TRƯỜNG ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN&TRUYỀN THÔNG

PHẠM VĂN THỦY

ĐÁNH GIÁ SỰ ẢNH HƯỞNG CỦA THAM SỐ ĐẾN KẾT QUẢ PHÂN TÁCH CỦA THUẬT TOÁN WHITESPACE

LUẬN VĂN THẠC SĨ

Thái Nguyên, tháng 06 năm 2017

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn cao học "**Đánh giá sự ảnh hưởng của tham số đến kết quả phân tách của thuật toán WhiteSpace"** *là công trình nghiên cứu của riêng tôi và hoàn thành dưới sự hướng dẫn khoa học của TS. Nguyễn Đức Dũng.*

Trong toàn bộ nội dung của luận văn, những phần được trình bày là của cá nhân tôi hoặc được tổ hợp từ nhiều nguồn tài liệu khác nhau. Tất cả các tài liệu, số liệu đều là trung thực có xuất xứ rõ ràng và được trích dẫn đúng theo quy định.

Tôi hoàn toàn chịu trách nhiệm với lời cam đoan của mình.

Học viên thực hiện luận văn

Phạm Văn Thủy

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn chân thành tới TS. Nguyễn Đức Dũng vì đã có những chỉ dẫn, động viên trong suốt quá trình thực hiện luận văn của tôi. Đồng thời tôi xin chân thành cảm ơn các thầy cô giáo trong Ban giám hiệu, phòng Đào tạo, các thầy cô giáo của trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên cùng các thầy cô giáo trong Viện Công nghệ Thông Tin - Viện Hàn lâm Khoa học Việt Nam đã quan tâm, tạo điều kiện thuận lợi, giảng dạy và hướng dẫn tôi trong suốt quá trình học tập và hoàn thiện luận văn.

Cuối cùng tôi xin cảm ơn mọi sự giúp đỡ từ người thân, đồng nghiệp những người đã luôn ủng hộ, hỗ trợ tôi trong suốt quá trình thực hiện luận văn của mình.

Mặc dù đã có nhiều cố gắng, tuy nhiên luận văn của tôi không thể tránh khỏi những thiếu sót, do đó tôi rất mong nhận được những ý kiến đánh giá, bổ sung để tôi có thể hoàn thiện luận văn của mình./.

Quảng Ninh, ngày tháng năm 2017

MỤC LỤC

DANH MỤC HÌNH ẢNH 7
PHẦN MỞ ĐẦU 10
1. Đặt vấn đề 10
2. Nội dung nghiên cứu chính 11
2.1. Mục tiêu chính của đề tài 11
2.2. Ý nghĩa khoa học của đề tài 12
2.3. Nhiệm vụ nghiên cứu 12
2.4. Phương pháp nghiên cứu 12
2.5. Phạm vi nghiên cứu 13
3. Bố cục của luận văn 13
CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH ẢNH TÀI LIỆU 14
1.1. Tổng quan về phân tích ảnh tài liệu 14
1.1.1. Giới thiệu về ảnh tài liệu 14
1.1.2. Hệ phân tích ảnh tài liệu 15
1.1.3 Quá trình thu nhận ảnh tài liệu 20
1.1.4. Vai trò của phân tích ảnh tài liệu. 21
1.2. Cấu trúc của ảnh tài liệu 23
1.2.1. Cấu trúc vật lý 23
1.2.2. Cấu trúc logic 24
1.3. Phân tích trang tài liệu 24
1.3.1. Tiền xử lý (preprocessing): 26
1.3.2. Phân tích cấu trúc vật lý 27
1.3.3. Phân tích cấu trúc logic: 29
1.4 Kết luận 30
CHƯƠNG 2: ĐÁNH GIÁ SỰ ẢNH HƯỞNG CỦA THAM SỐ ĐẾN KẾT QUẢ PHÂN TÁCH CỦA THUẬT TOÁN WHITESPACE 31

2.1. Các hướng tiếp cận và một số thuật toán phân tách trang tiêu biểu 31

2.1.1. Hướng tiếp cận Top-down 31

a) Tổng quan 31

c) Ưu điểm: 35

d) Nhược điểm: 35

2.1.2. Hướng tiếp cận Bottom-up 38

a) Tổng quan 38

c) Ưu điểm 42

d) Nhược điểm 42

2.1.3. Hướng tiếp cận theo phương pháp lai ghép (hybrid). 43

a) Tổng quan 43

b) Thuật toán tách và Nối thích nghi (Adaptive Split - and - Merge) 43

c) Ưu điểm 45

d) Nhược điểm 45

2.1.4. Đánh giá và lựa chọn thuật toán. 46

2.2. Thuật toán phân tích trang tài liệu Whitespace 47

2.2.1. Giới thiệu 47

2.2.2. Whitespace Cover 48

2.2.2.1. Định nghĩa bài toán 48

2.2.2.2. Thuật toán 49

2.3. Ảnh hưởng của tham số đến kết quả phân tách của thuật toán Whitespace 54

2.3.1. Tham số về tỉ lệ chồng lấp (giao nhau) của các hình chữ nhật trắng. 54

2.3.2. Tham số về khoảng trắng tối đa trong trang văn bản 56

2.4 Kết luận..... 68
CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH VÀ THỰC NGHIỆM PHÂN TÍCH TRANG TÀI LIỆU..... 71
3.1. Yêu cầu hệ thống..... 71
3.2. Giới thiệu chương trình 71
3.2.1. Giao diện chương trình..... 72
3.2.2. Chức năng 72
3.3. Thực nghiệm..... 73
3.3.1. Dữ liệu 73
3.3.2. Giới thiệu độ đo PSET 73
3.3.3. Kết quả thực nghiệm và thảo luận 76
TÀI LIỆU THAM KHẢO 88

DANH MỤC HÌNH ẢNH

Hình 1.1: Sơ đồ tổng quan quá trình tạo ảnh tài liệu 14
Hình 1.2: Ví dụ ảnh tài liệu 14
Hình 1.3: Sơ đồ khối liệt kê nhiệm vụ xử lý ảnh tài liệu được phân chia theo cấp bậc trong mỗi vùng của ảnh. 17
Hình 1.4: mô phỏng một chuỗi các bước trong phân tích hình ảnh tài liệu phổ biến. 19
Hình 1.5. Một hình ảnh nhị phân của chữ "e" được thực hiện lên ON và OFF các điểm ảnh, ON điểm ảnh được hiển thị ở đây là "X"[15]. 21
Hình 1.6: Sơ đồ OCR cơ bản 22
Hình 1.7: Cấu trúc vật lý: c, d-Cấu trúc logic của một tài liệu 23
Hình 1.8: Ví dụ loại tài liệu có bố cục phức tạp 25
Hình 1.9: Sơ đồ nguyên lý hệ thống xử lý tài liệu[15] 25
Hình 1.10: a - Ảnh gốc b - Ảnh sau khi tách nền. 27
Hình 1.11: Ví dụ một ảnh tài liệu bị nghiêng một góc 5 độ 28
Hình 1.12: Ví dụ một cây mô tả cấu trúc logic của một trang tài liệu[14] 29
Hình 2.1: Kết quả chiếu nghiêng theo phương ngang và phương thẳng đứng của một trang tài liệu 4 32
Hình 2.2: Phân tách cột dựa vào phép chiếu nghiêng theo phương ngang 33
Hình 2.3: Phép chiếu nghiêng theo phương ngang để phân đoạn ký tự hoặc từ 33
Hình 2.4: Kết quả thực hiện của thuật toán X-Y Cut 35
Hình 2.5: Lược đồ chiếu ngang của một dòng chữ nghiêng 36
- rất khó phân đoạn ký tự 36
Hình 2.6: Lược đồ chiếu đứng của trang tài liệu bị nghiêng 37
Hình 2.7: Lược đồ chiếu đứng của một bài báo 37

Hình 2.8: Phương pháp Dostrum cho phân tích định dạng trang (a) Một phần của nội dung văn bản gốc. (b) Các thành phần lân cận gần nhất được xác định. (c) Các hình chữ nhật tối thiểu tạo nên nhóm láng giềng gần nhất từ đó xác định được dòng văn bản. 39
Hình 2.9: Kết quả thực hiện của kỹ thuật Smearing 41
Hình 2.10: Mô tả thuật toán Tách và Nối thích nghi 44
Hình 2.11: Hình minh họa bước đệ quy của thuật toán Cover khoảng trắng phân nhánh - giới hạn. Xem giải thích ở nội dung văn bản. 49
Hình 2.12: Áp dụng thuật toán tìm kiếm dòng ràng buộc cho các biến thức mô phỏng của một trang. 52
Hình 2.13: Fig. 1.Mô tả thuật toán WCover [16]. (a) hình bao và các hình chữ nhật, (b) điểm chốt tìm được (c,d) các miền con trai/phải và trên/dưới 54
Hình 2.14: Mô hình dòng văn bản được sử dụng tìm kiếm dòng ràng buộc. . 58
Hình 2.15: Minh họa bài toán tìm kiếm dòng ràng buộc với những trở ngại. 59
Hình 2.16: Ví dụ về kết quả đánh giá khoảng trắng để phát hiện các ranh giới cột trong tài liệu có bố cục phức tạp (các tài liệu A00C, D050, và E002 từ cơ sở dữ liệu UW-III). Lưu ý rằng ngay cả các bố cục phức tạp cũng được mô tả bởi một tập nhỏ các dấu tách cột. 63
Hình 3.1: Giao diện chương trình 72
Hình 3.2: Giao diện chức năng chương trình 72
Hình 3.3: Minh họa các kiểu lỗi trong phân tích trang ảnh tài liệu 74
Hình 3.4: Ảnh số 0000085 trong tập ảnh UW-III 76
Hình 3.5: Giao diện và kết quả thực nghiệm 77
Hình 3.6: Kết quả phân tách hình 0000085 – UW-III 77
Hình 3.7: Bảng kết quả thực nghiệm 79
Hình 3.8: Ảnh hưởng của số lượng khoảng trắng tối đa đến kết quả của Wcuts và ageblock. 80

Hình 3.9: Ảnh hưởng của Max_results đến thời gian thực hiện chương trình 80
Hình 3.10: Độ chính xác của thuật toán với độ đo PSET sử dụng tham số khoảng trắng là 300 82
Hình 3.11: Vùng bị bỏ qua 83
Hình 3.12: Vùng bị phân tách thành các phần quá nhỏ 83
Hình 3.13: Độ chính xác của thuật toán với độ đo PSET sử dụng tham số tỉ lệ giao nhau là 95% 84

PHẦN MỞ ĐẦU

1. Đặt vấn đề

Hiện nay, hầu hết tài liệu của con người đều đã được số hóa và được lưu trữ trên máy tính, việc số hóa đảm bảo tính an toàn và thuận tiện hơn hẳn so với sử dụng tài liệu giấy. Tuy nhiên việc sử dụng giấy để lưu trữ tài liệu trong một số mục đích là không thể thay thế hoàn toàn được (như sách, báo, tạp chí, công văn,…). Hơn nữa, lượng tài liệu được tạo ra từ nhiều năm trước vẫn còn rất nhiều mà không thể bỏ đi được vì tính quan trọng của chúng.

Việc chuyển đổi tài liệu điện tử sang tài liệu giấy có thể thực hiện được dễ dàng bằng cách in hay fax, nhưng công việc ngược lại là chuyển từ tài liệu giấy sang tài liệu điện tử lại là một vấn đề không hề đơn giản. Chúng ta mong muốn có thể số hóa tất cả các tài liệu, sách, báo đó và lưu trữ chúng trên máy tính, việc tổ chức và sử dụng chúng sẽ thuận tiện hơn rất nhiều. Vậy nhưng giải pháp sẽ là gì?

Công nghệ đang phát triển một cách chóng mặt, các máy scan với tốc độ hàng nghìn trang một giờ, các máy tính với công nghệ xử lí nhanh chóng và chính xác một cách siêu việt. Vậy tại sao chúng ta không quét các trang tài liệu vào và xử lý, chuyển chúng thành các văn bản một cách tự động? Nhưng vấn đề là khi quét chúng ta chỉ thu được các trang tài liệu đó dưới dạng ảnh nên không thể thao tác, sửa chữa, tìm kiếm như trên các bản Office được, khi đó máy tính không phân biệt được đâu là điểm ảnh của chữ và đâu là điểm ảnh của đối tượng đồ họa.

Một giải pháp được đưa ra đó là xây dựng các hệ thống nhận dạng chữ trong các tấm ảnh chứa cả chữ và đối tượng đồ họa, sau đó chuyển thành dạng trang văn bản và có thể mở, soạn thảo được trên các trình soạn thảo văn bản.

Trong thực tế quá trình nhận dạng thì có rất nhiều tham số ảnh hưởng đến kết quả của các chương trình nhận dạng như nhiễu, Font chữ, kích thước