

XÂY DỰNG HỆ THỐNG PHÂN LOẠI TÀI LIỆU TIẾNG VIỆT

Trần Thị Thu Thảo, Vũ Thị Chinh

Khoa Công nghệ thông tin, trường Đại học Lạc Hồng

Liên hệ với tác giả: tranthithuthao23121990@gmail.com, chinhvu060690@gmail.com

TÓM TẮT

Bài báo trình bày cách thức phân loại tài liệu tiếng Việt theo chín chuyên ngành thuộc lĩnh vực Công nghệ thông tin dựa trên việc phân chia chuyên ngành của các hội thảo như hội thảo Fair, hội thảo @ Cần Thơ. Hệ thống phân loại được xây dựng dựa trên một quy trình chặt chẽ mà bài báo đề xuất với việc áp dụng phương pháp Naïve Bayes. Bước đầu thử nghiệm trên các bài báo khoa học thuộc lĩnh vực Công nghệ thông tin đã cho những kết quả có độ chính xác cao so với yêu cầu.

Từ khóa: Phân loại, văn bản, trọng số từ, quy trình, từ đặc trưng, Naïve Bayes.

1. GIỚI THIỆU

Trong thời đại bùng nổ Công nghệ thông tin hiện nay, phương thức sử dụng giấy tờ trong giao dịch đã dần được số hoá chuyển sang các dạng văn bản lưu trữ trên máy tính hoặc truyền tải trên mạng. Bởi nhiều tính năng ưu việt của tài liệu số như: cách lưu trữ gọn nhẹ, thời gian lưu trữ lâu dài, tiện dụng trong trao đổi đặc biệt là qua Internet, dễ dàng sửa đổi... nên ngày nay, số lượng văn bản số tăng lên một cách chóng mặt đặc biệt là trên world-wide-web. Cùng với sự gia tăng về số lượng văn bản, nhu cầu tìm kiếm văn bản cũng tăng theo. Với số lượng văn bản đồ sộ thì việc phân loại văn bản tự động là một nhu cầu bức thiết.

Do vậy, các phương pháp phân loại văn bản tự động đã ra đời để phục vụ cho nhu cầu chính đáng đó.

Đề tài tìm hiểu một số cách phân loại tài liệu và thử nghiệm một phương pháp phân loại áp dụng thuật toán Naïve Bayes để xây dựng chương trình dựa trên tập dữ liệu huấn luyện từ đó hướng đến việc phân loại các bài báo khoa học trong lĩnh vực Công nghệ thông tin nhằm tiết kiệm thời gian và công sức cho các nhà tổ chức trong các hội thảo như hội thảo Fair, hội thảo @ Cần Thơ.

1. ĐẶT VẤN ĐỀ

1.1 Tổng quan

Từ trước đến nay, phân loại văn bản tự động đã có rất nhiều công trình nghiên cứu và đạt được kết quả đáng khích lệ. Dựa trên các thống kê của Yang & Xiu (1999)[7] một số

phương pháp phân loại thông dụng hiện nay là: *Support Vector Machine* -Joachims, 1998[6], *Linear Least Squares Fit* -Yang and Chute, 1994[8]...vv. Các phương pháp trên đều dựa vào xác suất thống kê hoặc thông tin về trọng số của từ trong văn bản.

Vấn đề phân loại văn bản tiếng Việt được nhiều cơ sở nghiên cứu trong cả nước quan tâm trong những năm gần đây. Một số công trình nghiên cứu cũng đạt được những kết quả khả quan. Các hướng tiếp cận bài toán phân loại văn bản đã được nghiên cứu bao gồm: hướng tiếp cận bài toán phân loại bằng lý thuyết đồ thị [1], cách tiếp cận sử dụng lý thuyết tập thô [4], cách tiếp cận thống kê [5], cách tiếp cận sử dụng phương pháp học không giám sát và đánh chỉ mục [2, 3]. Nhìn chung, những cách tiếp cận này đều cho kết quả chấp nhận được. Tuy vậy để đi đến những biện pháp triển khai khả thi thì vẫn cần đẩy mạnh nghiên cứu

1.2 Mục Tiêu

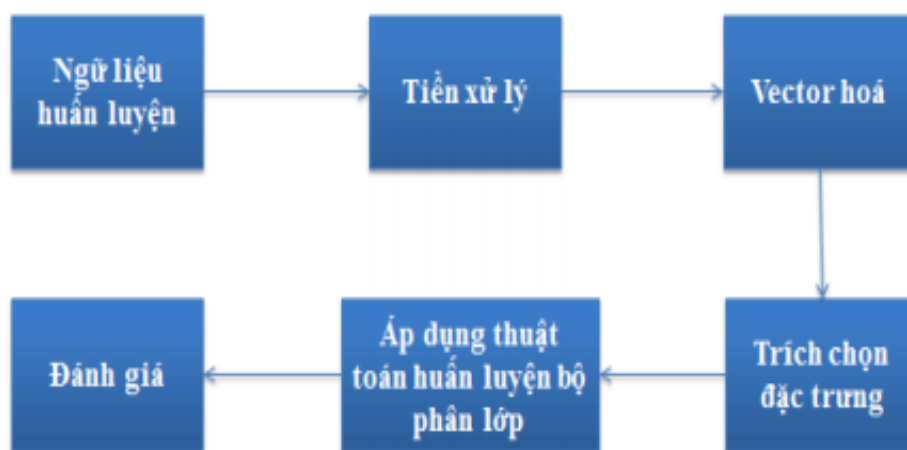
Tìm hiểu thuật toán Naïve Bayes ứng dụng vào xây dựng một chương trình phân loại văn bản tiếng Việt, bước đầu ứng dụng vào việc phân loại các bài báo khoa học điện tử thuộc lĩnh vực CNTT trong các hội thảo như: Hội thảo Fair, hội thảo @ Cần Thơ.

2. NỘI DUNG NGHIÊN CỨU

Đề tài này áp dụng phương pháp Naïve Bayes, thực hiện trên đối tượng là bài báo khoa học thuộc chín chuyên ngành trong lĩnh vực Công nghệ thông tin, nên đề tài sẽ tập trung khảo sát cấu trúc các loại tài liệu, đưa ra các số liệu thống kê về việc phân loại bài báo. Ngoài ra còn cho phép huấn luyện những đề tài theo ý muốn của người sử dụng khi có tập dữ liệu chuẩn. Kết nối cơ sở dữ liệu cho phép người dùng có thể thao tác thêm sửa xóa dữ liệu như: từ điển, từ phổ thông, chuyên ngành...vv.

2.1 Quy trình xử lý

Phân loại văn bản gồm các bước xử lý chung được thể hiện như sau:



Hình 1: Chi tiết giai đoạn huấn luyện

Trong đó:

- Ngữ liệu huấn luyện: kho ngữ liệu thu thập từ nhiều nguồn khác nhau.
- Tiền xử lý: chuyển đổi tài liệu trong kho ngữ liệu thành một hình thức phù hợp để phân loại.
- Vector hoá: mã hoá văn bản bởi một mô hình trọng số.
- Trích chọn đặc trưng: loại bỏ những từ (đặc trưng) không mang thông tin 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.
- Đánh giá: bước đánh giá hiệu suất (chất lượng) của bộ phân lớp.

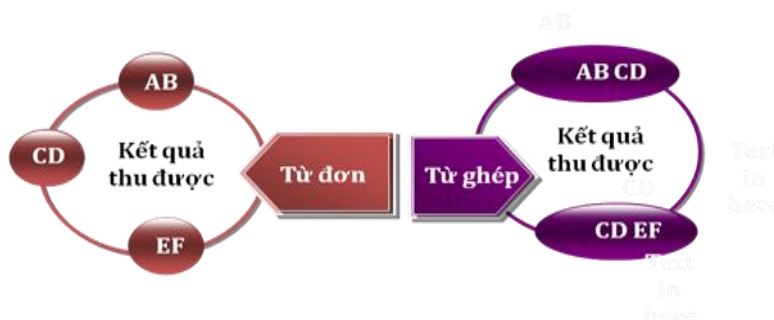
3.2 Phương pháp tách từ

Đề tài đã sử dụng phương pháp tách từ bằng N-gram

Hướng tiếp cận dựa trên nhiều ký tự là chia văn bản ra thành nhiều chuỗi, mỗi chuỗi gồm n từ.

Ưu điểm của phương pháp là tính đơn giản và dễ ứng dụng, ít tốn chi phí cho việc tạo chỉ mục và xử lý nhiều câu truy vấn Đề tài sử dụng mô hình n-gram với $n=2$.

Ví dụ minh họa: ta có một câu gồm 3 từ đơn “AB CD EF”



Hình 2: Quy trình tách từ

3.3 Phương pháp tính trọng số của từ Tf*idf weighting

Để tính trọng số của từ có rất nhiều phương pháp nhưng ở đây đề tài sử dụng phương pháp Tf*idf weighting.

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 (stopword)

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

Bước 3: Tính trọng số của từ theo công thức:

$$\text{Weight}_{wi} = \text{tf} * \text{idf}$$

Với:

$$\text{tf} = N_s(t) / \sum w$$

$$\text{idf} = \log(\sum d / (d:t \in d))$$

Trong đó:

- N_s : Số lần xuất hiện của từ t trong tài liệu f .

- $\sum w$: Tổng số các từ trong tài liệu f .

- $\sum d$ = tổng số tài liệu.

- $d:t \in d$: số tài liệu có chứa từ.

Ví dụ minh họa phương pháp tính trọng số:

Có một văn bản gồm 100 từ, trong đó từ “máy tính” xuất hiện 10 lần thì độ phổ biến:

$$\text{tf}(\text{“máy tính”}) = 10 / 100 = 0.1.$$

Bây giờ giả sử có 1000 tài liệu đã được huấn luyện, trong đó có 200 tài liệu chứa từ

“máy tính”. Lúc này ta sẽ tính được:

$$\text{idf}(\text{“máy tính”}) = \log(1000 / 200) = 0.699$$

Như vậy ta tính được độ đo:

$$\text{TF.IDF} = \text{tf} * \text{idf} = 0.1 * 0.699 = 0.0699$$

3.4 Phương pháp Naïve Bayes

Ta áp dụng phương pháp Naïve bayes vào chương trình phân loại với cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa 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. Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề.

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

- D : tập dữ liệu huấn luyện đã được vector hóa dưới dạng $\vec{x} = (x_1, x_2, \dots, x_n)$
- C_i : phân lớp i , với $i = \{1, 2, \dots, m\}$.

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

Theo định lý Bayes:

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

Theo tính chất độc lập điều kiện:

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

Trong đó:

- $P(C_i|X)$ là 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 .

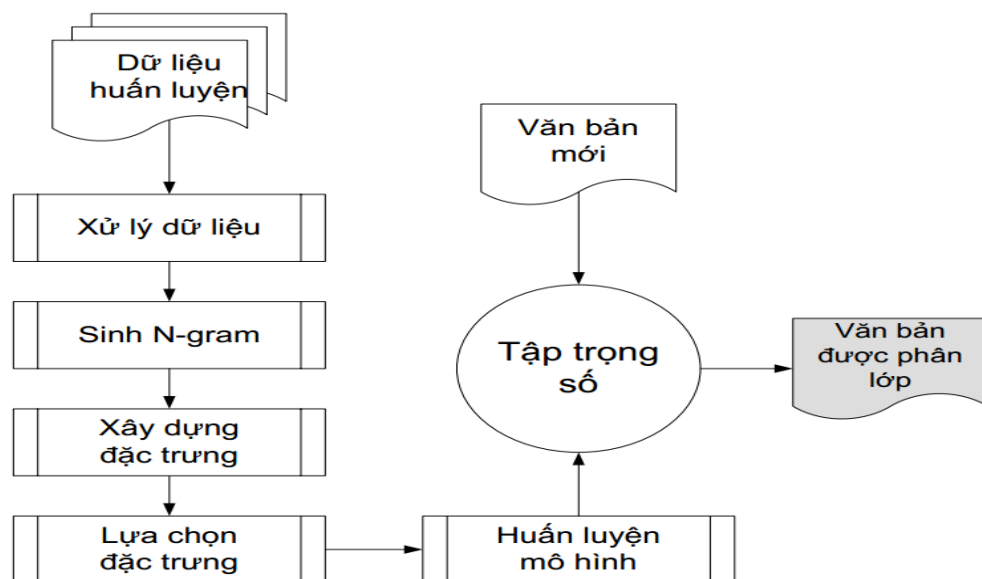
Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính $P(C_i)$ và $P(x_k|C_i)$

Bước 2: Phân lớp $X^{new} = (x_1, x_2, \dots, x_n)$, ta cần tính xác suất thuộc từng phân lớp khi đã biết trước X^{new} . X^{new} được gán vào lớp có xác suất lớn nhất theo công thức.

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

Mô hình tổng quát việc phân loại:



Hình 3. Mô tả bước xây dựng bộ phân lớp

Ví dụ minh họa: Cho hai tập dữ liệu gồm: Tài liệu 1 thuộc lớp Máy tính và Tài liệu 2 thuộc lớp Toán với từ được xét là: Công nghệ và Công thức.

Tập dữ liệu	Công nghệ	Công thức	Lớp
Tài liệu 1	2	1	Máy tính
Tài liệu 2	1	3	Toán

Ta có:

Tổng số từ thuộc lớp:	Máy tính	Toán
	3	4

Ước lượng xác suất	Bước Huấn C_i = “Máy tính”	Bước Huấn C_i = “Toán”
$P(C_1 = \text{“Máy tính”})$ $= 1/2$	$P(\text{“Công nghệ”} \text{Máy tính})$ $= 2/3$	$P(\text{“Công nghệ”} \text{Toán}) =$ $1/4$
$P(C_2 = \text{“Toán”})$ $= 1/2$	$P(\text{“Công thức”} \text{Máy tính})$ $= 1/3$	$P(\text{“Công thức”} \text{Toán}) =$ $3/4$


Khi ta xét một tập Tài liệu 3 với dữ liệu sau: hỏi xem Tài liệu 3 thuộc lớp nào trong hai lớp Máy tính và Toán

Tập dữ liệu	Công nghệ	Công thức	Lớp
Tài liệu 3	1	2	?

Dựa vào công thức tính ta có:

$$C_{nb} = P(\text{Máy tính}). [P(\text{“Công nghệ”} | \text{Máy tính}) * 1 * P(\text{“Công thức”} | \text{Máy tính}) * 2] = 0.222$$

$$C_{nb} = P(\text{Toán}). [P(\text{“Công nghệ”} | \text{Toán}) * 1 * P(\text{“Công thức”} | \text{Toán}) * 2] = 0.1875$$

 Tài liệu 3 thuộc lớp Máy tính

3. KẾT QUẢ THỬ NGHIỆM

3.1 Dữ liệu đầu vào

Để kiểm tra độ chính xác chương trình ta xét tập dữ liệu đã phân loại và được thu thập được của trường Đại học Khoa Học Tự Nhiên, hội thảo Fair, @ Cần Thơ gồm chuyên đề sau:

Bảng 1: Bảng số liệu xử lý theo con người

Thống kê bài báo		
Stt	Tập dữ liệu	Số lượng
1	Các hệ thống tính toán di động	23
2	Công nghệ đa phương tiện	34
3	Công nghệ phần mềm	32
4	Cơ sở toán học của công nghệ thông tin	25
5	Hệ thống thông tin	40
6	Khoa học máy tính	26
7	Mạng máy tính và truyền thông	31
8	Trí tuệ nhân tạo	28
9	Xử lý ngôn ngữ tự nhiên và tiếng nói	42

3.2 Kết quả phân lớp

Sau khi phân loại và so sánh với kết quả có sẵn ta thu được kết quả phần trăm trung bình là 87,37%

Bảng 2: Tỷ lệ (%) phân loại văn bản

Bảng đánh giá kết quả phân loại văn bản					
Stt	Tập dữ liệu	Phân loại bởi con người	Phân loại bằng máy	Phân loại sai chuyên ngành	Tỉ lệ (%)
1	Các hệ thống tính toán di động	23	20	3	86.95
2	Công nghệ đa phương tiện	34	30	4	88.23
3	Công nghệ phần mềm	32	28	4	87.5
4	Cơ sở toán học của công nghệ thông tin	25	22	3	88
5	Hệ thống thông tin	40	35	5	87.5
6	Khoa học máy tính	26	23	3	88.46
7	Mạng máy tính và truyền thông	31	27	4	87.09
8	Trí tuệ nhân tạo	28	23	5	82.14
9	Xử lý ngôn ngữ tự nhiên và tiếng nói	42	38	4	90.47
Phần trăm trung bình					87.37

Với cách tiếp cận như trên, bài toán phân loại văn bản tiếng Việt về cơ bản đã được giải quyết, đặc biệt là vấn đề phân loại văn bản theo chín chủ đề chuyên ngành Công nghệ thông tin.

4. KẾT LUẬN

Với các yêu cầu đặt ra về việc nắm bắt thuật toán Naïve Bayes để hiểu cách thức phân loại một tài liệu tiếng Việt từ đó áp dụng vào việc phân loại các bài báo khoa học trong lĩnh vực Công nghệ thông tin theo các chuyên ngành khác nhau dựa trên việc khảo sát một số hội thảo CNTT trong nước, chương trình cơ bản đã đáp ứng được các yêu cầu trên. Cùng với đó, chương trình cung cấp thêm một số chức năng giúp cho việc thêm sửa xóa, quản lý các bài báo một cách dễ dàng và thuận tiện.

Sau thời gian thực hiện đề tài chúng em đã hoàn thành được các công việc cụ thể sau:

- Xây dựng module tách từ theo mô hình n-gram.
- Khảo sát tài liệu thuộc chín chuyên ngành Công nghệ thông tin để tìm hiểu đặc trưng riêng.
- Tìm hiểu về các thuật toán tính trọng số từ đó áp dụng phương pháp Tf*Idf vào chương trình để xác định từ đặc trưng của chuyên ngành.
- Xây dựng form huấn luyện cho phép người dùng huấn luyện văn bản, có thể tạo ra một chuyên ngành mới khi có tập dữ liệu chuẩn.
- Không chỉ cho phép thao tác trên từng bài báo mà còn thao tác trên tập dữ liệu.
- Tìm hiểu sơ lược về các thuật toán phân loại văn bản, ưu nhược điểm của thuật toán Naïve Bayes so với các thuật toán khác.
- Phân tích nội dung và thiết lập cơ sở dữ liệu để xây dựng phần mềm.
- Xây dựng phần mềm phân loại các bài báo khoa học thuộc lĩnh vực Công nghệ thông tin.

Hướng phát triển của đề tài:

- Xây dựng danh sách hoàn thiện các từ phổ thông, ký tự đặc biệt nhằm loại bỏ các yếu tố gây nhiễu trong quá trình huấn luyện cũng như phân loại văn bản.
- Tiếp tục huấn luyện thêm dữ liệu để bộ từ đặc trưng của chuyên ngành được chính xác hơn.
- Thực hiện thử nghiệm trên số lượng lớn các bài báo chưa được phân loại.

TÀI LIỆU THAM KHẢO

1. Đỗ Bích Diệp, Phân loại văn bản dựa trên mô hình đồ thị, Luận văn cao học, Trường Đại học Tổng hợp New South Wales – Australia, 2004.

2. Đinh Thị Phương Thu, Hoàng Vĩnh Sơn, Huỳnh Quyết Thắng, Phương án xây dựng tập mẫu cho bài toán phân lớp văn bản tiếng Việt, nguyên lý, giải thuật, thử nghiệm và đánh giá kết quả, Tạp chí Khoa học và công nghệ, 2005.
3. Huỳnh Quyết Thắng, Đinh Thị Phương Thu, Tiếp cận phương pháp học không giám sát trong học có giám sát với bài toán phân lớp văn bản tiếng Việt và đề xuất cải tiến công thức tính độ liên quan giữa hai văn bản trong mô hình vector, Kỷ yếu Hội thảo ICT.rda'04, trang 251-261, Hà Nội 2005.
4. Nguyễn Ngọc Bình, Dùng lý thuyết tập thô và các kỹ thuật khác để phân loại, phân cụm văn bản tiếng Việt, Kỷ yếu hội thảo ICT.rda'04, Hà nội 2004
5. Nguyễn Duy Hải, Nguyễn Linh Giang, Mô hình thống kê hình vị tiếng Việt và ứng dụng, Các công trình nghiên cứu, triển khai Công nghệ Thông tin và Viễn thông, Tạp chí Bưu chính Viễn thông, số 1, trang 61-67, tháng 7-1999.
6. Joachims, Text Categorization with Support Vector Machines, Learning with Many Relevant Features, In European Conference on Machine Learning (ECML), 1998
7. Yang and Xin Liu, A re-examination of text categorization methods, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999.
8. Y. Yang and G.Chute, An example-based mapping method for text categorization and retrieval, ACM Transaction on Information Systems(TOIS), 12(3):252-277,1994.