Tạp chí Tin học và Điều khiển học, T.21, S.2 (2005), 151–161

THUẬT TOÁN PHÂN LỚP BAYES VÀ VẤN ĐỀ XÁC ĐỊNH NGƯỠNG PHÂN LỚP TRONG MÁY TÌM KIẾM

ĐẶNG THANH HẢI, NGUYỄN HƯƠNG GIANG, HÀ QUANG THỤY

Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Abstract. The problem of integrating a classification algorithm into search engines is likely to play an important role in the evolution to integrate web mining solutions into search engines. Some typical classification algorithms, such as Bayes, k-NN, FOIL ... are useful and Bayes algorithms are the most important. In this article, we propose a new mathematical formula which approximate to the posterior probability in the first Bayes algorithm. Moreover, the problem of determining the classification threshold is solved. By assuming that the classification threshold on Trebusep model, an algorithm to determine the classification threshold is proposed. Effect of proposed solutions is proved by experimental results, which are showed in this article.

Tóm tắt. Tích hợp thuật toán phân lớp vào máy tìm kiếm là một bài toán nổi bật trong xu thế tích hợp khai phá Web vào các máy tìm kiếm. Một số thuật toán phân lớp điển hình được sử dụng là Bayes, k-NN, FOIL, trong đó các thuật toán Bayes có một vị trí quan trọng. Trong báo cáo này, chúng tôi đề xuất một công thức tính toán mới xấp xỉ xác suất hậu nghiệm trong thuật toán phân lớp Bayes thứ nhất. Mặt khác, bài toán xác định ngưỡng phân lớp cũng được phân tích và giải quyết. Trên cơ sở giả thiết ngưỡng phân lớp tuân theo mô hình Trêbusep, thuật toán xác định ngưỡng phân lớp được đề xuất. Các kết quả thử nghiệm được trình bày trong báo cáo chứng tỏ tính hiệu quả của các giải pháp được đề xuất.

1. GIỚI THIỆU

Tích hợp giải pháp khai phá Web vào máy tìm kiếm đã trở thành lĩnh vực nghiên cứu nổi bật về khai phá dữ liệu trong những năm gần đây. Một trong những nội dung được đề cập đầu tiên là tích hợp giải pháp phân lớp vào máy tìm kiếm ([2, 4, 5, 6, 8, - 14]). Ngoài các yếu tố liên quan tới từ khóa, máy tìm kiếm còn cho phép người dùng đưa thêm yếu tố "lớp" vào câu hỏi tìm kiếm để sau đó máy tìm kiếm sẽ trả về cho người dùng các trang Web thỏa mãn yêu cầu của người dùng, bao gồm cả yếu tố lớp trong yêu cầu. Như đã biết, trong nhiều hệ thống quản lý văn bản, một số mô hình phân lớp điển hình như Bayes, cây quyết định, k-NN, FOIL thường được sử dụng ([2, 4, 5, 6, 9, 13, 14]).

Báo cáo này định hướng tới việc bổ sung thành phần phân lớp Bayes vào máy tìm kiếm tiếng Việt Vinahoo [15, 16]. Đầu tiên, chúng tôi nghiên cứu việc cải tiến công thức tính toán xác suất hậu nghiệm trong phương pháp Naive Bayes. Sau đó chúng tôi nghiên cứu việc đề xuất giải pháp đặt ngưỡng phân lớp trong thành phần phân lớp nói trên. Các kết quả thực nghiệm do chúng tôi thi hành trên máy tìm kiếm Vinahoo đã minh chứng được tính hiệu quả của thành phần phân lớp.

Mục ngay tiếp theo trình bày kết quả từ một số công trình nghiên cứu liên quan về giải pháp phân lớp Naive Bayes. Chất lượng phân lớp được đánh giá qua hai thông số là độ chính xác và độ hồi tưởng ([2, 4, 9, 10, 13, 14]). Mục 3 đề xuất một công thức phân lớp trong phương pháp Naive Bayes. Kết quả thực nghiệm thi hành trên máy tìm kiếm Vinahoo chỉ ra rằng chất lượng phân lớp theo công thức mới là cao hơn so với công thức nguyên thủy. Một bài toán quan trọng trong phương pháp phân lớp Naive Bayes là bài toán xác định ngưỡng phân lớp ([2, 9]). Mục 4 trình bày phương pháp xác định ngưỡng phân lớp do chúng tôi đề xuất và kết quả thực nghiệm thi hành trên máy tìm kiếm Vinahoo. Cuối cùng là một số ý kiến bàn luận và định hướng phát triển nghiên cứu.

2. MỘT SỐ CÔNG TRÌNH LIÊN QUAN

Trong [2], Dunja Mladenic trình bày những nội dung quan trọng trong lĩnh vực khai phá text và khai phá Web. Cùng với các phương pháp k-NN và FOIL, phương pháp phân lớp Nave Bayes được tác giả coi như một phương pháp phổ biến để giải quyết bài toán phân lớp văn bản. Phương pháp này cũng được nhiều tác giả đề cập ([4, 5, 8, 9, 10, 12, 13]). Cơ sở lý thuyết của phương pháp là định lý Bayes về xác suất có điều kiện. Quan hệ giữa xác suất $P(A|B)$ xuất hiện biến cố A theo điều kiện đã xuất hiện biến cố B với xác suất $P(B|A)$ xuất hiện biến cố B theo điều kiện đã xuất hiện biến cố A (giả thiết hai biến cố A, B là độc lập):

$$P(A|B) = P(B|A) \times P(A)/P(B) \tag{1}$$

được cụ thể khi tính giá trị xác suất $P(\boldsymbol{c}|\boldsymbol{d})$ phân tài liệu $\boldsymbol{d}$ vào lớp $\boldsymbol{c}$ theo phương pháp phân lớp văn bản Naive Bayes là:

$$P(c|d) = \frac{P(c) \times P(d|c)}{\sum\limits_{k} P(c_k) \times P(d|c_k)} \,. \tag{2}$$

Theo hình thức biểu diễn văn bản thông qua sự xuất hiện các từ khóa trong văn bản đó, công thức (2) được Dunja Mladenic trình bày dưới dạng:

$$P(c|d) = \frac{P(c) \times \prod\limits_{w_j \in d} P(w_j|c)^{TF(w_j)}}{\sum\limits_{k} P(c_k) \times \prod\limits_{w_j \in d} P(w_j|c_k)^{TF(w_j)}} \,. \tag{3}$$

Trong công thức (3), $P(\boldsymbol{c})$ hoặc $P(\boldsymbol{c_k})$ là xác suất xuất hiện lớp $\boldsymbol{c}$ hoặc lớp $\boldsymbol{c_i}$ tương ứng (chú ý là $\sum\limits_{k} P(c_k) = 1$); $P(\boldsymbol{w_j}|\boldsymbol{c})$ hoặc $P(\boldsymbol{w_j}|\boldsymbol{c_k})$ là xác suất xuất hiện từ khóa $\boldsymbol{w_j}$ trong điều kiện xuất hiện lớp $\boldsymbol{c}$ hoặc lớp $\boldsymbol{c_k}$ tương ứng; $TF(\boldsymbol{w_j})$ là số lần xuất hiện của từ khóa $\boldsymbol{w_j}$ trong văn bản cần phân lớp $\boldsymbol{d}$ (một số trường hợp, $TF(\boldsymbol{w_j})$ được chuẩn hóa bằng tỷ số giữa số lần xuất hiện từ khóa $\boldsymbol{w_j}$ với tổng số lần xuất hiện mọi từ khóa trong $\boldsymbol{d}$). Các xác suất $P(\boldsymbol{c})$ hoặc $P(\boldsymbol{c_k})$ và xác suất $P(\boldsymbol{w_j}|\boldsymbol{c})$ hoặc $P(\boldsymbol{w_j}|\boldsymbol{c_k})$ được tính dựa theo một tập cho trước các văn bản với thuộc tính lớp đã được xác định (tập ví dụ học). Các công trình [2, 4, 5, 8, 9, 10, 12, 13] đưa ra một số công thức tính toán các giá trị xác suất $P(\boldsymbol{w_j}|\boldsymbol{c_k})$, mà đa số trong đó là biến thể của công thức lý thuyết. Theo Dunja Mladenic [2], bài toán xác định ngưỡng phân lớp là bài toán khó, nhiều trường hợp cần có kinh nghiệm của người thi hành hệ thống phân lớp. Nhiều công trình đề cập tới các phương pháp biểu diễn trang Web theo tập từ khóa

([3, 4, 9, 11, 13, 14]). Sen Slattery [13] trình bày bốn phương pháp biểu diễn nội dung trang Web theo các từ khóa. Giá trị xác suất phân lớp được tính toán theo công thức (3).

Báo cáo này giải quyết hai vấn đề khi tích hợp thành phần phân lớp vào máy tìm kiếm Vinahoo ([15, 16]). Đầu tiên, chúng tôi đề xuất một công thức mới tính toán xác suất phân lớp nhằm nâng cao chất lượng phân lớp. Chúng tôi tiến hành thử nghiệm thi hành thành phần phân lớp trên máy tìm kiếm Vinahoo và cho chạy trong môi trường Internet thực. Việc đánh giá kết quả dựa trên việc xem xét hai thông số điển hình là độ chính xác macro (macro-precision) và độ hồi tưởng macro (macro-recall) [13]. Vấn đề thứ hai là đề xuất một giải pháp xác định ngưỡng phân lớp. Xuất phát từ giá trị thô ban đầu, ngưỡng phân lớp được xác định qua việc làm tinh dần theo hai thuật toán. Kết quả cài đặt giải pháp xác định ngưỡng trong máy tìm kiếm Vinahoo cho thấy giải pháp được đề xuất là thực sự hiệu quả.

3. CÔNG THỨC PHÂN LỚP NAIVE BAYES

Để ý trong công thức (3), các đại lượng $P(\boldsymbol{w}_j|\boldsymbol{c})$ và $P(\boldsymbol{w}_j|\boldsymbol{c}_k)$ thường rất nhỏ, lại được tính toán theo phép nhân và hàm mũ (các giá trị $\mathrm{TF}(\boldsymbol{w}_j)$ thường khá lớn) làm cho khối lượng tính toán lớn dẫn đến không chỉ tốn thời gian tính toán mà còn làm tăng sai sót tính toán trong kết quả cuối cùng. Trong mục này chúng tôi phân tích và biến đổi công thức (2) để nhận được một công thức tính xác suất hậu nghiệm có chất lượng hơn.

3.1. Công thức phân lớp Naive Bayes

Theo lý thuyết xác suất Bayes, xuất phát từ công thức (2), công thức tính toán giá trị xác suất hậu nghiệm dựa trên mô hình tham số θ là:

$$P(c|d,\theta) = \frac{P(c|\theta) \times P(d|c,\theta)}{\sum_i P(c_i|\theta) \times P(d|c_i,\theta)}, \tag{4}$$

trong đó θ là mô hình tham số cần được xây dựng. Từ đây trở đi ngầm định sẵn mô hình tham số θ, cho nên để đơn giản các công thức chúng ta không chỉ dẫn mô hình tham số θ nữa.

Theo mô hình vectơ biểu diễn văn bản ([2, 3, 4, 8, 10, 13, 14]), văn bản $\boldsymbol{d}$ được biểu diễn dưới dạng $\{(\boldsymbol{w}_1, n_1), (\boldsymbol{w}_2, n_2), ..., (\boldsymbol{w}_{|d|}, \boldsymbol{n}_{|d|})\}$. Ở đây, $\{\boldsymbol{w}_i\}$ là tập các từ khóa xuất hiện trong văn bản $\boldsymbol{d}$, còn n_i tương ứng là số lần xuất hiện từ khóa $\boldsymbol{w}_i$ trong $\boldsymbol{d}$. Như vậy, $\boldsymbol{d}$ được xem là sự kiện được xác định qua tập các từ khóa xuất hiện trong nó cùng với số lượt xuất hiện tương ứng của mỗi từ khóa. Sự kiện (văn bản) $\boldsymbol{d}$ được mô tả theo bộ $|d|+1$ các sự kiện (tính chất) như sau:

1. Số lượng từ khóa khác nhau trong $\boldsymbol{d}$ là $|d|$.
2. Từ khóa $\boldsymbol{w}_1$ xuất hiện n_1 lần.
3. Từ khóa $\boldsymbol{w}_2$ xuất hiện n_2 lần.

...

$|d|+1$. Từ khóa $\boldsymbol{w}_{|d|}$ xuất hiện $n_{|d|}$ lần.

Gọi X biến ngẫu nhiên biểu diễn số lượng từ khóa khác nhau trong một tài liệu, còn ω_i là biến ngẫu nhiên biểu diễn số lượng từ khóa $\boldsymbol{w}_i$ xuất hiện trong một tài liệu, chúng ta có:

$$P(d|c) = P(\{X = |d|,\ \omega_1 = n_1, \omega_2 = n_2, ..., \omega_{|d|} = n_{|d|}\}|c). \tag{5}$$

Thừa nhận giả thiết là số lượng từ khóa xuất hiện trong tài liệu là độc lập với ngữ nghĩa của tài liệu, chúng ta có:

$$P(d|c) = P(X = |d| \times P(\{\omega_1 = n_1, \omega_2 = n_2, ..., \omega_{|d|} = n_{|d|}\}|c). \quad (6)$$

Trong các hệ thống phân lớp, các biến cố $\omega_1, \omega_2, ..., \omega_{|d|}$ thường được coi là các biến cố độc lập nhau từng đôi một, vì vậy chúng ta nhận được:

$$P(d|c) = P(X = |d| \times P(\{\omega_1 = n_1|c) \times P(\omega_2 = n_2|c) \times \cdots \times P(\omega_{|d|} = n_{|d|}|c). \quad (7)$$

Đặt $N = \sum_{i=1}^{|d|} n_i$ là tổng số lượt xuất hiện các từ khóa trong tài liệu $\boldsymbol{d}$. Thực hiện lược đồ xác suất S có dạng như sau:

+ Chọn ngẫu nhiên một giá trị cho N.

+ Thực hiện N lần phép thử có tính chất là xác suất xuất hiện từ khóa $\boldsymbol{w}_i$ trong miền ngữ nghĩa của lớp c là hằng số ($P(\boldsymbol{w}_i|\boldsymbol{c}) = \text{const}$) và xác suất không xuất hiện từ khóa $\boldsymbol{w}_i$ tương ứng là $(1 - P(\boldsymbol{w}_i|\boldsymbol{c}))$.

Thấy rằng lược đồ S chính là lược đồ Becnulli, vì vậy chúng ta nhận được:

$$P(\omega_i = n_i|c) = P(N) \times C_N^{n_i} \times (P(w_i|c))^{n_i} \times (1 - P(w_i|c))^{N-n_i} \quad (8)$$

Từ các công thức (2), (7) và (8), chúng ta có:

$$P(c|d) = \frac{P(c) \times P(X = |d|) \times (P(N))^{|d|} \times \prod_{i=1}^{|d|} C_N^{n_i} \times \prod_{i=1}^{|d|} (P(w_i|c)^{n_i} \times (1 - P(w_i|c))^{N-n_i}}{\sum_k P(c_k) \times P(X = |d|) \times (P(N))^{|d|} \times \prod_{i=1}^{|d|} C_N^{n_i} \times \prod_{i=1}^{|d|} (P(w_i|c_k)^{n_i} \times (1 - P(w_i|c_k))^{N-n_i}}$$

là tương đương với

$$P(c|d) = \frac{P(c) \times \prod_{i=1}^{|d|} P(w_i|c)^{n_i} \times (1 - P(w_i|c))^{N-n_i}}{\sum_k P(c_k) \times \prod_{i=1}^{|d|} (P(w_i|c_k)^{n_i} \times (1 - P(w_i|c_k))^{N-n_i}}$$

hay

$$P(c|d) = \frac{P(c) \times \prod_{i=1}^{|d|} (1 - P(w_i|c))^N \times \left(\frac{P(w_i|c)}{(1 - P(w_i|c)}\right)^{n_i}}{\sum_k P(c_k) \times \prod_{i=1}^{|d|} (1 - P(w_i|c_k))^N \times \left(\frac{P(w_i|c_k)}{(1 - P(w_i|c_k)}\right)^{n_i}}. \quad (9)$$

Dù rằng trong công thức (9) các giá trị $(1 - P(\boldsymbol{w}_i|\boldsymbol{c}))$ và $\frac{P(w_i|c)}{(1-P(w_i|c)}$ nhìn chung đã không còn quá nhỏ như bản thân $P(\boldsymbol{w}_i|c)$, song trong công thức này vẫn còn phải tính các lũy thừa

của chúng, vì vậy cần làm giảm bớt các lũy thừa như vậy. Theo dạng thức của công thức (9), chúng tôi đề xuất công thức biến thể nó như dưới đây để tính xấp xỉ giá trị xác suất hậu nghiệm. Áp dụng phép chuẩn hóa $\frac{n_i-0}{N-0}(1-0)+0$ đưa giá trị n_i trong miền nguyên $[0, N]$ thành một giá trị trong miền thực $[0, 1]$, chúng ta nhận được (ở đây, $n'_i = \frac{n_i}{N}$):

$$P(c|d) = \frac{P(c) \times \prod_{i=1}^{|d|}(1 - P(w_i|c)) \times \left(\frac{P(w_i|c)}{1 - P(w_i|c)}\right)^{n'_i}}{\sum_k P(c_k) \times \prod_{i=1}^{|d|}(1 - P(w_i|c_k)) \times \left(\frac{P(w_i|c_k)}{1 - P(w_i|c_k)}\right)^{n'_i}} . \tag{10}$$

Hơn nữa, nếu một từ khóa xuất hiện trong nhiều lớp thì ý nghĩa phân lớp của từ khóa đó sẽ bị "loãng" đi có nghĩa là giá trị đóng góp của từ khóa đó vào xác suất hậu nghiệm phải được giảm đi. Gọi CF($\boldsymbol{w}_i$) là số lượng lớp mà miền ngữ nghĩa có chứa từ khóa $\boldsymbol{w}_i$ thì giá trị CF($\boldsymbol{w}_i$) là một yếu tố ảnh hưởng tới giá trị xác suất để tài liệu $\boldsymbol{d}$ thuộc vào một lớp vì vậy cần được tính đến. Chúng tôi đưa yếu tố đó vào trong công thức (10) dưới dạng bậc lũy thừa đối với thành phần tương ứng với $\boldsymbol{w}_i$. Khi đó, nhận được công thức:

$$P(c|d) = \frac{P(c) \times \prod_{i=1}^{|d|}\left[\left(1 - P(w_i|c)\right) \times \left(\frac{P(w_i|c)}{1 - P(w_i|c)}\right)^{n'_i}\right]^{CF(w_i)}}{\sum_k P(c_k) \times \prod_{i=1}^{|d|}\left[\left(1 - P(w_i|c_k)\right) \times \left(\frac{P(w_i|c_k)}{1 - P(w_i|c_k)}\right)^{n'_i}\right]^{CF(w_i)}} . \tag{11}$$

Tuy đòi hỏi một khối lượng tính toán bổ sung, nhưng công thức (11) cho phép đưa vào mối quan hệ giữa từ khóa với các lớp.

$$A(d|c) = \prod_{i=1}^{|d|}\left[\left(1 - P(w_i|c)\right) \times \left(\frac{P(w_i|c)}{1 - P(w_i|c)}\right)^{n'_i}\right]^{CF(w_i)} . \tag{12}$$

Công thức (11) có thể viết lại dưới dạng:

$$P(c|d) = \frac{P(c) \times A(d|c)}{\sum_k P(c_k) \times A(d|c_i)} .$$

Tương tự [2], chúng tôi đề nghị các công thức tính xác suất tiên nghiệm $P(c)$ và $P(w_i|c)$ (ký hiệu tương ứng là θ_c và $\theta_{c,i}$) trong mô hình tham số θ như sau:

$$\theta_c = \frac{1 + M_c}{K + M} \quad \text{và} \quad \theta_{c,i} = \frac{1 + \sum_{d \in c} n_{d,w_i}}{L + \sum_w \sum_{d \in c} n_{d,w}} . \tag{13}$$

Trong đó:
K là số lớp tài liệu trong hệ thống,
M là số lượng tài liệu trong tập huấn luyện,
M_c là số lượng tài liệu trong tập huấn luyện thuộc vào lớp c,
$n_{d,w_i}(n_{d,w})$ là số lượt xuất hiện từ khóa $\boldsymbol{w}_i$ ($\boldsymbol{w}$) trong tài liệu $\boldsymbol{d}$.

3.2. Kết quả thực nghiệm

Chúng tôi lấy ngẫu nhiên tập dữ liệu gồm 1722 trang Web từ trang chủ http://vnexpress.net thuộc 6 lớp Vi tính, Thể thao, Pháp luật, Sức khỏe, Văn hóa, Xã hội và chia tập này thành tập huấn luyện (với 802 trang Web) và tập kiểm tra (với 920 trang web) có phân bố lớp như trình bày trong Bảng 1.

Bảng 1. Tập dữ liệu thử nghiệm thứ nhất

Lớp / Tập	Vi tính	Thể thao	Pháp luật	Sức khoẻ	Văn hóa	Xã hội	Tổng
Huấn luyện	261	108	57	69	157	150	802
Kiểm tra	30	138	193	196	115	148	920
Tổng số	391	246	250	265	272	298	1722

Sau khi tính toán tham số mô hình dựa trên tập huấn luyện, tiến hành phân lớp tập dữ liệu kiểm tra lần lượt bằng các bộ phân lớp Naive Bayes dựa trên hai công thức (3) và (11). Bảng 2 và Bảng 3 tương ứng trình bày kết quả kiểm tra khi bộ phân lớp sử dụng công thức (3) và công thức (11).

Bảng 2. Kết quả phân lớp theo công thức (3)

P / R	Vi tính	Thể thao	Pháp luật	Sức khỏe	Văn hóa	Xã hội
Vi tính	130					
Thể thao	85	53				
Pháp luật	193		0			
Sức khỏe	195			0	1	
Văn hóa	3				112	
Xã hội	148				0	0
P_1 =(130/754 + 53/53 +0+0+112/113+0)/6=0,36 R_1=(130/130+ 53/138+0/193+0/196+112/115+0/148)/6=0,39						

Trong các bảng này, chỉ số hàng R để chỉ nhãn lớp thực sự của tài liệu, chỉ số cột P để chỉ nhãn lớp do bộ phân lớp gán cho tài liệu. Chẳng hạn, ở Bảng 2, số 85 tại ô giao của cột Vi tính và hàng Thể thao chỉ ra rằng có 85 tài liệu thực sự thuộc loại Thể thao song bộ phân lớp theo công thức (3) đã phân chúng vào lớp Vi tính.

Để đánh giá chất lượng phân lớp, chúng tôi sử dụng các thông số macro-precision và macro-recall [13]. Trong Bảng 2, P_1 là giá trị macro-precision còn R_1 là giá trị macro-recall. Tương tự, trong Bảng 3, P_2 là giá trị macro-precision còn R_2 là giá trị macro-recall.

Kết quả thực nghiệm trên cho thấy chất lượng của bộ phân lớp theo công thức (11) tốt hơn hẳn chất lượng bộ phân lớp theo công thức (3) (cả $P_2 >> P_1$ và $R_2 >> R_1$). Tuy còn

hạn chế về quy mô tính toán thực nghiệm do gặp khó khăn trong việc chạy Vinahoo trên môi trường Internet thực, song từ tính ngẫu nhiên của việc chọn tập dữ liệu thử nghiệm cho phép chúng tôi tin tưởng vào tính hiệu quả của công thức (11).

Bảng 3. Kết quả phân lớp theo công thức (11)

P / R	Vi tính	Thể thao	1	Sức khỏe	Văn hóa	Xã hội
Vi tính	**130**					
Thể thao	11	**126**			1	
Pháp luật	37		**126**		3	27
Sức khỏe	12			**103**	77	4
Văn hóa					**115**	
Xã hội					3	**145**
P_2=(130/190+126/126+126/126+103/103+115/199+145/176)/6=0,85 R_2=(130/130+126/138+126/203+103/196+115/115+145/148)/6=0,84						

4. GIẢI PHÁP XÁC ĐỊNH NGƯỠNG PHÂN LỚP

Với tài liệu mới $\boldsymbol{d}$ cần phân lớp, bộ phân lớp Naive Bayes thực hiện việc tính toán các giá trị $P(\boldsymbol{c}_i|\boldsymbol{d})$ đối với tất cả các lớp $\boldsymbol{c}_i$ và sau đó phân tài liệu $\boldsymbol{d}$ vào lớp $\boldsymbol{c}_i$ có giá trị $P(\boldsymbol{c}_i|\boldsymbol{d})$ cực đại. Để tăng độ mềm dẻo cho bộ phân lớp, người ta bổ sung tập giá trị ngưỡng phân lớp $\{\boldsymbol{CtgTsh}_i\}$, mỗi giá trị $\boldsymbol{CtgTsh}_i$ tương ứng với mỗi lớp $\boldsymbol{c}_i$ [2, 12, 13]. Bộ phân lớp phân tài liệu $\boldsymbol{d}$ vào lớp $\boldsymbol{c}_i$ nếu như xác suất hậu nghiệm $P(\boldsymbol{c}_i|\boldsymbol{d})$ vượt quá ngưỡng $\boldsymbol{CtgTsh}_i$ tương ứng. Tuy nhiên, bài toán xác định các giá trị ngưỡng phân lớp $\boldsymbol{CtgTsh}_i$ để bộ phân lớp hoạt động hiệu quả lại là một bài toán khó [2, 13]. Dưới đây là đề xuất của chúng tôi về giải pháp giải quyết vấn đề này.

4.1. Giải pháp xác định ngưỡng phân lớp

Gọi D_{learn} là tập các tài liệu dùng để huấn luyện bộ phân lớp. Quá trình xác định ngưỡng phân lớp được chia thành hai bước.

Bước thứ nhất thực hiện việc xác định giá trị thô cho ngưỡng, còn bước thứ hai thực hiện việc tinh chỉnh ngưỡng. Giá trị thô cho $\boldsymbol{CtgTsh}_i$ chính là giá trị cực tiểu của các giá trị xác suất $P(\boldsymbol{c}_i|\boldsymbol{d})$ đối với mọi tài liệu d có trong D_{learn} mà thuộc vào lớp $\boldsymbol{c}_i$.

Thuật toán 1 mô tả giải pháp xác định giá trị các ngưỡng. Trong thuật toán, để đơn giản công việc tính toán, chúng tôi sử dụng giá trị $A(d|c)$ thay cho $P(d|c)$. Với các giả thiết đã đặt ra về các biến cố liên quan thì các giá trị xác suất $P(\boldsymbol{c}_i|\boldsymbol{d})$ tuân theo phân bố xác suất Trêbusep, vì vậy giải pháp tinh chỉnh ngưỡng phân lớp trong thuật toán Naive Bayes được định hướng theo phân bố xác suất này. Nội dung tính toán chi tiết được trình bày tại Bước 3 của Thuật toán 1.

Thuật toán 1

1. Xây dựng mô hình tham số θ (tính các giá trị theo công thức (13)) đối với tập dữ liệu huấn luyện D_{learn};
2. For mỗi tài liệu $d \in D_{\text{learn}}$ do
   For mỗi tài liệu $c \in C$ do
   {
   Tính $P(c|d)$ theo công thức (11);
   Tính $A(d|c)$ theo công thức (12);
   }
3. For mỗi lớp $c \in C$ do
{
   othres $\leftarrow 1$;
   For mỗi tài liệu $d \in D_{\text{learn}}$ có nhãn là c do
   {
   Tính giá trị $P(c|d)$ theo công thức (11);
   if ($P(c|d) <$ othres) then othres $\leftarrow P(c|d)$;
   }
   tmp $\leftarrow \sum_{d \in c, d \in D_{\text{learn}}} P(c|d) \times A(d|c)$;
   tmpv $\leftarrow \sum_{d \in c, d \in D_{\text{learn}}} P^2(c|d) \times A(d|c)$;
   tmpv $\leftarrow SQRT(tmpv - tmp^2)$;
   $n \leftarrow 1$;
   while $((tmp - n \times tmpv) >$ othres) do
   {
   $CtgTsh_c \leftarrow tmp - n \times tmpv$;
   $n \leftarrow n + 1$;
   }
}

Như đã được chỉ ra trong [2, 12, 13], thuật toán phân lớp có sử dụng ngưỡng cần một số điều chỉnh nhỏ so với thuật toán không sử dụng ngưỡng để đảm bảo tính xác định của thuộc tính lớp khi thực hiện phân lớp. Khi mà có quá một lớp để cho xác suất hậu nghiệm vượt quá ngưỡng, chúng tôi chọn lớp làm cực đại tỷ số giữa giá trị xác suất hậu nghiệm so với ngưỡng. Trong trường hợp có quá một lớp có tỷ số đạt giá trị cực đại đó, chúng tôi chọn lớp có chỉ số nhỏ nhất trong chúng.

4.2. Kết quả thực nghiệm

Trong thực nghiệm thứ hai, chúng tôi sử dụng lại bộ dữ liệu huấn luyện của thực nghiệm thứ nhất, còn tập dữ liệu kiểm tra gồm 2434 trang Web được lấy ngẫu nhiên mà phân bố lớp được trình bày trong Bảng 4.

Bảng 4. Tập dữ liệu thử nghiệm thứ hai

Lớp / Tập	Vi tính	Thể thao	Pháp luật	Sức khoẻ	Văn hóa	Xã hội	Tổng
Huấn luyện	261	108	57	69	157	150	802
Kiểm tra	352	500	173	409	500	500	2434
Tổng số	613	608	230	478	657	650	3236

Bảng 5. Kết quả phân lớp theo công thức (3) không điều chỉnh ngưỡng

P / R	Vi tính	Thể thao	Pháp luật	Sức khỏe	Văn hóa	Xã hội
Vi tính	252	12				88
Thể thao	191	219				90
Pháp luật						173
Sức khỏe		14		210		85
Văn hóa		100		94	146	160
Xã hội		130	100	60		210
P_3=(252/443 + 219/475 +0/100+210/364+146/146+210/806)/6=0,48 R_3=(252/352 + 219/500 +0/173+210/409+146/500+210/500)/6=0,40						

Bảng 6. Kết quả phân lớp theo công thức (3) có điều chỉnh ngưỡng

P / R	Vi tính	Thể thao	Pháp luật	Sức khỏe	Văn hóa	Xã hội
Vi tính	263	21				68
Thể thao	182	218			30	70
Pháp luật			86			87
Sức khỏe		22		293		94
Văn hóa		112		82	227	79
Xã hội		83	60	4		353
P_4=(263/445 + 218/456 +86/144+293/379+227/257+353/751)/6=0,63 R_4=(263/352 + 218/500 +86/174+293/409+227/500+353/500)/6=0,59						

Bảng 5 và Bảng 6 trình bày kết quả phân lớp cho tập kiểm tra theo công thức (3) tương

ứng với các trường hợp ngưỡng điều chỉnh. Các giá trị macro-precision và macro-recall trong trường hợp có điều chỉnh ngưỡng ($P_4 = 0{,}63$, $R_4 = 0{,}59$) đều tốt hơn đáng kể so với trường hợp không điều chỉnh ngưỡng ($P_3 = 0{,}48$, $R_3 = 0{,}40$).

5. BÀN LUẬN

Để tích hợp được thành phần phân lớp Naive Bayes, chúng tôi đã bổ sung một số bảng dữ liệu, bổ sung và thay đổi một số môđun chương trình vào máy tìm kiếm Vinahoo. Kết quả nghiên cứu của chúng tôi cho thấy tính khả thi của việc phát triển máy tìm kiếm Vinahoo theo hướng tích hợp các yếu tố khai phá Web nói chung và khai phá text tiếng Việt nói riêng. Song song với việc tích hợp bộ phân lớp như trình bày trên đây, hiện nay Vinahoo đang được chúng tôi phát triển theo hướng nâng cấp thành phần tính hạng trang Web và song song hóa thành phần dò tìm. Chúng tôi đã định hướng tích hợp phương pháp biểu diễn trang Web thể hiện ngữ nghĩa trang Web đa ngôn ngữ [2, 3, 7, 9, 14], bổ sung chức năng xử lý web site [4, 8] vào Vinahoo nhằm xây dựng một máy tìm kiếm tiếng Việt tiên tiến.

TÀI LIỆU THAM KHẢO

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, Sriram Raghavan Searching the web, *Technical Report*, Computer Science Department, Stanford University, 2000.

[2] Dunja Mladenic, "Machine learning on non-homogeneous, distributed text data", Doctoral dissertation, University of Ljubljana, Slovenia, 1998.

[3] E. Herrera-Viedma, Modeling the retrieval process of an information retrieval system using an ordinal fuzzy lingguistic approach, *Journal of American Society for Information Science and Technology* (JASIS), **52** (6) (2001) 460-475.

[4] Ha Quang Thuy and Nguyen Tri Thanh, A web site representation method using concept vectors and web site classifications (Gửi đăng Tạp chí Tin học và Điều khiển học).

[5] Hwanjo Yu, Jiawei Han, Kevin Chen-Chuan, PEBL: Positive example based learning for web page classification using SVM, *Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aberta, Canada, July 23-26, 2002, 239–248.

[6] H. Yu, J. Han, and K.C.C. Chang, PEBL: Web page classification without negative examples, *IEEE Transaction on Knowledge and Data Engineering* **16** (1) (2004) 70–81.

[7] H.-Y. Kao, S.-H. Lin, J.-M. Ho, and M.-S. Chen, Mining web informative structures and contents based on entropy analysis, *IEEE Transaction on Knowledge and Data Engineering* **16** (1) (2004) 41–55.

[8] Martin Ester, Hans-Peter Kriegei, Matthias Schubert, Web Site Mining: A new way to spot competitors, customerrs and suppliers in the world wide web, *Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aberta, Canada, July 23-26 2002, 249–258.

[9] Nguyen Ngoc Minh, Nguyen Tri Thanh, Ha Quang Thuy, Luong Song Van, Nguyen Thi Van, A Knowledge discovery model in fulltext databases,*Proceedings of the First Workshop of International Joint Research "Parallel Computing, Data Mining and Optical Networks"*, Japan Advanced Institute of Science and Technology (JAIST), Tatsunokuchi, March 7, 2001, 59–68.

[10] P. Baldi, P. Frasconi, P. Smyth, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, Wiley, 2003.

[11] Pavel Calado, Marco Cristo, Edleno Moura, Nivio Ziviani, Berthier Ribeiro-Neto, Marcos Andre Goncalves, *Combining link-based and content-based methods for web document classification.*

[12] Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Publishers, 2003.

[13] Sen Slattery, "Hypertext classification", Doctoral dissertation (CMU-CS-02-142), School of Computer Science, Carnegie Mellon University, 2002.

[14] Son Doan, Susumu Horiguchi, A new text representation method using fuzzy concepts in text categorization, *JAIST Science Reports,* 2002.

[15] Http://www.fotech.vnu.edu.vn/vinahooo/

[16] Http://www.aspseek.com/

Nhận bài ngày 3 - 9 - 2004
Nhận lại sau sửa ngày 11 - 8 - 2005