NGHIÊN CỨU LÝ THUYẾT NAIVE BAYES VÀ ỨNG DỤNG PHÂN LOẠI TÀI LIỆU TIẾNG VIỆT TRONG THƯ VIỆN SỐ

Hoàng Anh Công*

Tóm tắt: *Hiện nay, khoa học công nghệ ngày càng phát triển. Các hệ thống thư viện điện tử, thư viện trực tuyến ngày càng được sử dụng rộng rãi, kèm theo đó là các vấn đề liên quan đến phân loại, tìm kiếm chia theo danh mục và gợi ý nội dung đọc Ebook cho người dùng. Với lượng thông tin đồ sộ, một yêu cầu lớn đặt ra là làm sao tổ chức và tìm kiếm thông tin có hiệu quả nhất. Phân loại thông tin là một trong những giải pháp hợp lý cho yêu cầu trên. Nhưng một thực tế là khối lượng thông tin quá lớn, việc phân loại dữ liệu thủ công là điều không tưởng. Hướng giải quyết là một chương trình máy tính tự động phân loại các thông tin trên.*

Từ khóa: *Thư viện số; Phân loại tài liệu tiếng Việt; Thuật toán Naïve Bayes; Lý thuyết Naïve Bayes.*

1. ĐẶT VẤN ĐỀ

Nghiên cứu lý thuyết Naive Bayes và ứng dụng trong phân loại tài liệu tiếng Việt trong thư viện điện tử nhằm tìm hiểu và thử nghiệm các phương pháp phân loại tài liệu áp dụng trên tiếng Việt. Phân loại văn bản (Text classification) là một trong những công cụ khai phá dữ liệu dạng văn bản một cách hữu hiệu, làm nhiệm vụ đưa những tài liệu có cùng nội dung chủ đề giống nhau về cùng một lớp có sẵn. Phân loại tài liệu giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin cần thiết đồng thời có thể lưu trữ các thông tin theo đúng chủ đề (topic) hay lớp (class) dựa trên các thuật toán phân loại.

* Thạc sĩ, Trường Đại học Văn hóa, Thể thao và Du lịch Thanh Hóa.

Trong bài viết này sẽ nhằm giải quyết một số vấn đề chính nâng cao hiệu năng của hệ thống phân loại tài liệu tiếng Việt tự động:

- Phương pháp phân loại tài liệu tiếng Việt tự động có kết hợp với giảm chiều nhằm giảm đi độ phức tạp tính toán, đồng thời tăng độ chính xác của phương pháp đã đề xuất.

- Có ý nghĩa thực tiễn cao trong cuộc sống, hệ thống thực nghiệm được xây dựng dựa trên phương pháp đề xuất mang lại tính ứng dụng hỗ trợ ngày một tốt hơn cho người dùng trên Internet.

2. PHÂN LOẠI TÀI LIỆU TIẾNG VIỆT DỰA TRÊN PHƯƠNG PHÁP NAIVE BAYES

2.1. Lý thuyết Naive Bayes

Trong học máy, phân loại Naive Bayes là một thành viên trong nhóm các phân loại có xác suất dựa trên việc áp dụng định lý Bayes khai thác mạnh giả định độc lập giữa các hàm, hay đặc trưng.

Mô hình Naive Bayes cũng được biết đến với nhiều tên khác nhau ví dụ: Simple Bayes hay independence Bayes hay phân loại Bayes.

Phân loại Naive Bayes được đánh giá cao khả năng mở rộng, đòi hỏi một số thông số tuyến tính trong số lượng các biến (các tính năng/tố dự báo) trong nhiều lĩnh vực khác nhau.

Khái niệm

Một phân loại Naive Bayes dựa trên ý tưởng nó là một lớp được dự đoán bằng các giá trị của đặc trưng cho các thành viên của lớp đó. Các đối tượng là một nhóm (group) trong các lớp nếu chúng có cùng các đặc trưng chung. Có thể có nhiều lớp rời rạc hoặc lớp nhị phân.

Các luật Bayes dựa trên xác suất để dự đoán chúng về các lớp có sẵn dựa trên các đặc trưng được trích rút. Trong phân loại Bayes, việc học được coi như xây dựng một mô hình xác suất của các đặc trưng và sử dụng mô hình này để dự đoán phân loại cho một ví dụ mới.

Biến chưa biết hay còn gọi là biến ẩn là một biến xác suất chưa được quan sát trước đó. Phân loại Bayes sử dụng mô hình xác suất trong đó phân loại là một biến ẩn có liên quan tới các biến đã được

quan sát. Quá trình phân loại lúc này trở thành suy diễn trên mô hình xác suất.

Trường hợp đơn giản nhất của phân loại Naive Bayes là tạo ra các giả thiết độc lập về các đặc trưng đầu vào và độc lập có điều kiện với mỗi một lớp đã cho. Sự độc lập của phân loại Naive Bayes chính là thể hiện của mô hình mạng tin cậy (belief network) trong trường hợp đặc biệt, và phân loại là chỉ dựa trên một nút cha duy nhất của mỗi một đặc trưng đầu vào. Mạng tin cậy này đề cập tới xác suất phân tán P(Y) đối với mỗi một đặc trưng đích Y và $P(X_i|Y)$ đối với mỗi một đặc trưng đầu vào X_i. Với mỗi một đối tượng, dự đoán bằng cách tính toán dựa trên các xác suất điều kiện của các đặc trưng quan sát được cho mỗi đặc trưng đầu vào.

Định lý Bayes: Giả sử A và B là hai sự kiện đã xảy ra. Xác suất có điều kiện A khi biết trước điều kiện B được cho bởi:

P(A|B) = P(B|A).P(A)/P(B)

- P(A): Xác suất của sự kiện A xảy ra.

- P(B): Xác suất của sự kiện B xảy ra.

- P(B|A): Xác suất (có điều kiện) của sự kiện B xảy ra, nếu biết rằng sự kiện A đã xảy ra.

- P(A|B): Xác suất (có điều kiện) của sự kiện A xảy ra, nếu biết rằng sự kiện B đã xảy ra.

Mô hình xác suất

Một cách trừu tượng, mô hình xác suất cho phân loại là một mô hình điều kiện $\rho(C|F_1,.., F_n)$

Trên một lớp biến C với số lượng nhỏ các đầu ra hoặc các lớp. Điều kiện trên một vài biến đặc trưng F_1 đến F_2. Vấn đề chính trong bài toán này là nếu số đặc trưng n là lớp hoặc một đặc trưng có thể có số lượng lớn các giá trị, thì một mô hình được tạo ra dựa trên các bảng xác suất là phù hợp trong điều kiện này. Lý thuyết Bayes có thể viết thành:

$$\rho(C|F_1,...,F_n = \frac{\rho(C)\rho(F_1,...,F_n|C)}{\rho(F_1,...,F_n|C)}$$

Một cách mô tả đơn giản cho công thức trên như sau:

$$\text{Hậu nghiệm} = \frac{\text{nghiệm trước} \times \text{khả năng}}{\text{Bằng chứng}}$$

Trên thực tế, chỉ cần quan tâm tới số các phân mảnh (fraction), bởi có một số đặc trưng không phụ thuộc vào C và các giá trị F_i đã cho, mô hình $\rho(C|F_1,.., F_n)$ có thể được viết lại như sau, sử dụng luật xích để lặp lại định nghĩa của xác suất điều kiện:

$$\begin{aligned}\rho(C,F_1,..., F_n) &= \rho(C)\,\rho(F_1,..., F_n|C)\\ &= \rho(C)\,\rho(F_1|C)\,\rho(F_2,..., F_n|C, F_1)\\ &= \rho(C)\,\rho(F_1|C)\,\rho(F_2|C, F_1)\,\rho(F_3,..., F_n|C, F_1,F_2)\\ &= \rho(C)\,\rho(F_1|C)\,\rho(F_2|C, F_1)\,...\rho(F_n|C, F_1,F_2, F_3, ... F_{n-1})\end{aligned}$$

Giả thiết của xác suất điều kiện: giả thiết rằng mỗi đặc trưng F_i là độc lập có điều kiện với các đặc trưng khắc F_j với $j \neq i$, trong lớp đã cho *C*. Điều đó có nghĩa rằng:

$$\rho(F_i|C, F_j) = \rho(F_i|C),$$

$$\rho(F_i|C, F_j, F_k) = \rho(F_i|C),$$

$$\rho(F_i|C, F_j, F_{k,} F_l) = \rho(F_i|C),$$

Với mọi trường hợp $i \neq j, k, l$. Từ đó, mô hình kết hợp được biểu diễn bởi

$$\rho(C| F_1,..., F_n)\ \alpha\ \rho(C, F_1,..., F_n)$$

$$\alpha\ \rho(C)\,\rho(F_1|C)\,\rho(F_2|C)\,\rho(F_3|C)...$$

$$\alpha\rho(C)\prod_{i=1}^{n}\rho(F_2|C)$$

Có nghĩa rằng dưới giả thiết độc lập trên, phân tán có điều kiện trên các lớp biến C là:

$$\rho(C|F_1,..., F_n) = \rho(C)\prod_{i=1}^{n}\rho(F_i|C)$$

Với $Z = \rho(F_1,..., F_n)$ được gọi là nhân tố độc lập trên $F_1,..., F_n$ và là một hằng nếu các giá trị của các biến đặc trưng là đã biết.

Xây dựng phân lớp từ mô hình xác suất

Phân lớp Bayes kết hợp với luật quyết định tạo ra phân loại Naive Bayes. Một luật thông thường đưa ra giả thuyết về khả năng nhất hay còn được xem như là cực đại hóa xác suất hậu nghiệm (maximum a posteriori). Bộ phân loại Bayes là một hàm phân loại được định nghĩa:

$$classify(f_1,...,f_n) = \text{argmax}\, p(C=c)\prod_{i=1}^{n} p(F_i=f_i|(C=c)$$

2.2. Bộ phân loại Naive Bayes

Naive Bayes là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học và nhiều lĩnh vực khác như trong các công cụ tìm kiếm, các bộ lọc mail.

Mục đích chính là làm sao tính được xác suất $Pr(C_j, d')$, xác suất để tài liệu d' nằm trong lớp C_j. Theo luật Bayes, tài liệu d' sẽ được gán vào lớp C_j nào có xác suất $Pr(C_j, d')$ cao nhất.

Công thức để tính $Pr(C_j, d')$ như sau:

$$H_{BAYES(d')} = \text{argmax}\left[\frac{\Pr(C_j)\times\prod_{i=1}^{|d'|}\Pr(w_i|C_j)}{\sum\Pr(c')\times\prod\prod^{|d'|}\Pr(w_i|C'}\right]$$

- $TF(w_i, d')$ là số lần xuất hiện của từ w_i trong tài liệu d'

- $|d'|$ là số lượng các từ trong tài liệu d'

- w_i là một từ trong không gian đặc trưng F với số chiều là $|F|$

- $Pr(C_j)$ được tính dựa trên tỷ lệ phần trăm của số tài liệu mỗi lớp tương ứng

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

trong tập dữ liệu huấn luyện

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

Ngoài ra còn có các phương pháp NB khác có thể kể ra như ML Naive Bayes, MAP Naive Bayes, Expected Naive Bayes. Nói chung, Naïve Bayes là một công cụ rất hiệu quả trong một số trường hợp.

Thuật toán Naive Bayes dựa trên nguyên 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 dạng $\vec{x} = (x_1, x_2, ..., x_n)$

- C_i: phân lớp im với i = {1, 2,.., 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 phân lớp i

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

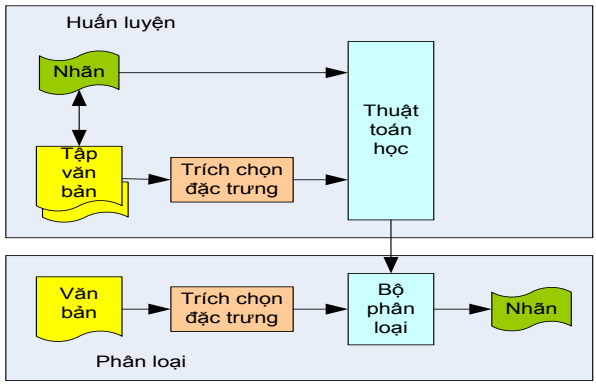
Các bước thực hiện thuật toán Naive Bayes

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

Bước 2: Phân lớp $X^{new} = (x_1, x_2, ..., x_n)$, 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

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

Mô hình tổng quát việc phân loại:



Hình 1. Mô tả bước xây dựng bộ phân lớp

2.3. Phân loại tài liệu tiếng Việt

2.3.1. Ứng dụng Naive Bayes trong phân loại tài liệu tiếng Việt

➢ Đặc điểm

Trong tất cả các ngôn ngữ, người ta thường phân chia dòng ngữ lưu thành các âm tiết. Âm tiết là đơn vị phát âm tối thiểu của lời nói. Nghiên cứu âm tiết tức là nghiên cứu sự tổ hợp các âm vị (phômen) trong dòng lưu ngữ, ví dụ như các thực từ.

Một điểm cơ bản nhất của các âm tiết tiếng Việt là ranh giới của âm tiết tiếng Việt trùng với ranh giới của hình vị (moocphem), tức là mỗi âm tiết đều đóng vai trò là dấu hiệu của một hình vị (moocphem), đơn vị có nghĩa dùng làm thành tố cấu tạo từ. Lời nói của con người là một chuỗi âm thanh được phát ra kế tiếp nhau trong không gian và thời gian. Việc phân tích chuỗi âm thanh ấy người ta nhận ra được các đơn vị của ngữ âm.

Đặc điểm thứ hai của âm tiết tiếng Việt là mỗi âm tiết tiếng Việt đều gắn liền với một trong sáu thanh điệu (không, huyền, ngã, hỏi, sắc, nặng) vì tiếng Việt là loại ngôn ngữ có thanh điệu khác với ngôn ngữ khác. Thanh điệu tham gia vào việc cấu tạo từ, làm chức năng phân biệt ý nghĩa của từ và làm dấu hiệu phân biệt từ. Thanh điệu có chức năng như một âm vị, nó gắn liền với âm tiết và biểu hiện trong toàn âm tiết [2].

Do đặc điểm trên mà âm tiết có vị trí rất quan trọng trong việc nghiên cứu âm tiếng Việt. Muốn xác định thành phần âm vị của ngôn ngữ, người ta thường xuất phát từ việc xác định các hình vị rồi từ các moocphem đó mà phân tích ra các âm vị, hình vị trong tiếng Việt trùng hợp với các âm tiết; chúng ta xuất phát từ việc phân tích các âm tiết để xác định các âm vị. Nếu như trong ngôn ngữ Ấn – Âu, âm tiết chỉ là vấn đề thuộc hàng thứ yếu so với âm vị và hình vị thì trong tiếng Việt, âm tiết là vấn đề hàng đầu của âm vị học.

➢ *Cấu trúc âm tiết*

Mỗi âm tiết tiếng Việt là một khối hoàn chỉnh trong phát âm. Trong ngữ cảm của người Việt, âm tiết tuy được phát âm liền một hơi, nhưng không phải là một khối bất biến mà có cấu tạo lắp ghép. Khối lắp ghép ấy có thể tháo rời từng bộ phận của âm tiết này để hoán vị với bộ phận tương ứng ở âm tiết khác.

Mỗi âm tiết tiếng Việt có 3 bộ phận: phụ âm đầu, vần và thanh điệu.

2.3.2. Rút trích đặc trưng

➢ **Giảm chiều đặc trưng**

Dữ liệu trong thế giới thực (real world data), chẳng hạn như tín hiệu tiếng nói, ảnh kỹ thuật số, ảnh scan MRI, thường có số chiều đặc trưng rất lớn. Để xử lý các dữ liệu này một cách đầy đủ, sẽ rất phức tạp và tốn thời gian. Do vậy, trong thực tế, ta có thể giảm chiều đặc trưng xuống một mức có thể, sau đó sẽ tính toán trên số chiều đặc trưng đã được giảm. Lý tưởng nhất, cần biểu diễn các chiều tương ứng với chiều nội tại của dữ liệu. Chiều nội tại của dữ liệu là số lượng đặc trưng tối thiểu nhất để có thể mô tả được thuộc tính của dữ liệu. Giảm chiều trở thành một bài toán ứng dụng trong nhiều lĩnh vực, những bài toán phức tạp trở nên đơn giản và dễ ứng dụng hơn trong cuộc sống.

Trong máy học và thống kê, giảm chiều hoặc giảm chiều là quá trình làm giảm số lượng các biến ngẫu nhiên được xem xét, và có thể được chia thành hai phần chính: lựa chọn đặc trưng (Feature selection) và trích rút đặc trưng (Feature extraction).

- Lựa chọn đặc trưng: Là cách tìm một tập hợp con của các biến ban đầu (còn gọi là tính năng hoặc các thuộc tính). Trong một số trường hợp, phân tích dữ liệu như hồi quy hoặc phân loại có thể được thực hiện trong không gian đã được giảm chiều chính xác hơn trong không gian ban đầu.

- Trích rút đặc trưng: Trích rút đặc trưng biến đổi các dữ liệu trong không gian có số chiều lớn (high dimensional space) tới một không gian có số chiều ít hơn. Việc chuyển đổi dữ liệu này có thể sử dụng phương pháp tuyến tính, như phân tích thành phần chính (PCA), hoặc có thể sử dụng những kỹ thuật giảm chiều phi tuyến tính. Đối với dữ liệu đa chiều, biểu diễn tensor có thể được sử dụng thông qua phương pháp học trong không gian con đa tuyến (multilinear subspace).

Đối với dạng dữ liệu văn bản, số lượng đặc trưng trở nên hàng nghìn, hàng trăm nghìn đặc trưng. Để xử lý các đặc trưng này, thường mất khá nhiều thời gian trong việc trích rút đặc trưng, và tính toán các đặc trưng. Do đó rất khó khăn khi xây dựng thành những hệ thống xử lý tài liệu ứng dụng trong thực tế.

Các phương pháp giảm chiều trong tài liệu hiện nay:

- Loại bỏ các từ dừng (stop words)

- Chỉ số ngữ nghĩa ẩn (Latent Semantic Indexing)

- Sử dụng từ loại danh từ

➢ **Giảm chiều đặc trưng bằng mô hình chủ đề**

Các tri thức hiện nay vẫn đang được số hóa và lưu trữ trong các trang tin tức, blog bài báo khoa học, các trang Web và các mạng xã hội,.. quá nhiều thông tin lưu trữ, do đó sẽ rất khó khăn để tìm kiếm và tổ chức dữ liệu, cũng như định nghĩa (define) một dữ liệu cụ thể. Do vậy, chúng ta cần những công cụ tính toán mới giúp tổ chức, tìm kiếm và hiểu (understand) những lượng lớn thông tin. Giả sử khi gõ vào ô tìm kiếm một từ khóa, kết quả trả về sẽ là một tập hợp tài liệu liên quan thông tin tới từ khóa đó.

Trong học máy và xử lý ngôn ngữ tự nhiên, một mô hình chủ đề là một loại mô hình thống kê để phát hiện ra các "chủ đề" trừu tượng xảy ra trong một bộ sưu tập các tài liệu. Một số phương pháp xây dựng mô hình

chủ đề như: *Xây dựng mô hình chủ đề dựa trên phân phối ẩn Dirichlet; Mô hình dựa trên mạng Bayesian; Mô hình chủ đề xây dựng dựa trên mô hình Markov ẩn*

➢ **Xây dựng mô hình chủ đề cho tiếng Việt**

Mô hình chủ đề cho tiếng Việt hiện nay vẫn chưa được xây dựng, các nghiên cứu cho tiếng Việt chủ yếu tập trung vào các vấn đề tách từ (word segmentation), nhận dạng từ loại (Pos tagging), phân tích cú pháp (syntax analysic),...

Một số các phương pháp xử lý văn bản đã có thường sử dụng công cụ tách từ để tách các từ trong văn bản và tính toán trọng số của các từ đó. Đối với những bài toán xử lý phân loại các đối tượng, việc quan trọng là xác định đặc trưng bởi hầu hết trong những bài toán này, số chiều đặc trưng là khá lớn. Bởi vậy, các nghiên cứu trước đây sẽ gặp phải những khó khăn sau:

- Thời gian tính toán lớn (do số chiều đặc trưng nhiều)
- Độ chính xác cũng như hiệu năng của hệ thống bị hạn chế.

Một khó khăn khác nữa trong cách xử lý phân loại tự động đối với các văn bản tiếng Việt, là độ khó trong xử lý ngôn ngữ, bởi ngôn ngữ tiếng Việt thuộc lớp ngôn ngữ đơn lập (single syllable language), các từ trong tiếng Việt có thể là từ đơn hoặc từ ghép, do vậy khó khăn trong việc tách từ. Bởi thế, trong luận văn đã tiếp cận bài toán theo hai bước: xử lý giảm đặc trưng và áp dụng lý thuyết Naive Bayes trong phân loại.

Xử lý giảm số chiều của đặc trưng bằng cách sử dụng mô hình chủ đề, do đó số lượng thuật ngữ trong mỗi văn bản sẽ giảm hơn nhiều so với số các từ trong một văn bản, mặt khác sẽ giải quyết bài toán tách từ tiếng Việt nhờ đó làm tăng độ chính xác của hệ thống, tiếp theo áp dụng lý thuyết Naive Bayes để phân loại các văn bản theo đúng chủ đề đã chọn [11].

2.3.3. Phân loại văn bản tiếng Việt dựa trên Naive Bayes

Sau khi xây dựng được tập từ chủ đề đối với mỗi một lớp chủ đề. Tiếp theo sử dụng phân loại Naive Bayes để xây dựng mô hình phân loại tự động.

Sử dụng luật cực đại hóa hậu nghiệm (Maximum a posteriori-MAP) có công thức sau:

$$c_{map} = \arg\max_{c \in C} (P(c|d)) = \arg\max_{c \in C} \left(P(c) \prod_{1 \le k \le n_d} P(t_k|c) \right) \quad (1)$$

Trong đó:

- T_k: các từ của tài liệu;
- C: chủ đề;
- $P(c|d)$: xác suất điều kiện của lớp c với tài liệu đã cho d;
- $P(c)$: xác suất tiền nghiệm của lớp c;
- $P(t_k|c)$: xác suất điều kiện của từ T_k với lớp c đã cho.

Sử dụng luật biến đổi Laplace cho công thức (1) chuyển thành

$$P(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'}) + B'} \quad (2)$$

Trong đó B' là tổng số tất cả các từ chủ đề, T_{ct} là số lần xuất hiện của thuật ngữ t trong các tài liệu huấn luyện thuộc lớp c.

4. KẾT LUẬN

Với các yêu cầu đặt ra về việc nắm bắt thuật toán Naive Bayes để hiểu cách thức phân loại tài liệu trong tiếng Việt từ đó áp dụng vào phân loại các tài liệu, bài báo trong thư viện điện tử hay trong các lĩnh vực công nghệ thông tin theo các chuyên ngành khác nhau.

Phương pháp phân loại tài liệu bằng thuật toán Naive Bayes thường được dùng trong phân loại tài liệu tiếng Anh, nay được áp dụng trong tiếng Việt. Nhờ tính đơn giản, các thông số không cần quá lớn như các phương pháp khác, khả năng linh hoạt đối với sự thay đổi về thông tin huấn luyện, thời gian phân loại phù hợp yêu cầu, Naive Bayes đã tỏ ra rất phù hợp với các yêu cầu đặt ra.

Bài viết này trình bày các kết quả nghiên cứu lý thuyết về Naive Bayes và quy trình phân loại tài liệu tiếng Việt, áp dụng các thuật toán Naive Bayes xử lí phân loại tài liệu tiếng Việt.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

1. Nguyễn Linh Giang, Nguyễn Mạnh Hiển, *Phân loại văn bản tiếng Việt với bộ phân loại vectơ hỗ trợ SVM,* 2002.

2. Nguyễn Hữu Quỳnh, *Ngữ pháp Tiếng Việt*, NXB Từ điển Bách Khoa, 2001.

Tài liệu tiếng Anh

3. C. Apte, F. Damerau, S. Weiss, *Automated Learning of Decision Rules for Text Categorization,* ACM Transactions on Information Systems, 12(3), pp. 233–251, 1994.

4. Novovicova J., Malik A., and Pudil P., *"Feature Selection Using Improved Mutual Information for Text Classification"*, SSPR&SPR 2004, LNCS 3138, pp. 1010–1017, 2004.

5. Aigars Mahinovs and Ashutosh Tiwari, *Text Classification Method Review*, Cranfield University, April 2007.

6. http://vlsp.vietlp.org:8080/.