu diễn bằng vector d_j , trong đó vector $d_j = \{w_{j1}, w_{j2}, w_{j3}, \dots, w_{jt}\}$.

2.6.1.1. Phương pháp rút trích đặc trưng [5]

Giả sử có một tập gồm m văn bản, mỗi văn bản được biểu diễn bằng một vector đặc trưng theo dạng $D = \{d_1, d_2, \dots, d_n\}$, trong đó d_i là trọng số của đặc trưng thứ i và n là số lượng các đặc trưng của văn bản D . Mỗi một đặc trưng tương ứng với một từ xuất hiện trong tập huấn luyện, sau khi loại bỏ các stop-word ra khỏi các văn bản.

Phương pháp 1:

Phương pháp phổ biến nhất để rút trích các đặc trưng là dựa vào tần suất xuất hiện của các từ riêng biệt trong các văn bản. Phương pháp này thực hiện thông qua hai bước sau:

- **Bước 1:** Loại bỏ các từ chung (ngữ nghĩa của các từ này không ảnh hưởng đến nội dung của văn bản) ra khỏi văn bản bằng cách sử dụng một từ điển đặc biệt, hoặc là sử dụng danh sách các từ tầm thường (stop-word).

- **Bước 2:** Xác định tần suất xuất hiện tf_{ij} của các từ t_i còn lại trong mỗi văn bản D_j . Sau đó dựa vào tần suất xuất hiện để tính giá trị trọng số cho các từ t_i . Khi đó, n từ t_i có giá trị trọng số lớn nhất sẽ được chọn làm n đặc trưng của văn bản D_j .

✚ **Phương pháp 2:**

Một phương pháp khác để rút trích các đặc trưng của văn bản là sự kết hợp tần suất xuất hiện của từ trong văn bản và tần suất xuất hiện ngược trong văn bản (TF-IDF). Như đã trình bày ở **Mục 2.3 (Trọng số của từ trong văn bản)**, ta có công thức tính giá trị trọng số cho từ t_i trong văn bản d_j như sau:

$$w_{ij} = tf_{ij} * \log\left(\frac{m}{df_i}\right)$$

Trong đó: df_i là số lượng văn bản có chứa từ khoá t_i trong tập m văn bản đang xét. Khi đó, n từ t_i có giá trị trọng số lớn nhất sẽ được chọn làm n đặc trưng của văn bản.

2.6.1.2. Phương pháp đặc trưng đề nghị sử dụng

Chúng ta sẽ sử dụng một phương pháp rút trích đặc trưng sao cho phù hợp với mục tiêu yêu cầu đặt ra của đề tài. Đề xuất lựa chọn phương pháp **TF*IDF weighting** để rút trích đặc trưng, vì các yếu tố sau:

- Phương pháp này không phụ thuộc vào tần suất xuất hiện của các từ trong văn bản.
- Phương pháp này cân bằng giữa yếu tố mức độ bao phủ và số lượng các đặc trưng được sử dụng để biểu diễn văn bản.

Chi tiết các bước thực hiện của phương pháp này:

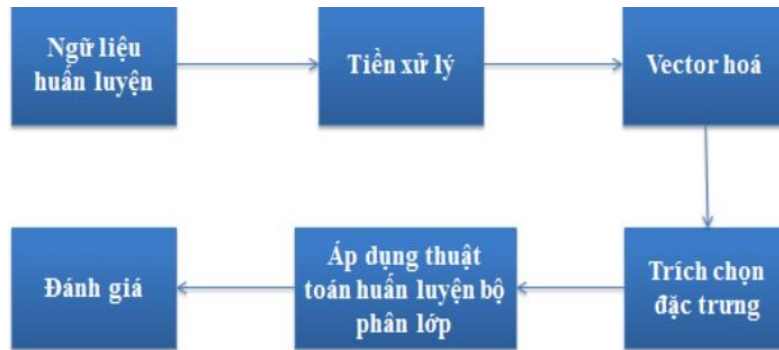
✚ Bước 1: Loại bỏ các từ tầm thường (stop-word).

✚ Bước 2: Đếm tần suất xuất hiện của các từ trong bước 1.

✚ Bước 3: Đặt lower = k, upper = k (tần suất xuất hiện của các từ - giả định ban đầu - và sẽ được xác định chính xác khi số lượng đặc trưng tìm được có mức độ phủ lớn hơn ngưỡng T, thông thường ngưỡng T được gán khoảng $0,95 \div 95\%$).

✚ Bước 4: Chọn tất cả các từ ở trên với tần suất xuất hiện nằm trong khoảng từ lower đến upper.

🚦 Bước 5: Kiểm tra mức độ phủ của các từ. Nếu mức độ phủ này lớn hơn ngưỡng T đã được định nghĩa trước thì dừng. Ngược lại thì đặt $\text{lower} = \text{lower} - 1$ và $\text{upper} = \text{upper} + 1$ rồi quay lại bước 4.



Hình 2.3. Chi tiết giai đoạn huấn luyện [5]

Ví dụ minh họa:

Một văn bản gồm 100 từ, trong đó từ “*máy tính*” xuất hiện 10 lần thì độ phổ biến là: $\text{tf}(\text{“máy tính”}) = 10 / 100 = 0,1$.

Giả sử có 1000 tài liệu đã được huấn luyện thuộc chuyên ngành *Hệ thống thông tin*, trong đó có 200 tài liệu chứa từ “*máy tính*”. Ta sẽ tính được:

$$\text{idf}(\text{“máy tính”}) = \log(1000 / 200) = 0,699$$

Như vậy ta tính được độ đo:

$$\text{TF.IDF} = \text{tf} * \text{idf} = 0,1 * 0,699 = 0,0699$$

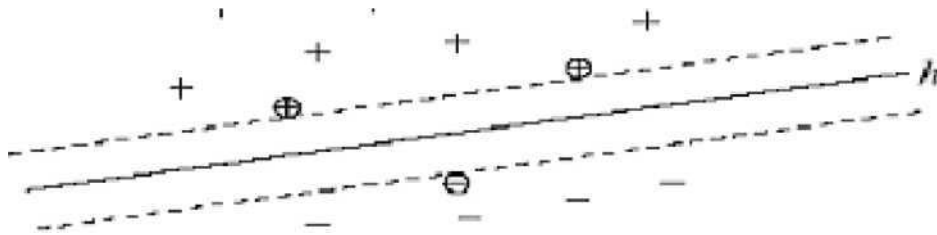
Dựa vào kết quả ta biết được trọng số của từ “*máy tính*” đối với chuyên ngành *Hệ thống thông tin*.

2.6.2. Sử dụng thuật toán SVM để phân loại văn bản

Trong lĩnh vực khai phá dữ liệu, bài toán phân loại văn bản đã được thực hiện dựa trên nhiều thuật toán như Naïve Bayes, K-Nearest Neighbor, Support Vector Machine... Những phương pháp này đã cho kết quả chấp nhận được và được sử dụng nhiều trong thực tế. Trong đó, phương pháp phân lớp sử dụng tập phân lớp vector hỗ trợ (SVM) những năm gần đây 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 (nhận dạng chữ viết tay, nhận dạng ảnh, ước lượng hồi quy...). Xét với các phương pháp phân lớp khác, khả năng phân lớp của SVM là khá tốt và hiệu quả [12].

Như đã trình bày ở **Mục 2.5.1** (*Thuật toán Support Vector Machine - SVM*), thực chất phương pháp phân lớp sử dụng thuật toán SVM là một bài toán tối ưu, mục tiêu là tìm ra một *siêu mặt phẳng* h phân tách các mẫu dương khỏi các mẫu âm với độ chênh lệch cực đại (*hình 2.4*). Mặt siêu phẳng này được gọi là *mặt siêu phẳng lề tối ưu*.



Hình 2.4. Mô hình SVM

Trong không gian đối tượng, mỗi siêu phẳng được viết dưới dạng một tập hợp các điểm thỏa mãn: $wx + b = 0$. Với w là vector trọng số, b là độ dịch.

Phương trình tổng quát của mặt siêu phẳng được tìm ra bởi SVM là:

$$f(x) = wx + b \quad (2.16)$$

Công thức này được sử dụng để tìm ra nhãn lớp cho các dữ liệu mới cần phân loại.

Huấn luyện SVM:

Huấn luyện SVM là việc giải bài toán quy hoạch toàn phương SVM. Các phương pháp số giải bài toán quy hoạch này yêu cầu phải lưu trữ một ma trận

có kích thước bằng bình phương của số lượng mẫu huấn luyện. Trong những bài toán thực tế, điều này là không khả thi vì thông thường kích thước của tập dữ liệu huấn luyện thường rất lớn (có thể lên tới hàng chục nghìn mẫu). Nhiều thuật toán khác nhau được phát triển để giải quyết vấn đề nêu trên. Những thuật toán này dựa trên việc phân rã tập dữ liệu huấn luyện thành những nhóm dữ liệu, giúp cho bài toán quy hoạch toàn phương sẽ được giải với kích thước nhỏ hơn. Sau đó, những thuật toán này kiểm tra các điều kiện KKT (Karush-Kuhn-Tucker) để xác định phương án tối ưu.

Một số thuật toán huấn luyện dựa vào tính chất: Nếu trong tập dữ liệu huấn luyện của bài toán quy hoạch toàn phương con cần giải ở mỗi bước có ít nhất một mẫu vi phạm các điều kiện KKT, thì sau khi giải bài toán này, hàm mục tiêu sẽ tăng. Như vậy, một chuỗi các bài toán quy hoạch toàn phương con với ít nhất một mẫu vi phạm các điều kiện KKT được đảm bảo hội tụ đến một phương án tối ưu. Do đó, ta có thể duy trì một tập dữ liệu làm việc đủ lớn có kích thước cố định và tại mỗi bước huấn luyện, ta loại bỏ và thêm vào cùng một số lượng mẫu.

Các ưu điểm của SVM trong phân loại văn bản:

Như đã biết, 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 lĩnh vực được xác định bởi một số tài liệu mẫu của lĩnh vực đó. Để thực hiện quá trình phân lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các tài liệu mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những tài liệu mới (chưa biết chủ đề).

Chúng ta có thể thấy 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 vector 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 vector 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...).

Phương pháp phân lớp sử dụng thuật toán SVM đã được nhiều tác giả nghiên cứu, so sánh với các phương pháp phân loại khác như Naïve Bayes, k-Nearest Neighbors... và đều chỉ ra SVM có nhiều ưu điểm, phù hợp hơn các phương pháp khác trong việc ứng dụng giải quyết bài toán 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 lớp cao và tỏ ra xuất sắc hơn so với các phương pháp phân loại văn bản khác [4]. Do vậy, luận văn lựa chọn phương pháp sử dụng thuật toán SVM để giải quyết bài toán phân loại văn bản tại chương sau.

2.7. Kết luận chương 2

Chương này trình bày chi tiết về bài toán phân loại văn bản tiếng Việt với các thuật toán phân loại và các khái niệm liên quan như: Các kỹ thuật cơ bản trong việc xử lý văn bản để phân loại như tách từ, đánh trọng số của từ trong văn bản, các mô hình biểu diễn văn bản, tính độ tương đồng văn bản... Nội dung của chương cũng đã tập trung phân tích, làm rõ một số giải pháp kỹ thuật liên quan, qua đó định hướng áp dụng trong việc giải quyết bài toán phân loại văn bản như phương pháp trích chọn đặc trưng, mô hình biểu diễn văn bản, phương pháp đánh trọng số của từ, thuật toán phân loại... Kết quả nghiên cứu của chương này là cơ sở để giải quyết bài toán phân loại văn bản tiếng Việt ở chương sau.

CHƯƠNG III. ÁP DỤNG THUẬT TOÁN SUPPORT VECTOR MACHINE PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT

3.1. Ứng dụng SVM vào bài toán phân loại văn bản hành chính tiếng Việt tại các cơ quan nhà nước tỉnh Bắc Kạn

Ở chương 2, luận văn đã tập trung giới thiệu một số thuật toán phân loại văn bản điển hình như Support Vector Machine (SVM), K-Nearest Neighbor (kNN) và Naïve Bayes (NB). Các thuật toán này có hướng tiếp cận khác nhau nhưng đều có một điểm chung, đó là sử dụng tập huấn luyện với các mẫu dữ liệu đã được gán nhãn để dự đoán giá trị của một hàm phân lớp cho một đối tượng đầu vào. Người ta gọi đây là các thuật toán học có giám sát. Nhiệm vụ của chương trình học có giám sát là huấn luyện khả năng dự đoán giá trị đầu ra cho hàm khi có một đối tượng đầu vào hợp lệ thông qua bộ dữ liệu huấn luyện. Chương trình học phải tiến hành tổng quát hóa từ các dữ liệu sẵn có để có thể đưa ra dự đoán những tình huống mới [4].

Trong phần này, luận văn sẽ giới thiệu một phương thức cải tiến của thuật toán SVM là bán giám sát SVM (Semi-Supervised Support Vector Machine - S^3VM) [4]. Bán giám sát SVM được đưa ra nhằm nâng SVM lên một mức cao hơn. Trong khi các thuật toán học có giám sát chỉ sử dụng dữ liệu huấn luyện đã gán nhãn thì học bán giám sát sử dụng cả dữ liệu đã gán nhãn kết hợp với dữ liệu chưa gán nhãn. Bài toán truyền dẫn sẽ dự đoán giá trị của một hàm phân lớp tới các điểm đã cho trong tập dữ liệu chưa gán nhãn.

Cho một tập huấn luyện gồm những dữ liệu đã gán nhãn (training set) và một tập các dữ liệu chưa gán nhãn (working set), S^3VM xây dựng một máy hỗ trợ vector sử dụng cả training set và working set. Mục đích là để gán các nhãn cho dữ liệu trong working set một cách tốt nhất có thể, sau đó sử dụng hỗn hợp dữ liệu huấn luyện đã gán nhãn cho trước (training set) và dữ liệu working set

vừa được gán nhãn để huấn luyện và phân lớp những dữ liệu mới. Nếu working set rỗng (toàn bộ dữ liệu đã được gán nhãn) thì bài toán này lại trở thành bài toán học có giám sát SVM. Ngược lại, nếu training set rỗng, tức là dữ liệu huấn luyện hoàn toàn chưa được gán nhãn, bài toán này sẽ trở thành một hình thể học máy khác gọi là học không giám sát. Học bán giám sát xảy ra khi cả training set và working set không rỗng.

Để hiểu một cách rõ ràng cụ thể về S^3VM , chúng ta cần hiểu về SVM đã được trình bày chi tiết ở phần trước. Trong luận văn này sẽ tìm hiểu về thuật toán S^3VM là bài toán phân lớp nhị phân.

Cho trước một tập huấn luyện gồm training set và working set 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 vector. Mỗi điểm i thuộc training set có một sai số là η_i và mỗi điểm j thuộc working set 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).

Nội dung thuật toán S^3VM [4],[11],[12]:

Đầu vào: Tập huấn luyện gồm cả dữ liệu có nhãn và chưa có nhãn:

$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^P, y_i \in \{-1, 0, 1\}, i = 1, 2, \dots, n\}$$

Tập dữ liệu đã gán nhãn trong D gồm l dữ liệu:

$$L = \{(x_i, y_i) \mid x_i \in \mathbb{R}^P, y_i \in \{-1, 1\}, i = 1, 2, \dots, l\}$$

Tập dữ liệu chưa có nhãn trong D gồm k dữ liệu:

$$K = \{(x_j, y_j) \mid x_j \in \mathbb{R}^P, y_j = 0, j = 1, 2, \dots, k\}$$

Đầu ra: Một siêu phẳng h phân chia dữ liệu trong D thành hai nhóm với sai số là nhỏ nhất.

Thực hiện thuật toán:

$$\text{Giải bài toán tối ưu: } \begin{cases} \text{Cực tiểu hóa } \frac{1}{2} \|w\|^2 \text{ theo } w, b, y_j \\ y_i(wx_i + b) \geq 1; i = 1, \dots, l \\ y_j(wx_j + b) \geq 1; j = 1, \dots, k \end{cases} \quad (3.1)$$

Cụ thể hơn, ta giải bài toán sau:

$$\min_{w, b, y_j} \left\{ \frac{\lambda}{2} \|w\|^2 + \frac{1}{2l} \sum_{i=1}^l \max(0, 1 - y_i(wx_i + b)) + \frac{\lambda'}{2k} \sum_{j=1}^k \max(0, 1 - y_j(wx_j + b)) \right\}$$

Vấn đề ở đây là ta cần phải xác định nhãn y_j của mỗi điểm j trong tập dữ liệu chưa được gán nhãn K . Ta thực hiện tìm kiếm một siêu phẳng w và ghi nhãn một trong những ví dụ không có nhãn, do đó hàm mục tiêu SVM được giảm thiểu, và bị ràng buộc bởi r phần nhỏ của dữ liệu không có nhãn được phân loại tích cực. Giá trị r được xác định theo công thức:

$$r = \frac{1}{k} \sum_{j=1}^k \max(0, \text{sign}(wx_j + b)) \quad (3.2)$$

Tập dữ liệu chưa gán nhãn (working set) 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 S^3VM có một siêu phẳng mới. Sau đó áp dụng siêu phẳng này để phân lớp các mẫu dữ liệu mới được đưa vào.

3.2. Áp dụng phân loại văn bản

Để áp dụng vào phân loại văn bản, thuật toán S^3VM xem mỗi tài liệu là một vector $f(d_1, d_2, \dots, d_n)$. Áp dụng phương trình tổng quát của siêu phẳng tìm được bởi thuật toán SVM (2.16):

$$f(x) = wx + b$$

hay còn có thể viết theo dạng sau:

$$f(x_1, x_2, \dots, x_n) = b + \sum_{i=1}^n w_i x_i \quad (3.3)$$

Thay thế mỗi văn bản tương ứng vào phương trình siêu phẳng này:

$$f(d_1, d_2, \dots, d_n) = b + \sum_{i=1}^n w_i d_i \quad (3.4)$$

Nếu: $f(d) \geq 0$, văn bản thuộc lớp +1, $f(d) < 0$ thì văn bản thuộc lớp -1.

Có thể thấy rằng quá trình áp dụng thuật toán S^3VM vào bài toán phân lớp văn bản chính là việc thay thế vector trọng số biểu diễn văn bản đó vào phương trình siêu phẳng của S^3VM , từ đó tìm ra được nhãn lớp của các văn bản chưa gán nhãn.

Như vậy, thực chất của quá trình phân lớp bán giám sát áp dụng đối với văn bản là: Tập dữ liệu huấn luyện là các văn bản, còn tập dữ liệu chưa gán nhãn (working set) là những văn bản được các văn bản đã có nhãn trong tập huấn luyện trở tới.

Giải thuật S^3VM chính là một phương pháp cải tiến của giải thuật SVM, giải thuật đã tận dụng được những ưu điểm của SVM là có độ chính xác cao, đồng thời tận dụng được nguồn dữ liệu huấn luyện không gán nhãn rất sẵn có nhằm giải quyết bài toán phân lớp một cách tối ưu.

3.3. Xây dựng chương trình thử nghiệm ứng dụng phân loại văn bản áp dụng vào máy tìm kiếm văn bản hành chính tiếng Việt

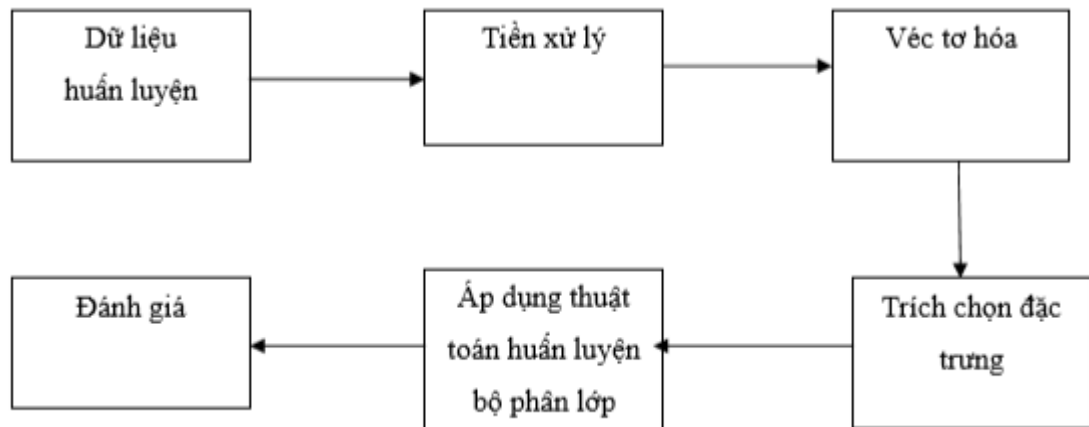
3.3.1. Mô tả bài toán

Cho n văn bản thuộc các lĩnh vực khác nhau. Yêu cầu đặt ra là cần phải xây dựng một ứng dụng thử nghiệm áp dụng một giải thuật phân lớp để phân loại n văn bản này theo các lĩnh vực khác nhau dựa vào các văn bản mẫu đã được huấn luyện theo các lĩnh vực khác nhau đó.

Như đã phân tích ở các phần trên, trong phạm vi đề tài này, luận văn sử dụng thuật toán SVM để xây dựng mô hình phân loại văn bản, bao gồm hai giai đoạn: Giai đoạn huấn luyện và giai đoạn phân lớp.

a. Giai đoạn huấn luyện:

Để xây dựng được mô hình ứng dụng thử nghiệm, cần có một tập huấn luyện với mỗi phần tử trong tập huấn luyện đã được xác định nhãn lớp (lĩnh vực) và được thể hiện bằng một mô hình mã hóa sử dụng không gian vector (đã được trình bày chi tiết ở Mục 2.3 - Các mô hình biểu diễn văn bản). Sau đó, chúng ta sẽ định nghĩa một lớp mô hình và một thủ tục huấn luyện, với lớp mô hình là họ các tham số của bộ phân loại, thủ tục huấn luyện với giải thuật được lựa chọn là SVM để chọn ra một họ các tham số tối ưu cho bộ phân loại. Chi tiết giai đoạn huấn luyện được mô tả như sơ đồ sau:



Hình 3.1. Chi tiết giai đoạn huấn luyện

Trong đó:

- + Dữ liệu huấn luyện: Kho dữ liệu thu thập được.
- + Tiền xử lý: Xử lý chuẩn hóa dữ liệu huấn luyện.
- + Véc tơ hóa: Mã hóa văn bản với một mô hình trọng số.
- + Trích chọn đặc trưng: Loại bỏ những từ (đặc trưng) không quan trọng (không chứa thông tin đặc trưng) khỏi tài liệu nhằm nâng cao hiệu suất phân loại và giảm độ phức tạp của thuật toán huấn luyện.
- + Thuật toán huấn luyện: Thủ tục huấn luyện bộ phân lớp để tìm ra họ các tham số tối ưu (sử dụng thuật toán SVM).

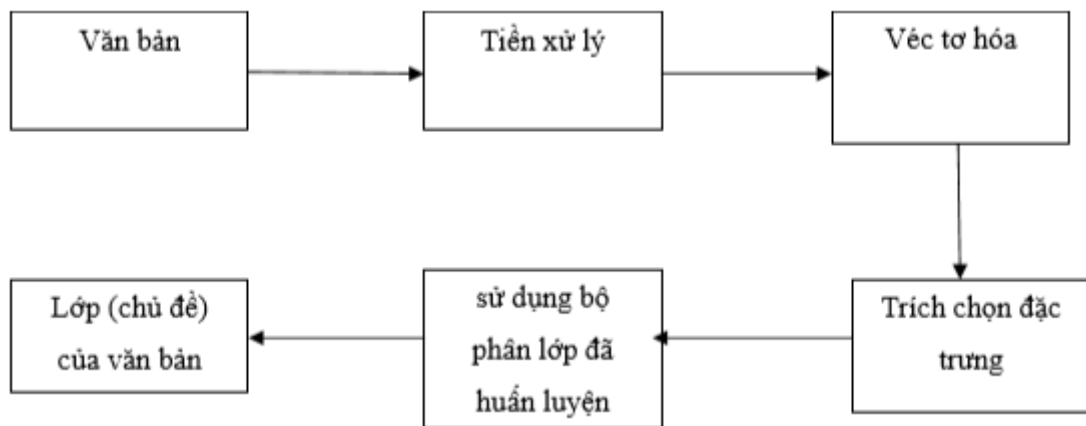
+ Đánh giá: Bước đánh giá hiệu suất (chất lượng) của bộ phân lớp.

Thủ tục huấn luyện sẽ được thực thi lặp lại nhiều lần để tìm họ các tham số tối ưu sau mỗi lần lặp.

b. Giai đoạn phân lớp:

Sau khi đã hoàn thành các giai đoạn huấn luyện, mô hình phân lớp sẽ được áp dụng cho các văn bản mới cần phân loại.

Chi tiết giai đoạn phân lớp được mô tả như sơ đồ sau:



Hình 3.2. Chi tiết giai đoạn phân lớp

3.3.2. Quá trình tiền xử lý văn bản

Văn bản trước khi được vector hóa, tức là trước khi đưa vào sử dụng bởi mô hình phân loại, cần phải được tiền xử lý. Quá trình tiền xử lý sẽ giúp nâng cao hiệu suất phân loại và giảm độ phức tạp của thuật toán huấn luyện. Tùy vào mục đích bộ phân loại mà chúng ta sẽ có những phương pháp tiền xử lý văn bản khác nhau, như:

- Chuyển văn bản về chữ thường;
- Loại bỏ các ký tự đặc biệt (ví dụ như: ~; @; #; \$; %; & *;...);
- Thực hiện tách từ: Sử dụng công cụ tách từ vnTokenizer, version 4.1.1

để phân tách ra các từ. Kết quả ta sẽ thu được file chứa các từ được phân tách (dấu “|” được sử dụng để ngăn cách giữa các từ).

- Loại bỏ các từ dừng hay từ tầm thường (stopword): Thực hiện loại bỏ các từ không có ý nghĩa sau khi tách từ dựa trên danh mục từ dừng có trước.

3.3.3. Vector hóa và trích chọn đặc trưng văn bản

Như đã trình bày ở các phần trên, trong mô hình không gian vector, một văn bản **d** được biểu diễn dưới dạng vector đặc trưng $f(d_1, d_2, \dots, d_n)$, trong đó n là số lượng đặc trưng hay số chiều của vector văn bản, d_i là trọng số của đặc trưng thứ i .

Để trích chọn đặc trưng văn bản ta sử dụng phương pháp TF*IDF đã giới thiệu tại Mục 2.6.1 Chương II.

Giả sử: Ta có m tài liệu thuộc lớp P ; trong đó n tài liệu có chứa từ A ($m \geq n$). Khi đó:

+ Độ phổ biến của từ A đối với tài liệu (văn bản) T chứa nó:

$$tf(A) = [\text{số lần xuất hiện của } A \text{ trong } T] / [\text{tổng số từ có trong } T]$$

+ Độ đo IDF của từ A trong m tài liệu mẫu thuộc lớp P , trong đó có n tài liệu chứa từ A : $idf(A) = \log(m/n)$

Từ đó ta tính được độ đo TF*IDF (chính là trọng số của từ A đối với lớp P): $TF*IDF(A) = tf(A)*idf(A)$.

3.3.4. Đánh giá bộ phân lớp

Sau khi đã tìm được họ các tham số tối ưu cho bộ phân lớp (hay có thể nói là bộ phân lớp đã được huấn luyện xong), nhiệm vụ tiếp theo là cần phải đánh giá (kiểm tra) bộ phân lớp đó cho kết quả như thế nào. Quá trình kiểm tra được thực hiện trên một tập dữ liệu khác với tập dữ liệu huấn luyện, gọi là tập dữ liệu kiểm tra. Để đơn giản, ta xét một bộ phân lớp nhị phân (phân hai lớp). Với các tham số:

+ a : Là số lượng đối tượng thuộc về lớp đang xét và được bộ phân lớp gán vào lớp;

+ b: Là số lượng đối tượng không thuộc về lớp đang xét nhưng được bộ phân lớp gán vào lớp;

+ c: Là số lượng đối tượng thuộc về lớp đang xét nhưng bị bộ phân lớp loại khỏi lớp;

+ d: Là số lượng đối tượng không thuộc về lớp đang xét và được bộ phân lớp loại khỏi lớp.

Để đánh giá chất lượng bộ phân lớp, có hai đơn vị đo lường quan trọng là độ đúng đắn (accuracy) được đo bằng công thức $\frac{a+d}{a+b+c+d}$ và độ sai lỗi (error) được tính bằng công thức $\frac{c+b}{a+b+c+d}$. Các độ đo này phản ánh đầy đủ chất lượng của bộ phân lớp. Tuy nhiên, khi đánh giá bộ phân lớp, thường người ta chỉ xét đến những đối tượng thuộc về lớp và được phân lớp đúng, còn những đối tượng không thuộc về lớp sẽ ít được quan tâm. Do đó, một số độ đo khác được định nghĩa như:

$$+ \text{Precision (độ chính xác): } \frac{a}{a+b} \quad (3.5)$$

$$+ \text{Recall (độ bao phủ, độ đầy đủ): } \frac{a}{a+c} \quad (3.6)$$

$$+ \text{Fallout (độ loại bỏ): } \frac{b}{b+d} \quad (3.7)$$

Tuy nhiên, trong một số trường hợp thực tế, nếu tính độ đo precision và độ đo recall riêng rẽ sẽ cho kết quả không cân đối. Do đó, để thuận tiện, người ta kết hợp hai độ đo này vào một đơn vị đo tổng quát duy nhất. Để thực hiện điều này, người ta sử dụng đơn vị đo lường F1 được định nghĩa như sau:

$$F1 = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} \quad (3.8)$$

Trong đó:

+ P: Là độ chính xác (Precision);

+ R: Là độ bao phủ (Recall);

+ α : Là hệ số xác định sự cân bằng của độ chính xác và độ bao phủ.

Giá trị $\alpha = 5$ thường được chọn cho sự cân bằng giữa P và R. Với giá trị này, độ đo được tính đơn giản là: $F1 = 2 * R * P / (R + P)$ (3.9)

3.3.5. Chương trình thực nghiệm

Chương trình thực nghiệm được xây dựng trên cơ sở sử dụng các công cụ mã nguồn mở có sẵn được chia sẻ tại thư viện LIBSVM, bộ công cụ lập trình Visual Studio 2013 và hệ quản trị CSDL Microsoft Access 2013.

Bộ dữ liệu huấn luyện bao gồm 43 tập văn bản, được gán nhãn phân loại thủ công vào 4 lĩnh vực: Giáo dục (ID=1); Kinh tế (ID=2); Thể thao (ID=3); Tin học (ID=4).

Bộ dữ liệu kiểm tra bao gồm 249 văn bản hành chính tiếng Việt thuộc 4 lĩnh vực nêu trên. Các văn bản được thu thập từ cơ sở dữ liệu văn bản hành chính đã được phát hành, đăng tải công khai trên hệ thống cổng thông tin điện tử của các cơ quan nhà nước.

Việc đánh giá bộ phân lớp dựa vào các chỉ số độ chính xác (precision), độ bao phủ (recall) và F1.

3.3.6. Kết quả thực nghiệm

Bảng 3.1. Bộ dữ liệu thử nghiệm

Tên lớp	Số mẫu huấn luyện	Số mẫu kiểm tra	Tổng số mẫu
Giáo dục	10	60	70
Kinh tế	10	58	68
Thể thao	12	45	57
Tin học	11	86	97
<i>Tổng cộng</i>	<i>43</i>	<i>249</i>	<i>292</i>

Bảng 3.2. Kết quả phân lớp bộ dữ liệu kiểm tra

Tên lớp	ID	1	2	3	4	Tổng số
Giáo dục	1	54	2	1	3	60
Kinh tế	2	2	52	3	1	58
Thể thao	3	2	2	41	0	45
Tin học	4	5	3	1	77	86

Bảng 3.3. Đánh giá hiệu suất phân lớp

Tên lớp	Precision	Recall	F1
Giáo dục	88,89%	93,33%	91,06%
Kinh tế	89,83%	91,38%	90,60%
Thể thao	93,18%	91,11%	92,13%
Tin học	95,18%	91,86%	93,49%
Trung bình			91,82%

Độ chính xác phân lớp các văn bản thuộc cả 4 lĩnh vực đều đạt tỷ lệ ~90%; độ bao phủ >90%. Kết quả thực nghiệm đã khẳng định tính hiệu quả của thuật toán SVM khi áp dụng vào bài toán phân lớp văn bản.

3.4. Kết luận chương 3

Chương này trình bày về thuật toán học bán giám sát S^3VM và áp dụng thuật toán trong việc phân loại văn bản tiếng Việt để xây dựng chương trình thử nghiệm đơn giản dựa trên ngôn ngữ lập trình Visual C# trong bộ công cụ lập trình Visual Studio 2013, hệ quản trị CSDL Microsoft Access 2013 và tiến hành chạy thử nghiệm chương trình với một số bộ dữ liệu đầu vào.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đánh giá kết quả thực hiện đề tài

Qua nghiên cứu và thực hiện, luận văn đã đạt được kết quả như sau:

- Trình bày bài toán phân loại văn bản và cơ sở lý thuyết của bài toán xây dựng hệ thống phân loại văn bản tiếng Việt.
- Giới thiệu các thuật toán phân loại văn bản như SVM, kNN, NB và nêu phương pháp sử dụng SVM để phân loại văn bản tiếng Việt.
- Thực hiện cài đặt thuật toán học bán giám sát SVM để xây dựng chương trình thử nghiệm phân loại văn bản tiếng Việt; tiến hành chạy thử nghiệm chương trình với một số bộ dữ liệu đầu vào đơn giản.

Tuy đã giải quyết được mục tiêu đề ra, nhưng luận văn mới chỉ đánh giá được phân loại văn bản dựa trên các bộ dữ liệu có sẵn trên cơ sở lý thuyết chứ chưa thực sự xây dựng được một ứng dụng hoàn thiện để đánh giá chính xác hơn về ưu, nhược điểm của hướng tiếp cận này. Chương trình thử nghiệm còn đơn giản, và mới chỉ dừng lại ở mức thực hiện được các thuật toán trên dữ liệu đầu vào là các file văn bản truyền thống có định dạng đơn giản (*.txt), chưa hỗ trợ việc đọc trực tiếp từ các file word, PDF,...

Hướng phát triển

Luận văn đã giải quyết được bài phân loại văn bản dựa trên nền tảng lý thuyết và các ứng dụng sẵn có. Để mở rộng tính thực tế cho luận văn cần tiếp tục xây dựng một ứng dụng cụ thể áp dụng giải pháp đã lựa chọn, ứng dụng cho việc xây dựng một hệ thống phân loại tự động văn bản tiếng Việt.

Nghiên cứu và áp dụng một số giải thuật tính toán độ tương đồng ngữ nghĩa trên mạng ngữ nghĩa để cải tiến mô hình phân loại văn bản tiếng Việt.

TÀI LIỆU THAM KHẢO

Tiếng Việt

- [1] Hà Quang Thụy (2009), *Giáo trình khai phá dữ liệu Web*, NXB Giáo dục, Hà Nội.
- [2] Ủy ban Khoa học Xã hội Việt Nam (1983), *Ngữ pháp tiếng Việt*, NXB Khoa học Xã hội, Hà Nội.
- [3] Nguyễn Thị Kim Anh, Trịnh Thị Ngọc Hương (2016), *Nghiên cứu kỹ thuật đánh giá độ tương đồng văn bản ứng dụng trong so sánh văn bản tiếng Việt*, Báo cáo nghiên cứu khoa học, Đại học Hàng hải Việt Nam, Hải Phòng.
- [4] Lê Hoàng Dương, Ngô Quốc Vinh (2016), *Nghiên cứu về thuật toán phân lớp sử dụng quá trình học máy bán giám sát, ứng dụng trong việc phân lớp trang web*, Báo cáo nghiên cứu khoa học, Đại học Hàng hải Việt Nam, Hải Phòng.
- [5] Trần Thị Thu Thảo, Vũ Thị Chinh (2012), *Xây dựng hệ thống phân loại tài liệu tiếng Việt*, Báo cáo nghiên cứu khoa học, Đại học Lạc Hồng, Đồng Nai.

Tiếng Anh

- [6] Jiawei Han, Micheline Kamber, Jian Pei (2012), *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers.
- [7] Steven Bird, Ewan Klein, Edward Loper (2009), *Natural language processing with Python*, O'Reilly Media, America.
- [8] Dinh Dien, Hoang Kiem, Nguyen Van Toan (2001), “Vietnamese Word Segmentation”, *The sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, pp. 749-756.
- [9] Eric Brill (1995), “Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging”, *Computational Linguistics*, 21(4), pp. 543–565.

- [10] T. Joachims (1997), “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization”, *Proceedings of International Conference on Machine Learning*, San Mateo, CA, pp. 143-151.
- [11] K. Bennett, A. Demiriz (1998), “Semi - Supervised Support Vector Machines”, *Advances in Neural information processing systems*, 12, p.368-374.
- [12] T. Joachims (1997), “Text Categorization with Support Vector Machine: Learning with Many Relevant Features”, Cornell Computer.
- [13] Alex Smola, S.V.N. Vishwanathan (2008), *Introduction to Machine Learning*, Departments of Statistics and Computer Science Purdue University, College of Engineering and Computer Science, Australian National University.