Tạp chí Tin học và Điều khiển học, T. 18, S. 1 (2002), 35–43

ỨNG DỤNG KHOẢNG CÁCH HAUSDORFF TRONG PHÂN TÍCH TRANG TÀI LIỆU

LƯƠNG CHI MAI, ĐỖ NĂNG TOÀN

Abstract. This paper dealts with a method for using Hausdorff distance to analyse the page layout based on bottom-up approach through Q_θ relation. Firstly, objects were isolated by out-contours. Then, the objects have the size smaller than a given tolerance would grouped by nearest Hausdorff distance to create a region. The other, which has smaller size, would be analysed as a document image.

Tóm tắt. Bài báo này đề cập đến phân tích trang văn bản hỗn hợp thành các thành phần theo tiếp cận dưới lên nhờ việc sử dụng khoảng cách Hausdorff giữa các đối tượng ảnh thông qua quan hệ Q_θ. Ban đầu các đối tượng ảnh được tách bởi chu tuyến ngoài. Sau đó, các đối tượng có kích thước hình chữ nhật phủ nhỏ hơn một ngưỡng nào đó sẽ được nhóm với nhau theo lân cận gần nhất dựa vào việc sử dụng khoảng cách Hausdorff thông qua quan hệ Q_θ để tạo ra các khối, còn các đối tượng ảnh còn lại sẽ được tiếp tục phân tích như là đối với một trang văn bản kích thước nhỏ hơn.

1. GIỚI THIỆU

Một trong những nhiệm vụ cơ bản của nhận dạng các trang văn bản nói chung và các trang văn bản có lẫn các đối tượng khác như ảnh, sơ đồ, biểu đồ v.v. (hình 1) là phải tách được chúng.

Trong bài báo này chúng tôi đề cập đến cách phân tích văn bản theo tiếp cận dưới lên [4, 5] nhờ việc sử dụng khoảng cách Hausdorff giữa các đối tượng ảnh [1]. Ban đầu các đối tượng ảnh sẽ


Thu MONDIAL

●DANH ĐỨC (Fax từ Paris)

Cảng Marseille khói lửa

9 giờ tối 14.6 trên đường về nhà, khi trời tứ hấm xe điện lên đã nghe trên một đài FM tin hooligans Anh quậy phá ở cảng Marseille, nơi sẽ diễn ra trận Anh - Tunisie ngày 15.6.98. Theo tin ban đầu, khoảng 300 hooligans Anh tụ tập ở Vieux Port (tức khu bến cảng cổ) trước một pub (quán rượu) mang tên "OM Café" nguyên là địa điểm tụ tập của các cổ động viên đội OM, giăng biểu ngữ và hô khẩu hiệu "đả đảo bọn Tunisie" ! Thông tin ban đầu cho biết cảnh sát đã can thiệp bằng lựu đạn cay song không giải tỏa được đám hooligans ngày càng đông hơn, và còn hứa hẹn "đêm nay sẽ nổi lửa" ở Marseille đấy. Sáng ra, tình hình Marseille chỉ còn là "ngổn ngang" những tổn thất, trong đó có 37 người bị thương nặng.

chiến, cứ ngồi yên trong các xe bít bùng, không chường mặt ra kẻo bị xem là khiêu khích hoặc giải tán sớm thì bị gọi là "quân phiệt" và càng làm cho bạo động sớm bùng nổ mà nguyên cớ sẽ đổ thừa cho cảnh sát. Khi vụ đụng độ đầu tiên xảy ra, cảnh sát phong tỏa ngay khu cảng cổ, bắt giữ ngay số người ẩu đả. Cảnh sát trưởng Marseille, ông Daniel Herbst vẫn còn tin rằng có thể nắm được tình hình, thông báo với cảnh

Holigans Anh đốt cờ Tunisie trên đường phố Marseille

trung. "Tiểu đoàn" ủng hộ viên Anh cuồng kích nhất tấn công đại đội dã chiến trang bị khiên, dùi cui, súng phóng lựu (đạn khói)... bằng các lon bia (bia mua trong

Hình 1. Trang văn bản có lẫn ảnh

được tách bởi chu tuyến ngoài [2, 3, 4], các đối tượng có kích thước hình chữ nhật phủ nhỏ hơn một ngưỡng nào đó sẽ được nhóm lại với nhau theo lân cận gần nhất dựa vào việc sử dụng khoảng cách Hausdorff để tạo ra các khối, còn các đối tượng ảnh còn lại sẽ được tiếp tục phân tích như là đối với một trang văn bản.

Nội dung của bài báo được thể hiện qua các phần tiếp theo như sau:

Phần 2 đưa ra các khái niệm và chứng minh một số tính chất liên quan đến chu tuyến. Phần 3 trình bày những tính chất cơ bản của không gian Hausdorff với khoảng cách Hausdorff và khoảng cách Hausdorff giữa các đối tượng ảnh. Phần 4 trình bày kĩ thuật phân tích trang văn bản theo tiếp cận dưới lên nhờ sử dụng khoảng cách Hausdorff giữa các đối tượng ảnh. Cuối cùng là những kết luận về ứng dụng khoảng cách Hausdorff trong phân trang tài liệu.

2. CHU TUYẾN CỦA MỘT ĐỐI TƯỢNG ẢNH

2.1. Một số khái niệm cơ bản

Ảnh và điểm ảnh

Ảnh là một mảng số thực 2 chiều (a_{ij}), kích thước $(m \times n)$, trong đó mỗi phần tử a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$ biểu thị mức xám của ảnh tại vị trí i, j tương ứng.

Một ảnh được gọi là nhị phân nếu các giá trị a_{ij} của nó chỉ nhận giá trị 0 hoặc 1.

Một ảnh bất kỳ có thể đưa về dạng nhị phân bằng phép cắt ngưỡng. Ta kí hiệu J là tập các điểm 1 (điểm vùng) và $\overline{J}$ là tập các điểm 0 (điểm nền).

Các điểm 4- và 8-láng giềng

Giả sử (i, j) là một điểm ảnh, các điểm 4-láng giềng là các điểm trực tiếp bên trên, dưới, trái, phải của điểm (i, j):

$N_4 = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$,

và những điểm 8-láng giềng gồm:

$N_8 = N_4 \cup$
$\{i-1, j-1), (i-1, j+1), (i+1, j-1), (i+1, j+1)\}$.

Ví dụ trong hình 2 các điểm 0, 2, 4, 6 là các 4-láng giềng của điểm P, còn các điểm 0, 1, 2, 3, 4, 5, 6, 7 là các 8-láng giềng của P.

3	2	1
4	*P*	0
5	6	7

Hình 2. Matrận 8-láng giềng của P

Đối tượng ảnh

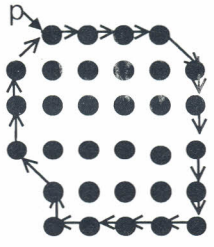
Hai điểm $P_1, P_2 \in E$, $E \subseteq J$ hoặc $\overline{J}$ được gọi là 8-liên thông (hay 4-liên thông) trong E nếu tồn tại tập các điểm được gọi là "đường đi" $(i_0, j_0) \ldots (i_n, j_n)$ sao cho $(i_0, j_0) = P_1$, $(i_n, j_n) = P_2$, $(i_r, j_r) \in E$ và (i_r, j_r) là 8-láng giềng (hay 4-láng giềng) của (i_{r-1}, j_{r-1}) với $r = 1, 2, \ldots, n$.

Quan hệ "k-liên thông trong E", $k = 4, 8$, là một quan hệ phản xạ, đối xứng và bắc cầu bởi vậy là một quan hệ tương đương. Về sau ta sẽ gọi mỗi lớp tương đương của nó là một đối tượng ảnh.

2.2. Chu tuyến của một đối tượng ảnh

Định nghĩa 2.1. [Chu tuyến]

Chu tuyến của một đối tượng ảnh là dãy các điểm của đối tượng ảnh $P_1, \ldots P_i, \ldots, P_n$ sao cho P_i và P_{i+1} là các 8-láng giềng của nhau $(i = 1, \ldots, n-1)$ và P_1 là 8-láng giềng của P_n, $\forall i\ \exists Q$ không thuộc đối tượng ảnh và Q là 4-láng giềng của P_i. Kí hiệu $\langle P_1 P_2 \ldots P_n \rangle$.



Hình 3. Ví dụ về chu tuyến của một đối tượng ảnh

Tổng các khoảng cách giữa hai điểm kế

tiếp nhau của chu tuyến là độ dài của chu tuyến và hướng P_iP_{i+1} là hướng chẵn (lẻ) nếu P_{i+1} là điểm 8-láng giềng chẵn (lẻ) của P_i. Kí hiệu độ dài của chu tuyến C là Len$C$. Hình 3 biểu diễn chu tuyến của ảnh, P là điểm khởi đầu chu tuyến.

Định nghĩa 2.2. [Chu tuyến đối ngẫu]

Hai chu tuyến $C = \langle P_1P_2 \dots P_i \dots P_n \rangle$ và $C^{\perp} = \langle Q_1Q_2 \dots Q_j \dots Q_m \rangle$ được gọi là đối ngẫu của nhau nếu và chỉ nếu $\forall i\ (i = 1, \dots, n-1)\ \exists j\ (j = 1, \dots, m)$, $\exists k\ (k = 1, \dots, m)$ sao cho:

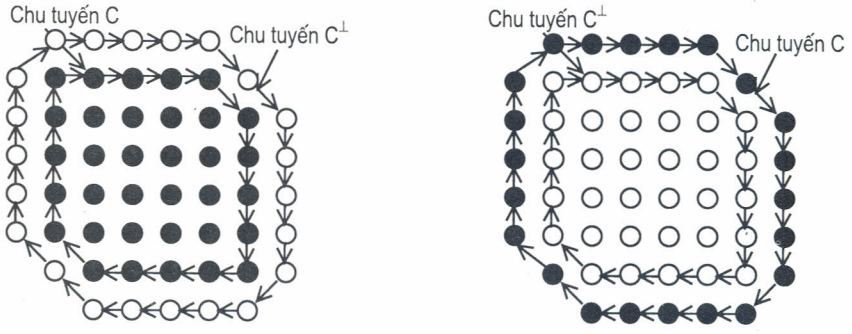
1. P_i và Q_j là 4-láng giềng của nhau.
2. P_{i+1} và Q_k là 4-láng giềng của nhau.
3. Q_j và Q_k là 8-láng giềng của nhau.
4. Các điểm P_i là vùng thì Q_j, Q_k là nền và ngược lại.

Định nghĩa 2.3. [Chu tuyến ngoài]

Chu tuyến C được gọi là chu tuyến ngoài (hình 4a) nếu và chỉ nếu độ dài của chu tuyến C nhỏ hơn độ dài chu tuyến đối ngẫu $C^{\perp}$ của nó.

Định nghĩa 2.4. [Chu tuyến trong]

Chu tuyến C được gọi là chu tuyến trong (hình 4b) nếu và chỉ nếu độ dài chu tuyến C lớn hơn độ dài chu tuyến đối ngẫu $C^{\perp}$ của nó.



a) Chu tuyến ngoài b) Chu tuyến trong

Hình 4. Chu tuyến trong, chu tuyến ngoài

Định lý 2.1. *Giả sử $E \subseteq J$ là một đối tượng ảnh và C là chu tuyến ngoài của E. Khi đó C là duy nhất.*

Chứng minh. Ta kí hiệu in(Q, C) để chỉ điểm Q nằm trong chu tuyến C, và out(Q, C) để chỉ điểm Q nằm ngoài chu tuyến C. $\forall x \in E$, ta chứng minh in(x, C_E). Thật vậy, giả sử out(x, C_E), vì $x \in E$ nên tồn tại một dãy $x_i \in E$ $(i = 1, \dots, m)$ sao cho x_i, x_{i+1} là các 8-láng giềng của nhau, x_m là 8-giáng giềng của x và in(x_1, C_E). Vì x nằm ngoài C_E nên $\exists k$ sao cho out(x_i, C_E) $(\forall i > k)$, khi đó hoặc $x_i \in C_E$, hoặc in(x_i, C_E). Vì C_E là chu tuyến ngoài của E gọi C_{EN} là chu tuyến láng giềng tương ứng của C_E, C_E nằm trong C_{EN} nên trong cả hai trường hợp ta có in(x_i, C_{EN}). Mặt khác, out(x_{i+1}, C_E) nên out(x_{i+1}, C_{EN}).

Do đó theo điều kiện Jordan về điểm trong thì x_ix_{i+1} sẽ cắt C_E tại một số lẻ lần (≥ 1). Như vậy giữa x_i và x_{i+1} sẽ có một số điểm (≥ 1) xen giữa, nhưng x_i và x_{i+1} là 2 điểm láng giềng của nhau điều đó dẫn đến mâu thuẫn. Vậy in(x, C_E).

Giả sử tồn tại chu tuyến C'_E cũng là chu tuyến ngoài của E ta đi chứng minh $C_E \equiv C'_E$. Thật vậy, giả sử tồn tại $x \in C'_E$ mà $x \notin C_E$, vì $C'_E \subseteq E$ mà C_E là chu tuyến ngoài nên theo chứng minh trên ta có in(x, C_E) từ đó suy ra in(x, C_E) $(\forall x \in C'_E)$, tương tự ta cũng có in$(x, C'_E)(\forall x \in C_E)$, điều đó dẫn đến mâu thuẫn.

Vậy C_E là duy nhất.

3. KHOẢNG CÁCH HAUSDORFF GIỮA CÁC ĐỐI TƯỢNG ẢNH

3.1. Khoảng cách Hausdorff

Định nghĩa 3.1. [Khoảng cách từ một điểm đến một tập]

(X, d) là không gian metric đầy đủ, kí hiệu $H(X)$ là tập các tập con compact của X. Gọi $x \in X$ và $B \in H(X)$, khi đó khoảng cách từ điểm x tới tập B được định nghĩa là: $d(x, B) = \min\{d(x, y) : y \in B\}$.

Định nghĩa 3.2. [Khoảng cách giữa 2 tập hợp]

(X, d) là không gian metric đầy đủ, $A, B \in H(X)$, khi đó khoảng cách từ tập A tới tập B được định nghĩa bởi: $d(A, B) = \max\{d(x, B) : x \in A\}$.

Định nghĩa 3.3. [Khoảng cách Hausdorff]

(X, d) là không gian metric đầy đủ. Khoảng cách Hausdorff giữa các điểm $A, B \in H(X)$ được xác định như sau: $h(A, B) = \max\{d(A, B), d(B, A)\}$.

Định lý 3.1. *h là metric trên $H(X)$.*

Chứng minh.

(i) $h(A, B) = \max\{d(A, B), d(B, A)\} = \max\{d(B, A), d(A, B)\} = h(B, A)$.

(ii) $A \neq B \in H(X) \Rightarrow$ có thể tìm được $a \in A$, $a \notin B$: $d(a, B) > 0 \Rightarrow h(A, B) \geq d(a, B) > 0$.

(iii) $h(A, A) = \max\{d(A, A), d(A, A)\} = d(A, A) = \max\{d(a, A) : a \in A\} = 0$.

(iv) $\forall a \in A$ ta có $d(a, B) = \min\{d(a, b) : b \in B\} \leq \min\{d(a, c) + d(c, b) : b \in B\}\ \forall c \in C$

$\Rightarrow d(a, B) \leq d(a, C) + \min\{d(c, b) : b \in B\}\ \forall x \in C$

$\Rightarrow d(a, B) \leq d(a, C) + \max\{\min\{d(c, b) : b \in B\} : c \in C\}$

$\Rightarrow d(a, B) \leq d(a, C) + d(C, B)$.

Do đó $d(A, B) = \max\{d(a, B) : a \in A\} \leq d(a, C) + d(C, B) \leq d(A, C) + d(C, B)$.

Tương tự có $d(B, A) \leq d(B, C) + d(C, A)$

$$\begin{aligned} h(A, B) &= \max\{d(A, B), d(B, A)\} \\ &\leq \max\{d(A, C) + d(C, B), d(B, C) + d(C, A)\} \\ &\leq \max\{d(A, C), d(C, A)\} + \max\{d(C, B), d(B.C)\} \\ &\leq h(A, C) + h(C, B). \end{aligned}$$

□

3.2. Khoảng cách Hausdorff giữa càc đối tượng ảnh

Mỗi đối tượng ảnh trong tập ảnh là tập k-liên thông và là tập hữu hạn điểm nên nó chính là tập compact trong không gian các điểm ảnh. Do vậy ta có thể áp dụng khoảng cách Hausdorff để tính khoảng cách giữa các đối tượng ảnh.

Việc tính khoảng cách Hausdorff giữa các đối tượng ảnh là phức tạp và tốn kém do các đối tượng này có thể chứa nhiều điểm khác nhau. Định lý sau giúp ta giảm bớt việc tính toán.

Bổ đề 3.1. *Giả sử $E \subseteq J$ là một đối tượng ảnh và C là chu tuyến ngoài của E, M_0 là một điểm nằm ngoài C ($M_0 \notin E$). Khi đó khoảng cách từ M_0 đến 1 điểm ảnh của E đạt cực trị tại C.*

Chứng minh. Gọi điểm đạt cực trị là P, cần phải chứng minh $P \in C$. Thật vậy, nếu $P \notin C$ thì do C là chu tuyến ngoài nên P là điểm trong của C. Ta xét các trường hợp:

+ P là điểm cực tiểu

Vì P là điểm trong của C nên PM_0 sẽ cắt C tại một số lẻ điểm. Giả sử N là một trong những giao điểm khi đó rõ ràng ta có:

$d(M_0, P) = d(M_0, N) + d(N, P)$.

Vì $P \neq N$ nên $d(M_0, N) < d(M_0, P)$.

Do đó P không phải là điểm cực tiểu. (*)

+ P là điểm cực đại

Vì P là điểm trong nên phần nửa đường thẳng M_0P kéo dài về phía P sẽ cắt C tại một số lẻ điểm. Giả sử N là một trong những giao điểm khi đó rõ ràng ta có:

$d(M_0, N) = d(M_0, P) + d(P, N)$.

Vì $P \neq N$ nên $d(M_0, N) > d(M_0, P)$.

Do đó P không phải là điểm cực đại. (**)

Từ (*) và (**) suy ra P không phải là điểm cực trị, điều này trái với giả thiết. Do đó bổ đề được chứng minh. □

Định lý 3.2. *Giả sử $U, V \subseteq J$ là các đối tượng ảnh và C_U là chu tuyến ngoài của U, C_V là chu tuyến ngoài của V. Khi đó $h(U, V) = h(C_U, C_V)$.*

Chứng minh. $\forall x \in U$, theo định nghĩa ta có $d(x, V) = \min\{d(x, y) : y \in V\}$. Vì U, V là 2 đối tượng ảnh khác nhau nên x nằm ngoài C_1 theo Bổ đề 3.2 ta có:

$$d(x, V) = \min\{d(x, y) : y \in Y\} = \min\{d(x, y) : y \in C_V\} = d(x, C_V).$$

Do đó

$$d(U, V) = \max\{d(x, V) : x \in U\} = \max\{d(x, C_V : x \in U\} = d(U, C_V. \quad (1)$$

Mặt khác, $\forall y \in C_V$, theo định nghĩa ta có $d(U, y) = \min\{d(x, y) : x \in U\}$, y nằm ngoài C nên theo Bổ đề 3.2 ta có:

$$d(U, y) = \min\{d(x, y) : x \in U\} = \min\{d(x, y) : x \in C\} = d(C, y).$$

Do đó

$$d(U, C_V) = \max\{d(U, y) : y \in C_V\} = \max\{d(C, y) : y \in C_V\} = d(C, C_V). \quad (2)$$

Từ (1) và (2) suy ra $d(U, V) = d(C, C_V)$.

Vậy:

$$h(U, V) = d(U, V) \vee d(V, U) = d(C, C_V) \vee d(C_V, C) = h(C, C_V).$$ □

4. ỨNG DỤNG KHOẢNG CÁCH HAUSDORFF TRONG PHÂN TÍCH TRANG TÀI LIỆU

4.1. Quan hệ Q_θ

Định nghĩa 4.1. [Liên kết Q_θ]

Cho trước ngưỡng θ, hai đối tượng ảnh $U, V \subseteq J$ hoặc $\overline{J}$ được gọi là liên kết theo θ và kí hiệu $Q_\theta(U, V)$ nếu tồn tại dãy các đối tượng ảnh $X_1, X_2, \ldots, X_n$ sao cho:

(i) $U \equiv X_1$,

(ii) $V \equiv X_n$,

(iii) $h(X_i, X_{i+1}) < \theta \ \forall i, \ 1 \leq i \leq n-1$.

Mệnh đề 4.1. *Quan hệ liên kết Q_θ là một quan hệ tương đương*

Chứng minh.

(i) Phản xạ: $U \subseteq J$ hoặc $\overline{J}$ ta có $h(U, U) = 0 < \theta$.

(ii) Đối xứng: Giả sử có $Q_\theta(U, V)$, cần phải chứng minh $Q_\theta(V, U)$.

Thật vậy, theo giả thiết tồn tại dãy đối tượng ảnh $X_1, X_2, \ldots, X_n$ sao cho:

$$U \equiv X_1, \ V \equiv X_n, \ h(X_i, X_{i+1}) < 0 \ \forall i, \ 1 \leq i \leq n-1.$$

Khi đó, với dãy đối tượng ảnh $Y_1, Y_2, \ldots, Y_n$ mà: $Y_i \equiv X_{n-i+1} \ \forall i, \ 1 \leq i \leq n$ ta có:

$$V \equiv Y_1, \ U \equiv Y_n, \ h(Y_i, Y_{i+1}) < \theta \ \forall i, \ 1 \leq i \leq n-1.$$

Suy ra $Q_\theta(V, U)$ (đpcm).

(iii) Bắc cầu: Giả sử ta có $Q_\theta(U,V)$ và $Q_\theta(V,T)$, ta cần chứng minh $Q_\theta(U,T)$.

Thật vậy, vì $Q_\theta(U,V)$ nên tồn tại dãy đối tượng ảnh $X_1, X_2, \ldots, X_n$ sao cho

$$U \equiv X_1\,,\ V \equiv X_n\,,\ h(X_i, X_{i+1}) < \theta\ \forall i,\ 1 \le i \le n-1.$$

$Q_\theta(V,T)$ nên tồn tại dãy đối tượng ảnh $Y_1, Y_2, \ldots, Y_m$ sao cho:

$$U \equiv Y_1\,,\ T \equiv Y_m\,,\ h(Y_i, Y_{i+1}) < \theta\ \forall i,\ 1 \le i \le m-1.$$

Khi đó, dãy các đối tượng ảnh $Z_1, Z_2, \ldots, Z_n, Z_{n+1}, \ldots, Z_{n+m}$ ở đây: $Z_i \equiv X_i\ \forall i,\ 1 \le i \le n$ và $Z_{n+i} \equiv Y_i\ \forall i,\ 1 \le i \le m$ có các tính chất:

$$U \equiv Z_1\,,\ T \equiv Z_{n+m}\,,\ h(Z_i, Z_{i+1}) < \theta\ \forall i,\ 1 \le i \le n+m-1.$$

Suy ra $Q_\theta(U,T)$ (đpcm).

4.2. Phân tích trang tài liệu

Thông thường, việc tiến hành phân tích định dạng trang thường được tiến hành sau khi ảnh được xác định góc nghiêng và quay về góc 0.

Phân tích định dạng trang có thể thực hiện từ dưới lên hay từ trên xuống. Với phân tích từ trên xuống, một trang được chia từ những phần lớn thành các phần con nhỏ hơn. Ví dụ nó có thể được chia thành một số cột văn bản. Sau đó mỗi cột có thể được chia thành các đoạn, mỗi đoạn lại được chia thành các dòng văn bản... Tiếp cận theo các hướng này có các phương pháp: sử dụng các phép chiếu nghiêng, gán nhãn chức năng, phân tích khoảng trống trắng v.v.. Ưu điểm lớn nhất của các phương pháp phân tích từ trên xuống là nó dùng cấu trúc toàn bộ trang để giúp cho phân tích định dạng được nhanh chóng. Đây là cách tiếp cận hiệu quả cho hầu hết các dạng trang. Tuy nhiên, với các trang không có các biên tuyến tính và có sơ đồ lẫn cả bên trong và quanh văn bản, các phương pháp này có thể không thích hợp. Ví dụ, nhiều tạp chí tạo văn bản quanh quanh một sơ đồ ở giữa, vì thế văn bản đi theo những đường cong của đối tượng trong sơ đồ chứ không theo đường thẳng.

Phân tích định dạng từ dưới lên bắt đầu với những phần nhỏ và nhóm chúng vào những phần lớn hơn kế tiếp tới khi mọi khối trên trang được xác định. Tuy nhiên không có một phương pháp tổng quát nào điển hình cho một kĩ thuật phân tích dưới lên. Trong phần nhỏ này, ta mô tả một cách tiếp cận được coi là dưới lên nhưng sử dụng những phương pháp trực tiếp rất khác nhằm đạt cùng mục đích. Phần này cũng đưa ra ý tưởng về hệ thống phần mềm hoàn chỉnh để phân tích định dạng trang.

Dưới đây chúng tôi đặc tả bằng ngôn ngữ RAISE (Rigorous Approach Industrial Software Engineering) thuật toán pageANALYSIS phân tích trang tài liệu theo tiếp cận dưới lên nhờ sử dụng quan hệ Q_θ đã nêu ở mục trên. Để tiến hành đặc tả bằng RAISE chúng tôi dùng các kiểu cơ bản như **Nat** - số tự nhiên, **Unit** - kiểu rỗng, **Bool** - kiểu logic, **Point** - kiểu điểm trừu tượng, **Point-list** - kiểu danh dách và **Orient** - kiểu các số tự nhiên nhỏ hơn 8.

Các biến sử dụng trong thuật toán

StartPT, NextPT	Điểm cuất phát và điểm tiếp
StartDir, NextDir	Hướng khởi tạo và hướng tiếp theo chiều xét duyệt chu tuyến
nWhite, nBlack	Độ dài của chu tuyến và chu tuyến láng giềng
ArayDest	Mảng lưu giữ chu tuyến trong (tập hợp các điểm *NextPT*)
nCount	Số các điểm của chu tuyến trong thu được
fLag	Cờ xác định xem đối tượng hình có phải là đối tượng tách được hay không.

Các hàm sử dụng trong thuật toán

Init	Thiết lập các tham số ban đầu
FindNext	Tìm điểm kế tiếp và hướng trong chu tuyến
LenWhite	Tính độ dài của chu tuyến láng giềng đến điểm kế tiếp
LenBlack	Tính độ dài của chu tuyến đến điểm kế tiếp
PutDest	Lưu giữ chu tuyến vào một mảng khác dùng các thủ tục *IsolateOBJECT* và *Simplification*
IsolateOBJECT	Hàm cô lập các đối tượng trong ảnh bằng cách dò theo các chu tuyến trong và ngoài của đối tượng.
Classification	Phân đối tượng vừa tách vào nhóm đã có nhờ quan hệ Q_θ. Trường hợp không phân được, tạo ra lớp mới và bổ sung đối tượng vừa tìm được vào lớp đó
pageANALYSIS	Các bước của thuật toán pageANALYSIS được tiến hành như sau: Khởi tạo các tham số bởi thủ tục Init, rồi cô lập các đối tượng hình học bằng thủ tục isolateOBJECT, sau đó phân đối tượng vừa tách vào nhóm đã có nhờ quan hệ Q_θ. Trường hợp không phân được, tạo ta lớp mới và bổ sung đối tượng vừa tìm được vào lớp đó.

Thuật toán được xác định trong sơ đồ sau bằng ngôn ngữ RAISE

```
scheme PAGEANALYSIS =
Class
type Oreint={|n:Nat:-(0 ≤ n) ∧ (n < 8)|},
     Point, Object,
     Area = Object-set
     Point=Nat><Nat, Object,
     Area=Object-set,
     Image,
     PageStruct
variable
     StarPT    : Point := (0,0), NextPT      : Point := (0,0),
     StartDir  : Orient := 0, NextDir: Orient := 0,
     nWhite    : Real := 0.0, nBlack: Real := 0.0,
     ArayDest  : Area-list := ⟨...⟩,
     nCoint    : Nat := 0,
     Im        : Image,
     PgStruct  : PageStruct
channel I: Image, PgStruct_c: PageStruct
value
     Init: Unit → in I
          read Im, StrarPT, NexPT, StartDir,
               NextDir, nWhete, nBlack, ArayDest, nCount
          write StarPT, NextPT, StarDir, NextDir, nWhite,
               nBlack, ArayDest, nCount
          Unit,
     FindNext: Unit → write NextPT, NextDir Unit,
     LenWhite, Lenblack: Point → Real,
     PutDest: Unit → write NextPT Unit,
     Classification: Unit → write ArayDest, nCount Unit,
```

```
    isolateOBJECT: Unit → in I
                read StartPT, NextPT, StarDir,
                    NextDir, nWhite, nBlack, ArayDest, nCount, Im
                    write StartPT, NextPT, StartDir, NexDir,
            nWhite, nBlack, ArayDest, nCount, Im Unit
    isolateOBJECT() is
        Im := I?;
        do
            FindNext(); nWhite := nWhite + LenWhite (NextPT);
            nBlack := nBlack + LenBaack(NextPT); PutDest()
        until (NextPT=StartPR∧NextDir=StartDir)
        end,
    pageANALYSIS: Unit →
            in I
            read StartPT, NextPT, StarDir, NextDir, nWhite,
                nBlacjk, ArayDest, nCount, Im
            out PgStruct_c
            write StartPT, NextPT, StarDir, NextDir,
                nWhite, nBlack, ArayDest, nCount, Im
            Unit
axiom
    pageANALYSIS() is Im := I?;          /*Đọc ảnh vào*/
                      Init();            /*Khởi tạo tham số*/
                      isolateOBJECT();   /*Cô lập các đối tượng*/
                      Classification ();  /*Phân loại tài liệu*/
                      PgStruct_c!PgStruct /*In cấu trúc trang*/
end
```

Mệnh đề 4.2. *Thuật toán pageANALYSIS gồm các bước cô lập các đối tượng, phân lớp các đối tượng dựa vào khoảng cách Hausforff theo quan hệ Q_θ dừng và cho kết quả đúng.*

Chứng minh. Vì số điểm của chu tuyến và đối tượng xác định bởi chu tuyến là hữu hạn nên bước xét duyệt chu tuyến là dừng do đó bước cô lập các đối tượng sẽ dừng. Số các đối tượng thu được là hữu hạn nên việc phân lớp các đối tượng dựa vào khoảng cách Hausdorff theo quan hệ Q_θ cũng dừng và do vậy thuật toán pageANALYSIS là dừng.

Bước phân lớp các đối tượng dựa vào khoảng cách Hausdorff theo quan hệ Q_θ sẽ cho ta kết quả là các lớp đối tượng tượng mà trong đó các đối tượng thuộc cùng một lớp sẽ có khoảng cách giữa chúng nhỏ hơn ngưỡng θ cho trước. Q_θ là một quan hệ tương đương, từ Mục 4.1 ta thấy tính đúng đắn của thuật toán.

Tổng hợp các bước ở trên ta có thuật toán pageANALYSIS là dừng và cho kết quả đúng. □

5. KẾT LUẬN

Trong bài báo này chúng tôi đề cập đến cách phân tích văn bản theo tiếp cận dưới lên nhờ việc sử dụng khoảng cách Hausdorff giữa các đối tượng ảnh. Ban đầu các đối tượng ảnh sẽ được tách bởi chu tuyến ngoài. Các đối tượng có kích thước hình chữ nhật phủ nhỏ hơn một ngưỡng nào đó sẽ được nhóm với nhau theo lân cận gần nhất dựa vào việc sử dụng khoảng cách Hausdorff để tạo ra các khối, còn các đối tượng ảnh còn lại sẽ được tiếp tục phân tích như là đối với một trang văn bản.

Định lý 3.2 đã chỉ ra rằng khoảng cách hausdorff giữa hai đối tượng ảnh chính là khoảng cách hai chu tuyến ngoài của các đối tượng. Hơn nữa, Định lý 2.1 còn chỉ ra rằng tồn tại duy nhất một

chu tuyến ngoài cho mỗi đối tượng ảnh. Việc sử dụng chu tuyến ngoài sẽ giảm đáng kể thời gian cho phân tích trang tài liệu theo tiếp cận dưới lên.

Lời cảm ơn. Chúng tôi xin chân thành cảm ơn GS TSKH Bạch Hưng Khang đã tận tình giúp đỡ trong công việc nghiên cứu. Chúng tôi cũng bày tỏ lòng biết ơn đến TS Ngô Quốc Tạo đã đóng góp những ý kiến quí báu giúp cho chúng tôi hoàn thành bài báo này một cách nhanh chóng.

TÀI LIỆU THAM KHẢO

[1] Bạch Hưng Khang, Đỗ Năng Toàn, Ứng dụng khoảng cách Hausforff trong đánh giá chuyển đổi các biểu diễn Raster và Vector, *Tạp chí Tin học và Điều khiển học* **16** (4) (2000) 52–58.

[2] Đỗ Năng Toàn, Một thuật toán phát hiện vùng và ứng dụng của nó trong trình vectơ hóa tự động, *Tạp chí Tin học và Điều khiển học* **16** (1) (2000) 45–51.

[3] Đỗ Năng Toàn, Ngô Quốc Tạo, Tách các đối tượng hình học trong phiếu điều tra dạng dấu, chuyên san *Các công trình nghiên cứu và triển khai Công nghệ thông tin và viễn thông, Tạp chí Bưu chính viễn thông*, số 2 (1999) 69–76.

[4] L. O'Gorman, *The Document Spectrum for Page Layout Analysis*, IEEE Trans, Pattern Analysis and Machine Intelligence, Nov. 1993, 1162–1173.

[5] Lawrence O'Gorman and Rangachar Kasturi, *Document Image Analysis*, IEEE Computer Society Press, 10662 Los Vaqueros Circle, 1998, 165–173.

[6] Nguyễn Ngọc Kỷ, "Biểu diễn và đồng nhất tự động ảnh đường nét", Luận án Phó tiến sĩ Toán - Lý, Hà Nội, 1992.

[7] S. Mao and T. Kanungo, Empirical performace evaluation of page segmentation algorithms, *Processings of the SPIE Conference on Document Recognition and Retrieval*, (2000) 303–314.

[8] Song Mao, Tapas Kanungo, Empirical pertformance evaluation methodology and its application to Page segmentation algotithms, *IEEE Trans, Pattern Analysis and Machine Intelligence* **23** (3) (2001) 242–256.

Nhận bài ngày 1 - 3 - 2001
Nhận lại sau khi sửa ngày 20 - 2 - 2002

Viện Công nghệ thông tin