# TÓM TẮT VÀ TRÍCH RÚT TÀI LIỆU VĂN BẢN TRONG THƯ VIỆN SỐ

ĐỖ QUANG VINH
Bộ môn Công nghệ Thông tin - Trường Đại học Văn hoá Hà Nội

## 1. MỞ ĐẦU

Hiện nay, thư viện số là một trong những hướng nghiên cứu chính về công nghệ thông tin trên thế giới. Bài toán tóm tắt và trích rút tài liệu văn bản trong thư viện số đang được nhiều nhà nghiên cứu về các ngành khoa học khác nhau như tin học, toán học và ngôn ngữ học quan tâm. Mục tiêu của bài báo là nhận được một số phương pháp có thể lập trình trên máy tính, như vậy, máy tính sau khi được cung cấp một tài liệu văn bản, sẽ sản xuất một tóm tắt giàu thông tin. Nhưng bài toán tóm tắt tổng quát gặp phải khó khăn lớn vì nó bao hàm các bài toán khác, xây dựng các câu mới. Một cách tóm tắt hạn chế hơn là trích rút các câu quan trọng nhất.

Tất nhiên, chúng ta còn cách khá xa một giải pháp thỏa đáng ngay cả đối với bài toán đơn giản hơn về trích rút tài liệu. Ở đây, chúng tôi trình bày một số kết quả nghiên cứu lý thuyết về bài toán. Cách tiếp cận của chúng tôi chủ yếu là áp dụng các phương pháp lấy mẫu và ước lượng thống kê tài liệu văn bản trong thư viện số.

## 2. TÓM TẮT TỐI ƯU

Cho T là một văn bản cho trước và A là một tóm tắt của T. Cho I(T) và I(A) tương ứng là thông tin chứa trong T và A, L(T) và L(A) là độ dài của T và A. Ở đây, bài toán đánh giá I và L không được xét và được thảo luận ở mục 4. Bây giờ, chúng ta có thể yêu cầu A chứa một phần thông tin định rõ chứa trong T. Điều này cực tiểu hoá độ dài trong tất cả tóm tắt thoả mãn yêu cầu trên, sau đó có thể được coi là tối ưu. Như một lựa chọn, chúng tôi yêu cầu độ dài của A là một phần định rõ của tóm tắt về T và xác định là tối ưu chứa lượng thông tin cực đại. Chính xác hơn, chúng tôi có:

### Định nghĩa 1

Một tóm tắt $A_L$ của một văn bản cho trước T được gọi là một tóm tắt có độ dài cực tiểu chứa $\alpha$ lượng thông tin liên quan nếu $I(A_L) = \alpha \cdot I(T)$ và $L(A_L) \le L(A)$ đối với mọi tóm tắt của A về T sao cho $I(A) = \alpha \cdot I(T)$. Một tóm tắt $A_I$ của T được gọi là tóm tắt thông tin cực đại có độ dài $\beta$ liên quan, nếu $L(A_I) = \beta \cdot L(T)$ và $I(A_I) \ge I(A)$ sao cho $L(A) = \beta \cdot L(T)$.

### Nhận xét 1

Có thể xuất hiện các yêu cầu có thể là $I(A) \ge \alpha \cdot I(T)$ và $L(A) \le \beta \cdot L(T)$ tương ứng. Nhưng một tóm tắt A đối với $I(A) > \alpha \cdot I(T)$ chẳng hạn, không thể là loại có độ

dài cực tiểu. Bởi vì L(A) tương ứng có thể bị giảm bằng cách loại bỏ lượng thông tin thừa I(A) - $\alpha \cdot$ I(T).

Bài toán tổng quát nhận được độ dài cực tiểu hoặc các tóm tắt thông tin cực đại rõ ràng là khó giải quyết, vì nó yêu cầu xây dựng các câu mới. Tiếp theo, sự đưa vào công thức hạn chế hơn dựa vào trích rút câu được định nghĩa và khảo sát. Cho s là câu bất kỳ trong T và I(s) và L(s) là thông tin chứa trong s và độ dài của s tương ứng. Chúng tôi giả thiết I(s) $\geq$ 0 và L(s) > 0 đối với mọi s trong T và nếu E là một trích rút, nghĩa là, một tập câu của T thì I(E) và L(E), thông tin chứa trong E và độ dài E là như sau:

$$I(E) = \sum_{s \in E} I(s) \quad \text{và} \quad L(E) = \sum_{s \in E} L(s) \qquad (1)$$

Ở các ứng dụng thực tế, chúng tôi cho rằng giả thiết trên liên quan đến L(E) là hợp lý. Mặt khác, giả thiết về I(E) phải được coi chỉ là một xấp xỉ. Thực tế, nó được đưa vào rộng rãi để tiện thao tác toán học. Tuy nhiên, nó cũng nên chú ý rằng bằng câu chúng tôi không cần thiết hiểu là một câu theo nghĩa truyền thống. Nó có thể là một nhóm câu hoặc một đoạn v.v... Nếu có các trường hợp, giả thiết về I(E) có khả năng được thoả mãn hơn.

Để nhận được một đoạn trích rút, chúng tôi lựa chọn một phần nhất định trong số câu từ T và loại bỏ các câu còn lại. Ở đây, chỉ có hai khả năng, nghĩa là, lựa chọn và loại bỏ. Tuy nhiên, đối với một số lý do kỹ thuật được giải thích sau đây, chúng tôi đưa xác suất vào quá trình lựa chọn.

**Định nghĩa 2**

Một hàm trích rút ngẫu nhiên F(s) là một hàm định nghĩa đối với mọi s $\in$ T sao cho 0 $\leq$ F(s) $\leq$ 1. Để sản xuất một đoạn trích rút bằng cách dùng một F(s) cho trước, chúng tôi tiến hành như sau. Đối với mọi s $\in$ T, kiểm tra giá trị của F(s). Nếu F(s) = 1, s được lựa chọn. Nếu F(s) = 0 , s bị loại bỏ. Nếu 0 < F(s) < 1, chúng tôi thực hiện một thử nghiệm ngẫu nhiên và lựa chọn s với xác suất F(s). Loại kỹ thuật ngẫu nhiên này hay được sử dụng trong thống kê và có các ưu điểm nhất định. Mục đích của chúng tôi là thiết lập định lý 1. Vì có khả năng bị bao hàm, F(s) giống nhau, áp dụng cho lần thứ 2, có thể không sinh ra trích rút như nhau như lần 1. Nhưng lượng thông tin trung bình I(F) và độ dài L(F) có thể được định nghĩa, trong đó

$$I(F) = \sum_{s \in T} F(s)\ I(s), \quad L(F) = \sum_{s \in T} F(s)\ L(s) \qquad (2)$$

## 3. HÀM TRÍCH RÚT TỐI ƯU

Bây giờ, chúng tôi đưa vào hai loại hàm trích rút tối ưu.

**Định nghĩa 3**

Một hàm trích rút $F^*(s)$ được coi là loại có độ dài cực tiểu, tương ứng với một $\alpha$ cho trước, $0 \le \alpha \le 1$, nếu $I(F^*) = \alpha \cdot I(T)$ và $L(F^*) \le L(F)$ đối với mọi hàm trích rút F sao cho $I(F) = \alpha \cdot I(T)$. $F^*(s)$ được coi là loại thông tin cực đại, tương ứng với một $\beta$ cho trước, $0 \le \beta \le 1$, nếu $L(F^*) = \beta \cdot L(T)$ và $I(F^*) \ge I(F)$ đối với mọi F sao cho $L(F) = \beta \cdot L(T)$.

Các đoạn trích rút sản xuất bởi hàm trích rút độ dài cực tiểu hoặc thông tin cực đại được đặt tên phù hợp.

Ở đây, chúng tôi chứng minh

**Định lý 1**

Cho $F_c(s)$ là một hàm trích rút sao cho

$$\begin{aligned} F_c(s) &= 1 \text{ nếu } I(s) > c \cdot L(s), \\ &= p \text{ nếu } I(s) = c \cdot L(s), \\ &= 0 \text{ nếu } I(s) < c \cdot L(s), \end{aligned} \qquad (3)$$

trong đó $c \ge 0$, $0 \le p \le 1$ và $p = 0$ nếu $c = 0$. Nếu $I(F_c) = \alpha \cdot I(T)$ thì $F_c$ một hàm trích rút có độ dài cực tiểu tương ứng với $\alpha$. Nếu $L(F_c) = \beta \cdot L(T)$ thì $F_c$ là một hàm trích rút có độ dài cực đại tương ứng với $\beta$.

*Chứng minh*:

Cho F là hàm trích rút bất kỳ sao cho $I(F) = \alpha \cdot I(T)$. Cho $T_1$, $T_2$ và $T_3$ tương ứng là các tập của tất cả s thuộc T thoả mãn $I(s) > c \cdot L(s)$, $I(s) = c \cdot L(s)$, $I(s) < c \cdot L(s)$. Sau đó,

$$L(F_c) - L(F) = \sum_{s \in T} [F_c(s) - F(s)] L(s) = \sum_{s \in T_1} + \sum_{s \in T_2} + \sum_{s \in T_3} \qquad (4)$$

Xét trường hợp trong đó $c > 0$. Đối với $s \in T_1$, $F_c(s) - F(s) \ge 0$ và $L(s) < I(s)/c$.

Do đó,

$$\sum_{s \in T_1} [F_c(s) - F(s)]\, L(s) \le (1/c) \sum_{s \in T_1} [F_c(s) - F(s)]\, I(s)$$

Suy diễn tương tự cho $\sum_{s \in T_2}$ và $\sum_{s \in T_2}$, chúng ta nhận được từ (4)

$\sum_{s \in T_1} [F_c(s) - F(s)]\, L(s) \le (1/c) \sum_{s \in T_1} [F_c(s) - F(s)]\, I(s) = 0$. Bây giờ, giả sử $c = 0$. $T_1$ và $T_2$ tương ứng là các tập của tất cả s đối với $I(s) > 0$ và $I(s) = 0$ tương ứng và $T_3$ rỗng.

Vì $I(F) = \sum_{s \in T_1} F(s)\, I(s) = I(F_0) = \sum_{s \in T_1} F_0(s)\, I(s) = I(T)$, chúng ta nhận thấy $F(s) = 1$.

Do đó, $\sum_{s\in T_1}[F_c(s)-F(s)]\,L(s)=0$. Hơn nữa, vì p = 0, $F_0(s) \geq F(s)$ đối với mọi $s \in T_2$ và $\sum_{s\in T_1}[F_c(s)-F(s)]\,L(s)\leq 0$. Vì vậy, chúng ta lại có $L_0(F)-L(F)\leq 0$. Cuối cùng, theo cách tương tự chúng ta chỉ ra $I(F_c) \geq I(F)$ đối với mọi F sao cho $L(F) = \beta \cdot L(T)$.

**Nhận xét 2**

Định lý trên phát biểu một câu s được trích rút chỉ nếu $I(s)/L(s) \geq c$. Định lý tương tự với Bổ đề Neyman Pearson nổi tiếng trong lý thuyết thống kê về kiểm định giả thuyết ([4], [7]).

Bây giờ, chúng tôi chỉ ra đối với α và β cho trước, tồn tại c và p sao cho F tương ứng của (3) là một hàm trích rút có độ dài cực tiểu tương ứng với α hoặc một hàm trích rút thông tin cực đại tương ứng với β. Chúng tôi cũng chỉ ra c và p có thể được xác định hoặc ước lượng chính xác như thế nào.

**Định lý 2**

Đối với $0 \leq \alpha, \beta \leq 1$, tồn tại một $F_{c_\alpha}$ và một $F_{c_\beta}$ có dạng cho trước bởi (3) sao cho $I(F_{c_\alpha}) = \alpha \cdot I(T)$ và $L(F_{c_\beta}) = \beta \cdot L(T)$.

*Chứng minh*:

Chúng ta sẽ chỉ ra tồn tại F sao cho $I(F_c) = \alpha \cdot I(T)$. Nếu c = 0 thì $I(F_c) = I(T)$. Cho c' > c. Bằng định nghĩa về $F_{c'}$, $F_{c'} \neq 0$ chỉ nếu $I(s) \geq c'$. L(s) đưa đến $I(s) \geq c \cdot L(s)$. Do đó, $F_{c'}(s) \neq 0$ chỉ nếu $F_{c'}(s) = 1$, hoặc $F_{c'}(s) \leq F_c(s)$ đối với mọi $s \in T$. Tiếp theo $I(F_{c'}) \leq I(F_c)$, hoặc $F_c$ là hàm không tăng của c (không quan tâm đến giá trị p). Hơn nữa, vì T là một tập hữu hạn và L(s) > 0, tồn tại các số $K_1$ và $K_2$ dương sao cho $I(s) < K_1$ và $L(s) > K_2$ đối với mọi $s \in T$. Bây giờ, đối với c đủ lớn, $K_1 < cK_2$. Do đó, đối với c' như thế, tập s đối với nó $I(s) \geq c \cdot L(s)$ là rỗng và $I(F_c) = 0$ đối với $F_c$ tương ứng bất kỳ. Như vậy, chúng ta nhận thấy vì c tăng từ 0 đến ∞, $I(F_c)$ giảm từ 1 đến 0, không quan tâm đến giá trị p.

Cho $F_c^1(s)$ và $F_c^2(s)$ là các hàm trích rút đối với chúng p =1 và 0 tương ứng đối với mọi $s \in T$ và mọi c thực $c \geq 0$. Mệnh đề trình bày ở cuối mục trước đưa ra đối với $F_c^1$. Bây giờ, đối với một α cho trước, $0 \leq \alpha \leq 1$, cho $c_\alpha$ là cận dưới lớn nhất của mọi c thực $c \leq 0$ sao cho $I(F_c^1) \leq \alpha \cdot I(t)$. Sau đó, $I(F_c^1) \leq \alpha \cdot I(t)$ nếu $c > c_\alpha$ và $I(F_c^1) \geq \alpha \cdot I(t)$ nếu $c < c_\alpha$. Chúng ta nhận thấy $I(F_{c_\alpha}^1) \geq \alpha \cdot I(t)$ và $I(F_{c_\alpha}^2) \leq \alpha \cdot I(t)$ ([4]). Cho $T_1$ và $T_2$ là các tập của tất cả $s \in T$ sao cho $I(s) \leq c_\alpha \cdot L(s)$ và $I(s) = c_\alpha \cdot L(s)$ tương ứng. Vì $I(F_{c_\alpha}^1) = \sum_{s\in T_1+T_2} I(s)$ và $I(F_{c_\alpha}^2) = \sum_{s\in T_1} I(s)$, chúng ta nhận thấy $\sum_{s\in T_1} I(s) + p_\alpha \sum_{s\in T_2} I(s) = \alpha$, trong đó $p_\alpha = \left[\alpha - \sum_{s\in T_1} I(s)\right] / \sum_{s\in T_2} I(s)$, nếu $\sum_{s\in T_2} I(s) > 0$ và $p_\alpha = 0$ nếu khác. Cho $F_{c_\alpha}$ được định nghĩa

bởi $c = c_\alpha$ và $p = p_\alpha$ thì $I(F_{c_\alpha}) = \alpha \cdot I(T)$. Bằng cách tương tự, chúng ta chỉ ra tồn tại một $F_{c_\beta}$ sao cho $L(F_{c_\beta}) = \beta \cdot L(T)$.

Định lý trên chỉ ra đối với $\alpha$ và $\beta$ cho trước, tồn tại các hàm trích rút tối ưu $F_{c_\alpha}$ và $F_{c_\beta}$. Bây giờ, chúng ta xét bài toán xác định và ước lượng $F_{c_\alpha}$ và $F_{c_\beta}$ . Bài toán ước lượng tăng lên khi xác định chính xác bao hàm quá nhiều công việc. Để xác định $F_{c_\alpha}$ hoặc $F_{c_\beta}$ , chúng tôi có thể tính giá trị của I(s)/ L(s) đối với mỗi một $s \in T$ và sắp xếp tất cả câu theo thứ tự giảm dần của I(s)/ L(s). Sau đó, các câu được trích rút lần lượt, $s_1$. $s_2$ , ... , $s_n$ , ... bắt đầu từ các câu có các giá trị lớn nhất của I(s)/ L(s), cho đến khi tổng tích luỹ của I(s) hoặc L(s) của các câu trích rút bằng hoặc vượt $\alpha \cdot I(T)$ hoặc $\beta \cdot L(T)$ đối với lần thứ nhất. Giả sử

$$\sum_{i=1}^{n} I(s_i) < \alpha \cdot I(T), \quad \sum_{i=1}^{n+1} I(s_n) > \alpha \cdot I(T)$$

và $I(s_{n+2})/ L(s_{n+2}) < I(s_{n+1})/ L(s_{n+1}) < I(s_n)/ L(s_n)$

Sau đó,

$$c_\alpha = I(s_{n+1})/ L(s_{n+1}) \text{ , } p_\alpha = \left[ \alpha \cdot I(T) - \sum_{i=1}^{n} I(s_i) \right] / I(s_{i+1})$$

và $F_{c_\alpha}$ được xác định. Các trường hợp khác có thể được giải quyết theo cách tương tự. Chú ý rằng không có nhu cầu tự sắp xếp thực sự các câu. Mỗi một câu được cho một khoá hoặc số định danh và các khoá được sắp xếp theo thứ tự giảm dần của I(s)/ L(s) tương ứng. Sau đó, chúng ta trích rút các khoá và câu tương ứng.

Phương pháp trên xác định chính xác các hàm trích rút tối ưu tương ứng với $\alpha$ và $\beta$. Nhưng phương pháp có một khiếm khuyết trong đó sắp xếp câu của T theo dãy có thể mất khá nhiều thời gian. Hơn nữa, từ quan điểm thực hành, không có nhu cầu thực nào xác định chính xác $F_{c_\alpha}$ và $F_{c_\beta}$, vì với $\alpha$ và $\beta$ hầu như được chọn tuỳ ý. Do đó, chúng ta đi đến bài toán tìm kiếm cách ước lượng $F_{c_\alpha}$ và $F_{c_\beta}$. Tiếp theo, chúng tôi đề xuất hai phương pháp dựa vào lý thuyết ước lượng thống kê.

Phương pháp thứ nhất chúng tôi thảo luận dựa vào giả thiết phân bố I(s) và L(s) của s trong T là siêu bội hoặc đa thức. Cho một mẫu ngẫu nhiên n câu được lấy từ văn bản cho trước T chứa tổng cộng N câu. Đối với mục đích thực hành, lấy mẫu hệ thống hoặc nhóm có thể được xem xét ([4]). Chúng tôi sử dụng lấy mẫu ngẫu nhiên chỉ để minh hoạ ý tưởng. Cho $T_n$ là tập hợp tất cả câu trong mẫu. Bây giờ áp dụng phương pháp trước để nhận được các trích rút tối ưu từ $T_n$. Cho $F^n_{c_\alpha}$ và $F^n_{c_\beta}$ là các hàm trích rút có độ dài cực tiểu và thông tin cực đại tương ứng với $\alpha$ và $\beta$. Chúng tôi sẽ chỉ ra $F^n_{c_\alpha}$ và $F^n_{c_\beta}$ là tối ưu theo ngữ nghĩa của định lý tiếp theo, khi sử dụng như các hàm trích rút đối với văn bản cho trước T.

### Định lý 3

Vì kích thước mẫu n tăng vô hạn $I(F^n_{c_\alpha})$ và $L(F^n_{c_\beta})$ tiến tới $\alpha \cdot I(T)$ và $\beta \cdot L(T)$. tương ứng theo xác suất. Tăng vô hạn nghĩa là tăng tới N đối với lấy mẫu không có thay thế và tăng tới $\infty$ đối với lấy mẫu có thay thế.

*Chứng minh*:

Cho $x_i$ và $y_j$ trong đó i, j = 1, 2, ..., là các giá trị riêng biệt của I(s) và L(s) tương ứng được giả thiết phù hợp với mọi s thuộc T. Cho $p(x_i, y_j)$ và $p_n(x_i, y_j)$ là mật độ và phân bố xác suất mẫu của I(s) và L(s), nghĩa là, tỷ lệ câu s thuộc T và $T_n$ mà I(s) và L(s) của chúng bằng $x_i$ và $y_j$ , i, j = 1, 2, ... Cho $F_c(x_i, y_j)$ được xác định trong giới hạn của $F_c(s)$, nghĩa là, $F_c(x_i, y_j)$ =1 nếu $x_i > cy_j$ v.v... Sau đó, đối với $T_n$ ,

$$I_n(F^n_{c_\alpha}) = n\sum x_i F^n_{c_\alpha}(x_i, y_j) p_n(x_i, y_j), \qquad (5)$$

trong đó lấy tổng đối với tất cả giá trị có thể của i và j. Do đó, đối với T,

$$I_n(F^n_{c_\alpha}) - \alpha \cdot I(T) = N\sum x_i F^n_{c_\alpha}(x_i, y_j)\left[p(x_i, y_j) - p_n(x_i, y_j)\right] + (N/n)\left[I_n(F^n_{c_\alpha}) - \alpha \cdot I(T_n)\right] + N\alpha\left[I(T_n)/n - I(T)/N\right] \qquad (6)$$

Thành phần thứ hai ở vế phải của (6) bằng 0. Bằng luật số lớn đối với các biến ngẫu nhiên độc lập và phụ thuộc ([8]) vì n tăng vô hạn, $I(T_n)/ n$ tiến tới $I(T)/ N$ theo xác suất và đối với $x_i$ và $y_j$ cố định, $p_n(x_i, y_j)$ tiến tới $p(x_i, y_j)$ theo xác suất. Vì chỉ có một số $x_i$ và $y_j$ riêng biệt hữu hạn, thành phần thứ nhất ở vế phải của (6) tiến tới 0 về xác suất. Như vậy, $I_n(F^n_{c_\alpha})$ tiến tới $\alpha \cdot I(T)$ về xác suất. Bằng cách tương tự chúng ta thiết lập mệnh đề về $L(F^n_{c_\beta})$.

### Nhận xét 3

Vì $F^n_{c_\alpha}$ và $F^n_{c_\beta}$ có dạng (3), đối với mỗi một n chúng có các hàm trích rút tối ưu khi áp dụng vào T. Định lý 3 phát biểu nếu n đủ lớn, chúng ta có thể kỳ vọng hầu như chắc chắn $I_n(F^n_{c_\alpha})$ và $L(F^n_{c_\beta})$ gần tới $\alpha \cdot I(T)$ và $\beta \cdot L(T)$. Một người nào đó có thể hỏi tại sao không định nghĩa $\lambda = \alpha N/ n$, $\mu = \beta N/ n$ và dùng các đoạn trích rút nhận được bằng cách áp dụng $F^n_{c_\alpha}$ và $F^n_{c_\beta}$ cho $T_n$ , vì $I_n(F^n_{c_\lambda}) = \lambda \cdot I(T_n)$, $L(F^n_{c_\mu}) = \mu \cdot L(T_n)$ và $\lambda \cdot I(T_n)$ và $\mu \cdot L(T_n)$ tiến tới $\alpha \cdot I(T)$ và $\beta \cdot L(T)$ về xác suất. Câu trả lời là trừ khi $F^n_{c_\alpha}$ và $F^n_{c_\beta}$ được áp dụng vào T, các đoạn trích rút được sản xuất không phải là tối ưu đối với T.

Một sự lựa chọn cho bài toán ước lượng là giả thiết rằng p(x,y) có thể được xấp xỉ bởi một hàm liên tục hai chiều $f(x,y,\theta)$ trong đó $\theta$ là một tham số (vô hướng hoặc vector). Các yêu cầu $I(F_{c_\alpha}) = \alpha \cdot I(T)$ và $L(F_{c_\beta}) = \beta \cdot L(T)$ trở thành

$$\iint x\, F_{c_\alpha}(x,y)\, f(x,y,\theta)\, dxdy = \alpha \iint x\, f(x,y,\theta)\, dxdy$$
$$\iint x\, F_{c_\beta}(x,y)\, f(x,y,\theta)\, dxdy = \beta \iint x\, f(x,y,\theta)\, dxdy \qquad (7)$$

Đối với $\alpha$ và $\beta$ cho trước, $c_\alpha$ và $c_\beta$ là hàm của $\theta$, tức là $c_\alpha(\theta)$ và $c_\beta(\theta)$. Bây giờ, $\theta$ có thể được ước lượng bằng cách lấy một mẫu câu ngẫu nhiên. Cho $\bar{\theta}$ là một ước lượng của $\theta$, sau đó $c_\alpha(\bar{\theta})$ và $c_\beta(\bar{\theta})$ là ước lượng của $c_\alpha(\theta)$ và $c_\beta(\theta)$ tương ứng. Bài toán ước lượng $p_\alpha$ và $p_\beta$ không tăng lên ở trường hợp này, vì xác suất của $x = cy$ bằng 0. Ở đây, dường như hợp lý giả sử p(x,y) có thể được xấp xỉ bằng một phân bố chuẩn hai chiều. Để đơn giản hoá tính toán, chúng tôi giả sử tương quan giữa x và y bằng 0, hoặc

$$f(x,y,\theta) = (1/\sqrt{2\pi}\sigma_1)e^{-(x-\mu_1)^2/\sigma_1^2} \cdot (1/\sqrt{2\pi}\sigma_2)e^{-(x-\mu_2)^2/\sigma_2^2} \qquad (8)$$

Ở đây, $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ là một vector 4 chiều.

### Bổ đề

Đối với $\alpha$ và $\beta$ cho trước, $0 \le \alpha, \beta \le 1$, cho $c_\alpha$ và $c_\beta$ thoả mãn (7) trong đó f(x, y, F) được cho bởi (8). Sau đó,

$$\left[\sigma_1^2/\sqrt{\sigma_1^2 + c_\alpha^2\sigma_2^2}\right] g\left[(c_\alpha\mu_2 - \mu_1)/\sqrt{\sigma_1^2 + c_\alpha^2\sigma_2^2}\right] + \mu_1 G\left[(c_\alpha\mu_2 - \mu_1)/\sqrt{\sigma_1^2 + c_\alpha^2\sigma_2^2}\right] = \alpha\mu_1 \qquad (9.1)$$

$$\left[c_\beta\sigma_2^2/\sqrt{\sigma_1^2 + c_\beta^2\sigma_2^2}\right] g\left[(\mu_1 - c_\beta\mu_2)/\sqrt{\sigma_1^2 + c_\beta^2\sigma_2^2}\right] + \mu_2 G\left[(\mu_1 - c_\beta\mu_2)/\sqrt{\sigma_1^2 + c_\beta^2\sigma_2^2}\right] = (1-\beta)\mu_2 \qquad (9.2)$$

trong đó:

$$g(x) = (1/\sqrt{2\pi})e^{-x^2/2} \quad \text{và} \quad G(x) = \int_x^\infty g(t)\, dt \qquad (9.3)$$

Hơn nữa, nếu $0 < \alpha, \beta < 1$, $\mu_1 >> \sigma_1$ và $\mu_2 >> \sigma_2$, nghĩa là, $\mu_1$ và $\mu_2$ lớn hơn nhiều $\sigma_1$ và $\sigma_2$ thì $c_\alpha, c_\beta > 0$

Từ bổ đề, chúng ta nhận thấy nói chung không thể tìm được $c_\alpha$ và $c_\beta$ trong phạm vi của $\theta$ rõ ràng, dù đối với $\theta$ cho trước, các giá trị tương ứng của $c_\alpha$ và $c_\beta$ có thể nhận được bằng tích phân số. Tuy nhiên, các xấp xỉ đối với $c_\alpha$ và $c_\beta$ có thể nhận được dưới các điều kiện hợp lý chung. Chúng ta có

### Định lý 4

Nếu $0 \le \alpha, \beta \le 1$ và $c_\alpha$ và $c_\beta$ thoả mãn (9.1) và (9.2) trong bổ đề thì

$c_\alpha \le \mu_1/\mu_2$ hoặc $c_\alpha \ge (\mu_1 + d_\alpha\sigma_1)/\mu_2$

và $c_\beta \ge \mu_1/\mu_2$ hoặc $c_\alpha \le (\mu_1 - d_{1-\beta}\sigma_1)/\mu_2$,

trong đó $G(d_\alpha) = \alpha$, $G(d_{1-\beta}) = 1 - \beta$ và G(x) được cho bởi (9.3).

Nếu $\mu_1 >> \sigma_1 >> \sigma_2$ , nghĩa là, $\mu_1$ lớn hơn nhiều $\sigma_1$ và $\sigma_1$ lớn hơn nhiều $\sigma_2$ thì cận dưới đối với $c_\alpha$ và cận trên đối với $c_\beta$ có thể được sử dụng như xấp xỉ đối với $c_\alpha$ và $c_\beta$ tương ứng.

*Chứng minh*:

Nếu $c_\alpha\mu_2 - \mu_1 \le 0$ thì $c_\alpha \le \mu_1/\mu_2$. Mặt khác, từ (9.1), chúng ta nhận thấy nếu $c_\alpha\mu_2 - \mu_1 \le 0, \alpha \ge G\left[(c_\alpha\mu_2 - \mu_1)/\sqrt{\sigma_1^2 + c_\alpha^2\sigma_2^2}\right] \ge G[(c_\alpha\mu_2 - \mu_1)/\sigma_1]$

Do đó, $(c_\alpha\mu_2 - \mu_1)/\sigma_1 \ge d_\alpha$ và $c_\alpha \ge (\mu_1 + d_\alpha\sigma_1)/\mu_2$ . Hơn nữa, nếu $\mu_1 >> \sigma_1 >> \sigma_2$ , chúng ta có thể xấp xỉ $\sigma_1^2 + c_\alpha^2\sigma_2^2$ bằng $\sigma_1^2$ và xoá thành phần thứ nhất ở vế trái của (9.1). Tiếp theo, $(\mu_1 + d_\alpha\sigma_1)/\mu_2$ là một xấp xỉ đối với $c_\alpha$ . Bằng cách tương tự chúng ta chỉ ra mệnh đề liên quan đến $c_\beta$ .

**Nhận xét 4**

Điều kiện $\mu_1 >> \sigma_1 >> \sigma_2$ , nghĩa là sự thay đổi về độ dài câu nhỏ hơn nhiều so với sự thay đổi về nội dung thông tin, mà nó lại nhỏ hơn nhiều so với nội dung thông tin trung bình của các câu. Các điều kiện dường như hợp lý đối với các ứng dụng thực tế. Các loại xấp xỉ khác đối với $c_\alpha$ và $c_\beta$ cũng có thể nhận được.

Bây giờ, chúng ta đi đến bài toán ước lượng. Để ước lượng $\mu$ trung bình và độ lệch chuẩn $\sigma$ của một phân bố chuẩn, các phương pháp khác nhau có sẵn ([4], [7]). Giả sử một mẫu n câu ngẫu nhiên được rút ra bằng phép thay thế và các giá trị I(s) và L(s) là $x_i$ và $y_i$ , i = 1, 2, ..., n. Cho $\bar{x} = \sum_{i=1}^{n} x_i/n$, $\bar{y} = \sum_{i=1}^{n} y_i/n$ và $S = a_n\sum_{i=1}^{n}(x_i - \bar{x})^2/n$ ,

trong đó $a_n = \sqrt{n/2}\Gamma((n-1)/2)/\Gamma(n/2)$ và $\Gamma(x)$ là hàm Gamma.

Chúng ta có:

**Định lý 5**

Nếu $\mu_1 >> \sigma_1 >> \sigma_2$ thì $\bar{c}_\alpha = (\bar{x} + d_\alpha S_1)/y$ và $\bar{c}_\beta = (\bar{x} - d_{1-\beta}S_1)/y$ là các ước lượng vững của $c_\alpha$ và $c_\beta$ theo nghĩa là $\bar{c}_\alpha$ và $\bar{c}_\beta$ tiến tới $c_\alpha$ và $c_\beta$ về xác suất khi $n \to \infty$.

*Chứng minh*:

$\bar{x}$, $\bar{y}$ và $S_1$ là các ước lượng vững của $\mu_1$, $\mu_2$ và $\sigma_1$ tương ứng ([4]). Theo định lý Slutsky, chúng ta nhận thấy $\bar{c}_\alpha$ và $\bar{c}_\beta$ tiến tới $c_\alpha$ và $c_\beta$ về xác suất.

## 4. ĐÁNH GIÁ VỀ THÔNG TIN VÀ ĐỘ DÀI

Ở các mục trước, chúng tôi định nghĩa và đề xuất một số phương pháp để nhận được các tóm tắt tối ưu. Tuy nhiên, trước khi áp dụng các phương pháp này, chúng ta

phải biết cách đánh giá thông tin và độ dài của một câu. Tiếp theo, chúng tôi duyệt lại một số phương pháp nổi tiếng nhằm đánh giá các đại lượng và đề xuất một số phương pháp mới cùng với một phương pháp ước lượng cho đánh giá thông tin.

Độ dài L(s) của câu s dường như đánh giá tương đối dễ. Chẳng hạn, chúng tôi có thể định nghĩa L(s) là số từ hoặc chữ chứa trong s. Hình thành các cách đánh giá L(s) khác là khá khó khăn, dù cho xác suất không nên bị loại bỏ không có khảo sát sâu hơn. Mặt khác, thông tin I(s) chứa trong s rõ ràng không đến mức dễ đánh giá. Nói ngắn gọn, đề xuất được mô tả sau đây:

1. I(s) là một hàm thông tin I(w) chứa trong từ w của s.
2. I(w) có thể được định nghĩa là tích F(w) và G(w), trong đó F(w) là tần suất xuất hiện tương đối của w trong văn bản cho trước T và G(w) là trọng số của w.

Hàm F(w) được định nghĩa là tỷ số của tần suất xuất hiện của w trong T với tần suất của w trong tất cả tài liệu, hoặc hạn chế hơn, tất cả tài liệu của loại nhất định thuộc về T. Như vậy, các từ thông thường như the, and, v.v... không có tần suất tương đối cao, vì chúng xuất hiện khắp nơi với khoảng tần suất như nhau. Chúng cũng không chứa nhiều thông tin. Mặt khác, nếu từ tóm tắt xuất hiện thường xuyên trong một tài liệu nhất định, chỉ thị tài liệu hầu như chắc chắn liên quan gần với tóm tắt. Do đó, dường như hợp lý giả thiết I(w) tỉ lệ với F(w). Khái niệm tần suất tương đối là do Edmundson và Wylls đưa ra và là sự cải tiến về khái niệm tần suất của Luhn. Trọng số của một từ chắc chắn là một đánh giá về ý nghĩa thực chất của nó. Chẳng hạn, nó được đề xuất bởi Edmundson và Wylls, nếu một từ mang tiêu đề hoặc chỉ thị tóm tắt (như là tóm tắt, kết luận, v.v... ), nên được cho một trọng số tương đối cao, dù cho nó có thể xuất hiện chỉ ít lần trong văn bản. Nó được coi là một loại trọng số chủ quan, nên được đưa vào. Chẳng hạn, nếu người nào đó quan tâm đến tập hợp tất cả định lý đã chứng minh trong một bài báo toán học, anh ta nên gán trọng số cao cho từ định lý, như vậy, anh ta có thể tin chắn tất cả định lý sẽ được trích rút. Mặt khác, nếu anh ta chỉ muốn một tóm tắt ngắn, sự mô tả về một định lý có thể là quá dài để được trích rút. Ở trường hợp này, không có một trọng số cao nào cần được gán cho từ định lý.

Bây giờ, chúng tôi đi đến bài toán đánh giá I(s): I(s) nên là một hàm của I(w). Nhưng loại hàm gì ? Luhn đề xuất I(s) nên là một hàm của phân bố của các từ có nghĩa, tức là, với I(w) bên trong câu. Như vậy, các câu chứa các từ biệt lập có nghĩa không được coi là có nghĩa. Một câu s có nghĩa và I(s) tương ứng nên lớn chỉ nếu nó chứa cụm từ có nghĩa. Người ta khó phát biểu loại quan hệ hàm gì thực sự tồn tại giữa I(s) và I(w) trong đó $w \in s$. Từ quan điểm lý thuyết, chúng ta đưa ra các mẫu sau đây về đánh giá I(s):

$$I(s) = \operatorname{Max}_{w \in s} I(w), \quad I(s) = \operatorname{Min}_{w \in s} I(w), \quad \text{và} \quad I(s) = \sum_{w \in s} I(w)$$

Nếu công thức thứ nhất được dùng, một câu có ý nghĩa nếu một trong số từ của nó có ý nghĩa. Nếu công thức thứ hai được dùng, một câu có ý nghĩa chỉ nếu tất cả từ của nó có ý nghĩa. Nếu công thức cuối cùng được dùng và L(s) được định nghĩa là số từ chứa trong s, thì I(s)/ L(s) là thông tin trung bình chứa trong một từ của s. Tới một mức độ nhất định, đại lượng này tương thích với đánh giá ý nghĩa một câu của Luhn.

Để kết luận, chúng tôi cho một phương pháp ước lượng tần suất xuất hiện p của từ w trong một văn bản T cho trước. Sau đó, dựa vào ước lượng p, chúng tôi có thể ước lượng I(w) vì tần suất xuất hiện w trong tất cả tài liệu và trọng số G(w) của w có thể giả sử biết rõ. Hơn nữa, chúng tôi có thể ước lượng I(s) đối với một đánh giá I(s) cho trước. Các phương pháp ước lượng giá trị p nên quan tâm thực hành, vì tìm giá trị thực có thể mất thời gian. Cho tổng số câu và từ chứa trong T là N và M tương ứng. Giả sử một mẫu ngẫu nhiên có n câu được rút ra có hoặc không có thay thế. Đối với một từ cho trước w, cho $x_i$ bằng số xuất hiện w trong câu thứ i ở mẫu. Định nghĩa $x = \sum_{i=1}^{n} x_i$ . Dễ dàng nhận thấy Nx/ nM là một ước lượng không chệch p. Đối với E(Nx/ nM) = (N/ M) $E(x_i)$ = p ([4], [7]).

## 5. KẾT LUẬN

Ở mục 2, chúng tôi đưa vào khái niệm tóm tắt và trích rút tối ưu. Hai loại tóm tắt và trích rút tối ưu được định nghĩa, nghĩa là, độ dài cực tiểu và tóm tắt và trích rút thông tin cực đại (định nghĩa 1 và 3). Để nhận được trích rút tối ưu, sử dụng hàm trích rút ngẫu nhiên được đề xuất ở định nghĩa 2. Ở định lý 1, chúng tôi trình bày các hàm trích rút tối ưu phải có một dạng nhất định. Ở định lý 2, chúng tôi trình bày đối với $\alpha$ và $\beta$ cho trước, $0 \le \alpha, \beta \le 1$, tồn tại độ dài cực tiểu và các hàm trích rút thông tin cực đại sinh ra đoạn trích rút mà nội dung thông tin và độ dài liên quan của nó về trung bình bằng $\alpha$ và $\beta$ tương ứng. Ở mục 3, trước tiên chúng tôi thảo luận cách xác định chính xác các hàm trích rút tối ưu tương ứng với $\alpha$ và $\beta$ theo ngữ cảnh trên. Tiếp theo, chúng tôi thảo luận cách tiết kiệm thời gian và cố gắng tìm kiếm các hàm trích rút tối ưu chỉ tương ứng gần đúng với $\alpha$ và $\beta$. Hai loại phương pháp được đề xuất, phụ thuộc vào bản chất của phân bố thông tin chứa đựng các câu của văn bản cho trước và độ dài. Nếu phân bố là siêu bội hoặc đa thức, chúng tôi trình bày ở định lý 3 tồn tại các hàm trích rút mẫu hội tụ xác suất tới hàm trích rút tối ưu thực vì kích thước mẫu tăng lên. Nếu phân bố có thể được xấp xỉ bởi một phân bố chuẩn, chúng tôi trình bày ở định lý 4 và 5 có thể xác định và ước lượng bằng cách lấy mẫu ngẫu nhiên, các hằng số xác định các hàm trích rút tối ưu thực. Dưới các điều kiện chung hợp lý, các công thức xấp

xỉ đơn giản nhận được. Ở mục 4, chúng tôi thảo luận các phương pháp đánh giá thông tin chứa trong một câu và độ dài câu.

## TÀI LIỆU THAM KHẢO


[1] Arms W.Y., *Digital Libraries*, MIT Press, Cambridge, 2003.

[2] Chen H., Houston A.L., Digital Libraries: social issues and technological advances, *Advanced in Computers* 48, 1999, pp. 257-314.

[3] Chowdhary G.G., Digital Library Research: major issues and trends, *Journal of Documentation* 55(4), 1999, pp. 409-448.

[4] Cramér H., *Phương pháp toán học trong thống kê*, 2 tập, Nxb Khoa học và kỹ thuật, Hà Nội, 1970.

[5] Nguyễn Đức Dân, Đặng Thái Ninh, *Nhập môn thống kê ngôn ngữ học*, Nxb Giáo dục, Hà Nội, 1998.

[6] Nguyễn Đức Dân, Đặng Thái Ninh, *Thống kê ngôn ngữ học – một số ứng dụng*, Nxb Giáo dục, Hà Nội, 1999.

[7] Trần Tuấn Điệp, Lý Hoàng Tú, *Lý thuyết xác suất và thống kê toán học*, xuất bản lần 3, Nxb Giáo dục, Hà Nội, 1999.

[8] Feller W., *An Introduction to Probability Theory and Its Applications*, vol.1, 3rd Edition, John Wiley, New York, 1971.

[9] Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như, *Thống kê toán học*, Nxb Đại học và trung học chuyên nghiệp, Hà Nội, 1984.

[10] Fox E.A., *Advanced Digital Libraries*, Virginia Polytechnic Institue and State University, 2000.

[11] Journal of Network and Computer Applications, *Special Issue of JNCA on Digital Libraries* 20 (1-2), 1997.

[12] Lesk M., *Practical Digital Libraries*, Morgan Kaufmann, San Francisco, 1997.

[13] Mendelhall W., Sincich T., *Statistics for the Engineering and Computer Science*, 2nd Edition, Collier Macmillan, London, 1989.

[14] Ross S.M., *Probability Models for Computer Science*, Harcourt Academic Press, San Diego, 2002.

[15] Sun Microsystems, *Digital Library Technology Trends*, 2002.

## SUMMARY

## ABSTRACTING AND EXTRACTING TEXT DOCUMENTS IN DIGITAL LIBRARIES

This article presents some results of a theoretical study of abstracting and extracting text documents in digital libraries. Our approach is the use of statistical sampling and estimation of text document. We introduce the concept of optimal abstracting and extracting. Two types of optimal abstracts and extracts are defined: minimum lenth and maximum information abstracts and extracts (definitions 1 and 3). Next we suggest a randomized extracting function to obtain optimal extracs (definition 2). In theorem 1, we show that extracting function must have a certain form. In theorem 2, we show that for given $\alpha$ and $\beta$, $0 \leq \alpha, \beta \leq 1$, there exist minimum length and maximum information extracting function. In section 3, we dicuss how to determine exactly optimal extracting functions corresponding to $\alpha$ and $\beta$. Then we discuss how to save some time and effort by finding optimal extracting functions which correcpond approximately to $\alpha$ and $\beta$. We suggest two types of methods. If the distribution is hypergeometric or multinomial, we show that in theorem 3 there exist example extracting functions which converge in probability to actual optimal extracting functions. If the distribution may be approximated by a normal distribution, we show in theorem 4 and 5 that it is possible to determine and estimate the contants by random sampling. Finally, we discuss methods for measuring the information contained in and the length of a sentence.