BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

NGUYỄN NƯƠNG QUỲNH

XÂY DỰNG ỨNG DỤNG PHÂN LOẠI VĂN BẢN

Chuyên ngành : Khoa học máy tính
Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. VÕ TRUNG HÙNG

Phản biện 1: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 2: **PGS.TS. LÊ MẠNH THẠNH**

Luận văn được bảo vệ tại Hội đồng chấm luận văn tốt nghiệp Thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 28 tháng 12 năm 2013.

** Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Trong thời đại bùng nổ công nghệ thông tin ngày nay, phương thức số hóa các văn bản trong các giao dịch dần thay thế các giao dịch sử dụng giấy theo truyền thống. Với nhiều tính năng mà phương thức này mang lại như: cách lưu trữ gọn nhẹ, thời gian lưu trữ lâu dài, dễ dàng cập nhật, sửa đổi, tiện dụng trong trao đổi,… do đó số lượng văn bản số ngày càng tăng. Cùng với sự gia tăng của các văn bản số thì nhu cầu quản lý và tìm kiếm văn bản cũng tăng theo. Với số lượng văn bản lớn thì việc phân loại văn bản tự động là một nhu cầu bức thiết.

Ngày nay, công nghệ thông tin được ứng dụng vào hầu hết các lĩnh vực hoạt động của đời sống xã hội và có những ảnh hưởng nhất định đến chất lượng và hiệu quả hoạt động của các đơn vị. Đối với công tác văn thư việc ứng dụng các công cụ, các sản phẩm công nghệ thông tin vào quá trình hoạt động ngày càng trở nên quan trọng và cần thiết để phân loại và xử lý văn bản.

Hằng năm, tại bộ phận văn thư của Trường Đại học Quảng Bình nhận và chuyển đi số lượng lớn công văn các loại trong đó có số lượng lớn văn bản số. Tại bộ phận này đã ứng dụng công nghệ thông tin vào quá trình hoạt động của mình, tuy nhiên chưa có công cụ phân loại văn bản để thực hiện việc phân loại văn bản để giúp cho việc lưu trữ công văn thuận lợi, giúp chọn lọc văn bản phải xử lý một cách dễ dàng và giúp tìm kiếm văn bản một cách nhanh chóng, thuận lợi và hiệu quả.

Chính vì vậy, để hỗ trợ các cán bộ tại bộ phận văn thư của Trường Đại học Quảng Bình có được một công cụ quản lý công văn một cách thuận tiện, chính xác, tiết kiệm thời gian cũng như ứng dụng công nghệ thông tin vào công tác quản lý công văn, tôi thực hiện đề tài: *"Xây dựng ứng dụng phân loại văn bản"*.

2. Mục tiêu nghiên cứu

Mục tiêu chung là tự động hóa công tác phân loại văn bản tại Trường Đại học Quảng Bình.

Mục tiêu cụ thể là xây dựng phần mềm phân loại được cho 3 loại công văn (công văn Đoàn thanh niên, công văn Công đoàn và công văn Đảng), với các chức năng chính là: huấn luyện và phân loại văn bản.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng

- Hệ thống công văn của Trường Đại học Quảng Bình.
- Các phương pháp phân loại văn bản tiếng Anh.
- Các phương pháp phân loại văn bản tiếng Việt.
- Phương pháp phân loại văn bản sử dụng mạng Nơ-ron kết hợp cây quyết định áp dụng cho tiếng Việt.
- Ngôn ngữ lập trình Java sử dụng để xây dựng ứng dụng phân loại công văn.

Phạm vi nghiên cứu

Trong khuôn khổ của luận văn, tôi nghiên cứu và xây dựng ứng dụng phân loại công văn thành các loại theo các bộ phận chức năng như: công văn Đảng, công văn Công đoàn, công văn Đoàn

thanh niên. Ứng dụng được phát triển chạy trên môi trường máy tính đơn.

4. Phương pháp nghiên cứu

Khi triển khai nghiên cứu đề tài, chúng tôi đã sử dụng hai phương pháp chính là nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

Đối với phương pháp nghiên cứu lý thuyết, chúng tôi tập trung nghiên cứu cơ sở lý thuyết về phân loại văn bản, các phương pháp phân loại văn bản tiếng Anh, các phương pháp phân loại văn bản tiếng Việt; nghiên cứu tài liệu phân loại văn bản tiếng Anh sử dụng phương pháp mạng Nơ-ron kết hợp cây quyết định; nghiên cứu các tài liệu về ngôn ngữ lập trình Java.

Đối với phương pháp nghiên cứu thực nghiệm, chúng tôi tập trung vào việc sử dụng ngôn ngữ lập trình Java để xây dựng ứng dụng phân loại công văn tiếng Việt.

5. Bố cục đề tài

CHƯƠNG 1: NGHIÊN CỨU TỔNG QUAN.

Chương này trình bày khái quát về phân loại văn bản, các phương pháp phân loại văn bản đối với tiếng Anh và các phương pháp tách từ tiếng Việt.

CHƯƠNG 2: ĐỀ XUẤT GIẢI PHÁP.

Chương này trình bày phân loại văn bản tiếng Anh sử dụng phương pháp mạng Nơ-ron kết hợp phương pháp cây quyết định và áp dụng phương pháp mạng Nơ-ron kết hợp cây quyết định để phân loại cho văn bản tiếng Việt.

CHƯƠNG 3: PHÁT TRIỂN ỨNG DỤNG.

Sử dụng ngôn ngữ lập trình Java để viết chương trình demo việc phân loại công văn bằng tiếng Việt sử dụng phương pháp mạng Nơ-ron kết hợp phương pháp cây quyết định.

6. Tổng quan tài liệu tham khảo

Khi triển khai nghiên cứu đề tài, tôi đã tham khảo các tài liệu về khái niệm văn bản [7] [25], công văn [26], đặc điểm của tiếng Anh và tiếng Việt [8], các phương pháp phân loại văn bản [1] [3] [4] [5] [6] [8] [12] [14] [15] [17] [19] [21] [23] [25], tài liệu về các phương pháp tách từ tiếng Việt [2] [9] [11] [13] [18] [20] [24] và nghiên cứu phương pháp phân loại văn bản sử dụng mạng Nơ-ron kết hợp cây quyết định [14]. Tài liệu tham khảo được sử dụng trong luận văn gồm các từ điển, các luận văn, các bài báo trong các tạp chí tiếng Anh.

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN

Trong chương này sẽ trình bày các khái niệm văn bản, phân loại văn bản; phân biệt sự giống nhau và khác nhau của hai ngôn ngữ đó là tiếng Anh và tiếng Việt; sau đó trình bày một số thuật toán phân loại văn bản và một số thuật toán tách từ tiếng Việt.

1.1. PHÂN LOẠI VĂN BẢN

1.1.1. Văn bản

1.1.2. Phân loại văn bản

1.2. SO SÁNH ĐẶC ĐIỂM TIẾNG ANH VÀ TIẾNG VIỆT

1.3. CÁC PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN

1.3.1. Phương pháp Naïve Bayes

1.3.2. Phương pháp k-Nearest Neighbors

1.3.3. Phương pháp Support Vector Machine

1.3.4. Phương pháp cây quyết định

1.3.5. Phương pháp mạng Nơ-ron

1.4. CÁC PHƯƠNG PHÁP TÁCH TỪ TIẾNG VIỆT

1.4.1. Phương pháp Maximum Matching

1.4.2. Phương pháp Transformation-based Learning

1.4.3. Phương pháp Weighted Finite-State Transducer

CHƯƠNG 2
MÔ TẢ ỨNG DỤNG

Chương này là chương quan trọng của đề tài. Trong chương này trình bày bài toán cần giải quyết của đề tài là phân loại công văn và các thuật toán được sử dụng để xây dựng ứng dụng phân loại công văn.

2.1. GIỚI THIỆU BÀI TOÁN

Công văn là giấy tờ trao đổi, liên hệ công việc của cơ quan nhà nước [26]. Công văn là hình thức văn bản hành chính dùng phổ biến trong các cơ quan, tổ chức, doanh nghiệp.

Ngày nay, việc giao dịch giữa các đơn vị bằng công văn cũng đã dần chuyển sang phương thức giao dịch bằng băn bản được số hóa. Do đó, đòi hỏi ứng dụng công nghệ thông tin để quản lý, lưu trữ và xử lý công văn cho công tác văn thư.

Hiện nay, tại Trường Đại học Quảng Bình đã ứng dụng công nghệ thông tin vào công tác văn thư cho việc quản lý, lưu trữ và xử lý công văn. Tuy nhiên, việc ứng dụng công nghệ thông tin này vẫn còn có một số công việc vẫn chưa thực hiện hoàn toàn bằng máy tính, vẫn còn có sự tác động của con người.

Quy trình quản lý công văn tại Trường Đại học Quảng bình được thực hiện như sau:

Đối với quản lý công văn, văn thư quản lý 2 loại công văn là công văn đi và công văn đến.

Công văn đến được chia thành 2 loại: loại 1 bao gồm tất cả văn bản, tài liệu, thư từ do Trường nhận được từ bên ngoài gửi đến; loại 2 bao gồm các văn bản, tài liệu do các cá nhân, phòng, ban, đơn

vị trong Trường gửi trình Ban Giám hiệu. Văn bản đến khi nhận được, cán bộ văn thư đọc nội dung của công văn và phân công văn thành các loại sơ bộ, nếu là công văn gửi Ban giám hiệu thì trình Ban giám hiệu xử lý, sau khi nhận sự phản hồi cho các công văn từ Ban giám hiệu, cùng với các công văn đến các khoa, phòng bộ phận văn thư sẽ chuyển công văn đó đến các khoa, phòng để xử lý và thực hiện. Công văn được chuyển đến nơi cần xử lý thì vẫn còn lưu ở văn thư và được lưu vào thư mục chứa loại công văn đó.

Công văn đi là tất cả văn bản, tài liệu của các cá nhân, phòng, ban, đơn vị, bộ môn trong Trường gửi tới các cá nhân, đơn vị trong và ngoài Trường. Mọi văn bản đi của Trường đều phải trình ký lên Ban Giám hiệu thông qua bộ phận thư ký. Khi bộ phân văn thư nhận văn bản trình ký Ban Giám hiệu đều thực hiện kiểm tra các phần và thể thức văn bản đã đúng quy định và phải kiểm tra thủ tục hành chính. Nếu phát hiện sai sót thì báo với người có trách nhiệm để sửa chữa, bổ sung cho hoàn chỉnh trước khi trình Ban Giám hiệu. Văn bản sau khi được phê duyệt chuyển xuống bộ phận văn thư để vào sổ đăng ký rồi đóng dấu làm các thủ tục gửi đi. Văn bản đi khi được gửi đi cũng được lưu một bản ở bộ phận văn thư và một bản ở đơn vị soạn văn bản. Văn bản lưu sẽ được lưu cả bản in và văn bản số, văn bản số sẽ được lưu vào thư mục chứa văn bản cùng loại để có thể tìm lại sau này.

Việc phân công văn đi và đến thành các loại khác nhau và lưu ở các thư mục riêng nhằm giúp cho việc quản lý, xử lý và tìm kiếm văn bản sau này khi cần một cách dễ dàng. Tuy nhiên, việc phân loại văn bản và lưu văn bản vào đúng thư mục chứa văn bản cùng loại trong hệ thống được thực hiện dựa vào việc đọc nội dung công văn

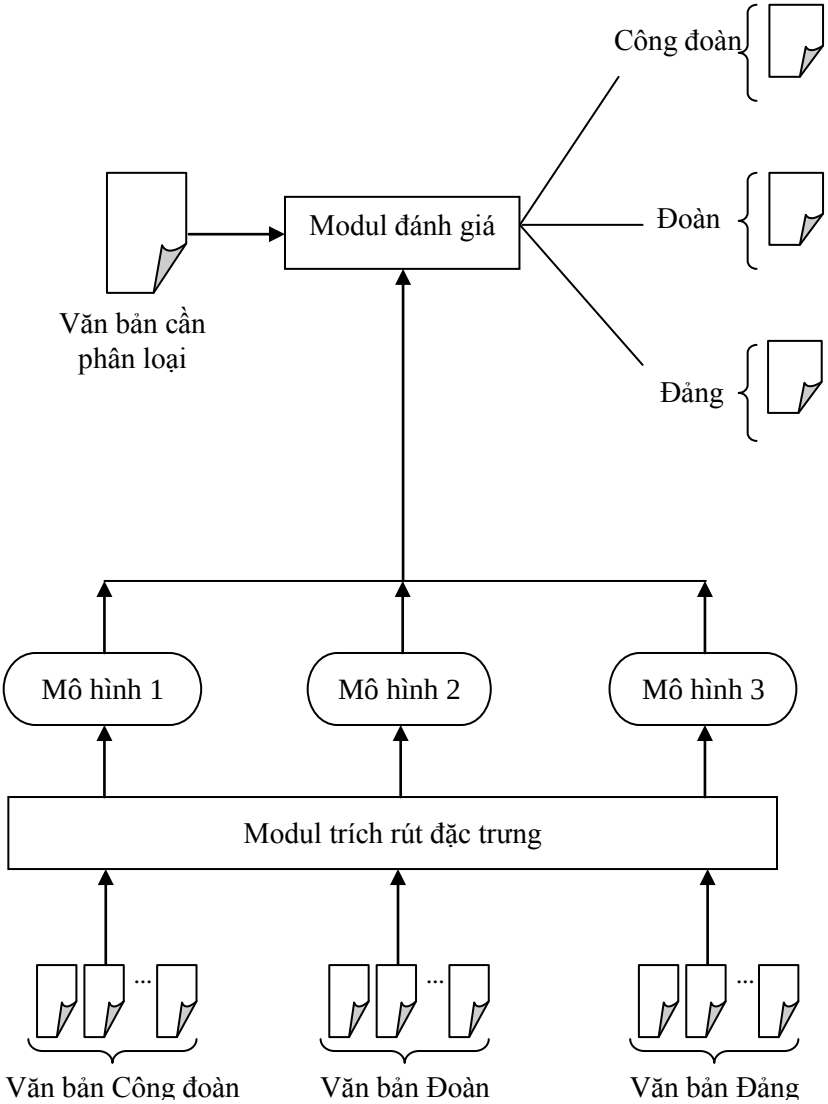
của cán bộ văn thư để phân loại và văn thư thực hiện việc di chuyển văn bản đó vào vị trí lưu trữ. Hai thao tác này vẫn chưa thực hiện một cách tự động.

Tại Trường Đại học Quảng Bình có nhu cầu xây dựng và tích hợp mô-đun phân loại văn bản tự động vào hệ thống quản lý công văn. Giúp cho việc lưu trữ công văn thuận lợi, giúp chọn lọc văn bản phải xử lý một cách dễ dàng và giúp tìm kiếm văn bản một cách nhanh chóng, thuận lợi và hiệu quả.

Đầu vào của hệ thống là công văn đã được số hóa. Đầu ra báo cho người sử dụng biết công văn đó thuộc loại nào và chuyển công văn đó đến thư mục lưu trữ của loại công văn đó.

2.2. MÔ HÌNH ĐỀ XUẤT

Trên cơ sở khảo sát việc phân loại văn bản tại trường Đại học Quảng Bình, chúng tôi đề xuất mô hình xử lý như sau:



Hình 2.1. Mô hình đề xuất

Quá trình phân loại văn bản trong mô hình trên gồm hai bước chính: rút trích đặc trưng để xây dựng mô hình phân loại và sử dụng mô hình đó để phân loại văn bản.

Bước rút trích đặc trưng: Để xây dựng mô hình, trong bước này được chia thành 2 bước nhỏ sau:

- Tiền xử lý: là quá trình xử lý và biểu diễn văn bản thành một dạng biểu diễn lô-gic có thể xử lý được. Trong bước này thực hiện các công việc: Tách từ tiếng Việt, Loại bỏ từ dừng, Biểu diễn văn bản.
- Huấn luyện: Sử dụng thuật toán phân loại văn bản để xây dựng mô hình từ dữ liệu đã qua bước tiền xử lý.

Bước phân loại văn bản: Sử dụng mô hình được xây dựng ở bước trước để phân loại văn bản.

2.3. GIẢI PHÁP ĐỀ XUẤT

2.3.1. Tách từ tiếng Việt

Trong phần này, chúng tôi sử dụng Công cụ JVNSegmenter trong phần mềm JVnTextPro-v.2.0 để thực hiện tách từ tiếng Việt. Đây là phần mềm mã nguồn mở được phát triển bởi hai tác giả Nguyễn Cẩm Tú và Phan Xuân Hiếu (Trường đại học Công nghệ, Đại học Quốc gia Hà Nội).

2.3.2. Loại từ dừng

Ta thấy trong ngôn ngữ tự nhiên có nhiều từ chỉ dùng để biểu diễn cấu trúc câu chứ không biểu đạt nội dung của nó, như các giới từ, liên từ,... Những từ như vậy xuất hiện nhiều trong các văn bản mà không liên quan gì tới chủ đề hoặc nội dung của văn bản, những từ như vậy được gọi là những từ dừng. Việc loại bỏ các từ này, đồng

nghĩa với việc giảm số chiều của văn bản, tăng độ chính xác và tốc độ xử lý văn bản.

2.3.3. Chọn thuộc tính

Mặc dù ta đã thực hiện việc loại bỏ từ dừng nhưng tập các từ trong tập dữ liệu huấn luyện thường vẫn còn rất lớn. Trong đó nhiều từ không mang thông tin, không ảnh hưởng đến nội dung của văn bản, hoặc nhập nhằng, hoặc bị nhiễu, có thể ảnh hưởng không tốt đến kết quả của việc phân loại văn bản. Lựa chọn thuộc tính là quá trình chọn ra những từ mang nhiều thông tin, ảnh hưởng đến nội dung của văn bản, trong tập dữ liệu huấn luyện và loại bỏ các từ gây nhiễu.

Phương pháp lọc (filtering) là phương pháp này sử dụng kỹ thuật tính toán thống kê để loại bỏ các từ không thích hợp được xây dựng để áp dụng cho quá trình lựa chọn thuộc tính [19]. Trong luận văn, độ đo thông tin qua lại được sử dụng để lựa chọn tập thuộc tính dựa vào mô hình thống kê.

2.3.4. Biểu diễn văn bản

Mô hình tần suất, ma trận $W = \{w_{ij}\}$ được xác định dựa trên tần số xuất hiện của từ t_i trong văn bản d_j hoặc tần số xuất hiện của từ t_i trong toàn bộ cơ sở dữ liệu. Trong mô hình này có 3 phương pháp phổ biến là phương pháp dựa trên tần số từ (TF – Term Frequency), phương pháp dựa trên nghịch đảo tần số văn bản (IDF – Inverse Document Frequency) và phương pháp TF × IDF.

Phương pháp TF × IDF là tổng hợp của hai phương pháp TF và IDF, giá trị của ma trận trọng số được tính như sau:

$$w_i = \begin{cases} \left[1 + \log(f_{ij})\right] \log\left(\frac{m}{h_i}\right) & \text{nếu } f_{ij} \geq 1 \\ 0 \end{cases} \qquad (2.21)$$

2.3.5. Phương pháp mạng Nơ-ron khởi tạo với cây quyết định

Trong luận văn này trình bày phương pháp lai cây quyết định và mạng Nơ-ron cho vấn đề phân loại văn bản và áp dụng phương pháp này vào việc xây dựng ứng dụng phân loại công văn.

Thuật toán xây dựng cây quyết định

Gọi T là tập hợp các trường hợp liên quan tại nút. Tần số xuất hiện freg(C_i, T) được tính của các trường hợp T có lớp là C_i, i ∈ [1, N]. Nếu tất cả các trường hợp trong T thuộc về cùng một lớp C_i (hoặc số các trường hợp trong T nhỏ hơn một giá trị nhất định) thì nút đó là nút lá, thuộc lớp C_i.

Nếu T_1, ..., T_s là tập con của T và T chứa các trường hợp thuộc hai hoặc nhiều hơn các lớp, sau đó thông tin đạt được của mỗi thuộc tính được tính:

$$I = H(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \times H(T_i) \qquad (2.22)$$

Trong đó:

$$H(T) = -\sum_{j=1}^{n} \frac{freg(C_j, T)}{|T|} \times \log_2 \left(\frac{freg(C_j, T)}{|T|}\right) \qquad (2.23)$$

là hàm entropy.

Trong khi có một tùy chọn để chọn thông tin thu được, theo mặc định, tuy nhiên, C4.5 gồm xem xét tỷ lệ thông tin thu được của các tập con T_1, ..., T_s đó là tỷ lệ tăng thông tin của nó:

$$S(T) = -\sum_{i=1}^{s} \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{T}\right) \qquad (2.24)$$

Đào tạo mạng Nơ-ron đa lớp

Các Nơ-ron trong các lớp đầu vào chỉ hành động như bộ đệm để phân phối các tín hiệu đầu vào x_i cho các Nơ-ron trong lớp ẩn. Mỗi Nơ-ron j trong lớp ẩn tổng hợp các tín hiệu đầu vào x_i của nó sau khi đánh trọng số chúng với những thế mạnh của các kết nối tương ứng w_{ji} từ lớp đầu vào và tính đầu ra của nó y_j như là một hàm f của tổng như sau:

$$y_j = f(\sum w_{ij} x_i) \qquad (2.25)$$

trong đó f có thể là một hàm ngưỡng đơn giản, một hàm tuyến hoặc một hàm sigmoid.

Có nhiều thuật toán học, thuật toán lan truyền ngược được chấp nhật với thuật toán huấn luyện mạng Nơ-ron [27]. Thuật toán lan truyền ngược sử dụng kỹ thuật mô tả độ dốc để chấp nhận các trọng số mạng Nơ-ron để giảm đến mức tối thiểu độ sai lệch bình phương trung bình giữa đầu ra của mạng Nơ-ron nhận tạo và đầu ra mong đợi. Lỗi bình phương trung bình của Nơ-ron đầu ra trên tất cả các ví dụ n là độ sai lệch giữa mục tiêu t và đầu ra mạng thực tế a:

$$mse = \frac{1}{n}\sum_{i}^{n}(t(i) - a(i))^2 \qquad (2.26)$$

Thay đổi trong trọng số $\Delta w_{ij}(k)$ giữa các Nơ-ron i và j như sau:

$$\Delta w_{ij}(k) = \eta \delta_j x_i + \alpha \Delta w_{ij}(k-1) \qquad (2.27)$$

Trong đó η là một tham số được gọi là hệ số học, α là hệ số động lực, và δ_j là một yếu tố phụ thuộc vào sản lượng Nơ-ron hoặc Nơ-ron ẩn.

$$\delta_j = \frac{\partial f}{\partial net_j}(y_j - y_{net_j}) \qquad (2.28)$$

Trong đó $net_j = \sum_i x_i w_{ij}, y_j, y_{net_j}$ là đích và các Nơ-ron đầu ra cho Nơ-ron j.

Với các Nơ-ron ẩn:

$$\delta_j = \frac{\partial f}{\partial net_j} \sum_q w_{qj} \delta_q \qquad (2.29)$$

Như không có đầu ra mục tiêu cho các Nơ-ron ẩn trong (2.25), sự khác biệt giữa đầu ra mục tiêu và Nơ-ron đầu ra thực tế của Nơ-ron ẩn j được thay thế bởi tổng các trọng số của giới hạn δq (terms) đã đạt được cho các Nơ-ron q kết nối đến đầu ra của j. Do đó, một cách lặp đi lặp lại, bắt đầu với lớp ra, giới hạn δ được tính cho các Nơ-ron trong tất cả các lớp và trọng số cập nhật được xác định hci các tất cả các kết nối theo (2.26).

Huấn luyện mạng Nơ-ron bằng cách lan truyền ngược với thuật toán huấn luyện (momentum) để tính y liên quan đến việc trình bày nó một cách tuần tự với các tập huấn luyện khác nhau. Sự khác biệt giữa đầu ra mục tiêu và đầu ra thực tế của mạng Nơ-ron được truyền ngược thông qua mạng để làm lại cho thích hợp các trọng số của nó sử dụng (2.26) - (2.28). Việc sửa lại cho hợp được thực hiện sau trình bày của mỗi tập. Mỗi lần huấn luyện được hoàn thành sau khi tất cả các mẫu trong tập huấn luyện đã được áp dụng cho các mạng.

Mạng Nơ-ron khởi tạo với cây quyết định

Nếu chúng ta so sánh cây quyết định với mạng Nơ-ron chúng ra có thể thấy rằng các lợi thế và các hạn chế của chúng gần như được bổ sung cho nhau. Ví dụ chúng ta dễ dàng hiểu được biểu diễn tri thức của cây quyết định, mà không phải là trường hợp cho mạng Nơ-ron. Cây quyết định gặp khó khăn trong việc đối phó nhiễu trong

dữ liệu huấn luyện, một lần nữa không phải là trường hợp của mạng Nơ-ron, cây quyết định học rất nhanh và mạng Nơ-ron học tương đối chậm,... Ý tưởng của phương pháp này là kết hợp cây quyết định và mạng Nơ-ron để kết hợp các lợi thế của chúng.

Trước hết chúng ta xây dựng một cây quyết định mà sau đó được sử dụng để khởi tạo một mạng Nơ-ron. Như thế một mạng sau đó được huấn luyện sử dụng các đối tượng huấn luyện tương tự.

Cây quyết định nguồn được chuyển sang dạng chuẩn rời rạc, là một tập các luật chuẩn hóa. Sau đó dạng chuẩn rời rạc phục vụ như nguồn để xác định cấu trúc liên kết và các trọng số của mạng Nơ-ron. Mạng Nơ-ron có hai lớp ẩn. Số lượng các Nơ-ron trên mỗi lớp ẩn phụ thuộc vào các luật trong dạng chuẩn rời rạc.

Số lượng các Nơ-ron trong lớp đầu ra phụ thuộc vào bao nhiêu kết quả có thể xảy ra trong tập huấn luyện. Việc chuyển đổi được mô tả trong các bước tiếp theo:

1. Xây dựng cây quyết định sử dụng từ (2.22) – (2.24).
2. Mỗi đường đi từ gốc của cây đến mỗi lá được thể hiện như một luật.
3. Tập các luật nếu được chuyển thành dạng chuẩn rời rạc, đó là đại diện của tập luật ban đầu.
4. Trong lớp đầu vào tạo các Nơ-ron là các thuộc tính trong tập huấn luyện.
5. Đối với mỗi chữ trong dạng chuẩn rời rạc có một Nơ-ron được tạo ra trong lớp ẩn đầu tiên của một mạng Nơ-ron.
6. Tập trọng số cho mỗi Nơ-ron trong lớp chữ (lớp ẩn 1), đại diện cho một chữ trong dạng (thuộc tính > giá

trị) với $w_0 = -\sigma * value$, với mỗi chữ dạng (attribute <= value) với $w_0 = \sigma * value$. Đặt tất cả các trọng số còn lại cho +β hoặc −β với xác suất bằng nhau. Hằng σ =5 và hằng β =0.025. Những giá trị này được xác định bằng cách qua xác nhận trên bộ dữ liệu nhân tạo.

7. Với mọi liên kết của các chữ (literal) tạo ra một Nơ-ron trong lớp ẩn thứ hai (lớp liên tục).

8. Thiết lập trọng số mà liên kết mỗi Nơ-ron trong lớp liên tục với Nơ-ron thích hợp trong lớp chữ (literal) để $w_0 = \sigma * (2n - 1)/2$, trong đó n là số chữ (literal) trong liên kết. Đặt tất cả các trọng số còn lại bằng + β hoặc −β với xác suất bằng nhau.

9. Với mỗi lớp có thể tạo một Nơ-ron trong lớp đầu ra (lớp phân biệt).

10. Thiết lập trọng số mà liên kết mỗi Nơ-ron trong lớp phân biệt với Nơ-ron thích hợp trong lớp liên tục $w_0 = -\sigma * (1/2)$. Đặt tất cả các trọng số còn lại vào + β hoặc -β với xác suất bằng nhau.

11. Huấn luyện mạng Nơ-ron với các đối tượng đào tạo tương tự như đã được sử dụng để huấn luyện cây quyết định, sử dụng (2.25) – (2.29).

Như thế mạng sau đó được huấn luyện sử dụng lan truyền ngược. Nghĩa là sai số trung bình bình phương của mạng như trên hội tụ hướng về 0 nhanh hơn nó trong trường hợp các trọng số thiết lập ngẫu nhiên trong mạng.

CHƯƠNG 3
PHÁT TRIỂN ỨNG DỤNG

Chương này sẽ thực hiện xây dựng chương trình ứng dụng phân loại công văn với những thuật toán đã được trình bày ở chương 2. Sau khi xây dựng chương trình sẽ thực hiện huấn luyện và phân loại văn bản.

3.1. CÔNG CỤ PHÁT TRIỂN

3.1.1. Công cụ phát triển chương trình

Chương trình được xây dựng trên nền JDK 1.7.0 với công cụ hỗ trợ lập trình NetBean 7.1.1.

Chương trình sử dụng công cụ JVNSegmenter thuộc phần mềm JVnTextPro-v.2.0 để thực hiện tách từ tiếng Việt. Đây là phần mềm mã nguồn mở được phát triển bởi hai tác giả Nguyễn Cẩm Tú và Phan Xuân Hiếu (Trường đại học Công nghệ, Đại học Quốc gia Hà Nội).

Ngoài công cụ trên, chương trình được xây dựng với các mô-dun xử lý sau:

- Mô-dun Tách câu dùng để tách các đoạn văn bản khi gặp dấu câu.
- Mô-dun Loại từ dừng để loại bỏ các từ dừng, từ không ảnh hưởng đến nội dung của văn bản, chuỗi số và các ký hiệu đặc biệt.
- Mô-dun Lựa chọn đặc trưng để tạo các tập tin chứa các từ đặc trưng của các văn bản.
- Mô-dun Biểu diễn văn bản để tạo vector trọng số của các từ đặc trưng của từng nhóm văn bản.

- Mô-dun Xây dựng Cây quyết định và Huấn luyện Mạng Nơ-ron để phục vụ quá trình trích rút đặc trưng.
- Mô-dun Phân loại văn bản dùng để phân loại các văn bản.

3.1.2. Chuẩn bị dữ liệu

Chuẩn bị dữ liệu huấn luyện và kiểm tra gồm:

Bảng 3.1. Dữ liệu huấn luyện và kiểm tra

Chủ đề	Số tập tin huấn luyện	Số tập tin kiểm tra
Đoàn thanh niên	100	40
Công đoàn	100	40
Đảng	100	40

Các file dữ liệu huấn luyện và kiểm tra đều có font Unicode và định dạng file văn bản UTF-8 có phần mở rộn tên file là .txt.

3.2. PHÁT TRIỂN CÁC MODULE CHƯƠNG TRÌNH

3.2.1. Tách câu

Đầu vào là nội dung văn bản. Đầu ra là các câu đã được tách.

Trong phần này luận văn thực hiện tách câu bằng cách đơn giản là duyệt từ đầu đến cuối văn bản, gặp các dấu câu như {, . ? ! ; : + / ( ) *} thì loại các dấu đó khỏi văn bản và tách đoạn văn bản từ đầu đến dấu câu đó thành một câu và bắt đầu lại với đoạn còn lại của văn bản sau khi tách đoạn câu đầu tiên.

3.2.2. Tách từ

Đầu vào là các câu trong văn bản đã được tách ở mô-dun tách câu. Đầu ra là tập các từ đã tách. Như đã trình bày ở chương 2, luận văn sử dụng Công cụ JVNSegmenter trong phần mềm JVnTextPro-v.2.0 để thực hiện tách từ tiếng Việt.

3.2.3. Loại bỏ từ dừng

Đầu vào là tập các từ của văn bản đã được tách ở mô-dun tách từ. Đầu ra là tập các từ đã được loại bỏ bớt các từ dừng không ảnh hưởng đến nội dung của văn bản. Mô-dun loại bỏ từ dừng kiểm tra và loại bỏ các từ có tính chất kết nối, mô tả không có ý nghĩa trong văn bản và chỉ giữ lại những từ có ý nghĩa nhất. Loại bỏ từ dừng bằng cách xây dựng một từ điển từ dừng, ta duyệt từ đầu văn bản nếu từ đó có trong từ điển từ dừng thì ta loại từ đó ra khỏi văn bản.

3.2.4. Biểu diễn văn bản

Đầu vào là tập các từ đặc trưng của văn bản cần biểu diễn. Đầu ra là tập trọng số của các từ. Trong luận văn này chọn mô hình không gian vector biểu diễn giá trị bằng mô hình tấn suất được tính bằng phương pháp TF × IDF.

3.3. HUẤN LUYỆN VÀ PHÂN LOẠI

3.3.1 Huấn luyện

Trong phần này các văn bản huấn luyện đã được thực hiện bước tiền xử lý, được lưu theo các folder thuộc loại huấn luyện. Việc huấn luyện chúng ta thực hiện theo 2 bước.

Bước 1: Xây dựng Cây quyết định: Đầu vào của bước này là tập dữ liệu huấn luyện, tập các thuộc tính. Đầu ra của bước này là cây quyết định được xây dựng từ tập huấn luyện, tập các thuộc tính và tập luật được sinh ra từ cây quyết định.

Bước 2: Huấn luyện với mạng Nơ-ron: Đầu vào của bước này là tập luật được xây dựng bởi cây quyết định ở bước 1. Đầu ra là mạng Nơ-ron được xây dựng từ tập luật của bước 1. Mạng Nơ-ron được xây dựng này được sử dụng để phân lớp cho các văn bản cần

phân loại. Kết quả của bước này là một mạng Nơ-ron được xây dựng với các trọng số của các Nơ-ron hợp lý.

3.3.2. Phân loại văn bản

Đầu vào là văn bản cần phân loại. Đầu ra thông báo cho người dùng văn bản thuộc loại nào và lưu văn bản vào thư mục chứa loại văn bản đó. Phân loại một văn bản mới ta thực hiện lần lượt các việc sau: Đọc nội dung của văn bản, tách câu, tách từ, loại từ dừng, và thực hiện phân loại.

3.4. KẾT QUẢ THỬ NGHIỆM

Chương trình ứng dụng sau khi đã huấn luyện, ta thực hiện thử nghiệm với tập văn bản kiểm tra để nhận dạng cho các băn bản đó.

Bảng 3.2. Dữ liệu kiểm tra

STT	Nhóm	Số VB được phân loại
1.	Đoàn thanh niên	40
2.	Công đoàn	40
3.	Đảng	40

Kết quả kiểm tra với nhóm văn bản Đoàn thanh niên:

Bảng 3.3. Kết quả thử nghiệm văn bản Đoàn thanh niên

Nội dung	Số lượng tập tin
Số tập tin huấn luyện	100
Số tập tin kiểm tra	40
Số tập tin phân loại đúng	32
Số tập tin phân loại sai	5
Số tập tin không nhận dạng được	3

Ta thấy nhóm văn bản Đoàn thanh niên phân loại được với tỷ lệ 32/40 đạt 80%.

Kết quả kiểm tra với nhóm văn bản Công đoàn:

Bảng 3.4. Kết quả thử nghiệm văn bản Công đoàn

Nội dung	Số lượng tập tin
Số tập tin huấn luyện	100
Số tập tin kiểm tra	40
Số tập tin phân loại đúng	34
Số tập tin phân loại sai	3
Số tập tin không nhận dạng được	2

Ta thấy nhóm văn bản Đoàn thanh niên phân loại được với tỷ lệ 34/40 đạt 85%.

Kết quả kiểm tra với nhóm văn bản Đảng:

Bảng 3.5. Kết quả thử nghiệm văn bản Đảng

Nội dung	Số lượng tập tin
Số tập tin huấn luyện	100
Số tập tin kiểm tra	40
Số tập tin phân loại đúng	33
Số tập tin phân loại sai	3
Số tập tin không nhận dạng được	4

Ta thấy nhóm văn bản Đoàn thanh niên phân loại được với tỷ lệ 33/40 đạt 82.5%.

3.5. ĐÁNH GIÁ

Kết quả thực nghiệm như trên chưa cao tuy nhiên với kết quả này chúng ta chấp nhận được. Điều này có thể hoàn toàn giải thích được do các nguyên nhân sau:

- Số lượng văn bản trong tập huấn luyện chưa nhiều.

- Ba chủ đề dùng để phân loại có sự giao thoa về dữ liệu do đó có sự nhập nhằng trong quá trình phân loại.

KẾT LUẬN VÀ KIẾN NGHỊ

Luận văn đã trình bày khái quát về bài toán phân loại văn bản, khái quát về một số phương pháp phân loại văn bản như Naïve Bayes, k Nearest Neighbor, Support Vector Machine,…, khái quát một số phương pháp tách từ tiếng Việt. Trong đó luận văn chú trọng nghiên cứu phương pháp phân loại văn bản sử dụng mạng Neural kết hợp với Cây quyết định.

Từ việc nghiên cứu thuật toán phân loại văn bản bằng phương pháp sử dụng mạng Neural kết hợp cây quyết định, tác giả đã áp dụng xây chương trình dựng ứng dụng phân loại công văn với ba loại công văn là: công văn của Đảng, công văn của công đoàn và công văn của Đoàn thanh niên.

Tuy nhiên, với thời gian hạn chế nên chương trình được xây dựng với các tính năng còn thô sơ để mô phỏng thuật toán chứ chưa đưa ra được một chương trình ứng dụng hoàn hảo. Mặc dù kết quả phân loại với xác suất phân loại còn thấp, chưa thật sự nổi trội hơn so với các chương trình khác, với dữ liệu của ba loại công văn này có sự giao thoa về dữ liệu do đó vẫn còn sự nhập nhằng trong phân loại nhưng với kết quả này vẫn chấp nhận được.

Hạn chế của chương trình:

Chương trình chỉ nhận dạng và phân loại cho văn bản định dạng file text .TXT, chưa nhận dạng và phân loại cho các văn bản có định dạng khác như định dạng .DOC, PDF. Tập dữ liệu huấn luyện còn hạn chế, do đó độ chính xác của chương trình còn chưa cao. Các tính năng sử dụng của chương trình còn thô sơ, chưa hợp lý và chưa khoa học. Chương trình mới chỉ có thể áp dụng cho 3 nhóm chủ đề xác định trước, chưa mở rộng cho các nhóm chủ đề mới.

Các hướng cải tiến chương trình:

Xây dựng thêm mô-dun chuyển đổi văn bản từ định dạng word và định dạng pdf sang dạng text để thực hiện phân loại.

Xây dựng bộ dữ liệu huấn luyện nhiều hơn để có độ chính xác cao hơn. Nghiên cứu và xây dựng bộ phân tích ngữ nghĩa tiếng Việt để tăng mức độ chính xác cho việc tách từ và rút trích từ đặc trưng nhằm tăng độ chính xác của việc phân loại. Chương trình chỉ mới ứng dụng phân loại cho 3 loại công văn, tuy nhiên trong công việc hằng ngày ở đơn vị nhận và gửi đi nhiều chủ đề công văn hơn. Hướng nghiên cứu tiếp theo là xây dựng chương trình mở rộng để có thể phân loại công văn theo nhiều chủ đề hơn và tăng cường tính tiện dụng cho chương trình.