**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

---


**NGUYỄN THỊ THANH TÂM**

# TIẾP CẬN KHAI PHÁ DỮ LIỆU VĂN BẢN VÀ THỬ NGHIỆM ỨNG DỤNG PHƯƠNG PHÁP NAIVE BAYES TRONG BỘ LỌC THƯ RÁC TỰ ĐỘNG

**Chuyên ngành: Truyền dữ liệu và Mạng máy tính**
**Mã số: 60.48.15**

**Người hướng dẫn khoa học: PGS.TS NGUYỄN BÁ TƯỜNG**

**TÓM TẮT LUẬN VĂN THẠC SỸ KỸ THUẬT**

HÀ NỘI – 2010

# MỞ ĐẦU

Ngày nay sự phát triển không ngừng của công nghệ thông tin, đặc biệt là sự ra đời của Internet đã đưa con người lên một tầm cao mới. Sự việc đó dẫn đến bùng nổ thông tin làm cho những nhà quản lý rơi vào tình trạng "ngập lụt thông tin" trong đó một lượng thông tin, tri thức có ích bị che dấu. Khai phá dữ liệu trong đó có lĩnh vực khai phá dữ liệu văn bản là một lĩnh vực khoa học liên ngành mới xuất hiện gần đây nhằm đáp ứng nhu cầu này. Nhiều kỹ thuật khai phá dữ liệu văn bản đã được nghiên cứu và phát triển như Naïve Bayes, Cây quyết định, phương pháp Support vector machine,…trong đó, phương pháp Naïve Bayes thu hút nhiều quan tam nghiên cứu và ứng dụng.

Sự ra đời của các dịch vụ trên Internet làm cho nhu cầu trao đổi thông tin, tìm kiếm thông tin của con người được đáp ứng một cách tốt nhất và nhanh nhất.

Tốc độ phát triển của các dịch vụ thư điện tử ngày nay và những lợi ích mà nó mang lại cho chúng ta là rất lớn. Tuy nhiên nó cũng có thể gây ra những thiệt hại to lớn nếu không biết cách loại bỏ và phòng chống nó. Một trong những vấn đề nghiêm trọng cần giải quyết hiện nay trong các thư điện tử đó là nạn thư rác hay còn gọi là "spam". Với lý do đó, dưới sự hướng dẫn của thầy giáo hướng dẫn, Đại tá, PGS.TS Nguyễn Bá Tường, tôi nhận đề tài **"** Tiếp cận khai phá dữ liệu văn bản và thử nghiệm ứng dụng phương pháp Naive Bayse trong bộ lọc thư rác tự động".

# CHƯƠNG 1
# TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VĂN BẢN

## 1.1. Phát hiện tri thức trong cơ sở dữ liệu và khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình phát hiện những tri thức hữu ích ẩn chứa trong cơ sở dữ liệu hay các kho chứa thông tin khác. Khai phá dữ liệu là một bước trong quy trình phát hiện tri thức trong CSDL (Knowledge Discovery in Dabases - KDD). Theo nhiều tài liệu khác nhau thì tiến trình KDD nói chung đều bao gồm 5 bước cơ bản sau đây:

- Trích lọc dữ liệu
- Tiền xử lý dữ liệu
- Biến đổi dữ liệu
- Khai phá dữ liệu
- Đánh giá và biểu diễn tri thức

## 1.2. Khai phá dữ liệu văn bản

- Khai phá dữ liệu văn bản là việc trích ra, lấy ra các thông tin có ích, chưa được biết đến còn tiềm ẩn trong các kho dữ liệu văn bản lớn.

- Khai phá dữ liệu văn bản là việc thu thập và phân tích dữ liệu bằng các công cụ tự động hoặc bán tự động từ các nguồn tài liệu đã có khác nhau để có được các tri thức mới, chưa được biết đến trước đó.

## 1.3. Các bài toán trong lĩnh vực khai phá dữ liệu văn bản

### 1.3.1. Phát hiện xu hướng văn bản

Đây là bài toán phát hiện các xu hướng, các luật chưa được biết đến trong các CSDL text lớn.

### 1.3.2. Tìm kiếm văn bản

Tìm kiếm văn bản là quá trình tìm các văn bản trong một kho dữ liệu theo các yêu cầu của người dùng. Ở đây, các yêu cầu là các truy vấn và thường được biểu diễn dưới dạng thuật ngữ hay biểu thức logic giữa các thuật ngữ.

### 1.3.3. Phân loại văn bản

Phân loại văn bản tức là gán văn bản vào một hoặc một số nhóm văn bản đã được biết trước. Phân loại văn bản có hai dạng là phân loại nhị phân và phân loại theo cấp độ.

### 1.3.4. Lập nhóm văn bản

Lập nhóm văn bản là bài toán tự động lập ra các nhóm văn bản từ một tập các văn bản sao cho các văn bản trong cùng một nhóm thì tương tự với nhau nhiều hơn so với các văn bản ở các nhóm khác nhau. Người sử dụng có thể chỉ định số nhóm cần lập hoặc hệ thống tự động tính số nhóm sao cho phù hợp nhất.

### 1.3.5. Tóm tắt văn bản

Tóm tắt văn bản là bài toán tìm ra thể hiện nội dung của một văn bản thông qua một vài đoạn văn bản, hoặc thông qua các câu quan trọng nhất của văn bản đó.

### 1.3.6. Dẫn đường văn bản

Bài toán dẫn đường văn bản là sự tổ hợp giữa bài toán tìm kiếm văn bản và phân loại văn bản. Giống như phân loại văn bản, bài toán dẫn đường đưa các văn bản về các nhóm khác nhau. Tuy nhiên nó cũng giống bài toán tìm kiếm, mỗi nhóm văn bản được gán với các thông tin cần thiết của một hay nhiều nhóm người dùng.

### 1.3.7. Trích chọn từ khóa

Bài toán trích chọn từ khoá, thực hiện việc trích ra được các từ khoá quan trọng nhất của văn bản, thể hiện đặc thù về chuyên môn của văn bản đó.

## 1.4. Các khó khăn trong khai phá dữ liệu văn bản

*Tính đa chiều (high dimensonality):* Số thuật ngữ trong một văn bản lớn dẫn đến số chiều của không gian vector sẽ rất lớn.

*Tính khả cỡ (scability):* Các CSDL lớn thường chứa hàng trăm nghìn văn bản

*Tính chính xác (accuracy):* Bất kỳ ngôn ngữ nào cũng đều có sự nhập nhằng.

*Tri thức tiên nghiệm*: Trong nhiều bài toán chẳng hạn như bài toán lập nhóm văn bản thì người sử dụng phải xác định trước một số tham số đầu vào như số nhóm văn bản cần lập.

## 1.5. Các bước tiền xử lý văn bản

Quá trình tiền xử lý đóng vai trò quan trọng trong việc ảnh hưởng đến hiệu năng và độ chính xác của các giải thuật khai phá dữ liệu. Các công việc chính trong quá trình tiền xử lý là tách thuật ngữ và giảm số chiều thuật ngữ.

### 1.5.1. Tách thuật ngữ

Tách thuật ngữ có thể được hiểu là quá trình phân tách chuỗi ký tự trong văn bản thô ban đầu thành các từ có nghĩa.

***Các giải thuật tách thuật ngữ Tiếng Việt***

*Bài toán:* Nhập vào một câu tiếng Việt bất kỳ, hãy tách câu đó thành những đơn vị từ vựng (từ), hoặc chỉ ra những âm tiết nào không có trong từ điển (phát hiện đơn vị từ vựng mới).

***a) Tách thuật ngữ theo độ dài từ dài nhất***

Đây là phương pháp tách thuật ngữ đơn giản và dễ cài đặt. Phương pháp này sử dụng một từ điển từ vựng để làm cơ sở phân tách các thuật ngữ.

***b) Tách thuật ngữ tiếng Việt bằng phương pháp đồ thị***

Phương pháp tách thuật ngữ bằng đồ thị quy việc phân tách câu về việc tìm đường đi trên một đồ thị có hướng, không có trọng số.

Như đã nói ở trên, cách phân tách câu đúng đắn nhất tương ứng với đường đi qua ít cung nhất trên đồ thị. Do đó ta có thể quy bài toán liệt kê các phương án phân tách câu về bài toán tìm tất cả những đường đi ngắn nhất từ đỉnh 0 đến đỉnh n của đồ thị phân tách câu.

**1.5.2. Giảm chiều cho tập thuật ngữ**

Có rất nhiều kỹ thuật để giảm chiều của tập thuật ngữ bao gồm:

- *Tìm gốc từ*
- *Sử dụng từ điển đồng nghĩa*
- *Loại bỏ các từ dừng*
- *Chỉ trích chọn một phần văn bản*
- *Loại bỏ những thuật ngữ có trọng số thấp nhất*
- *Các kỹ thuật dựa trên lý thuyết thông tin*

# CHƯƠNG 2
# MỘT SỐ CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN

## 2.1 Giới thiệu bài toán phân loại văn bản

### 2.1.1 Sự cần thiết phải phân loại văn bản

Nhiều năm trở lại đây, các loại thông tin đã phát triển không ngừng về cả số lượng và chất lượng. Việc bùng nổ thông tin cũng làm cho vấn đề tổ chức, quản lí, phân loại thông tin ngày càng có vai trò quan trọng. Để đáp ứng được yêu cầu này thì trước tiên phải tiến hành phân loại văn bản.

### 2.1.2 Định nghĩa phân loại văn bản

Phân loại văn bản là sự phân loại không cấu trúc các tài liệu văn bản dựa trên một tập hợp của một hay nhiều loại văn bản đã được định nghĩa trước. Quá trình này thường được thực thi bằng một hệ thống tự động gán cho các tài liệu văn bản một loại nào đó.

## 2.2 Tiến trình phân loại văn bản

Đưa ra một tập tài liệu mẫu D, cần được phân bổ thành một số loại tài liệu nhất định - mỗi tài liệu đó cần được gán cho một loại văn bản nào đó. Nhiệm vụ của chúng ta là tìm một hệ thống phân hoạch, mà nó sẽ cung cấp cho ta một nhãn y phù hợp cho một số tài liệu trong D vừa được đưa vào từ nguồn tài nguyên giống nhau như các văn bản mẫu.

**Các bước trong tiến trình phân loại văn bản:**

- Lựa chọn các đặc trưng văn bản
- Biểu diễn văn bản
- Học một bộ phân loại văn bản

- Tiến hành phân loại văn bản

**2.3 Đặc trưng văn bản và cách lựa chọn các đặc trưng văn bản**

**2.3.1 Tần suất tài liệu**

Tần suất tài liệu DF là là số tài liệu có sự xuất hiện của một từ. Người ta đã tính toán tần suất tài liệu cho một từ đơn trong tập văn bản mẫu. Cốt lõi của phương pháp này là phải tìm ra được một không gian các từ đặc trưng. Cách xác định DF là kĩ thuật đơn giản nhất để làm giảm bớt vốn từ có trong văn bản.

**2.3.2 Lượng tin tương hỗ**

Lượng tin tương hỗ là giá trị logarit của nghịch đảo xác suất xuất hiện của một từ thuộc vào lớp văn bản c nào đó. Đây là một tiêu chí thể hiện sự phụ thuộc của từ t với loại văn bản c. Lượng tin tương hỗ giữa từ t và lớp c được tính như sau:

$$MI(t,c) = \sum_{t\in\{0,1\}} \sum_{c\in\{0,1\}} P(t,c) \log \frac{P(t,c)}{P(t)P(c)}$$

Trong đó:

P(t, c) là xác suất xuất hiện đồng thời của từ t trong lớp c;

P(t) là xác suất xuất hiện của từ t và P(c) là xác suất xuất hiện của lớp c.

Độ đo MI toàn cục (tính trên toàn bộ tập tài liệu huấn luyện) cho từ t được tính như sau:

$$MI_{avg}(t) = \sum_{i} P(c_i) MI(t,c_i)$$

$$MI_{\max}(t) = \max_{i=1}^{m} \{MI(t,c_i)\} \qquad (2.4)$$

## 2.4 Các mô hình biểu diễn văn bản

### 2.4.1 Mô hình không gian vector

Bản chất của mô hình không gian vector là mỗi văn bản được biểu diễn thành một vector mà mỗi thành phần là một thuật ngữ riêng biệt trong tập văn bản gốc và được gán một giá trị trọng số w biểu thị mức độ quan trọng của từng thuật ngữ đối với văn bản. Có nhiều cách tính trọng số cho thuật ngữ, sau đây là một số cách tính trọng số thuật ngữ điển hình.

*2.4.1.1. Các phương thức tính trọng số thuật ngữ*

- Tính trọng số theo mô hình Boolean
- Tính trọng số theo mô hình tần suất – TF
- Tính trọng số theo mô hình nghịch đảo tần số văn bản - IDF
- Tính trọng số theo mô hình kết hợp TFxIDF

*2.4.1.2. Phép tính độ tương tự giữa hai vector*

Trong mô hình không gian vector có sử dụng tới phép tính độ tương tự giữa 2 vector văn bản và phép tính độ tương tự giữa 2 nhóm văn bản. Phép tính độ tương tự không chỉ quan trọng đối mô hình không gian vector mà còn cả với các mô hình khác nữa.

*2.4.1.3. Biểu diễn nhóm văn bản*

Xét một nhóm văn bản C, khi đó vector trọng tâm c của nhóm C được tính thông qua vector tổng Sum, Sum = $\sum_{d_i \in C} d_i$ của các văn bản trong nhóm c:

$$c = \frac{sum}{|C|}$$

Ở đó |C| là số phần tử của nhóm văn bản C.

Trong các bài toán xử lý văn bản thì vector trọng tâm được dùng để làm đại diện cho cả nhóm văn bản. Độ tương tự giữa hai nhóm $C_1$, $C_2$ được tính bằng độ tương tự giữa hai vector trọng tâm c1, c2 :

$$S(C_1, C_2) = S(c_1, c_2)$$

### 2.4.2 Mô hình dựa trên tập mờ

Giả sử có 1 tập các văn bản $D = \{d_1, d_2, \ldots, d_M\}$. Khi đó ta có một tập các thuật ngữ $T = \{t_1, t_2, \ldots, t_N\}$. Sự liên quan của các từ khoá tới một văn bản được xác định tương ứng bằng cách sử dụng một phương pháp đánh chỉ số nào đó đã biết:

$$\mu(T) = \{\mu_T(t_1), \mu_T(t_2), \ldots, \mu_T(t_N)\}$$

Thực hiện chuẩn hoá các giá trị của $\mu(T)$ vào [0, 1].

**Đinh nghĩa 2:** Hàm tích hợp khái niệm mờ

Hàm $F: [0, 1]^n \rightarrow [0, 1]$ được gọi là hàm tích hợp mờ nếu thoả mãn các tính chất của hàm tích hợp, tức là:

1. $0 \leq F(\mu_T(t_1), \mu_T(t_2), \ldots, \mu_T(t_m)) \leq 1$
2. $F(\mu_T(t_1), \mu_T(t_2), \ldots, \mu_T(t_m)) \leq F(\mu_T(t'_1), \mu_T(t'_2), \ldots, \mu_T(t'_m))$

với $\mu_T(t_i) \leq \mu_T(t'_i)$; $i = 1 \div m$

Trong đó $\mu_T(t_i)$ và $\mu_T(t'_i)$ biểu diễn mức độ quan trọng của các thuật ngữ. Về mặt ngữ nghĩa, trong hai khái niệm, khái niệm nào có nhiều thuật ngữ liên quan đến văn bản hơn thì khái niệm đó được xác đinh rõ ràng hơn và ngược lại.

Khi đó một văn bản d có thể được biểu diễn dưới dạng:

$$d = \{\mu(k_1), \mu(k_2), \ldots, \mu(k_i)\}$$

Như vậy khái niệm mờ có thể giải quyết vấn đề từ đồng nghĩa trong xử lý văn bản.

### 2.4.3 Mô hình dựa trên tập thô

Bất cứ một tập nào chứa các đối tượng không phân biệt được với nhau thì được gọi là một tập cơ sở (elementary set). Hợp của các tập cơ sở được gọi là một tập chính xác, ngược lại thì tập đó được gọi là tập thô (không chính xác). Nếu các tập con của tập vũ trụ được coi là các khái niệm thì các khái niệm nhập nhằng, tương ứng với các tập thô, không thể mô tả bởi thông tin về các thành viên của chúng. Bởi vậy, theo cách tiếp cận của tập thô, mỗi khái niệm nhập nhằng được thay thế bởi một cặp khái niệm chính xác gọi là xấp xỉ dưới và xấp xỉ trên của khái niệm nhập nhằng đó. Xấp xỉ dưới bao gồm các đối tượng chắc chắn thuộc vào khái niệm còn xấp xỉ trên chứa các đối tượng có thể thuộc vào khái niệm. Mô hình tập thô ban đầu sử dụng quan hệ tương đương với các tính chất phản xạ đối xứng, bắc cầu. Tuy nhiên tính chất bắc cầu tỏ ra quá cứng nhắc đối với trường hợp nghĩa của các từ và không thích hợp trong xử lý văn bản.

## 2.5 Các phương pháp phân loại văn bản

### 2.5.1 Nguyên mẫu

Nguyên mẫu (prototype) có thể là phương pháp đơn giản nhất được áp dụng trong phân loại văn bản. Mỗi văn bản đầu vào là một vector $\overrightarrow{D_i}$ ($w_1$, $w_2$ ,… $w_k$ ) trong đó mỗi chiều $w_i$ đặc trưng cho một từ loại (term). Một tập tài liệu mẫu sẽ được phân chia làm các lớp văn bản khác nhau và được đặc trưng bởi đại lượng $c_j$ (categorization). Có thể có nhiều tài liệu $D_i$ trong một lớp tài liệu $c_j$, tuy nhiên để đơn giản người ta xác định trong $c_i$ một vector trung bình ($\overline{D_i}$). Và sử dụng cosin của góc tạo bởi hai vector (một vector biểu diễn văn bản cần phân loại D, một vector biểu diễn lớp văn bản $c_i$) làm độ đo sự phù hợp giữa văn bản D với loại văn bản $c_i$.

D sẽ được xác định thuộc vào loại văn bản $c_i$ nào mà cosin($\overrightarrow{D}, \overrightarrow{\overrightarrow{D_i}}$) là lớn nhất.

### 2.5.2 Mô hình xác suất Naive Bayes

Cơ sở của phương pháp phân loại văn bản Naive Bayes là chủ yếu dựa trên các giả định của Bayes. Với mỗi văn bản D (document), người ta sẽ tính cho mỗi loại một xác suất mà tài liệu D có thể thuộc vào lớp tài liệu đó bằng việc sử dụng luật Bayes.

Xác suất $P(C_i| D)$ gọi là xác suất mà tài liệu D có khả năng thuộc vào lớp văn bản $C_i$ được tính toán như sau:

$$P(C_i \mid D) = \frac{P(C_i) * P(D \mid C_i)}{P(D)} \qquad (2.13)$$

Theo giả định của Naive Bayes xác suất của mỗi từ trong tài liệu D là độc lập với ngữ cảnh xuất hiện các từ đồng thời cũng độc lập với vị trí của các từ trong tài liệu. Xác suất $P(D|C_i)$ được tính toán từ tần suất xuất hiện của các từ đơn $w_j$ (word) trong D

$$P(D \mid C_i) = \prod_{1<j<l} P(w_j \mid C_i) \qquad (2.14)$$

l là tổng số từ w trong tài liệu D

Giá trị lớn nhất của xác suất $P(C_i \mid D)$ được đưa ra bởi người làm công tác phân loại. Tài liệu D sẽ được gán cho loại văn bản nào có xác suất hậu nghiệm cao nhất nên được biểu diễn bằng công thức:

$$\text{Class of D} = \arg \max_{1 \le i \le N} \left\{ P(C_i \mid D) \right\} = \arg \max_{1 \le i \le N} \frac{P(C_i) * P(D \mid C_i)}{P(D_i)} \qquad (2.15)$$

trong đó N là tổng số tài liệu.

### 2.5.3 Phương pháp dựa trên cây quyết định

Đây là phương pháp học xấp xỉ các hàm mục tiêu có giá trị rời rạc. Cây quyết định này được tổ chức như sau: Các nút trung gian được gán nhãn bởi các thuật ngữ, nhãn của các cung tương ứng với

trọng số của thuật ngữ trong tài liệu mẫu, nhãn của các lá tương ứng với nhãn của các lớp. Cho một tài liệu $d_j$, ta sẽ thực hiện so sánh các nhãn của cung xuất phát từ một nút trung gian (tương ứng với một thuật ngữ nào đó) với trọng số của thuật ngữ này trong $d_j$, để quyết định nút trung gian nào sẽ được duyệt tiếp. Quá trình này được lặp từ nút gốc của cây, cho tới khi nút được duyệt là một lá của cây. Kết thúc quá trình này, nhãn của nút lá sẽ là nhãn của lớp được gán cho văn bản.

Các giải thuật ID3 và cải tiến của nó là C45 được đánh giá là hiệu quả và được sử dụng phổ biến nhất.

### 2.5.4 Phương pháp phân loại văn bản K-NN (K – Nearest Neighbor)

Tư tưởng chính của giải thuật này là tính toán độ phù hợp của văn bản đang xét với từng nhóm chủ đề dựa trên K văn bản mẫu có độ tương tự gần nhất. Giải thuật này còn được sử dụng trong bài toán tìm kiếm văn bản và bài toán tóm tắt văn bản.

### 2.5.5 Phương pháp Support Vector Machine

Giả sử dữ liệu huấn luyện bao gồm $n$ mẫu được cho dưới dạng $< \vec{x}_i, y_i >$, i=1…$n$, trong đó $\vec{x}_i \in \Re^m$ là véctơ bao gồm $m$ phần tử chứa giá trị của $m$ thuộc tính hay đặc trưng và $y_i$ là nhãn phân loại có thể nhận giá trị +1 hoặc -1. Có thể hình dung dữ liệu như các điểm trong không gian ơclit $m$ chiều và được gán nhãn. SVM được xây dựng trên cơ sở hai ý tưởng chính.

*Ý tưởng thứ nhất* là ánh xạ dữ liệu gốc sang một không gian mới gọi là *không gian đặc trưng* với số chiều lớn hơn sao cho trong không gian mới có thể xây dựng một siêu phẳng cho phép phân chia dữ liệu thành hai phần riêng biệt, mỗi phần bao gồm các điểm có cùng nhãn phân loại.

*Ý tưởng thứ hai* là trong số những siêu phẳng như vậy cần lựa chọn siêu phẳng có lề lớn nhất. Lề ở đây là khoảng cách từ siêu phẳng tới các điểm gần nhất nằm ở hai phía của siêu phẳng (mỗi phía tương ứng với một nhãn phân loại). Lưu ý rằng siêu phẳng nằm cách đều các điểm gần nhất với nhãn khác nhau.

Ta sử dụng một phương pháp gọi là *thủ thuật nhân* bằng cách tìm một *hàm nhân* (kernel function) *K* sao cho:

$$K(\vec{a},\vec{b}) = <\vec{a},\vec{b}>$$

Sử dụng phương pháp nhân tử Lagrăng và thay thế tích vô hướng của hai vectơ bằng giá trị hàm nhân

Quá trình huấn luyện SVM là quá trình xác định $\alpha_i$. Sau khi huấn luyện xong, giá trị nhãn phân loại cho một ví dụ mới $\vec{x}$ sẽ được tính bởi:

$$f(\vec{x}) = sign(\sum_{i=1}^{n} y_i \alpha_i K(\vec{x}_i, \vec{x}) + b)$$

Đối với bài toán phân loại thư điện tử, $\vec{x}_i$ là vectơ đặc trưng biểu diễn cho nội dung thư như trong phần phân loại Bayes và $y_i$ là nhãn phân loại đối với dữ liệu huấn luyện. Thư mới được phân loại theo công thức: giá trị âm là thư bình thường, trong khi giá trị dương tương ứng với thư rác.

**2.6 Bài toán phân loại thư rác**

# CHƯƠNG 3
# ỨNG DỤNG PHƯƠNG PHÁP NAIVE BAYES TRONG BỘ LỌC THƯ RÁC TỰ ĐỘNG

## 3.1 Các công nghệ lọc thư rác hiện nay

Một số công nghệ lọc thư rác hiện nay:

- DNS Blacklist
- SURBL List
- Chặn IP
- Kiểm tra địa chỉ
- Sử dụng bộ lọc Bayesian
- Sử dụng danh sách Black/white list
- Sử dụng Challenge/Response
- Kiểm tra header
- Report Spam Email

Một số công nghệ chống spam thú vị đang được nghiên cứu:

- Tem cho e-mail- Cài mật mã
- Khai báo thông tin
- Lọc email qua nội dung
- Lọc theo danh sách website chuyển tiếp

## 3.2 Quá trình hoạt động của bộ lọc thư rác Bayes

Ở đây mỗi mẫu mà ta xét chính là một email, tập các lớp mà mỗi email có thể thuộc về là C = {spam,non-spam}

Khi ra nhận được một email, sử dụng phương pháp Naives Bayes huấn luyện tập mẫu (email) ban đầu, sau đó sẽ sử dụng các xác suất này ứng dụng vào phân loại một mẫu (email) mới.

Giả thiết mỗi một thư được đại diện bởi một vectơ thuộc tính đặc trưng $\vec{X} = (x_1, x_2, \ldots, x_n)$, trong đó $x_1, x_2, \ldots, x_n$ là giá trị của thuộc tính $X_1, X_2, \ldots, X_n$ tương ứng trong không gian vector đặc trưng $\vec{X}$.

Theo M Sahami et al ta sử dụng giá trị nhị phân, $X_i = 1$ nếu các đặc điểm của $X_i$ có trong email, ngược lại $X_i=0$

Ta tính giá trị tương hỗ MI(X,C) mà mỗi một đại diện của X thuộc về loại C như sau:

$$MI(X,C) = \sum_{x\in\{0,1\}} P(X=x,C=c).\log\frac{P(X=x,C=c)}{P(X=x)P(C=c)} \qquad (3.1)$$

Sau đó to chọn các thuộc tính có giá trị MI cao nhất. Các xác suất P(X), P(C), P(X,C) được tính dựa trên dữ liệu học.

Dựa vào công thức xác suất Bayes và công thức xác suất đầy đủ ta có được xác suất của một thư với vector đặc trưng $\vec{x}$,

$$P(C=c \mid \vec{X}=\vec{x}) = \frac{P(C=c).P(\vec{X}=\vec{x} \mid C=c)}{\sum_{k\in\{spam,non-spam\}} P(C=k).P(\vec{X}=\vec{x} \mid C=k)} \qquad (3.2)$$

Thực tế thì rất khó tính được xác suất $P(\vec{X} \mid C)$ bởi Naïve Bayes giả thiết rằng $X_1$, $X_2$, …,$X_n$ là những biến cố độc lập, do đó chúng ta có thể tính được xác suất ở trên như sau:

$$P(C=c \mid X=x) = \frac{P(C=c).\prod_{i=1}^{n} P(X_i=x_i \mid C=c)}{\sum_{k\in\{spam,non-spam\}} P(C=k).\prod_{i=1}^{n} P(X_i=x_i \mid C=k)}$$

Với $P(X_i|C)$ và P(C) được tính dựa trên dữ liệu học, việc tính này dựa vào tập huấn luyện ban đầu. Từ xác suất này, ta so sánh với một giá trị ngưỡng t mà ta cho là ngưỡng để phân loại thư rác hay không, nếu xác suất này lớn hơn t, ta cho là thư đó là thư rác ngược lại thì không phải là thư rác.

### 3.3 Sự hoạt động của các bộ lọc thư rác thực tế

Phương pháp Bayes tiếp cận với các thư rác một cách có hiệu quả cao. Tháng 5/2003 một bài báo BBC cho biết kết quả của việc tìm kiếm thư rác trong bộ lọc đạt 99.7% có thể hoàn thành với một số thấp các sai sót.

### 3.4 Các ưu điểm của bộ lọc thư rác Bayes

Phương pháp Bayes nhận dạng một thư điện tử dựa vào các mô tả. Nhiều thông minh hơn bởi vì nó kiểm tra tất cả các khía cạnh của tin nhắn.Bộ lọc Bayes giải quyết và thích nghi với các công nghệ lọc thư rác kiểu mới. Bộ lọc thư rác sử dụng thuật toán Naive Bayes cung cấp một chức năng lọc thư tự rác tự động

### 3.5 Các bước xây dựng bộ lọc thư rác sử dụng thuật toán Naive Bayes

**Tạo một cơ sở dữ liệu từ Bayes thích hợp**

Trước khi lọc thư cần áp dụng phương thức này, một người sử dụng có thể cung cấp một cơ sở dữ liệu với tập các từ và các tokens (ví dụ $, địa chỉ IP của các vùng...) tập hợp các mẫu thư được coi là các thư rác (spam) và tập mẫu thư được coi là thư hợp lệ.

#### 3.5.1 Lựa chọn các đặc trưng

Để xem xét nội dung thư và lựa chọn các đặc trưng chúng tôi dùng khái niệm "token". Chúng ta lựa chọn các đặc trưng bằng việc sử dụng phương pháp sử dụng trong lĩnh vực phân loại văn bản. Với mỗi từ xuất hiện trong nội dung của các thư điện tử trong tập văn bản mẫu, chúng ta sẽ đưa vào một đặc trưng thích hợp.

#### 3.5.2 Biểu diễn các thư điện tử

Chúng ta tiến hành biểu diễn thư điện tử thành vector $\vec{x} = (x_1, x_2, \ldots, x_n)$, trong đó $x_1, x_2, \ldots, x_n$ là giá trị của thuộc tính $X_1, X_2, \ldots, X_n$ tương ứng trong không gian vector đặc trưng. Trong trường hợp đơn

giản nhất, chúng tôi chọn thuộc tính là 1 từ đơn như vậy $X_i$=1 nếu thư chứa từ đó, ngược lại $X_i$ =0. Nhưng thay vì $X_i$ nhận giá trị 0 và 1, tôi tính xác suất từ đó là thư rác có giá trị trong đoạn [0,1]

### 3.5.3 Xác định ngưỡng

Xác định rõ ngưỡng dựa vào công thức (3.3) để loại bỏ tất cả các thư điện tử mà xác suất của chúng lớn hơn xác suất này.

## 3.6 Thử nghiệm ứng dụng Naive Bayes trong bộ lọc thư rác tự động.

### 3.6.1 Thử nghiệm với kho dữ liệu PU

*3.6.1.1 Vài nét về kho PU*

Tôi sử dụng kho dữ liệu trong kho PU [10] để học và kiểm thử. PU là một kho dữ liệu email chuẩn, gồm bốn kho nhỏ hơn là PU1, PU2, PU3, PUA. Mỗi token sẽ được thay thế tương ứng bằng một con số duy nhất như minh hoạ

*3.6.1.2 Xác định công thức theo Paulgraham*

*3.6.1.3 Kết quả thử nghiệm*

Thử nghiệm với kho ngữ liệu pu. Bởi vì kho dữ liệu học và kiểm thử là số, do đó tôi thay đổi về cách lấy token, ở đây tôi xem token là các con số và dấu hiệu tách token là các khoảng trắng.

Tôi thử nghiệm với non-spam w=2. Với mỗi w, tôi thử nghiệm với λ lần lượt với các giá trị 1, 9 và 999. Tương ứng với mỗi giá trị λ và w tôi thực hiện tính xác suất spam theo công thức 3.5. Số token lấy lần lượt là 10, 15, 20.

Tôi kiểm tra với kho dữ liệu pu, tôi cho học từ part1 đến part9, sau đó chúng tôi thử nghiệm phân loại trên part10 chứa những email chưa được học.

### 3.6.2 Minh hoạ thuật toán phân loại thư rác Naive Bayes

Bài toán phân loại thư rác thực chất là bài toán phân loại văn bản hai lớp, trong đó: tập tài liệu mẫu ban đầu là các thư rác (spam) và các thư hợp lệ (ham), các văn bản cần phân lớp là các Email được gửi đến client. Kết quả đầu ra của quá trình phân loại này là hai lớp văn bản: Spam(thư rác), Ham (thư hợp lệ). Mô hình phân loại thư rác tổng quát có thể mô tả như sau:

Mô tả dữ liệu bài toán: chương trình cài đặt ở mức đơn giản, với dữ liệu gồm 100 dấu hiệu non-spam và 100 dấu hiệu spam là các từ đơn được lưu trữ trong một bảng.

Chương trình minh hoạ

### 3.6.3 Giới thiệu phần mềm lọc thư Spam Reader 3.0

Spam Reader 3.0 là một add-on chống thư rác mạnh mẽ, dễ sử dụng được tích hợp vào MS Outlook và có mức đề phòng cao đối với các email không mong muốn. Spam Reader. Phần mềm sử dụng cách tiếp cận đáng tin cậy nhất để lọc spam-bộ lọc Bayes, tự động điều chỉnh lọc theo nhu cầu người sử dụng và phát hiện chính xác đến 98%,download phần mềm tại địa chỉ http://www.spam-reader.com/

Spam Reader tích hợp đầy đủ vào MS Outlook nên bạn không cần chạy một chương trình bên ngoài. Sau khi cài đặt nó, bạn sẽ thấy một thanh công cụ mới và một mục mới vào trình đơn chính của Outlook.

**III. Kết luận và hướng phát triển**

Luận văn ***" Tiếp cận khai phá dữ liệu văn bản và thử nghiệm ứng dụng phương pháp Naive Bayse trong bộ lọc thư rác tự động"*** đã trình bày một số kết quả sau đây:

- Những nghiên cứu về khai phá dữ liệu văn bản và các bài toán ứng dụng.

- Khai phá dữ liệu văn bản có nhiều hướng tiếp cận: Naïve Bayes, Cây quyết định, Phương pháp Support vector machine, mạng nơron…Trong đó, tập trung tìm hiểu thuật toán Naïve Bayes.

- Thử nghiệm ứng dụng Naive Bayes trong hệ thống lọc thư rác với kho dữ liệu PU. Giới thiệu phần mềm lọc thư rác tự động Spam Reader 3.0

**Hướng phát triển tiếp theo của luận văn:**

- Xây dựng một Email Client với khả năng lọc thư rác tự động bằng việc ứng dụng phương pháp phân loại văn bản Naive Bayes ứng dụng trong trường Cao đẳng kinh tế - kỹ thuật Thương mại và một số dịch vụ mail khác.

- Hiện nay, dữ liệu được lưu trữ ngày một tăng, để ứng dụng khai phá dữ liệu vào các bài toán này cần tiếp tục nghiên cứu các phương pháp xử lý cho bài toán có dữ liệu lớn. Xem xét, nghiên cứu một số ứng dụng khác của khai phá dữ liệu văn bản nõi riêng cũng như khai phá dữ liệu nói chung