



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0036265

(51)⁷ G06F 17/27; G06K 9/62

(13) B

(21) 1-2019-06870

(22) 05/12/2019

(30) 10201904554T 21/05/2019 SG

(45) 25/07/2023 424

(43) 25/11/2020 392A

(73) Advanced New Technologies Co., Ltd. (KY)

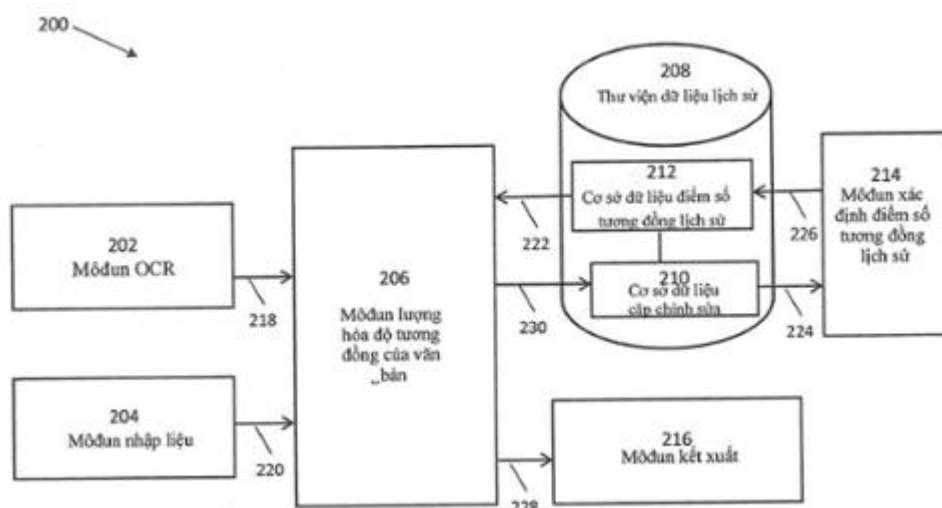
Cayman Corporate Centre, 27 Hospital Road, George Town, Grand Cayman KY1-9008, Cayman Islands

(72) Ruoyu Li (CN).

(74) Công ty TNHH Tầm nhìn và Liên danh (VISION & ASSOCIATES CO.LTD.)

(54) PHƯƠNG PHÁP VÀ THIẾT BỊ LƯỢNG HÓA ĐỘ TƯƠNG ĐỒNG CỦA VĂN BẢN

(57) Sáng chế đề cập đến phương pháp và thiết bị lượng hóa độ tương đồng của văn bản. Theo một phương án, sáng chế đề xuất thiết bị lượng hóa độ tương đồng của văn bản bao gồm: bộ xử lý; và bộ nhớ bao gồm mã chương trình máy tính. Bộ nhớ và mã chương trình máy tính được tạo cấu hình để, cùng với bộ xử lý, khiến thiết bị: thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chỉnh sửa quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa; xác định các điểm số tương đồng, mỗi trong số các điểm số tương đồng này tương ứng với một trong số các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất; và chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập theo nghĩa rộng, nhưng không giới hạn, đến phương pháp và thiết bị lượng hóa độ tương đồng của văn bản.

Tình trạng kỹ thuật của sáng chế

Nhận dạng ký tự quang học, thường được viết tắt là OCR, là phương pháp để nhận dạng các văn bản trong các hình ảnh hoặc các tài liệu cứng và chuyển đổi chúng thành các văn bản mã hóa bằng máy. Các văn bản mã hóa bằng máy này thường được gọi là các văn bản OCR.

Xác minh danh tính điện tử (eKYC - electronic-Know Your Customer) là quy trình thẩm định kỹ thuật số được thực hiện bởi doanh nghiệp để xác minh danh tính của khách hàng và đánh giá rủi ro tiềm tàng của các mục đích bất hợp pháp nhằm đến quan hệ doanh nghiệp (ví dụ như rửa tiền).

Khi thực hiện eKYC, OCR có thể được sử dụng để nhận nội dung văn bản, ví dụ như tên, địa chỉ, mã số định danh, v.v. trong các chứng từ nhận dạng (ID), ví dụ thẻ nhận dạng, hộ chiếu v.v.. Thông thường, các văn bản OCR được thể hiện trên giao diện người dùng để người dùng xác nhận hoặc chỉnh sửa. Nếu người dùng sửa các văn bản OCR bằng các văn bản chỉnh sửa, thì sẽ cần phải xác minh xem các văn bản chỉnh sửa có hợp lý hay không, là hành động để kiểm soát rủi ro, trước khi việc chỉnh sửa này được chấp thuận. Do đó, bước này trở nên hữu dụng để đánh giá độ tương đồng giữa các văn bản OCR và các văn bản chỉnh sửa của người dùng để xác minh xem các văn bản chỉnh sửa có hợp lý hay không.

Hiện nay, phương pháp khoảng cách chỉnh sửa và phương pháp khoảng cách Jaccard được sử dụng để đánh giá độ tương đồng giữa hai chuỗi văn bản. Tuy nhiên, phương pháp khoảng cách chỉnh sửa đánh giá độ tương đồng dựa trên số lượng nhỏ nhất của các thao tác cần thiết để biến đổi một chuỗi văn bản thành chuỗi văn bản khác, nhưng lại bỏ qua độ tương đồng trực quan giữa hai chuỗi văn bản; mặt khác, phương pháp

khoảng cách Jaccard bỏ qua cả độ tương đồng trực quan và thứ tự của các ký tự trong hai chuỗi văn bản.

Các ví dụ về cách mà độ tương đồng giữa 2 chuỗi văn bản được đánh giá theo phương pháp khoảng cách chỉnh sửa được thể hiện trong Bảng 1 dưới đây.

Bảng 1		
Ví dụ	Chuỗi văn bản A $\rightarrow$ Chuỗi văn bản B	Khoảng cách chỉnh sửa
1	B0B $\rightarrow$ BOB	1
2	BOB $\rightarrow$ 1OB	1

Như được thể hiện trong Bảng 1, trong phương pháp khoảng cách chỉnh sửa, khoảng cách chỉnh sửa được sử dụng để biểu thị số lượng nhỏ nhất của các thao tác cần thiết để biến đổi Chuỗi văn bản A thành Chuỗi văn bản B. Trong ví dụ 1, cần có một thao tác thay thế số “0” bằng chữ cái “O” để biến đổi “B0B” thành “BOB”. Trong ví dụ 2, cần có một thao tác thay thế chữ cái “B” bằng số “1” để biến đổi “BOB” thành “1OB”. Các khoảng cách chỉnh sửa trong các ví dụ 1 và 2 là giống nhau, tức là mỗi khoảng cách đều bằng 1. Như vậy, độ tương đồng giữa “B0B” và “BOB” và độ tương đồng giữa “BOB” và “1OB” được xem là giống nhau trong phương pháp khoảng cách chỉnh sửa. Sự đánh giá tương đồng như vậy chỉ xem xét số lượng nhỏ nhất của các thao tác nhưng lại bỏ qua độ tương đồng trực quan giữa cặp ký tự trong mỗi thao tác, xem xét một cách không chính xác rằng “0” và “O” với “B” và “1” là giống nhau.

Do đó cần có phương pháp và thiết bị cung cấp cơ chế tổng thể và toàn diện để cải thiện sự đánh giá độ tương đồng của văn bản không chỉ bằng cách đánh giá số lượng nhỏ nhất của các thao tác cần thiết để biến đổi một chuỗi văn bản thành chuỗi văn bản khác

mà còn cả bằng mối tương quan giữa hai chuỗi văn bản này, tức là khả năng các thao tác này có thể xảy ra giữa hai chuỗi văn bản này là như thế nào.

Nhờ có cơ chế đánh giá độ tương đồng của văn bản được cải thiện, nên phương pháp và thiết bị theo sáng chế được tạo cấu hình để không chỉ đánh giá (tức là để xác định xem “có tương đồng hay không”) mà còn lượng hóa (tức là để xác định xem “tương đồng như thế nào”) độ tương đồng của văn bản giữa hai chuỗi văn bản, mà hỗ trợ các hệ thống eKYC và các hệ thống điện tử bất kỳ khác mà triển khai phương pháp OCR để thực hiện chức năng với độ chính xác và độ tin cậy được tăng cường của việc xác minh tự động việc chỉnh sửa thủ công các văn bản OCR.

Bản chất kỹ thuật của sáng chế

Theo phương án thứ nhất, sáng chế đề xuất thiết bị lượng hóa độ tương đồng của văn bản, thiết bị bao gồm: bộ xử lý; và bộ nhớ bao gồm mã chương trình máy tính. Bộ nhớ và mã chương trình máy tính được tạo cấu hình để, cùng với bộ xử lý, ít nhất làm cho thiết bị có thể: thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử; tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó khi xác định điểm số tương đồng lịch sử, thì thiết bị được tạo cấu hình để: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0 theo công thức sau: $\text{điểm}(p) = 1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$, trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p), trong đó thiết bị còn được tạo cấu hình để: thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng ký tự quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa; xác định các điểm số tương

đồng, mỗi trong số các điểm số tương đồng này tương ứng với một trong số các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất, trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Theo phương án thứ hai, sáng chế đề xuất phương pháp được thực hiện bằng máy tính để lượng hóa độ tương đồng của văn bản, phương pháp này bao gồm các bước: thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử; tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó bước tính toán điểm số tương đồng lịch sử bao gồm các bước: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0 theo công thức sau: $\text{điểm}(p) = 1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$, trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p), trong đó phương pháp này còn bao gồm các bước: thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chỉnh sửa quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa; xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất, trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và chọn điểm số nhỏ nhất

trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Theo phương án thứ ba, sáng chế đề xuất hệ thống lượng hóa độ tương đồng của văn bản, hệ thống này bao gồm: thiết bị OCR được tạo cấu hình để thu được chuỗi văn bản OCR từ hình ảnh kỹ thuật số; thiết bị nhập liệu được tạo cấu hình để thu được chuỗi chỉnh sửa của người dùng; thiết bị lượng hóa độ tương đồng của văn bản để lượng hóa độ tương đồng của văn bản, trong đó thiết bị lượng hóa độ tương đồng của văn bản này bao gồm thiết bị xác định điểm số tương đồng lịch sử, trong đó thiết bị lượng hóa độ tương đồng của văn bản này được tạo cấu hình để: thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử; tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử bằng thiết bị xác định điểm số tương đồng lịch sử, trong đó thiết bị xác định điểm số tương đồng lịch sử này được tạo cấu hình để: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1 theo công thức sau: $\text{điểm}(p) = 1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$, trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p), trong đó thiết bị lượng hóa độ tương đồng của văn bản này còn được tạo cấu hình để: thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng ký tự quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa; xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất, trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa

độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa; và thiết bị kết xuất được tạo cấu hình để kết xuất kết quả lượng hóa độ tương đồng của văn bản.

Mô tả vắn tắt các hình vẽ

Các phương án và ví dụ thực hiện chỉ được đưa ra bằng cách ví dụ, và được hiểu dễ hơn và rõ ràng hơn đối với người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan từ phần mô tả sau đây, kết hợp với hình vẽ kèm theo, trong đó:

Fig.1 là sơ đồ giản lược của thiết bị lượng hóa độ tương đồng của văn bản, theo một phương án của sáng chế.

Fig.2 là sơ đồ giản lược của hệ thống lượng hóa độ tương đồng của văn bản, theo một phương án của sáng chế.

Fig.3 là lưu đồ minh họa phương pháp lượng hóa độ tương đồng của văn bản, theo một phương án của sáng chế.

Fig.4 thể hiện sơ đồ ma trận minh họa phương án về phép tính đường thao tác ngắn nhất để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa. Ví dụ được minh họa theo phương pháp khoảng cách chỉnh sửa.

Fig.5 thể hiện sơ đồ dạng bảng minh họa phương án 502 về bước thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa như được thể hiện trên Fig.4 và phương án 504 về bước xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, theo ví dụ thực hiện của sáng chế.

Fig.6 là sơ đồ dạng bảng thể hiện điểm số tương đồng lịch sử của các cặp chỉnh sửa theo một phương án của sáng chế.

Fig.7 là lưu đồ minh họa phương pháp lượng hóa độ tương đồng của văn bản theo một phương án khác.

Fig.8 thể hiện sơ đồ khối của hệ thống máy tính phù hợp để sử dụng làm thiết bị ví dụ được thể hiện trên Fig.1 và trong hệ thống làm ví dụ như được thể hiện trên Fig.2 để thực hiện ít nhất một số bước của phương pháp lượng hóa độ tương đồng của văn bản theo các phương án được thể hiện trên các hình vẽ từ Fig.3 đến Fig.7.

Người có hiểu biết trong lĩnh vực liên quan sẽ hiểu rằng các phần tử trên các hình vẽ được minh họa cho mục đích làm đơn giản và làm rõ và không cần thiết phải được mô tả theo tỷ lệ. Ví dụ, kích thước của một số phần tử trong phần minh họa, sơ đồ khối hoặc lưu đồ có thể được phóng đại so với các phần tử khác để giúp cải thiện việc hiểu các phương án.

Mô tả chi tiết sáng chế

Các phương án sẽ được mô tả, chỉ bằng ví dụ, với tham chiếu đến các hình vẽ. Các số chỉ dẫn và các ký tự giống nhau trên các hình vẽ biểu thị các bộ phận giống nhau hoặc tương đương.

Một số phần của phần mô tả sau đây được thể hiện rõ ràng hoặc hoàn toàn theo các thuật toán và biểu diễn dạng chức năng hoặc biểu tượng của các thao tác trên dữ liệu trong bộ nhớ máy tính. Các phần mô tả dạng thuật toán và biểu diễn dạng chức năng hoặc biểu tượng là các phương tiện được sử dụng bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật xử lý dữ liệu để truyền tải được hiệu quả nhất bản chất công việc của họ đến những người có hiểu biết trung bình khác trong lĩnh vực kỹ thuật này. Thuật toán thường được tạo ra để làm chuỗi các bước tự thống nhất tạo ra kết quả mong muốn. Các bước này cần đến sự điều khiển vật lý các thực thể vật lý, chẳng hạn như tín hiệu điện, từ hoặc quang mà có thể được lưu trữ, được truyền, được tích hợp, được so sánh, và ngoài ra được điều khiển.

Trừ khi có quy định khác, và rõ ràng từ phần sau đây, cần hiểu rằng trong suốt bản mô tả, phần thảo luận sử dụng các thuật ngữ chẳng hạn như “nhận”, “biểu hiện”, “xác định”, “chọn”, “lấy ra”, “tính toán”, “thu thập”, “thực hiện”, “chuẩn hóa”, “bổ sung”, “cập nhật”, “chỉnh sửa”, “duy trì”, “kết xuất”, “nhận được”, “nhận” hoặc dạng tương tự, đề cập đến hành động và các quy trình của hệ thống máy tính, hoặc thiết bị điện tử tương tự, mà điều khiển và biến đổi dữ liệu được biểu diễn dưới dạng các thực thể vật lý trong hệ thống máy tính thành dữ liệu tương tự khác được biểu diễn dưới dạng các thực thể vật lý trong hệ thống máy tính hoặc bộ nhớ thông tin, thiết bị truyền phát hoặc hiển thị khác.

Sáng chế còn bộc lộ thiết bị thực hiện các thao tác của các phương pháp. Các thiết bị này có thể được cấu tạo đặc biệt cho mục đích theo yêu cầu, hoặc có thể bao gồm máy tính hoặc thiết bị khác được kích hoạt hoặc được tạo cấu hình lại một cách có lựa chọn bởi chương trình máy tính được lưu trữ trong máy tính. Các thuật toán và cách thể hiện được biểu diễn ở đây không có liên hệ với bất kỳ máy tính hoặc thiết bị cụ thể nào khác. Các loại máy khác nhau có thể được sử dụng với các chương trình theo các phần đã nêu ở đây. Theo cách khác, cấu tạo của thiết bị chuyên dụng để thực hiện các bước theo yêu cầu của phương pháp có thể là thích hợp. Cấu trúc của máy tính phù hợp để thực thi các phương pháp/quy trình khác nhau được mô tả ở đây sẽ được nêu ở phần mô tả sau đây. Theo sáng chế, các thuật ngữ “thiết bị” và “mô đun” được sử dụng thay thế được cho nhau để đề cập đến các thiết bị thực hiện các thao tác của các phương pháp.

Ngoài ra, sáng chế cũng bộc lộ rõ ràng chương trình máy tính, trong đó sẽ rõ ràng với người có hiểu biết trung bình trong lĩnh vực kỹ thuật rằng các bước độc lập của phương pháp được mô tả ở đây có thể được thực hiện bằng mã máy tính. Chương trình máy tính không nhằm giới hạn các ngôn ngữ lập trình cụ thể và phương án thực hiện bất kỳ của nó. Cần hiểu rằng các ngôn ngữ lập trình và mã hóa của nó có thể được sử dụng để thực hiện phần bộc lộ của phần mô tả ở đây. Ngoài ra, chương trình máy tính không nhằm giới hạn luồng điều khiển cụ thể bất kỳ. Có nhiều biến thể khác của chương trình máy tính, mà có thể sử dụng các luồng điều khiển khác nhau mà không tách rời khỏi bản chất và phạm vi của sáng chế.

Ngoài ra, một hoặc nhiều bước của chương trình máy tính có thể được thực hiện song song thay vì lần lượt. Chương trình máy tính này có thể được lưu trữ trên vật ghi đọc được bằng máy tính bất kỳ. Vật ghi đọc được bằng máy tính có thể bao gồm thiết bị lưu trữ như là đĩa từ hoặc đĩa quang, chip nhớ, hoặc các thiết bị lưu trữ phù hợp khác để giao tiếp với máy tính. Vật ghi đọc được bằng máy tính cũng có thể bao gồm vật ghi dùng dây ví dụ như trong hệ thống Internet, hoặc vật ghi không dây ví dụ như trong công nghệ điện thoại di động dùng GSM. Chương trình máy tính khi được nạp và thực thi trên máy tính làm cho thiết bị có thể thực hiện các bước của các phương pháp ưu tiên.

Các phương án của sáng chế đưa ra phương pháp điều khiển bằng dữ liệu mà sử dụng tần suất của các thao tác từ dữ liệu lịch sử để đánh giá và lượng hóa độ tương đồng của văn bản giữa hai chuỗi văn bản tốt hơn.

Các thao tác bao gồm các thao tác chèn, các thao tác xóa và/hoặc các thao tác thay thế mà có thể thực hiện được cho các ký tự của chuỗi văn bản thứ nhất để sửa chuỗi văn bản thứ nhất bằng chuỗi văn bản thứ hai. Các ký tự bao gồm số, chữ cái, biểu tượng, v.v.. Trong ngữ cảnh OCR, chuỗi văn bản thứ nhất bao gồm chuỗi văn bản OCR được nhận diện từ nội dung văn bản của hình ảnh hoặc tài liệu cứng. Chuỗi văn bản OCR có thể là tên, địa chỉ, ngày sinh, mã số định danh, v.v.. Chuỗi văn bản thứ hai bao gồm chuỗi chỉnh sửa của người dùng mà người dùng thực hiện để sửa chuỗi văn bản OCR.

Theo các phương án khác nhau của sáng chế, các cặp chỉnh sửa được sử dụng để biểu thị các thao tác nêu trên. Ví dụ, cặp chỉnh sửa (“”, “a”) biểu thị thao tác chèn ký tự “a”; cặp chỉnh sửa (“a”, “”) biểu thị thao tác xóa ký tự “a”; và cặp chỉnh sửa (“a”, “b”) biểu thị thao tác thay thế ký tự “a” bằng ký tự “b”.

Fig.1 minh họa sơ đồ giản lược của thiết bị 100 lượng hóa độ tương đồng của văn bản theo một phương án của sáng chế. Thiết bị 100 bao gồm ít nhất là môđun xử lý 102 và môđun lưu trữ 104. Môđun xử lý 102 và môđun lưu trữ 104 được nối thông với nhau. Môđun lưu trữ 104 bao gồm mã chương trình máy tính (không được thể hiện trên Fig.1). Môđun lưu trữ 104 và mã chương trình máy tính này được tạo cấu hình để, cùng với môđun xử lý 102, khiến thiết bị 100 thực hiện các bước lượng hóa độ tương đồng của văn bản như được mô tả theo sáng chế. Chi tiết về các bước này sẽ được mô tả dưới đây với tham chiếu đến Fig.3, mà minh họa lưu đồ minh họa phương pháp 300 lượng hóa độ tương đồng của văn bản theo một phương án của sáng chế.

Thiết bị 100 để lượng hóa độ tương đồng của văn bản có thể được thực hiện dưới dạng môđun lượng hóa độ tương đồng của văn bản 206 trong hệ thống làm ví dụ 200 để lượng hóa độ tương đồng của văn bản như được thể hiện trên Fig.2, theo một phương án của sáng chế.

Như được thể hiện trên Fig.2, ngoài môđun lượng hóa độ tương đồng của văn bản 206, hệ thống 200 có thể bao gồm môđun OCR 202, môđun nhập liệu 204, thư viện dữ liệu lịch sử 208, môđun xác định điểm số tương đồng lịch sử 214, và môđun kết xuất 216.

Môđun OCR 202 có thể bao gồm môđun tiếp nhận và môđun chuyển đổi hình ảnh, cả hai trong số này đều không được thể hiện trên Fig.2, môđun tiếp nhận có thể ở dạng máy quét quang học được tạo cấu hình để thu được hình ảnh kỹ thuật số của tài liệu cứng. Môđun chuyển đổi hình ảnh, cùng với phần mềm OCR phù hợp, có thể được tạo cấu hình để biến đổi ít nhất một phần của nội dung văn bản của hình ảnh kỹ thuật số của tài liệu cứng thành văn bản (hoặc mã) đọc được bằng máy tương ứng. Các văn bản (hoặc mã) đọc được bằng máy được đề cập đến theo cách hoán đổi được cho nhau là chuỗi văn bản OCR theo sáng chế. Theo cách này, có thể thu được chuỗi văn bản OCR bằng môđun OCR 202.

Môđun nhập liệu 204 có thể bao gồm bàn phím, chuột, bút cảm ứng, giao diện người dùng nhận diện giọng nói, v.v. để thu được dữ liệu nhập từ người dùng. Khi chuỗi văn bản OCR được hiển thị đến môđun kết xuất 216 để người dùng xác nhận hoặc chỉnh sửa, thì người dùng có thể sử dụng môđun nhập liệu 204 để xác nhận chuỗi văn bản OCR. Nếu người dùng sử dụng môđun nhập liệu 204 để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa, thì có thể thu được chuỗi văn bản chỉnh sửa bằng môđun nhập liệu 204.

Theo một số phương án, khi thu được chuỗi văn bản chỉnh sửa bằng môđun nhập liệu 204, hệ thống có thể ra lệnh (không được thể hiện trên Fig.2 cho đơn giản) cho môđun OCR 202 và môđun nhập liệu 204 truyền chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa đến môđun lượng hóa độ tương đồng của văn bản 206 tương ứng.

Môđun lượng hóa độ tương đồng của văn bản 206 có thể nhận 218 chuỗi văn bản OCR từ môđun OCR 202 và nhận 220 chuỗi văn bản chỉnh sửa từ môđun nhập liệu 204, và lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa theo các phương pháp của sáng chế. Khi lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa theo các phương pháp của sáng chế, thì môđun lượng hóa độ tương đồng của văn bản 206 có thể truyền thông 222 với thư viện dữ

liệu lịch sử 208 để lấy ra điểm số tương đồng lịch sử của các cặp chỉnh sửa nhất định cần để lượng hóa độ tương đồng của văn bản.

Các ví dụ về các phương pháp cho mô đun lượng hóa độ tương đồng của văn bản 206 để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa được mô tả trên các hình vẽ từ Fig.3 đến Fig.6. Như được thể hiện trên Fig.3, khi lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa, thì mô đun lượng hóa độ tương đồng của văn bản 206 được tạo cấu hình để thực hiện ít nhất là các bước sau:

302: thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa;

304: xác định các điểm số tương đồng, mỗi trong số các điểm số tương đồng này tương ứng với một trong số các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất; và

306: chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Chi tiết về các bước trên sẽ được mô tả trong phần mô tả sau với tham chiếu đến Fig.3, Fig.4 và Fig.5.

Tham khảo lại Fig.2, thư viện dữ liệu lịch sử 208 có thể bao gồm cơ sở dữ liệu cặp chỉnh sửa 210 để lưu trữ các cặp chỉnh sửa, và cơ sở dữ liệu điểm số tương đồng lịch sử 212 để lưu trữ điểm số tương đồng lịch sử của các cặp chỉnh sửa mà được lưu trữ trong cơ sở dữ liệu cặp chỉnh sửa 210, Cơ sở dữ liệu cặp chỉnh sửa 210 có thể được tạo cấu hình để tính toán tần suất của các cặp chỉnh sửa được lưu trữ trong đó trong khoảng thời gian nhất định. Ví dụ, khoảng thời gian nhất định có thể là một tháng, sáu tháng hoặc khoảng thời gian bất kỳ phụ thuộc vào nhu cầu thực tế.

Sau đó, tần suất của các cặp chỉnh sửa có thể được truyền 224 từ cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208 đến môđun xác định điểm số tương đồng lịch sử 214. Sau khi nhận tần suất của các cặp chỉnh sửa, môđun xác định điểm số tương đồng lịch sử 214 được tạo cấu hình để xác định điểm số tương đồng lịch sử của các cặp chỉnh sửa dựa trên tần suất của các cặp chỉnh sửa nhận được từ cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208. Sau đó, điểm tương đồng lịch sử của các cặp chỉnh sửa được truyền 226 đến cơ sở dữ liệu điểm số tương đồng lịch sử 212 của thư viện dữ liệu lịch sử 208 và được lưu trữ trong đó.

Tùy thuộc vào sự lượng hóa độ tương đồng của văn bản, sau đó, môđun lượng hóa độ tương đồng của văn bản 206 có thể kết xuất 228 kết quả lượng hóa độ tương đồng của văn bản đến môđun kết xuất 216. Kết quả lượng hóa độ tương đồng của văn bản này có thể bao gồm các tin nhắn thông báo hệ thống biểu thị việc chấp thuận hoặc từ chối các chuỗi văn bản chỉnh sửa hoặc cảnh báo hệ thống biểu thị rủi ro tiềm tàng hoặc việc khóa tài khoản người dùng.

Nếu chuỗi văn bản chỉnh sửa được chấp thuận, thì môđun lượng hóa độ tương đồng của văn bản 206 có thể bổ sung 230 các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất vào cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208. Để đáp lại việc bổ sung các cặp chỉnh sửa này, thư viện dữ liệu lịch sử 208 có thể được tạo cấu hình để tương tác với môđun xác định điểm số tương đồng lịch sử 214 để cập nhật điểm số tương đồng lịch sử của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử 208.

Theo một số phương án thay thế, môđun xác định điểm số tương đồng lịch sử 214 có thể được thực hiện dưới dạng môđun nằm trong thư viện dữ liệu lịch sử 208.

Theo một số phương án thay thế, thư viện dữ liệu lịch sử 208 và/hoặc môđun xác định điểm số tương đồng lịch sử 214 có thể được thực hiện dưới dạng các môđun có trong môđun lượng hóa độ tương đồng của văn bản 206.

Theo một số phương án thay thế, cơ sở dữ liệu cặp chỉnh sửa 210 và cơ sở dữ liệu điểm số tương đồng lịch sử 212 có thể được thực hiện dưới dạng một cơ sở dữ liệu trong

thư viện dữ liệu lịch sử 208 mà lưu trữ cả các cặp chỉnh sửa và các điểm số tương đồng lịch sử tương ứng của chúng.

Như được mô tả ở trên, Fig.3 minh họa lưu đồ minh họa phương pháp 300 để lượng hóa độ tương đồng của văn bản, theo một phương án của sáng chế.

Ở bước 302, thiết bị 100 hoặc mô đun lượng hóa độ tương đồng của văn bản 206 thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa. Mỗi trong số các đường thao tác ngắn nhất bao gồm một hoặc nhiều cặp chỉnh sửa. Mỗi một trong một hoặc nhiều cặp chỉnh sửa biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa.

Các đường thao tác ngắn nhất có thể thu được theo nhiều cách khác nhau. Trong một phương án, các đường thao tác ngắn nhất có thể thu được bằng cách thực hiện phép tính khoảng cách chỉnh sửa để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa. Theo phương án này, thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa có thể là một trong số thao tác chèn, thao tác xóa, hoặc thao tác thay thế. Sẽ được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng các phương pháp khác cũng có thể được sử dụng để thu được các đường thao tác ngắn nhất. Ví dụ, nhiều thao tác hơn có thể được xác định, chẳng hạn như thao tác hoán đổi hai ký tự liên kề nhau, mà coi việc chuyển đổi chuỗi văn bản “AB” thành chuỗi văn bản “BA” khác là một thao tác.

Theo một phương án, khoảng cách chỉnh sửa giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa lấy ra được bằng cách sử dụng phương pháp khoảng cách chỉnh sửa để thu được các đường thao tác ngắn nhất.

Dựa trên phương pháp khoảng cách chỉnh sửa, cho hai chuỗi văn bản: chuỗi X có chiều dài n và chuỗi Y có chiều dài m , khoảng cách chỉnh sửa giữa các ký tự i thứ nhất của chuỗi X và các ký tự j thứ nhất của chuỗi Y được xác định là $D(i, j)$, và khoảng cách chỉnh sửa giữa chuỗi X và chuỗi Y được xác định là $D(n, m)$.

Khởi tạo $D(i, 0) = i$ và $D(0, j) = j$,

với mỗi $i = 1 \dots N$ và mỗi $j = 1 \dots M$,

$D(i, j)$ có thể được lấy ra dựa trên công thức (1) sau:

Công thức (1)

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + 1; \text{ if } X(i) \neq Y(j) \\ D(i-1, j-1) + 0; \text{ if } X(i) = Y(j) \end{cases}$$

Dựa trên công thức (1), khoảng cách chỉnh sửa $D(N, M)$ giữa chuỗi X và chuỗi Y có thể được lấy ra theo đó. Fig.4 thể hiện phương án minh họa sơ đồ ma trận 400 lấy ra khoảng cách chỉnh sửa giữa chuỗi văn bản OCR 402, tức là “Azaha” và chuỗi văn bản chỉnh sửa 404, tức là “Aziz”, dựa trên công thức (1). Chuỗi văn bản làm ví dụ trong phương án này liên quan đến tên người dùng. Cần được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa cũng có thể liên quan đến địa chỉ, mã số định danh, ngày sinh, v.v..

Như được thể hiện trong ô 406 của sơ đồ ma trận 400, khoảng cách chỉnh sửa là 3 giữa chuỗi văn bản OCR 402, tức là “Azaha” và chuỗi văn bản chỉnh sửa 404, tức là “Aziz”. Việc này có nghĩa là số lượng nhỏ nhất của các thao tác cần để chỉnh sửa “Azaha” thành “Aziz” là 3.

Dựa trên khoảng cách chỉnh sửa 3, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 thu được, ở bước 302, tất cả các đường khả thi mà có thể biến đổi “Azaha” thành “Aziz” trong 3 bước thao tác. Các đường khả thi này được xem là các đường thao tác ngắn nhất.

Phương án của bước 304 mà thu được các đường thao tác ngắn nhất được minh họa trong Bảng 500 trên Fig.5. Như được minh họa trong bảng phụ 502, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 thu được ba đường thao tác ngắn nhất để sửa chuỗi văn bản OCR 402, tức là “Azaha” thành chuỗi văn bản chỉnh sửa 404, tức là “Aziz”. Ba đường thao tác ngắn nhất bao gồm: đường thao tác ngắn nhất 1, đường thao tác ngắn nhất 2, và đường thao tác ngắn nhất 3.

Như được thể hiện trong bảng phụ 502, đường thao tác ngắn nhất 1 bao gồm ba thao tác, tức là thao tác 1 thay thế ký tự “a” bằng ký tự “i”, thao tác 2 thay thế ký tự “h”

bằng ký tự “z”, và thao tác 3 xóa ký tự “a”. Các thao tác 1, 2, 3 trong đường thao tác ngắn nhất 1 được biểu thị là các cặp chỉnh sửa (“a”, “i”), (“h”, “z”), và (“a”, “”) tương ứng.

Đường thao tác ngắn nhất 2 bao gồm ba thao tác, tức là thao tác 1 xóa ký tự “a”, thao tác 2 thay thế ký tự “h” bằng ký tự “i”, và thao tác 3 thay thế ký tự “a” bằng ký tự “z”. Các thao tác 1, 2, 3 trong đường thao tác ngắn nhất 2 được biểu thị là các cặp chỉnh sửa (“a”, “”), (“h”, “i”), và (“a”, “z”) tương ứng.

Đường thao tác ngắn nhất 3 bao gồm ba thao tác, tức là thao tác 1 thay thế ký tự “a” bằng ký tự “i”, thao tác 2 xóa ký tự “h”, và thao tác 3 thay thế ký tự “a” bằng ký tự “z”. Các thao tác 1, 2, 3 trong đường thao tác ngắn nhất 3 được biểu thị là các cặp chỉnh sửa (“a”, “i”), (“h”, “”), và (“a”, “z”) tương ứng.

Như được thể hiện theo một phương án trên, mỗi trong số ba đường thao tác ngắn nhất bao gồm ba cặp chỉnh sửa. Mỗi trong số các cặp chỉnh sửa biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR 402, tức là “Azaha”, trong quá trình chỉnh sửa thành chuỗi văn bản chỉnh sửa 404, tức là “Aziz”.

Sẽ được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng số lượng đường thao tác ngắn nhất và số lượng cặp chỉnh sửa được có trong đường thao tác ngắn nhất có thể thay đổi dựa trên chiều dài của chuỗi văn bản và/hoặc số lượng thao tác cần thiết. Ví dụ, khi chỉnh sửa chuỗi văn bản OCR “B0B” thành chuỗi văn bản chỉnh sửa “BOB”, thì khoảng cách chỉnh sửa là 1, có nghĩa là chỉ cần một thao tác. Trong ví dụ này, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể thu được chỉ một đường thao tác ngắn nhất để sửa chuỗi văn bản OCR “B0B” thành chuỗi văn bản chỉnh sửa “BOB”. Một đường thao tác ngắn nhất bao gồm một thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR “B0B”, tức là thay thế ký tự “0” bằng ký tự “O”. Thao tác này được biểu thị bởi cặp chỉnh sửa (“0”, “O”).

Ở bước 304, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định các điểm số tương đồng cho các đường thao tác ngắn nhất. Mỗi trong số các điểm số tương đồng tương ứng với một trong số các đường thao tác ngắn nhất như được xác định trong bước 302. Mỗi trong số các điểm số tương đồng được xác định bằng cách

cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất.

Một phương án của bước 304 được minh họa trong bảng phụ 504 trên Fig.5. Như được thể hiện trong bảng phụ 504, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định ba điểm số tương đồng tương ứng với ba đường thao tác ngắn nhất tương ứng. Tức là, với mỗi trong số đường thao tác ngắn nhất 1, đường thao tác ngắn nhất 2, và đường thao tác ngắn nhất 3, thì điểm số tương đồng được xác định. Theo phương án này, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định mỗi trong số ba điểm số tương đồng bằng cách cộng các điểm số tương đồng lịch sử của các cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất.

Ví dụ, đối với đường thao tác ngắn nhất 1, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định điểm số tương đồng bằng cách cộng các điểm số tương đồng lịch sử của các cặp chỉnh sửa (“a”, “i”), (“h”, “z”), và (“a”, “”) tương ứng.

Đối với đường thao tác ngắn nhất 2, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định điểm số tương đồng bằng cách cộng các điểm số tương đồng lịch sử của các cặp chỉnh sửa (“a”, “”), (“h”, “i”), và (“a”, “z”) tương ứng.

Đối với đường thao tác ngắn nhất 3, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 xác định điểm số tương đồng bằng cách cộng các điểm số tương đồng lịch sử của các cặp chỉnh sửa (“a”, “i”), (“h”, “”), và (“a”, “z”) tương ứng.

Trong một ví dụ, khi cộng điểm số tương đồng lịch sử của các cặp chỉnh sửa ở bước 403, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể lấy ra 222 điểm số tương đồng lịch sử của các cặp chỉnh sửa từ thư viện dữ liệu lịch sử 208, như được minh họa trên Fig.2.

Trong một ví dụ khác, khi cộng điểm số tương đồng lịch sử của các cặp chỉnh sửa ở bước 403, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể lấy ra điểm số tương đồng lịch sử của các cặp chỉnh sửa từ cơ sở dữ liệu lịch sử. Cơ sở dữ liệu lịch sử có thể được thực hiện bằng môđun lưu trữ 104 có thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Trong một ví dụ khác, thư viện dữ liệu lịch sử 208 có thể được thực hiện dưới dạng môđun có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Trong ví dụ này, khi cộng điểm số tương đồng lịch sử của các cặp chỉnh sửa ở bước 403, thì thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể lấy ra điểm số tương đồng lịch sử của các cặp chỉnh sửa từ môđun có trong đó.

Theo một số phương án, việc tạo ra điểm số tương đồng lịch sử của các cặp chỉnh sửa có thể được thực hiện bởi thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể bao gồm cơ sở dữ liệu cặp chỉnh sửa 210 và cơ sở dữ liệu điểm số tương đồng lịch sử 212 như được mô tả trên Fig.2. Theo một phương án về việc tạo ra điểm số tương đồng lịch sử, thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 được tạo cấu hình để thu thập tất cả các loại cặp chỉnh sửa được thực hiện bởi người dùng trong khoảng thời gian định trước. Ví dụ, tất cả các loại cặp chỉnh sửa có thể được thu thập bằng thiết bị 100 hoặc thiết bị lượng hóa độ tương đồng của văn bản 206 để đáp lại việc nhận chuỗi văn bản OCR của các hình ảnh kỹ thuật số thu được bởi thiết bị OCR/môđun và nhận chuỗi văn bản chỉnh sửa thu được bằng thiết bị nhập liệu/môđun tương ứng với chuỗi văn bản OCR. Các cặp chỉnh sửa này có thể được khởi tạo với mỗi cặp có giá trị bằng 1,0 và được lưu trữ vào cơ sở dữ liệu cặp chỉnh sửa 210.

Khoảng thời gian định trước có thể là một tháng, sáu tháng hoặc khoảng thời gian bất kỳ phụ thuộc vào nhu cầu thực tế. Trong khoảng thời gian định trước, tần suất của các cặp chỉnh sửa được lưu trữ trong cơ sở dữ liệu cặp chỉnh sửa 210 có thể được tính toán. Các cặp chỉnh sửa được lưu trữ càng thường xuyên, thì số lần các thao tác được biểu thị bởi các cặp chỉnh sửa xảy ra càng nhiều, và gần như là các thao tác được biểu thị bởi các cặp chỉnh sửa có thể sẽ xảy ra lần nữa. Theo đó, tần suất của các cặp chỉnh sửa được lưu trữ trong cơ sở dữ liệu cặp chỉnh sửa 210 được sử dụng để xác định điểm số tương đồng lịch sử của các cặp chỉnh sửa được lưu trữ trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư

viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Theo một số phương án, tần suất của các cặp chỉnh sửa có thể được cung cấp cho môđun xác định điểm số tương đồng lịch sử 214 để xác định điểm số tương đồng lịch sử của các cặp chỉnh sửa được lưu trữ trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Môđun xác định điểm số tương đồng lịch sử 214 có thể là môđun độc lập, như được mô tả trên Fig.2, truyền thông được với thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Ngoài ra, môđun xác định điểm số tương đồng lịch sử 214 có thể được thực hiện dưới dạng môđun có trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Theo một số phương án, môđun xác định điểm số tương đồng lịch sử 214 có thể được tạo cấu hình để thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Sau đó, phép tính tần suất log của các cặp chỉnh sửa trong cơ sở dữ liệu cặp chỉnh sửa 210 có thể được chuẩn hóa bằng môđun xác định điểm số tương đồng lịch sử 214 về khoảng từ 0,0 đến 1,0. Phép tính tần suất log được chuẩn hóa được xem là điểm số tương đồng lịch sử của các cặp chỉnh sửa trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Theo một số phương án, điểm số tương đồng lịch sử của các cặp chỉnh sửa có thể được truyền từ môđun xác định điểm số tương đồng lịch sử 214 đến cơ sở dữ liệu điểm số tương đồng lịch sử 212 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206, và được lưu trữ trong đó. Theo một số phương án, môđun xác định điểm số tương đồng 214 và cơ sở dữ liệu điểm số tương đồng lịch sử 212 có thể được thực hiện dưới dạng một môđun.

Theo một phương án, phép tính tần suất log có thể được chia thành hai bước. Thứ nhất, với mỗi cặp chỉnh sửa p , $\log$ (tần suất của p) được tính toán. Sau đó, với mỗi điểm số được lấy ra từ phép tính log (tần suất của p), điểm số này được chuẩn hóa về khoảng từ 0,0 đến 1,0, có nghĩa là điểm số lớn nhất của phép tính log (tần suất của p) được ánh xạ về 0,0, và điểm số nhỏ nhất của phép tính log (tần suất của p) được ánh xạ về 1,0. Theo cách này, điểm số tương đồng lịch sử cho mỗi cặp chỉnh sửa được xác định, ví dụ, theo công thức (2) sau:

Công thức (2)

$$\text{Điểm}(p) = 1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$$

Sẽ được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng điểm số tương đồng lịch sử của các cặp chỉnh sửa có thể được xác định theo cách khác nhờ sáng chế, ví dụ, mà không cần phép tính tần suất log.

Một phương án được thể hiện trong sơ đồ dạng bảng 600 trên Fig.6 để minh họa điểm số tương đồng lịch sử của các cặp chỉnh sửa mà được lưu trữ trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Theo một phương án được thể hiện trong bảng 600, mỗi trong số các cặp chỉnh sửa của đường thao tác ngắn nhất 1, đường thao tác ngắn nhất 2, và đường thao tác ngắn nhất 3 có thể có điểm số tương đồng lịch sử tương ứng. Ví dụ, mỗi trong số các cặp chỉnh sửa (“a”, “”), (“a”, “i”), (“a”, “z”), (“h”, “”), (“h”, “i”) và (“h”, “z”) có thể có điểm số tương đồng lịch sử bằng 0,6, 0,5, 0,4, 0,6, 0,4, và 0,8 tương ứng. Điểm số tương đồng lịch sử càng nhỏ, thì các thao tác được biểu thị bởi cặp chỉnh sửa xảy ra càng nhiều, và gần như và sẽ hợp lý hơn rằng thao tác được biểu thị bởi cặp chỉnh sửa có thể xảy ra lần nữa.

Nếu không thể tìm thấy điểm số tương đồng lịch sử cho cặp chỉnh sửa nhất định, thì có thể có nghĩa là thao tác được biểu thị bởi cặp chỉnh sửa nhất định chưa bao giờ xảy ra trước sự thực hiện lập tức của người dùng. Theo đó, thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể trả lại giá trị bằng 1,0 cho thiết bị 100 hoặc môđun

lượng hóa độ tương đồng của văn bản 206 làm điểm số tương đồng lịch sử cho cặp chỉnh sửa đó.

Dựa trên các ví dụ của bảng 600 và theo một phương án được thể hiện trong bảng phụ 504 trên Fig.5, điểm số tương đồng của đường thao tác ngắn nhất 1 có thể được xác định là $0,5 + 0,8 + 0,6$, tức là bằng 1,9. Điểm số tương đồng của đường thao tác ngắn nhất 2 có thể được xác định là $0,6 + 0,4 + 0,4$, tức là bằng 1,4. Tương tự, điểm số tương đồng của đường thao tác ngắn nhất 3 có thể được xác định là $0,5 + 0,6 + 0,4$, tức là bằng 1,5.

Bảng 600 có thể bao gồm điểm số tương đồng lịch sử của các cặp chỉnh sửa khác. Để cho đơn giản, các điểm số tương đồng lịch sử này không được thể hiện trên Fig.6.

Được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng điểm số tương đồng lịch sử của các cặp chỉnh sửa có thể được lưu trữ dưới các định dạng khác trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác bao gồm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Ở bước 306, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 chọn điểm số nhỏ nhất trong số các điểm số tương đồng được xác định cho các đường thao tác ngắn nhất. Điểm số tương đồng nhỏ nhất được chọn được sử dụng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Trong một số ví dụ, nếu điểm số nhỏ nhất trong số các điểm số tương đồng nhỏ hơn ngưỡng định trước, thì thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa. Ngưỡng định trước có thể được xác định bằng các yêu cầu nhất định đối với các ứng dụng khác nhau. Ví dụ, có thể đặt ngưỡng định trước bằng 1,5 để làm ví dụ theo các phương án, hoặc 0,5 có nghĩa là cần tổng số lượng thao tác phải đủ nhỏ, và các thao tác phải có tần suất cao.

Theo một phương án, đối với ba điểm số tương đồng của ba đường thao tác ngắn nhất được xác định như được thể hiện trong bảng phụ 504 trên Fig.5, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể chọn điểm số tương đồng nhỏ nhất là 1,4 trong số ba điểm số tương đồng 1,9, 1,4 và 1,5 làm mức lượng hóa của độ tương đồng của văn bản giữa chuỗi văn bản OCR 402, tức là “Azaha”, và chuỗi văn bản chỉnh sửa 404, tức là “Aziz”. Nếu ngưỡng định trước bằng 1,5, thì mức lượng hóa của độ

tương đồng của văn bản trong phương án này nhỏ hơn ngưỡng định trước, do đó thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể sửa chuỗi văn bản OCR “Azaha” thành chuỗi văn bản chỉnh sửa “Aziz”.

Theo phương án này, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể còn bổ sung các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất, tức là các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) trong đường thao tác ngắn nhất 2 mà có điểm số tương đồng nhỏ nhất là 1,4, vào cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Ngoài ra, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể cập nhật điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử. Việc cập nhật các điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử bao gồm một hoặc nhiều bước sau:

tính toán tần suất của các cặp chỉnh sửa trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất bằng cách: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0.

Ví dụ, theo một phương án, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể bổ sung các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) trong

đường thao tác ngắn nhất 2 mà có điểm số tương đồng nhỏ nhất là 1,4 vào cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206, và cập nhật điểm số tương đồng lịch sử cho các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) trong cơ sở dữ liệu cặp chỉnh sửa 210 của thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 bằng các bước sau: tính toán tần suất của các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) và xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) bằng cách thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa (“a”, “”), (“h”, “i”) và (“a”, “z”) và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0.

Trong một số ví dụ, nếu điểm số nhỏ nhất trong số các điểm số tương đồng lớn hơn ngưỡng định trước, thì thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể duy trì chuỗi văn bản OCR.

Ngoài ra hoặc theo cách khác, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 có thể xác định kết quả lượng hóa độ tương đồng của văn bản và kết xuất kết quả lượng hóa độ tương đồng của văn bản này đến môđun kết xuất 216. Kết quả lượng hóa độ tương đồng của văn bản có thể bao gồm các tin nhắn thông báo hệ thống mà chỉ báo “chấp thuận” (nếu điểm số nhỏ nhất trong số các điểm số tương đồng nhỏ hơn ngưỡng định trước) hoặc “từ chối” (nếu điểm số nhỏ nhất trong số các điểm số tương đồng lớn hơn ngưỡng định trước) đối với chuỗi văn bản chỉnh sửa hoặc cảnh báo hệ thống chỉ báo rủi ro tiềm tàng hoặc việc khóa tài khoản người dùng.

Fig.7 là lưu đồ minh họa phương pháp 700 để lượng hóa độ tương đồng của văn bản theo một phương án khác.

Ở bước 702, các cặp chỉnh sửa điểm số tương đồng được khởi tạo bằng thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 cho mỗi cặp có giá trị bằng 1,0.

Ở bước 704, các cặp chỉnh sửa được thu thập vào thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Ở bước 706, tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 được tính toán dựa trên dữ liệu thu thập được.

Ở bước 708, điểm số tương đồng lịch sử của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 được tính toán bằng các phép tính log (tần suất) và được chuẩn hóa về khoảng từ 0,0-1,0.

Ở bước 710, các cặp chỉnh sửa điểm số tương đồng được cập nhật làm điểm số tương đồng lịch sử.

Ở bước 712, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 nhận chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa khi người dùng thực hiện chỉnh sửa, và tìm tất cả đường thao tác ngắn nhất bằng phương pháp khoảng cách chỉnh sửa cho hai chuỗi văn bản.

Ở bước 714, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 tính toán điểm số tương đồng của các đường thao tác ngắn nhất bằng cách cộng điểm số tương đồng lịch sử của các cặp chỉnh sửa và chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Ở bước 716, thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 quyết định chấp thuận chỉnh sửa của người dùng bằng cách so sánh điểm số nhỏ nhất của điểm số tương đồng với ngưỡng định trước.

Ở bước 718, nếu điểm số nhỏ nhất của điểm số tương đồng nhỏ hơn ngưỡng định trước, thì chỉnh sửa của người dùng được chấp thuận bởi thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206. Thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 còn bổ sung các cặp chỉnh sửa của đường thao tác ngắn nhất có điểm số

tương đồng nhỏ nhất vào thư viện dữ liệu lịch sử 208, môđun lưu trữ 104, hoặc các môđun khác có trong thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206.

Fig.8 thể hiện sơ đồ khối của hệ thống máy tính phù hợp để sử dụng làm thiết bị 100 hoặc môđun lượng hóa độ tương đồng của văn bản 206 được thể hiện trên Fig.1 và Fig.2, hoặc hệ thống 200 như được thể hiện trên Fig.2 để thực hiện ít nhất một số bước của phương pháp lượng hóa độ tương đồng của văn bản theo các phương án được thể hiện trên các hình vẽ từ Fig.3 đến Fig.7.

Phần mô tả sau đây về hệ thống máy tính / thiết bị tính toán 800 được đưa ra chỉ để ví dụ và không nhằm để giới hạn.

Như được thể hiện trên Fig.8, thiết bị tính toán làm ví dụ 800 bao gồm bộ xử lý 804 để thực thi các phần mềm tiện ích. Mặc dù chỉ một bộ xử lý được thể hiện cho rõ ràng, nhưng thiết bị tính toán 800 cũng có thể bao gồm hệ thống có nhiều bộ xử lý. Bộ xử lý 804 được nối với hạ tầng truyền thông 806 để truyền thông với các bộ phận khác của thiết bị tính toán 800. Hạ tầng truyền thông 806 có thể bao gồm, ví dụ, bus truyền thông, mạng chuyển mạch, hoặc mạng.

Thiết bị tính toán 800 còn bao gồm bộ nhớ sơ cấp 808, chẳng hạn như bộ nhớ truy cập ngẫu nhiên (RAM - Random Access Memory), và bộ nhớ thứ cấp 810, Bộ nhớ thứ cấp 810 có thể bao gồm, ví dụ, ổ đĩa cứng 812 và/hoặc ổ đĩa lưu trữ tháo lắp được 814, mà có thể bao gồm ổ băng từ, ổ đĩa quang, hoặc dạng tương tự. Ổ đĩa lưu trữ tháo lắp được 814 đọc từ và/hoặc ghi vào khối lưu trữ tháo lắp được 818 theo cách đã biết. Khối lưu trữ tháo lắp được 818 có thể bao gồm băng từ, đĩa quang, hoặc dạng tương tự, được đọc bằng và ghi vào ổ đĩa lưu trữ tháo lắp được 814. Sẽ được hiểu bởi người có hiểu biết trung bình trong lĩnh vực kỹ thuật liên quan rằng, khối lưu trữ tháo lắp được 818 bao gồm vật ghi lưu trữ đọc được bằng máy tính lưu trữ mã lệnh chương trình thực thi được bằng máy tính và/hoặc dữ liệu.

Theo phương án thực hiện thay thế, bộ nhớ thứ cấp 810 có thể còn hoặc ngoài ra bao gồm các phương tiện tương tự khác để cho phép các chương trình hoặc lệnh máy tính được nạp vào thiết bị tính toán 800, Các phương tiện này có thể bao gồm, ví dụ, khối lưu trữ tháo lắp được 822 và giao diện 820, Các ví dụ về khối lưu trữ tháo lắp được 822 và

giao diện 820 bao gồm chip nhớ tháo lắp được (chẳng hạn như bộ nhớ chỉ đọc lập trình được và xóa được (EPROM - Erasable Programmable Read-Only Memory) hoặc bộ nhớ chỉ đọc lập trình được (PROM - Programmable Read-Only Memory)) và chân nối tích hợp, và các khối lưu trữ tháo lắp được khác 822 và các giao diện 820 cho phép phần mềm và dữ liệu được chuyển từ khối lưu trữ tháo lắp được 822 đến hệ thống máy tính 800,

Thiết bị tính toán 800 còn bao gồm ít nhất một giao diện truyền thông 824. Giao diện truyền thông 824 cho phép phần mềm và dữ liệu được truyền giữa thiết bị tính toán 800 và các thiết bị ngoại vi thông qua đường truyền thông 826. Theo các phương án khác nhau, giao diện truyền thông 824 cho phép dữ liệu được truyền giữa thiết bị tính toán 800 và mạng truyền thông dữ liệu, chẳng hạn như mạng truyền thông dữ liệu công cộng và dữ liệu riêng. Giao diện truyền thông 824 có thể được sử dụng để trao đổi dữ liệu giữa các thiết bị tính toán 800 khác nhau mà các thiết bị tính toán 800 này tạo thành một phần của mạng máy tính liên kết với nhau. Các ví dụ về giao diện truyền thông 824 có thể bao gồm môđem, giao diện mạng (chẳng hạn như tám mạch Ethernet), cổng truyền thông, ăng ten với hệ mạch liên kết và tương tự. Giao diện truyền thông 824 có thể là giao diện dùng dây hoặc có thể là giao diện không dây. Phần mềm và dữ liệu được chuyển đổi thông qua giao diện truyền thông 824 có dạng các tín hiệu mà có thể là các tín hiệu điện, điện từ, quang hoặc tín hiệu khác có khả năng nhận được bằng giao diện truyền thông 824. Các tín hiệu này được cung cấp cho giao diện truyền thông thông qua đường truyền thông 826.

Tùy ý là, thiết bị tính toán 800 còn bao gồm giao diện hiển thị 802 để thực hiện các thao tác để kết xuất các hình ảnh đến màn hình 830 và giao diện âm thanh 832 tích hợp để thực hiện các thao tác chơi nội dung âm thanh thông qua các loa liên kết 834.

Như được sử dụng ở đây, thuật ngữ "sản phẩm chương trình máy tính" có thể đề cập một phần đến khối lưu trữ tháo lắp được 818, khối lưu trữ tháo lắp được 822, ổ cứng được lắp đặt trong ổ đĩa cứng 812, hoặc sóng mang truyền tải phần mềm qua đường truyền thông 826 (kết nối không dây hoặc dùng cáp) đến giao diện truyền thông 824. Vật ghi lưu trữ đọc được bằng máy tính đề cập đến vật ghi lưu trữ hữu hình lâu dài bất kỳ để cung cấp các lệnh và/hoặc dữ liệu đã ghi cho thiết bị tính toán 800 để thực thi và/hoặc xử lý. Các ví dụ về vật ghi lưu trữ này bao gồm đĩa mềm, băng từ, đĩa CD-ROM, đĩa đa năng

số DVD, đĩa Blu-ray™, ổ đĩa cứng, ROM hoặc hệ mạch liên kết, bộ nhớ USB, đĩa quang từ, hoặc tấm mạch đọc được bằng máy tính chẳng hạn như tấm mạch của hiệp hội quốc tế thẻ nhớ máy tính cá nhân PCMCIA (Personal Computer Memory Card International Association card) và tương tự, bất kể các thiết bị này là thiết bị nội bộ hay ngoại vi của thiết bị tính toán 800. Các ví dụ về vật ghi truyền thông ngắn hạn hoặc không hữu hình đọc được bằng máy tính mà cũng có thể tham gia vào việc tạo ra phần mềm, các chương trình ứng dụng, các lệnh và/hoặc dữ liệu cho thiết bị tính toán 800 bao gồm các kênh truyền dẫn dùng sóng vô tuyến hoặc hồng ngoại cũng như kết nối mạng đến máy tính khác hoặc thiết bị được nối mạng khác, và Internet hoặc mạng cục bộ bao gồm việc truyền email và thông tin được ghi trên các trang web và dạng tương tự.

Các chương trình máy tính (còn được gọi là mã chương trình máy tính) được lưu trữ trong bộ nhớ sơ cấp 808 và/hoặc bộ nhớ thứ cấp 810, Các chương trình máy tính có thể còn nhận được thông qua giao diện truyền thông 824. Các chương trình máy tính này, khi được thực thi, cho phép thiết bị tính toán 800 thực hiện một hoặc nhiều dấu hiệu của các phương án được thảo luận trong bản mô tả. Theo các phương án khác nhau, các chương trình máy tính, khi được thực thi, cho phép bộ xử lý 804 thực hiện các dấu hiệu của các phương án nêu trên. Theo đó, các chương trình máy tính này đại diện cho các bộ điều khiển của hệ thống máy tính 800.

Phần mềm có thể được lưu trữ trong sản phẩm chương trình máy tính và được nạp vào thiết bị tính toán 800 bằng cách sử dụng ổ đĩa lưu trữ tháo lắp được 814, ổ đĩa cứng 812, hoặc giao diện 820. Theo cách khác, sản phẩm chương trình máy tính có thể được tải về hệ thống máy tính 800 qua đường truyền thông 826. Phần mềm, khi được thực thi bằng bộ xử lý 804, khiến thiết bị tính toán 800 thực hiện các chức năng của các phương án được mô tả trong bản mô tả.

Cần hiểu rằng phương án trên Fig.8 được thể hiện chỉ để làm ví dụ. Do đó, theo một số phương án, một hoặc nhiều chức năng của thiết bị tính toán 800 có thể được bỏ qua. Ngoài ra, theo một số phương án, một hoặc nhiều chức năng của thiết bị tính toán 800 có thể được kết hợp với nhau. Ngoài ra, theo một số phương án, một hoặc nhiều chức

năng của thiết bị tính toán 800 có thể được chia thành một hoặc nhiều phần của các bộ phận.

Theo một phương án, sáng chế đề xuất thiết bị lượng hóa độ tương đồng của văn bản. Thiết bị bao gồm bộ xử lý (chẳng hạn như bộ xử lý 404) và bộ nhớ (chẳng hạn như bộ nhớ thứ cấp 410) bao gồm mã chương trình máy tính. Bộ nhớ và mã chương trình máy tính được tạo cấu hình để, cùng với bộ xử lý, ít nhất làm cho thiết bị có thể thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chính sửa quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa; xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất; và chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

Các phương pháp được mô tả trong bản mô tả tạo ra một hoặc nhiều hiệu quả kỹ thuật khác nhau. Cụ thể, sáng chế không chỉ đánh giá số lượng nhỏ nhất của các thao tác cần thiết để biến đổi một chuỗi văn bản thành chuỗi văn bản khác, mà còn lượng hóa mối tương quan giữa hai chuỗi văn bản, tức là khả năng các thao tác này có thể xảy ra giữa hai chuỗi văn bản là như thế nào. Đặc biệt, sáng chế sử dụng tần suất của các cặp chỉnh sửa để xác định điểm số tương đồng lịch sử của các cặp chỉnh sửa và xem xét rằng các cặp chỉnh sửa được lưu trữ càng thường xuyên, thì số lần các thao tác được biểu thị bởi các cặp chỉnh sửa xảy ra càng nhiều, và gần như các thao tác được biểu thị bởi các cặp chỉnh sửa có thể xảy ra lần nữa. Theo đó, tần suất của các thay đổi lịch sử giữa các cặp của các ký tự được tính đến theo sáng chế để lượng hóa các thay đổi trong tương lai giữa các cặp của các ký tự giống nhau. Tức là, theo các phương án, tần suất của các cặp chỉnh sửa lịch sử được sử dụng tạo mô hình các độ tương đồng trực quan tương ứng của chúng, mà tạo ra độ chính xác cao của việc xác minh tự động việc chỉnh sửa thủ công các văn bản OCR.

Sẽ được hiểu bởi người có hiểu biết trong lĩnh vực kỹ thuật rằng các biến thể và/hoặc cải biến khác nhau có thể được thực hiện đối với sáng chế như được thể hiện trong các phương án cụ thể mà không tách rời khỏi bản chất và phạm vi của sáng chế như được mô tả theo nghĩa rộng. Do đó, các phương án trong bản mô tả được xem xét theo tất cả các khía cạnh là chỉ để minh họa và không nhằm giới hạn.

YÊU CẦU BẢO HỘ

1. Thiết bị lượng hóa độ tương đồng của văn bản, trong đó thiết bị này bao gồm:

bộ xử lý; và

bộ nhớ bao gồm mã chương trình máy tính;

bộ nhớ và mã chương trình máy tính được tạo cấu hình để, cùng với bộ xử lý, khiến thiết bị:

thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử;

tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó khi xác định điểm số tương đồng lịch sử, thì thiết bị này được tạo cấu hình để:

thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0 theo công thức sau:

điểm (p) = $1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$,

trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p),

trong đó thiết bị này còn được tạo cấu hình để:

thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chỉnh sửa quang học (OCR - optical correction recognition) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa;

xác định các điểm số tương đồng, mỗi trong số các điểm số tương đồng này tương ứng với một trong số các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất, trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và

chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

2. Thiết bị theo điểm 1, trong đó thiết bị này còn được tạo cấu hình để:

bổ sung các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất vào thư viện dữ liệu lịch sử; và

cập nhật điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó khi cập nhật điểm số tương đồng lịch sử, thì thiết bị được tạo cấu hình để:

tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất bằng cách: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0.

3. Thiết bị theo điểm bất kỳ trong số các điểm trên, trong đó khi thu được các đường thao tác ngắn nhất, thì thiết bị này được tạo cấu hình để:

thực hiện phép tính khoảng cách chỉnh sửa để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa, trong đó thao tác có thể thực hiện được đối với ký tự của chuỗi văn

bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa là một trong số thao tác chèn, thao tác xóa, hoặc thao tác thay thế.

4. Thiết bị theo điểm bất kỳ trong số các điểm trên, trong đó thiết bị này còn được tạo cấu hình để:

nếu điểm số nhỏ nhất trong số các điểm số tương đồng nhỏ hơn ngưỡng định trước, thì sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa.

5. Thiết bị theo điểm bất kỳ trong số các điểm trên, trong đó thiết bị này còn được tạo cấu hình để:

nếu điểm số nhỏ nhất trong số các điểm số tương đồng lớn hơn ngưỡng định trước, thì duy trì chuỗi văn bản OCR.

6. Thiết bị theo điểm bất kỳ trong số các điểm trên, trong đó thiết bị này còn được tạo cấu hình để:

quét hình ảnh kỹ thuật số để nhận chuỗi văn bản OCR, và

thu được chuỗi văn bản chỉnh sửa.

7. Phương pháp được thực hiện bằng máy tính để lượng hóa độ tương đồng của văn bản, trong đó phương pháp này bao gồm các bước:

thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử;

tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó bước tính toán điểm số tương đồng lịch sử bao gồm các bước:

thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0 theo công thức sau:

điểm (p) = $1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$,

trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p),

trong đó phương pháp này còn bao gồm các bước:

thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chỉnh sửa quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa;

xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất, trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và

chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa.

8. Phương pháp theo điểm 7, trong đó phương pháp này còn bao gồm các bước:

bổ sung các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất vào thư viện dữ liệu lịch sử; và

cập nhật điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử, trong đó bước cập nhật điểm số tương đồng lịch sử bao gồm các bước:

tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất bằng cách: thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh

sửa trong thư viện dữ liệu lịch sử tương ứng với các cặp chỉnh sửa trong đường thao tác ngắn nhất có điểm số tương đồng nhỏ nhất và chuẩn hóa phép tính tần suất log về khoảng từ 0,0 đến 1,0.

9. Phương pháp theo điểm bất kỳ trong số các điểm từ 7 đến 8, trong đó bước thu được các đường thao tác ngắn nhất bao gồm bước thực hiện phép tính khoảng cách chỉnh sửa để sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa, và trong đó thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa là một trong số thao tác chèn, thao tác xóa, hoặc thao tác thay thế.

10. Phương pháp theo điểm bất kỳ trong số các điểm từ 7 đến 9, trong đó phương pháp này còn bao gồm bước:

nếu điểm số nhỏ nhất của điểm số tương đồng nhỏ hơn ngưỡng định trước, thì sửa chuỗi văn bản OCR thành chuỗi văn bản chỉnh sửa.

11. Phương pháp theo điểm bất kỳ trong số các điểm từ 7 đến 10, trong đó phương pháp này còn bao gồm bước:

nếu điểm số nhỏ nhất của điểm số tương đồng lớn hơn ngưỡng định trước, thì duy trì chuỗi văn bản OCR.

12. Phương pháp theo điểm bất kỳ trong số các điểm từ 7 đến 11, trong đó phương pháp này còn bao gồm các bước:

quét hình ảnh kỹ thuật số để thu được chuỗi văn bản OCR, và
thu được chuỗi văn bản chỉnh sửa.

13. Hệ thống lượng hóa độ tương đồng của văn bản, trong đó hệ thống này bao gồm:

thiết bị OCR được tạo cấu hình để nhận được chuỗi văn bản OCR từ hình ảnh kỹ thuật số;

thiết bị nhập liệu được tạo cấu hình để thu được chuỗi chỉnh sửa của người dùng;

thiết bị lượng hóa độ tương đồng của văn bản để lượng hóa độ tương đồng của văn bản, trong đó thiết bị lượng hóa độ tương đồng của văn bản này bao gồm thiết bị xác định

điểm số tương đồng lịch sử, trong đó thiết bị lượng hóa độ tương đồng của văn bản được tạo cấu hình để:

thu thập các cặp chỉnh sửa vào thư viện dữ liệu lịch sử;

tính toán tần suất của các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

xác định điểm số tương đồng lịch sử cho các cặp chỉnh sửa trong thư viện dữ liệu lịch sử bằng thiết bị xác định điểm số tương đồng lịch sử, trong đó thiết bị xác định điểm số tương đồng lịch sử này được tạo cấu hình để:

thực hiện phép tính tần suất log cho mỗi trong số các cặp chỉnh sửa trong thư viện dữ liệu lịch sử; và

chuẩn hóa phép tính tần suất log nằm trong khoảng từ 0,0 đến 1 theo công thức sau:

điểm (p) = $1,0 - (\log(p) - \text{điểm số nhỏ nhất}) / (\text{điểm số lớn nhất} - \text{điểm số nhỏ nhất})$,

trong đó p biểu diễn mỗi trong số các cặp chỉnh sửa, $\log(p)$ biểu diễn phép tính log (tần suất của p), điểm số lớn nhất biểu diễn điểm số lớn nhất của phép tính log (tần suất của p), và điểm số nhỏ nhất biểu diễn điểm số nhỏ nhất của phép tính log (tần suất của p),

trong đó thiết bị lượng hóa độ tương đồng của văn bản này còn được tạo cấu hình để:

thu được các đường thao tác ngắn nhất để sửa chuỗi văn bản nhận dạng chỉnh sửa quang học (OCR) thành chuỗi văn bản chỉnh sửa, trong đó mỗi trong số các đường thao tác ngắn nhất này bao gồm một hoặc nhiều cặp chỉnh sửa, mỗi trong số một hoặc nhiều cặp chỉnh sửa này biểu thị thao tác có thể thực hiện được đối với ký tự của chuỗi văn bản OCR trong quá trình chỉnh sửa bằng chuỗi văn bản chỉnh sửa;

xác định các điểm số tương đồng cho các đường thao tác ngắn nhất, trong đó mỗi trong số các điểm số tương đồng được xác định bằng cách cộng các điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa của mỗi trong số các đường thao tác ngắn nhất,

trong đó điểm số tương đồng lịch sử của một hoặc nhiều cặp chỉnh sửa được lấy ra từ thư viện dữ liệu lịch sử; và

chọn điểm số nhỏ nhất trong số các điểm số tương đồng để lượng hóa độ tương đồng của văn bản giữa chuỗi văn bản OCR và chuỗi văn bản chỉnh sửa; và

thiết bị kết xuất được tạo cấu hình để kết xuất kết quả lượng hóa độ tương đồng của văn bản.

14. Hệ thống theo điểm 13, trong đó thiết bị lượng hóa độ tương đồng của văn bản được tạo cấu hình để thu thập các cặp chỉnh sửa để đáp lại việc thu được chuỗi văn bản OCR của các hình ảnh kỹ thuật số và thu được chuỗi văn bản chỉnh sửa tương ứng với chuỗi văn bản OCR này.

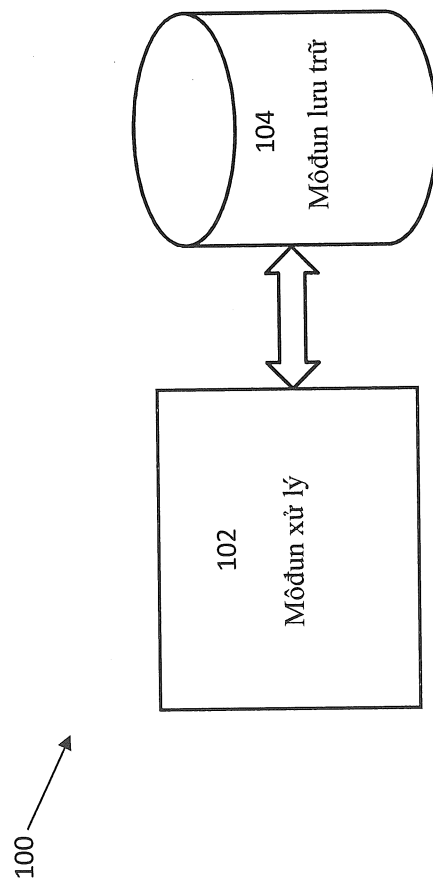


Fig.1

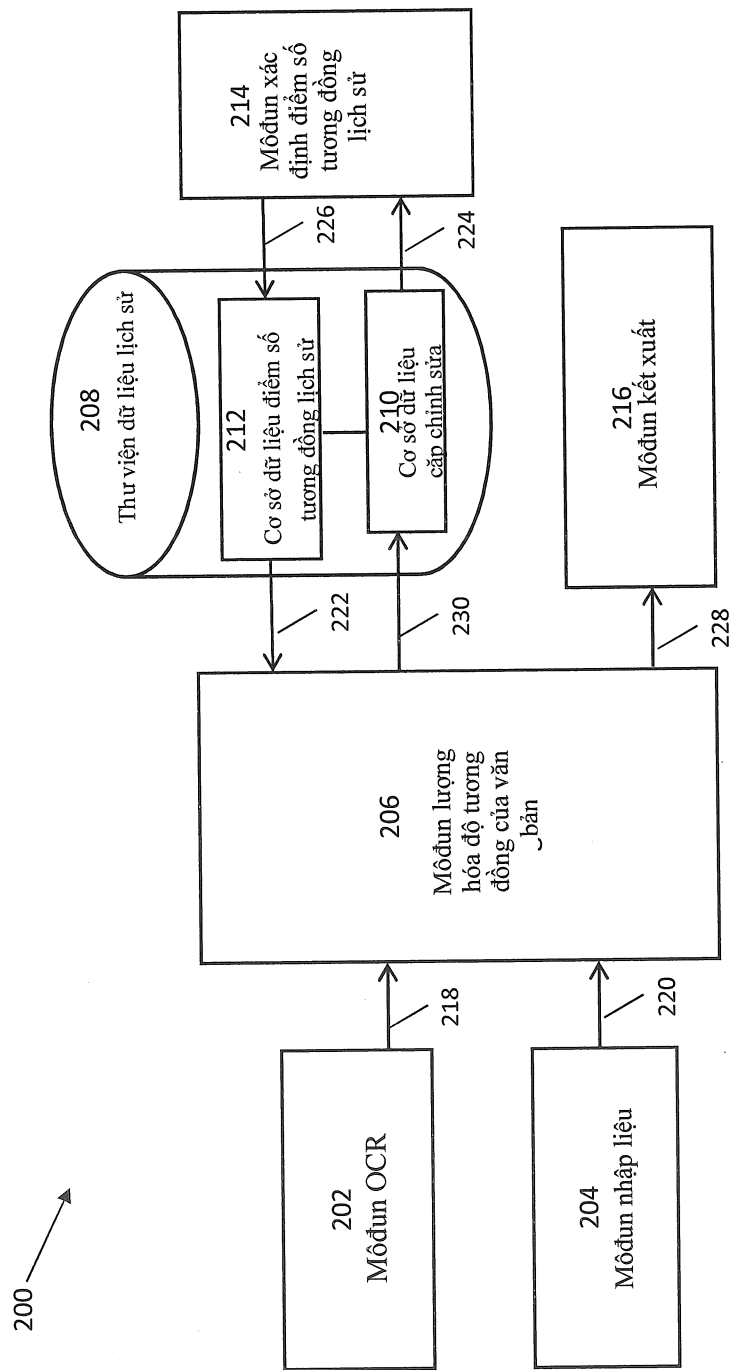


Fig.2

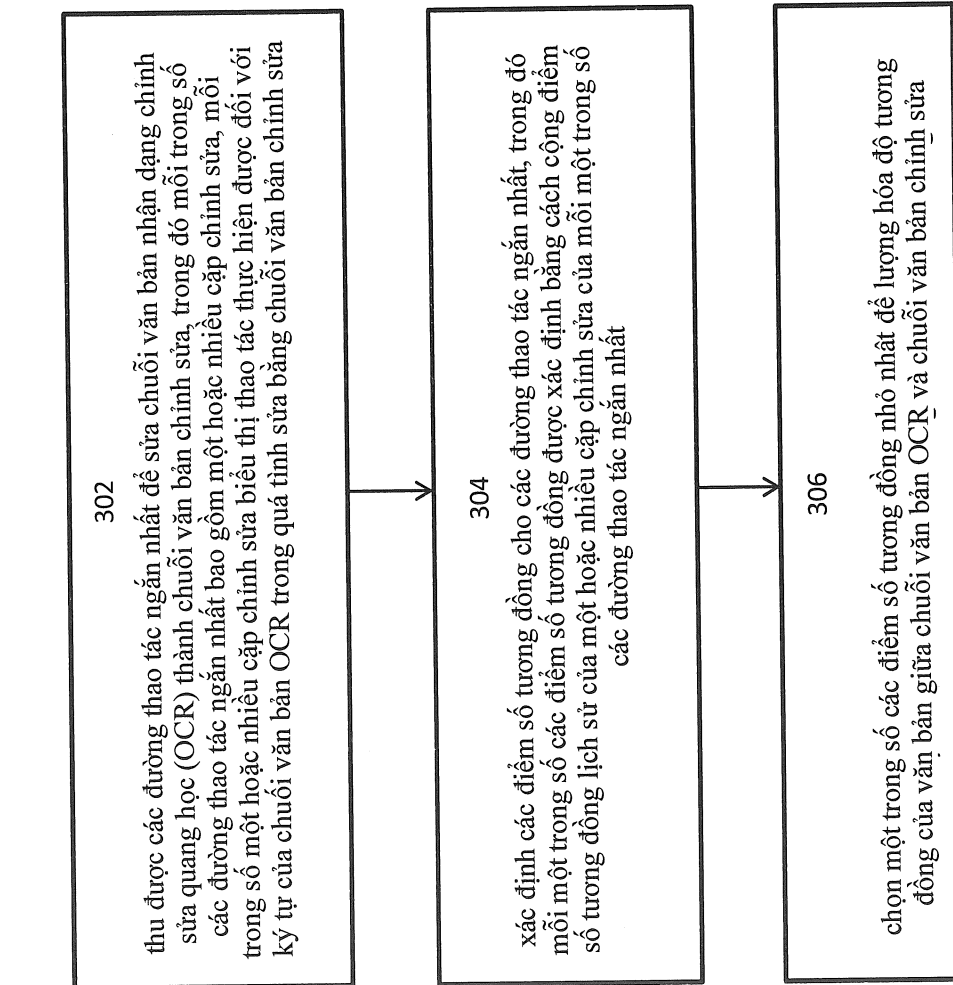


Fig.3

400

	#	A	z	a	h	a
#	0	1	2	3	4	5
A	1	0	1	2	3	4
z	2	1	0	1	2	3
i	3	2	1	1	2	3
z	4	3	2	2	2	3

402

404

406

Fig.4

500

"Azaha" --> "Aziz"	Thao tác 1	Thao tác 2	Thao tác 3	Điểm số tương đồng
Đường thao tác ngắn nhất 1	Thay thế ký tự "a" bằng ký tự "i", mà được biểu thị bằng cặp chỉnh sửa ("a", "i")	Thay thế ký tự "h" bằng ký tự "z", mà được biểu thị bằng cặp chỉnh sửa ("h", "z")	Xóa ký tự "a" mà được biểu thị bằng cặp chỉnh sửa ("a", " ")	Tổng bảng: Điểm số tương đồng lịch sử ("a", "i") + Điểm số tương đồng lịch sử ("h", "z") + Điểm số tương đồng lịch sử ("a", " "). Ví dụ, $0,5 + 0,8 + 0,6 = 1,9$
Đường thao tác ngắn nhất 2	Xóa ký tự "a" mà được biểu thị bằng cặp chỉnh sửa ("a", " ")	Thay thế ký tự "h" bằng ký tự "i", mà được biểu thị bằng cặp chỉnh sửa ("h", "i")	Thay thế ký tự "a" bằng ký tự "z", mà được biểu thị bằng cặp chỉnh sửa ("a", "z")	Tổng bảng: Điểm số tương đồng lịch sử ("a", " ") + Điểm số tương đồng lịch sử ("h", "i") + Điểm số tương đồng lịch sử ("a", "z"). Ví dụ, $0,6 + 0,4 + 0,4 = 1,4$
Đường thao tác ngắn nhất 3	Thay thế ký tự "a" bằng ký tự "i", mà được biểu thị bằng cặp chỉnh sửa ("a", "u")	Xóa ký tự "h" mà được biểu thị bằng cặp chỉnh sửa ("h", " ")	Thay thế ký tự "a" bằng ký tự "z", mà được biểu thị bằng cặp chỉnh sửa ("a", "z")	Tổng bảng: Điểm số tương đồng lịch sử ("a", "i") + Điểm số tương đồng lịch sử ("h", " ") + Điểm số tương đồng lịch sử ("a", "z"). Ví dụ, $0,5 + 0,6 + 0,4 = 1,5$

Fig.5

600 →

Cặp chỉnh sửa	Điểm số tương đồng lịch sử
("a", "u")	0,6
("a", "i")	0,5
("a", "z")	0,4
("h", "u")	0,6
("h", "i")	0,4
("h", "z")	0,8
...	...

Fig.6

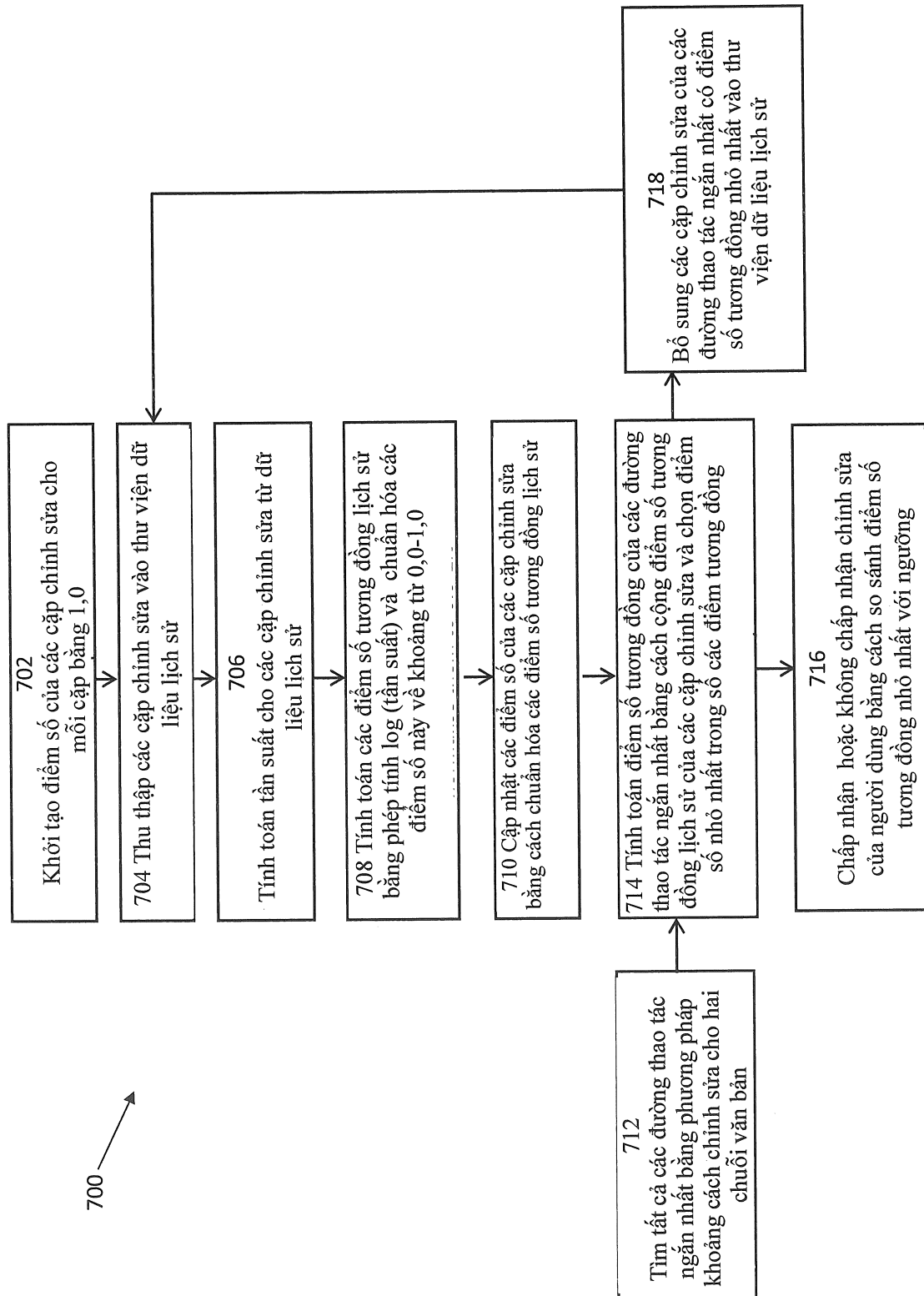


Fig.7

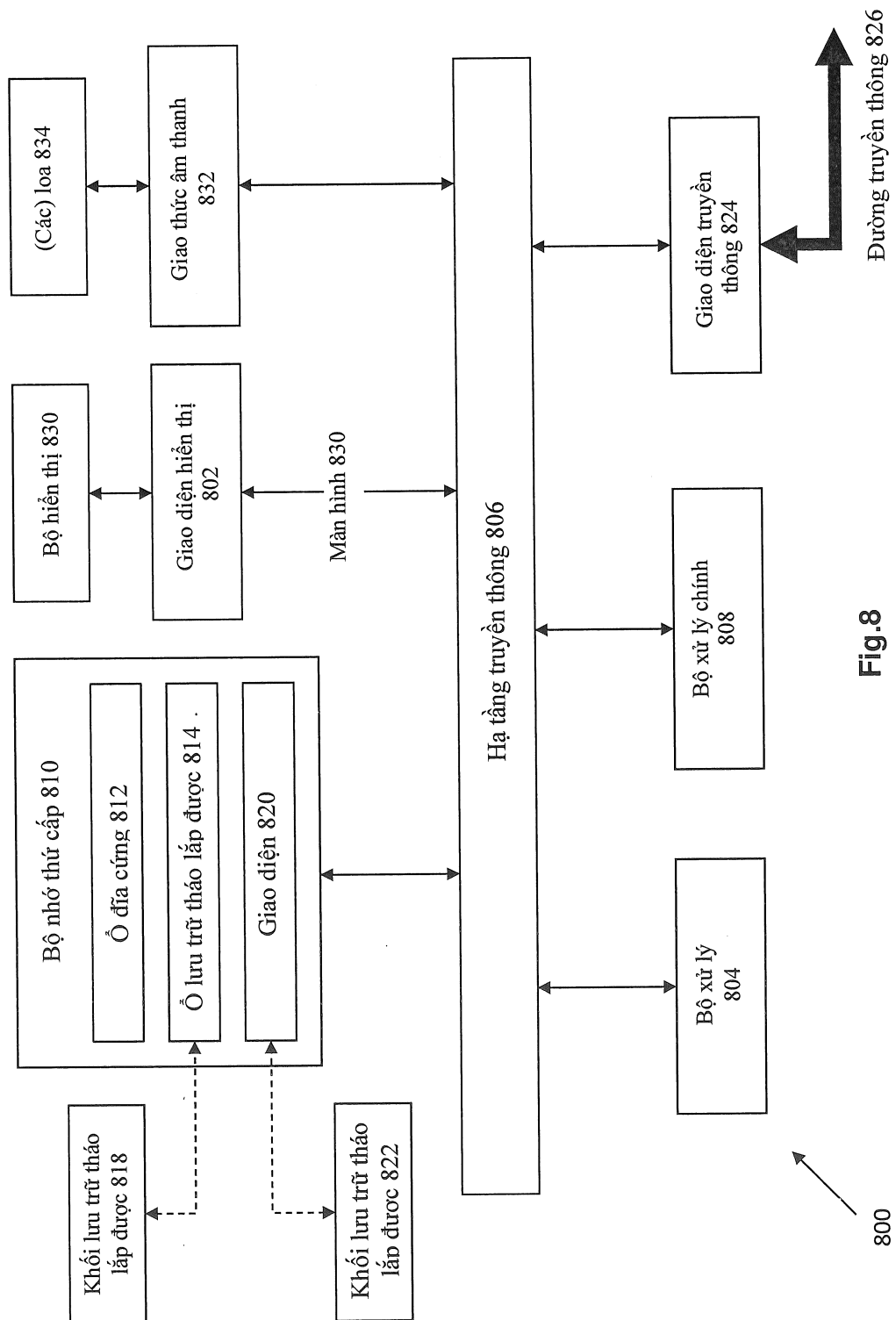


Fig.8