Tạp chí Tin học và Điều khiển học, T.21, S.1 (2005), 47–55

PHÂN TÍCH TRANG VĂN BẢN DỰA VÀO MẪU

ĐỖ NĂNG TOÀN, LƯƠNG CHI MAI

Viện Công nghệ thông tin

Abstract. Document analysis is an important step in optical characters recognition system. Analysing a document helps to understand structure of the document and then to output blocks of the document after having recognized correctly.

In fact, we often meet some basic types of documents which usually have fixed structure. This paper deals with a technique to calculate the error between two document structures. Then, we'll present an algorithm for page layout analysis based on the given templates.

Tóm tắt. Phân tích trang văn bản là một bước không thể thiếu trong các hệ thống nhận dạng văn bản. Việc phân tích trang văn bản là để xác định phạm vi, cấu trúc của các khối từ, từ đó mới có thể nhận dạng và trả lại cấu trúc của trang văn bản sau nhận dạng một cách chính xác.

Trong thực tế, thường gặp nhiều trường hợp, ở đó người ta phải làm việc với một số dạng văn bản nhất định, các văn bản này thường có cấu trúc xác định trước. Bài báo này đề cập đến một kỹ thuật đánh giá độ sai lệch của văn bản dựa vào cấu trúc. Qua đó, đề xuất một thuật toán phân tích trang văn bản dựa vào các văn bản mẫu đã có.

1. GIỚI THIỆU

Phân tích trang văn bản là một bước không thể thiếu trong các hệ thống nhận dạng văn bản [3, 5, 6, 7, 8]. Việc phân tích trang văn bản là để xác định phạm vi, cấu trúc của các khối từ đó mới có thể nhận dạng và trả lại cấu trúc của trang văn bản sau nhận dạng một cách chính xác.

Trong thực tế, đối với lĩnh vực ứng dụng chúng ta thường gặp một số dạng cấu trúc văn bản cố định. Việc phân tích cấu trúc một văn bản ở các trường hợp này sẽ được tiến hành thuận lợi hơn nhiều nếu ta sử dụng các thông tin về cấu trúc của các văn bản mẫu.

Trong các nghiên cứu trước, các tác giả đã đề xuất thuật toán phân tích trang văn bản hỗn hợp (pageANALYSIS [1]) thành các thành phần theo tiếp cận dưới lên nhờ việc sử dụng khoảng cách Hausdorff ([3]) giữa các đối tượng ảnh thông qua quan hệ Q_θ ([1]). Ban đầu các đối tượng ảnh sẽ được cô lập bởi chu tuyến ngoài ([2]) (đường biên kín nhỏ nhất chứa mọi điểm ảnh của đối tượng ảnh). 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 theo quan hệ Q_θ dựa vào việc sử dụng khoảng cách Hausdorff để tạo ra các khối, các đối tượng ảnh còn lại sẽ được tiếp tục phân tích như là đối với một trang văn bản. Trong đó, ngưỡng θ được lựa chọn theo kinh nghiệm của người sử dụng.

Trong bài báo này chúng tôi đề cập đến việc nâng cao chất lượng thuật toán phân tích trang văn bản pageANALYSIS dựa vào các văn bản mẫu đã có thông qua một kỹ thuật được đề xuất nhằm đánh giá độ sai lệch của khối văn bản dựa vào cấu trúc. Phần còn lại của bài

báo được thể hiện như sau: Mục 2 trình bày lựa chọn ngưỡng θ tự động dựa vào biểu đồ tần xuất của ảnh. Mục 3 trình bày kỹ thuật đánh giá độ sai lệch của khối văn bản so với văn bản mẫu và đề xuất thuật toán phân tích dựa vào mẫu. Cuối cùng là những kết luận về phương pháp xác định ngưỡng tự động trong phân tích trang văn bản.

Hình 1. Ảnh văn bản với các vùng

2. LỰA CHỌN NGƯỠNG TỰ ĐỘNG TRONG PHÂN TÍCH TRANG VĂN BẢN

2.1. Quan hệ Q_θ ([1])

Định nghĩa 2.1. (Liên kết Q_θ)

Cho trước ngưỡng θ, hai đối tượng ảnh $U, V \subseteq \Im$ hoặc $\overline{\Im}$ được gọi là liên kết theo θ và kí hiệu $Q_\theta(U, V)$ nếu tồn tại dãy các đối tượng ảnh $X_1, X_2, ..., X_n$ sao cho:
(i) $U \equiv X_1$,
(ii) $U \equiv X_n$,
(iii) $h(X_i, X_{i+1}) < \theta \ \forall i,\ 1 \leqslant i \leqslant n-1$.

Mệnh đề 2.1. ([1]) *Quan hệ liên kết Q_θ là một quan hệ tương đương.*

2.2. Phân tích trang văn bản nhờ khoảng cách Hausdorff bởi quan hệ Q_θ

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.

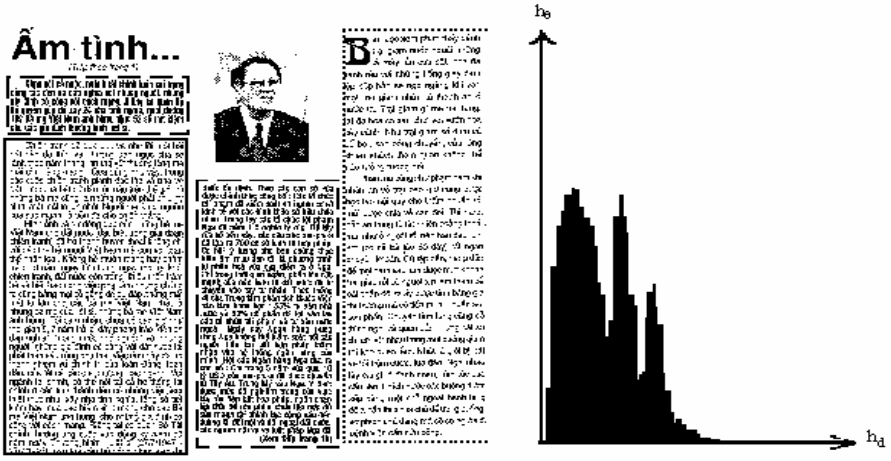
Phân tích định dạng trang có thể thực hiện từ dưới lên hay từ trên xuống [6, 7]. Với phân tích từ trên xuống, một trang được chia từ những phần lớn thành các phần con nhỏ hơn, ví dụ nó có thể được chia thành một số cột văn bản. Sau đó mỗi cột có thể được chia thành các đoạn, mỗi đoạn lại được chia thành các dòng văn bản... Tiếp cận theo hướng này có các phương pháp: sử dụng các phép chiếu nghiêng, gán nhãn chức năng, phân tích khoảng trống trắng... 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 mọi kỹ thuật phân tích dưới lên. Trong [1] các tác giả đã đề xuất thuật toán phân tích trang văn bản hỗn hợp thành các thành phần pageANALYSIS 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 sẽ được cô lập bởi chu tuyến ngoài (đường biên kín nhỏ nhất chứa mọi điểm ảnh của đối tượng ảnh). Các đối tượng có kích thước hình chữ nhật phủ nhỏ hơn một ngưỡng θ 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, các đối tượng ảnh còn lại sẽ được tiếp tục phân tích như là đối với một trang văn bản. Trong đó, ngưỡng θ thường được xác định theo kinh nghiệm người sử dụng. Dưới đây chúng tôi đề xuất việc lựa chọn ngưỡng một cách tự động dựa vào biểu đồ tần xuất (histogram).

2.3. Lựa chọn ngưỡng dựa vào histogram

Do thuật toán phân tích trang văn bản pageANALYSIS [1] dựa khoảng cách Hausdorff bởi quan hệ Q_θ là quá trình duyệt tìm các lớp tương đương theo khoảng cách θ. Các đối tượng ảnh trong cùng một khối văn bản có những đặc trưng tương đối giống nhau về kích thước và khoảng cách giữa chúng với các đối tượng lân cận. Hơn nữa, một trang văn bản lại thường có một vài dạng đối tượng chỉ đạo. Do đó, ta có thể lựa chọn ngưỡng θ ban đầu thông qua việc đánh giá biểu đồ tần xuất khoảng cách Hausdorff giữa các đối tượng ảnh (Hình 2).



Hình 2. Ảnh văn bản và biểu đồ tần xuất khoảng cách Hausdorff giữa các đối tượng ảnh

Từ biểu đồ tần xuất khoảng cách Hausdorff giữa các đối tượng ảnh của ảnh văn bản cần phân tích, ngưỡng θ được lựa chọn trong các giá trị h_θ tương ứng là các đỉnh trong biểu đồ tần xuất đó chính là các giá trị ứng với nhiều phần tử cùng loại nhất. Với ngưỡng θ đã chọn ta tiến hành phân vùng theo tiếp cận dưới lên nhờ việc sử dụng khoảng cách Hausdorff giữa các đối tượng ảnh thông qua quan hệ Q_θ. Kết quả thu được và tập hợp các hình chữ nhật

rời nhau thể hiện các vùng trong ảnh.

Việc lựa chọn ngưỡng θ phù hợp nhất sẽ được tiến hành thông qua việc đánh giá sự sai lệch của văn bản so với mẫu. Với mỗi ngưỡng θ, ta sẽ tìm được mẫu tương ứng có độ lệch nhỏ nhất. Ngưỡng θ và văn bản mẫu tương ứng với độ sai lệch nhỏ nhất trong số các độ lệch sẽ được lựa chọn. Nếu sai số nhỏ nhất chấp nhận được (nhỏ hơn một ngưỡng cho trước nào đó) thì số vùng của văn bản sẽ được xác định tương ứng với số vùng của văn bản mẫu được lựa chọn. Khi đó văn bản sẽ được phân tích trang dựa theo các thuộc tính của văn bản mẫu. Trong trường hợp ngược lại có thể xem văn bản không thuộc tập văn bản mẫu và do vậy có thể tiến hành bổ sung văn bản đang xét vào tập mẫu.

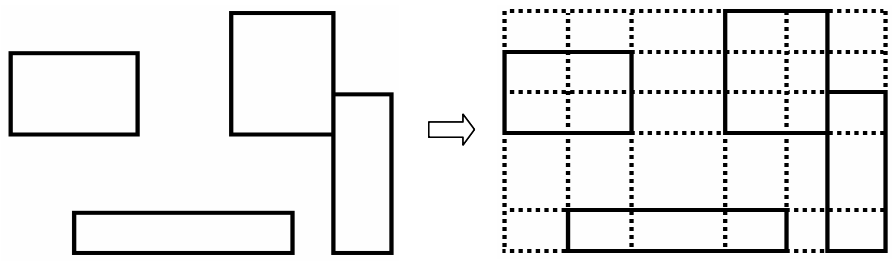
3. PHÂN TÍCH TRANG VĂN BẢN DỰA VÀO MẪU

3.1. Đánh giá độ lệch cấu trúc văn bản theo mẫu

Quá trình phân tích trang văn bản dựa theo các mẫu đã có sẽ được thực hiện thông qua việc đánh giá sự sai lệch của của văn bản so với văn bản mẫu. Văn bản mẫu có độ lệch nhỏ nhất so với văn bản cần hiệu chỉnh sẽ được lựa chọn. Nếu độ lệch nhỏ nhất tìm được nằm trong phạm vi cho phép thì có thể xem văn bản cần hiệu chỉnh thuộc trong số mẫu đã có, trường hợp ngược lại văn bản được xem như là mẫu mới và được bổ sung vào tập văn bản mẫu.

Việc đánh giá độ sai lệch của văn bản so với văn bản mẫu sẽ được tiến hành thông qua việc xây dựng lưới tựa các vùng chữ nhật cơ bản của mẫu và đánh giá độ lệch của vùng so với lưới. Độ lệch của văn bản so với mẫu sẽ được đánh giá dựa trên sự tương đồng của cả văn bản và mẫu so với lưới tương ứng.

Việc xây dựng lưới tựa các vùng hình chữ nhật tìm được trong văn bản thông qua việc chọn ngưỡng θ dựa vào biểu đồ tần xuất hay các vùng văn bản chữ nhật trong mẫu. Lưới là tập các tọa độ ngang dọc, Hình 3 thể hiện ví dụ minh họa việc xây dựng lưới từ tập các hình chữ nhật.



Hình 3. Xây dựng lưới tựa các hình chữ nhật

Độ lệch của một vùng c_k so với ô lưới $M_{\text{Grid}}(i,j)$ được tính bởi công thức:

$$\text{Intersec}(c_k, M_{\text{Grid}}(i,j)) = \begin{cases} 1 \text{ nếu } c_k \cap M_{\text{Grid}}(i,j) \neq \emptyset \\ 0 \text{ nếu ngược lại} \end{cases}$$

và độ lệch của một vùng c_k so với lưới M_{Grid} được xác định bởi tổng độ lệch của vùng so với các ô của của lưới M_{Grid}:

$$\text{Segments}(c_k, M_{\text{Grid}}(i,j)) = \sum_{i=1}^{n_h} \sum_{j=1}^{n_v} \text{Intersec}(c_k, M_{\text{Grid}}(i,j)).$$

Gọi tập hợp các vùng nằm trong lưới mà có độ lệnh khác 0 là $C_{M_{\text{Grid}}}$, ta có:

$$C_{M_{\text{Grid}}} = \{c_k \big| \text{Segments}(c_k, M_{\text{Grid}}) > 0\}.$$

Khi đó, độ lệnh của văn bản so với ô lưới (i,j) được xác định bởi công thức:

$$N_{M_{\text{Grid}}(i,j)} = \sum_{c_k \in C_{M_{\text{Grid}}}} \Big(\text{Intersec}(c_k, M_{\text{Grid}}(i,j)) \times \frac{1}{\text{Segments}(c_k, M_{\text{Grid}})} \Big)$$

và độ của văn bản so với lưới được xác định là tổng độ lệch của văn bản so với từng ô của lưới và bằng:

$$N_{M_{\text{Grid}}} = \sum_{i=1}^{n_h} \sum_{j=1}^{n_v} N_{M_{\text{Grid}}}(i,j)$$

Độ lệch của văn bản so với mẫu được đánh giá bằng tỷ số giữa tổng độ lệch của các vùng trong văn bản và mẫu đối với từng ô của lưới kết hợp giữa hai lưới được xây dựng từ các vùng của văn bản và mẫu trên tổng số vùng của văn bản và mẫu:

$$S = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{n_v} \left| N_{MG_{\text{Grid}}}(i,j) - N'_{MG_{\text{Grid}}}(i,j) \right|}{n_c + n'_c}$$

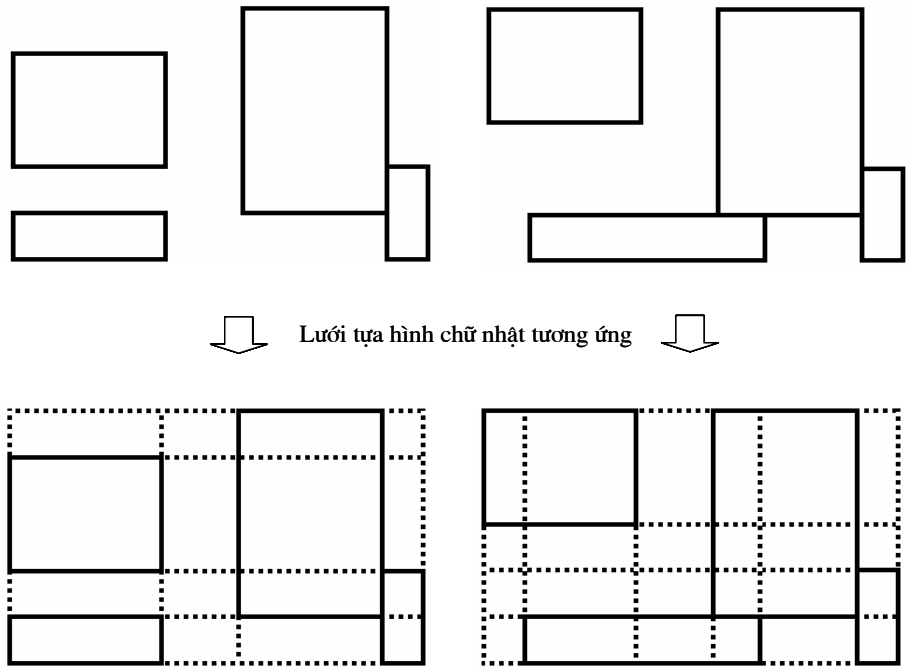
trong đó, MG_{Grid} là lưới kết hợp từ hai lưới được xây dựng từcác hình chữ nhật vùng của văn bản và mẫu;

$n_c,\ n'_c$ là số vùng của văn bản và số vùng của mẫu;

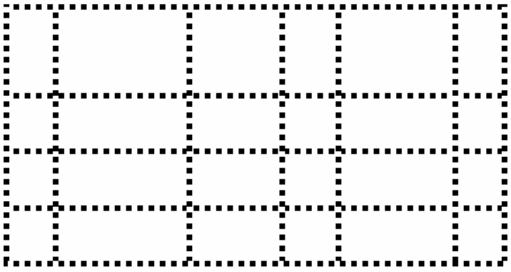
$N_{MG_{\text{Grid}}}(i,j),\ N_{MG_{\text{Grid}}}$ là độ lệnh của văn bản và mẫu so với ô lưới (i,j).

Ví dụ minh họa đánh giá độ lệch văn bản so với mẫu

Cấu trúc văn bản, cấu trúc mẫu và lưới tựa hình chữ nhật xây dựng tương ứng



Lưới xây dựng kết hợp từ các lưới tựa vùng chữ nhật văn bản và mẫu



Khi đó, giá trị độ lệch của văn bản và mẫu so với các ô lưới được tính theo công thức $N_{M_{G_{\text{Grid}}}}(i,j)$ là:

0	0	0	1/8	1/8	0
1/4	1/4	0	1/8	1/8	0
1/4	1/4	0	1/8	1/8	0
0	0	0	1/8	1/8	1/2
1/2	1/2	0	0	0	1/2

1/4	1/4	0	1/8	1/8	0
1/4	1/4	0	1/8	1/8	0
0	0	0	1/8	1/8	0
0	0	0	1/8	1/8	1/2
0	1/3	1/3	1/3	0	1/2

Hình 6

và do đó, độ lệch của văn bản so với mẫu được tính theo công thức là:

$$S = \frac{4 \times 1/4 + 1/2 + 1/6 + 2/3}{4 + 4} = \frac{5}{16} = 0,3125.$$

3.2. Thuật toán phân tích trang văn bản dựa vào mẫu

Dưới đây, chúng tôi trình bày thuật toán phân tích trang văn bản pageANALYSIS* dựa vào mẫu nhờ kỹ thuật phân tích trang văn bản pageANALYSIS [1] theo tiếp cận dưới lên nhờ sử dụng quan hệ Q_θ và việc đánh giá độ lệch cấu trúc văn bản theo mẫu ở mục trên.

Vào: Ảnh văn bản I cần phân tích,

Tập cấu trúc văn bản mẫu tempStructs,

Ngưỡng Tolerance.

Ra: Cấu trúc trang văn bản cầu phân tích pageStruct.

Phương pháp: Thuật toán gồm các bước cơ bản sau:

(1) Tính biểu đồ tần xuất theo khoảng cách Hausdorff:

+ Tách các đối tượng dựa vào chu tuyến ngoài,

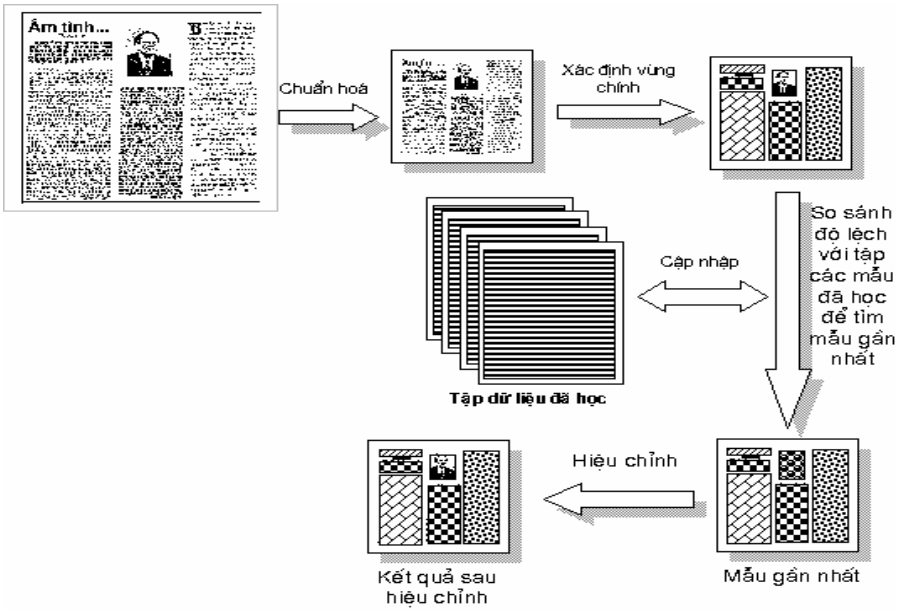
+ Tính khoảng cách Hausdorff giữa các đối tượng,

+ Xây dựng biểu đồ tần xuất theo khoảng cách đã tính.

(2) Với biểu đồ tần xuất đã xây dựng lựa chọn ngưỡng θ.

(3) Phân tích trang văn bản theo thuật toán pageANALYSIS theo quan hệ Q_θ với ngưỡng θ lựa chọn dựa vào biểu đồ tần xuất ở bước 2.

(4) Đánh giá lệch của cấu trúc trang văn bản vừa được phân tích ở bước 3 với các cấu trúc trang văn bản mẫu và tìm ra cấu trúc trang tương ứng có độ lệch nhỏ nhất.



Hình 4. Các bước tiến hành phân vùng và đối sánh mẫu

(5) Lặp lại bước 2 đến bước 4 chừng nào còn lựa chọn được θ theo các đỉnh biểu đồ tần xuất theo khoảng cách Hausdorff giữa các đối tượng ảnh.

(6) Chọn ra mẫu có độ lệch nhỏ nhất trong số các độ lệch nhỏ nhất tìm được trong bước 4 ứng với các θ lựa chọn.

(7) Kiểm tra nếu độ lệch nhỏ nhất tìm được trong bước 6 nhỏ hơn ngưỡng Tolerance thì có thể kết luận văn bản cần phân tích có dạng là mẫu có độ lệch nhỏ nhất tương ứng và cấu trúc trang phân tích thu được cấu trúc tương ứng thu được ở bước 2 sau bước phân tích theo thuật toán pageANALYSIS theo quan hệ Q_θ. Trong trường hợp ngược lại có thể kết luận văn bản không nằm trong các mẫu văn bản cho trước, để nâng cao chất lượng cho bước sau có thể bổ sung thêm văn bản với các cấu trúc tìm được tương ứng vào tập mẫu cấu trúc văn bản.

Các bước tiến hành phân phân vùng và đối sánh mẫu trong pageANALYSIS* như trên Hình 4.

Mệnh đề 3.1. *Thuật toán phân tích trang văn bản pageANALYSIS* dựa vào mẫu là 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 tính biểu đồ tần xuất theo khoảng cách Hausdorff là dừng. Do đó, các bước lựa chọn ngưỡng θ dựa vào các đỉnh của biểu đồ tần xuất là hữu hạn.

Vì thuật toán phân tích trang văn bản pageANALYSIS* dựa vào mẫu nhờ kỹ thuật phân tích trang văn bản pageANALYSIS theo tiếp cận dưới lên nhờ sử dụng quan hệ Q_θ và việc đánh giá độ lệch cấu trúc văn bản theo mẫu. Tính đúng đắn của thuật toán pageANALYSIS đã được chỉ ra trong [1] và từ mục 3.1 ta thấy tính đúng đắn của việc đánh giá độ lệch văn bản theo mẫu dẫn đến tính đúng đắn của thuật toán pageANALYSIS*.

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

4. ĐÁNH GIÁ VÀ KẾT LUẬN

Phân tích trang văn bản là một bước không thể thiếu trong các hệ thống nhận dạng văn bản. Việc phân tích trang văn bản nhằm xác định phạm vi, cấu trúc của các khối từ đó mới có thể nhận dạng và trả lại cấu trúc của trang văn bản sau nhận dạng một cách chính xác. Trong thực tế, đối với mỗi trường hợp cụ thể chúng ta thường gặp một số dạng cấu trúc văn bản cố định. Việc phân tích cấu trúc một văn bản ở các trường hợp này sẽ được tiến hành thuận lợi hơn nhiều nếu ta sử dụng các thông tin về cấu trúc của các văn bản mẫu.

Trong bài báo này chúng tôi đã đề xuất một kỹ thuật đánh giá độ sai lệch của văn bản dựa vào cấu trúc. Qua đó, đưa ra một thuật toán phân tích trang văn bản dựa vào các văn bản mẫu đã có. Thuật toán phân tích trang văn bản đề xuất pageANALYSIS* dựa vào thuật toán phân tích trang pageANALYSIS thông qua quan hệ Q_θ.

Hơn nữa, việc xác định ngưỡng θ ban đầu trong thuật toán phân tích trang văn bản đề xuất pageANALYSIS* được ước lượng một cách tự động thông qua việc tính biểu đồ tần xuất khoảng cách Hausdorff giữa các đối tượng ảnh đã tỏ ra ưu điểm hơn hẳn so với thuật toán pageANALYSIS được đề xuất trong [1] khi ngưỡng θ được ước lượng bởi người sử dụng. Do trong văn bản, khoảng cách giữa các dòng và khoảng cách giữa các từ trong dòng là khác nhau, nên trong các nghiên cứu tiếp theo chúng tôi sẽ xem θ là ngưỡng hai thành phần θ_x và

θ_y để nâng cao độ chính xác.

Qua thực tế chúng tôi thấy, việc đề xuất kỹ thuật đánh giá độ sai lệch của văn bản dựa vào cấu trúc giúp xác định loại văn bản cần phân tích trong số các văn bản mẫu, trên cơ sở đó chính xác hóa kết quả phân tích trang văn bản dựa do khuynh hướng của văn bản đã được xác định.

TÀI LIỆU THAM KHẢO

[1] Lương Chi Mai, Đỗ Năng Toàn, Ứng dụng khoảng cách Hausdorff trong phân tích trang tài liệu, *Tạp chí Tin học và Điều khiển học* **18** (1) (2002) 35–43.

[2] Đỗ Năng Toàn, Biên ảnh và một số tính chất, *Tạp chí Khoa học Công nghệ* **40** (ĐB) (2002) 41–48.

[3] Song Mao, Tapas Kanungo, Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms, *IEEE Trans, Pattern Anolysis and Machine Intelligence* **23** (3) (2001) 242–256.

[4] S. Mao, T. Kanungo, Empirical performance evaluation of page segmentation algorithms, *Proceedings of the SPIE Conference on Document Recognition and Retrieval*, January 2000, 303–314.

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

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

[7] Thomas Breuel, "High Performance Document Layout Analysis", Symposium on Document Image Understanding Technology Greenbelt Marriott, Greenbelt Maryland, Available at http://citeseer.nj.nec.com/568589.html, 2003.

[8] Thomas M. Breuel, An Algorithm for Finding Maximal Whitespace Rectangles at Arbitrary Orientations for Document Layout Analysis, *Proceedings of the ICDAR 2003* **1** (2003) 66–70.

Nhận bài ngày 20 - 12 - 2002
Nhận lại sau sửa ngày 29 - 7 - 2003