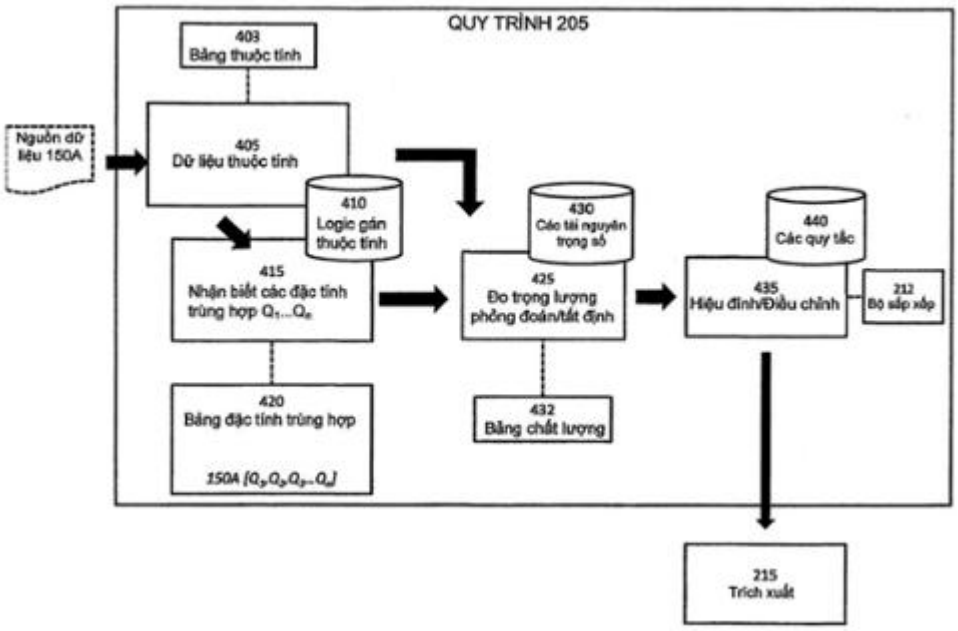
(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ
(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ
(11) **1-0029797**
(51)[7] **G06N 5/02**
(13) **B**

(21) 1-2017-01239
(22) 03/09/2015
(86) PCT/US2015/048322 03/09/2015
(87) WO/2016/036940 10/03/2016
(30) 62/045,398 03/09/2014 US
(45) 25/10/2021 403
(43) 26/06/2017 351A
(73) THE DUN & BRADSTREET CORPORATION (US)
103 JFK Parkway, Short Hills, NJ 07078, USA
(72) SCRIFFIGNANO, Anthony, J. (US); SUNBHANICH, Yiem (US); DAVIES, Robin, Fry (US); MATTHEWS, Warwick (AU).
(74) Công ty TNHH T&T INVENMARK Sở hữu trí tuệ Quốc tế (T&T INVENMARK CO., LTD.)

(54) PHƯƠNG PHÁP, HỆ THỐNG VÀ VẬT GHI DÙNG ĐỂ GÁN CÁC THUỘC TÍNH CHO NGUỒN DỮ LIỆU

(57) Sáng chế đề cập đến phương pháp bao gồm (a) tiếp nhận dữ liệu từ nguồn dữ liệu, (b) gán thuộc tính cho nguồn dữ liệu theo các quy tắc, do đó thu được thuộc tính, (c) phân tích dữ liệu để nhận dạng các đặc tính trùng hợp trong dữ liệu, (d) tính toán số đo định tính của thuộc tính, do đó thu được thuộc tính đã đo trọng lượng, (e) tính toán số đo định tính của đặc tính trùng hợp, do đó thu được đặc tính trùng hợp đã đo trọng lượng, (f) phân tích thuộc tính đã đo trọng lượng và đặc tính trùng hợp đã đo trọng lượng, để tạo ra bộ sắp xếp, (g) lọc dữ liệu theo bộ sắp xếp, do đó thu được dữ liệu đã trích xuất, và (h) chuyển dữ liệu đã trích xuất đến quy trình xuôi dòng. Sáng chế cũng đề cập đến hệ thống thi hành phương pháp này, và thiết bị lưu trữ chứa các lệnh để điều khiển bộ xử lý thực hiện phương pháp này.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến hệ thống áp dụng các quy trình gán thuộc tính và phân biệt mới, theo thực nghiệm, tức là khoa học và lặp lại được, cũng như đề cập đến các dung lượng, để tạo ra các thuộc tính mô tả và ngữ cảnh cho dữ liệu từ các nguồn được xử lý kém hoặc được cấu trúc kém, không được cấu trúc hoặc bán cấu trúc, và cụ thể là, các nguồn truyền thông xã hội. Tiếp theo, các thuộc tính này được dùng để xác định đặc điểm, hiệu đính, phân biệt và cuối cùng quyết định về việc sắp xếp hoặc xử lý dữ liệu một cách thích hợp nhất, bằng cách sử dụng các phương pháp vượt xa các quy trình và phương thức đệ quy-hoàn thành hiện có. Vấn đề cố hữu mà sáng chế giải quyết là hiện không thể hiệu đính, phân xử và nhập dữ liệu một cách chắc chắn có quy mô khi không có đủ bản thể học hoặc dạng chính tắc để cấu trúc quy trình nhập và xử lý dữ liệu.

Dung lượng được mô tả trong bản mô tả này có thể được sử dụng để xử lý dữ liệu thu được từ các tệp, tải trực tiếp từ các nguồn trực tuyến hoặc để đáp lại câu hỏi được khởi đầu bởi người dùng đầu cuối, hệ thống, ứng dụng, hoặc phương pháp khác bất kỳ cung cấp dữ liệu để nhập, xử lý và dùng cho mục đích nhất định. Trong trường hợp này, "xử lý và dùng cho mục đích nhất định" có thể là hệ thống hoặc hàm xuôi dòng khai thác dữ liệu, và sẽ có lợi từ dung lượng, nghĩa là, rút ra kết luận, hỗ trợ quan sát các mô hình, thực hiện tốt hơn, nhanh hơn, hiệu quả hơn hoặc theo cách hướng tới gia tăng giá trị của dữ liệu đó trong ngữ cảnh của hệ thống hoặc hàm đó.

Dung lượng có thể vận hành ở cấp độ ngữ cảnh, cấp độ tệp nguồn, hoặc cấp độ nội dung và có thể được thông báo bằng kinh nghiệm thu thập được của các bước lặp trước đó của chính quy trình. Việc gán thuộc tính "cấp độ ngữ cảnh" vận hành ở cấp độ các tình huống quanh việc thu thập và nhập nguồn dữ liệu. Việc gán thuộc tính "cấp độ tệp nguồn" thường, mà không duy nhất, vận hành ở cấp độ tệp dữ liệu khi được cấp bởi hoặc thu thập từ nguồn. Việc gán thuộc tính "cấp độ nội dung" vận hành ở cấp độ dữ liệu cơ sở và thường, mà không duy nhất, là trên cơ sở phân tích các phần tử dữ liệu riêng rẽ và/hoặc các mối quan hệ giữa chúng.

Ví dụ về việc gán thuộc tính "cấp độ ngữ cảnh" có thể là việc tạo ra siêu dữ liệu để mô tả tần suất mà tại đó dữ liệu từ nguồn cụ thể được phân phát và "thời hạn sử dụng" của dữ liệu trong nguồn đó, nghĩa là, dữ liệu sẽ thường được coi là "hiện tại" trong bao lâu. Ví dụ về việc gán thuộc tính "cấp độ tệp nguồn" có thể là kiểm tra siêu dữ liệu từ chính tệp đó, ví dụ, ngày tạo. Ví dụ về việc gán thuộc tính "cấp độ nội dung" có thể là việc phát hiện hệ thống viết được dùng để trình bày dữ liệu, ví dụ, Giản thể tự (Simplified Chinese).

Các ước tính công nghiệp chỉ ra rằng trên 80% dữ liệu mới tạo ra là không được cấu trúc. Để dẫn xuất giá trị đầy đủ từ dữ liệu mà sẽ ở các định dạng không được cấu trúc hoặc chỉ được hiểu một cách lỏng lẻo, hoặc ngược lại để tránh sự phát triển thành một khối của dữ liệu mà cuối cùng chứng tỏ là không chính xác, sai lệch hoặc có hại nếu được bổ sung vào kho ngữ liệu đã xử lý hiện có hoặc cấp liệu cho trường hợp sử dụng cụ thể, như hàm nghiệp vụ ra quyết định, điều quan trọng là có thể sàng lọc sơ bộ dữ liệu đó theo các tiêu chuẩn có ý nghĩa, mà không cần định trước, và/hoặc đo theo các chiều đã biết. Lợi ích của việc sàng lọc sơ bộ là ở chỗ dữ liệu mà không qua được các thử nghiệm nhất định hoặc không có điểm ở cấp độ chất lượng đủ cao sẽ bị loại bỏ, và nguy cơ tác dụng có hại giảm. Lợi ích bổ sung có thể là hỗ trợ hoặc thậm chí định hướng các nỗ lực xử lý khi tài nguyên hạn chế hoặc các lý do khác không cho phép nhập tất cả các nguồn dữ liệu mới khả dụng. Lưu ý rằng từ "chất lượng" được dùng trong bản mô tả này có nghĩa là số đo bất kỳ về sự phù hợp cho mục đích cụ thể và không cần chỉ ra giá trị vốn có cụ thể.

Có nhiều kỹ thuật để thực hiện các hàm xử lý nhập nhằng và phân biệt đối với dữ liệu không được cấu trúc bao gồm:

a) Trích xuất thực thể - dẫn xuất các thành phần riêng rẽ quan tâm từ văn bản, như danh từ, động từ, và từ bổ nghĩa.

b) Phân tích tình cảm - gán thuộc tính cho giọng nói và cảm xúc có chủ ý của nội dung.

c) Xử lý nhập nhằng ngữ nghĩa - làm giảm văn bản thành các cấu trúc dễ tính toán hơn (ví dụ, thẻ hóa).

d) Chuyển đổi ngôn ngữ - bao gồm chuyển tự, dịch, và thông dịch qua bộ xử lý

ngôn ngữ tự nhiên (natural language processing: NLP).

Nguy cơ nêu trên và nhu cầu giảm các ứng dụng một cách cụ thể khi chính dữ liệu là dữ liệu truyền thông xã hội, dữ liệu này luôn có thành phần chính không được cấu trúc, hoặc "văn bản tự do", có kích thước hạn chế, "có nguồn gốc đám đông", nghĩa là, có nguồn gốc từ tập hợp những người không xác định không liên kết, và có thể chứa một hoặc nhiều "đặc tính trùng hợp".

Một số ví dụ về các đặc tính trùng hợp này là:

a) Mỉa mai: Các từ hoặc vị ngữ được ghép theo cách sao cho truyền tải ẩn ý ngược với ý thông dịch nhanh.

Ví dụ: XYZ Oil Co. là một công ty tuyệt vời để hợp tác kinh doanh, nếu bạn thích phá hủy thiên nhiên (XYZ Oil Co. is an excellent company to do business with, if you like destroying nature).

b) Từ mới: Các từ và cụm từ vừa mới được cấu trúc và tập hợp để có nghĩa chung nhất định.

Ví dụ: - Thẻ băm (Hashtags)

c) Các biến đổi ngữ pháp hoặc văn bản được diễn đạt không chính xác: Cách dùng từ không đúng có chủ đích hoặc không có chủ đích, dẫn đến thông dịch nhập nhằng hoặc lộn xộn.

Ví dụ: FBI đang săn lùng những kẻ khủng bố bằng chất nổ (FBI is Hunting Terrorists With Explosives)

d) Dấu chấm câu: Cách dùng dấu chấm câu theo cách không chuẩn hoặc không nhất quán, hoặc thiếu dấu chấm câu, dẫn đến thông dịch nhập nhằng hoặc mâu thuẫn.

Ví dụ: "Ăn măng và lá (Eats shoots and leaves)" so với "Ăn, bắn và bỏ đi (Eats, shoots, and leaves)"

e) Dữ liệu đa ngôn ngữ: Chèn các từ và cụm từ từ một ngoại ngữ. Bao gồm các từ vay mượn, cụm từ vay mượn chính thức hoặc không chính thức, và dịch sao phỏng.

Ví dụ: Anh ta có một câu "Tôi không biết điều gì đó" ("I don't know what") khiến cho việc hiểu hoàn toàn ý nghĩa của anh ta trở nên khó khăn (He had a certain *je ne sais*

quoi that made it difficult to understand his meaning completely).

f) Cách viết chính tả: Cách viết chính tả hư cấu, không chính xác hoặc chấp nhận, dẫn đến việc thông dịch không nhất quán, không chính xác hoặc lộn xộn.

Ví dụ: Bạn có ở đó không (RU There?)

g) Che giấu/Mã hóa: Chuyển đổi có chủ đích dữ liệu thành suy luận hoặc thông dịch trùng hợp.

h) Ngữ cảnh: Gia tăng sự phụ thuộc vào tính liên tục từ bên ngoài hoặc ngữ cảnh phụ thuộc vào bên ngoài do thiếu ngữ cảnh được cung cấp trong chính dữ liệu.

Ví dụ: "Anh ấy đã có một lát tuyệt vời" [Bánh? Pizza? Pha bóng tennis?] ("He had an awesome slice!" [Cake?Pizza?Tennis shot?]

i) Đa phương tiện: Văn bản và các dạng phương tiện khác được kết hợp trong một thông điệp hoặc mảnh dữ liệu để tạo ra nghĩa nhập nhằng hoặc không biết được nếu không hiểu toàn cảnh.

Ví dụ: Hình ảnh kèm theo "Đây là những gì chúng tôi nghĩ về hương vị mới của XYZ Beverage Co.!" (Picture accompanied by "This is what we think of XYZ Beverage Co.'s new flavor!").

Tình trạng kỹ thuật của sáng chế

Các phương án được mô tả trong phần này là các phương án có thể được theo đuổi, song không nhất thiết là các phương án đã được nhận thức hoặc theo đuổi trước đó. Do đó, các phương án được mô tả trong phần này có thể không phải là tình trạng kỹ thuật cho yêu cầu bảo hộ trong đơn này và không được thừa nhận là tình trạng kỹ thuật bằng cách đưa vào trong phần này.

Các hệ thống hiện có có thể nỗ lực thực hiện các hàm nêu trên (trích xuất thực thể, phân tích tình cảm, xử lý nhập nhằng ngữ nghĩa, chuyển đổi ngôn ngữ, v.v.) và do đó đo và thử dữ liệu, song rất khó biết là sử dụng các thử nghiệm và chỉ số gì nếu không có tiên nghiệm với dữ liệu từ nguồn cụ thể. Do đó, để tạo ra mức độ hiệu quả và tái sản xuất đầy đủ trong việc phân biệt và ra quyết định, các hệ thống tìm cách nhập dữ liệu không được cấu trúc, truyền thông xã hội và dữ liệu tương tự khác có thể làm việc đó theo cách đệ

quy, trong đó hệ thống có thể được cấu trúc trên cơ sở tiên nghiệm. Các hệ thống này có thể cũng thi hành vòng khép kín, còn gọi là "thông tin phản hồi tới máy chủ", các tình huống, bằng cách sử dụng thông tin phản hồi chất lượng (ex post facto) để tác động tới kết quả tương lai. Tuy nhiên, các hệ thống này gặp phải các hạn chế về khả năng mở rộng và tự động do việc thi hành là lúc nào cũng thủ công, và thậm chí khi "học máy" được sử dụng, nó chỉ ở mức độ thực nghiệm cơ bản nhất, tức là trên cơ sở phân tích tần suất và ngữ nghĩa của chính dữ liệu chi tiết. Các hạn chế cũng tồn tại do tác động của đặc tính trùng hợp của ngôn ngữ được mô tả ở trên.

Bản chất kỹ thuật của sáng chế

Sáng chế đề xuất phương pháp bao gồm (a) tiếp nhận dữ liệu từ nguồn dữ liệu, (b) gán thuộc tính cho nguồn dữ liệu theo các quy tắc, do đó thu được thuộc tính, (c) phân tích dữ liệu để nhận dạng các đặc tính trùng hợp trong dữ liệu, (d) tính toán số đo định tính của thuộc tính, do đó thu được thuộc tính đã đo trọng lượng, (e) tính toán số đo định tính của đặc tính trùng hợp, do đó thu được đặc tính trùng hợp đã đo trọng lượng, (f) phân tích thuộc tính đã đo trọng lượng và đặc tính trùng hợp đã đo trọng lượng, để tạo ra bộ sắp xếp, (g) lọc dữ liệu theo bộ sắp xếp, do đó thu được dữ liệu đã trích xuất, và (h) chuyển dữ liệu đã trích xuất đến quy trình xuôi dòng. Sáng chế cũng đề xuất hệ thống thi hành phương pháp này, và thiết bị lưu trữ chứa các lệnh để điều khiển bộ xử lý thực hiện phương pháp này.

Các kỹ thuật được mô tả trong bản mô tả này bao gồm dung lượng chưa được tình trạng kỹ thuật giải quyết. Cụ thể, các kỹ thuật được mô tả trong bản mô tả này cung cấp phương pháp sử dụng các chiều gán thuộc tính mới, đến lượt các chiều này lại cho phép các thi hành tự động mới ra quyết định nhập dữ liệu, cho phép cấu trúc các hệ thống nhanh hơn, mở rộng hơn, linh động hơn và nhất quán hơn so với khả năng bằng cách sử dụng các phương án trên cơ sở tình trạng kỹ thuật.

Mô tả vắn tắt các hình vẽ

Fig.1 thể hiện sơ đồ khối dạng hàm của hệ thống nhập, gán thuộc tính, tạo các chiến lược sắp xếp và xuất nguồn dữ liệu qua phép gán thuộc tính thực nghiệm.

Fig.2 thể hiện sơ đồ khối dạng hàm của phương pháp được thực hiện bởi hệ thống theo Fig.1.

Fig.3 là đại diện đồ họa của các cấp độ gán thuộc tính nguồn và mối quan hệ phân cấp của chúng.

Fig.4 thể hiện sơ đồ khối dạng hàm của quy trình là một phần của phương pháp được thể hiện trên Fig.2.

Thành phần hoặc đặc điểm chung đối với nhiều hơn một hình vẽ sẽ được chỉ ra bằng cùng số chỉ dẫn trong mỗi hình vẽ.

Mô tả chi tiết sáng chế

Cần cải thiện các quy trình hiện có tìm cách phân tích và định tính nguồn dữ liệu trước khi nhập. Để đáp ứng yêu cầu này, sáng chế đề xuất hệ thống thực hiện phương pháp bao gồm (a) gán các thuộc tính cho dữ liệu vào, từ một nguồn, ở nhiều cấp độ, (b) tạo bộ quy tắc sắp xếp để trích xuất tập con định tính của dữ liệu, nếu có, từ nguồn, trên cơ sở các tiêu chuẩn đo các thuộc tính đã gán theo nhiều chiều, do đó thu được dữ liệu đã định tính, (c) nhập dữ liệu đã định tính, và (d) thu nhận thông tin phản hồi và tác động làm thay đổi trong hệ thống trên cơ sở thông tin phản hồi này.

Do đó, sáng chế bộc lộ hệ thống và phương pháp tự động gán các thuộc tính cho dữ liệu nguồn, ra các quyết định trên cơ sở, ví dụ, các thuộc tính, nhập dữ liệu, và thu nhận thông tin phản hồi trên cơ sở kinh nghiệm của hệ thống đối với việc nhập dữ liệu (kinh nghiệm này có thể được hệ thống ghi lại và lưu trữ dưới dạng các thuộc tính mới của chính quy trình). Phương pháp này được thực hiện mà không có sự can thiệp của con người, theo đó cho phép tạo ra sự nhất quán và khả năng mở rộng, và cho phép con người tập trung vào các tình huống trong đó việc nghiên cứu sâu hoặc bổ sung đòi hỏi phải tác động đến cương vị quản lý dữ liệu thích hợp. Thuật ngữ "khả năng mở rộng" có nghĩa là phương pháp này không được giới hạn ở công nghệ hoặc giải pháp kỹ thuật cụ thể.

Trong một số đoạn sau đây, có định nghĩa về một số thuật ngữ được sử dụng trong bản mô tả này.

Thuộc tính: Khi được sử dụng dưới dạng động từ, thuật ngữ này có nghĩa là tính

toán và tổ hợp siêu dữ liệu (tức là dữ liệu mô tả) hoặc dữ liệu khác (ví dụ, dữ liệu thực nghiệm) đối với dữ liệu hiện có. Dữ liệu đính kèm theo cách này là "các thuộc tính".

Khối ngữ liệu: Phần nội dung của một thứ, như tệp dữ liệu, để phân biệt với dữ liệu về thứ đó như ngày tạo ra nó. Khối ngữ liệu, trừ khi rõ ràng từ ngữ cảnh, chỉ toàn bộ thứ đó.

Xử lý: Phân loại, chuyển đổi, lưu trữ và quản lý một thứ, tức là dữ liệu theo sáng chế.

Nhập: Thu nhận và lưu trữ dữ liệu. Quy trình nhập thường bao gồm chuyển đổi hoặc tái cấu trúc thành định dạng đích hoặc phân loại.

Gán thuộc tính thực nghiệm: Gán các thuộc tính trên cơ sở phương pháp khoa học. Trong trường hợp của sáng chế, là các quy trình thuật toán và toán học.

Phương pháp luận:

1. Chọn số nguồn dữ liệu trên cơ sở các tiêu chuẩn đã định cần thiết lập bằng cách xem xét các yếu tố như:

a. Tính khả dụng của dữ liệu bao gồm chi phí và sử dụng cho phép;

b. Sự phong phú của nội dung, khả năng quan sát các ví dụ đầy đủ để đưa ra các kết luận thực nghiệm;

c. Mức độ trùng lắp với các nguồn đã có sẵn bao gồm trong nghiên cứu; và

d. Các thành kiến đã biết trong nguồn dữ liệu.

2. Cấu trúc thử nghiệm A/B/C hỗn hợp/tự động hoặc thủ công để đo:

a. Sự tồn tại;

b. Sự gán thuộc tính sắp xếp; và

c. Mức độ quan sát qua tổng thể ngoại suy.

3. Thi hành thử nghiệm và đánh giá kết quả bao gồm:

a. Thống kê mô tả đơn giản; và

b. Trực quan hóa cơ sở.

4. Đo các thành kiến như sự lạc quan/bi quan của người đánh giá.

5. Đưa ra kết luận về mức độ mà mỗi giả thiết được quan sát và tác động đến toàn bộ đánh giá so với phần còn lại của tổng thể không biểu hiện các tiêu chuẩn giả thiết.

Đánh giá kết quả:

a. Đánh giá tác động của mỗi giả thiết đối với các mẫu được chọn.

b. Giả sử chúng ta chứng minh được sự thích đáng, phát triển hệ thống tính điểm để đánh giá các nguồn khác nhau theo các chiều giả thiết.

Có thể có các khía cạnh trùng hợp bổ sung phát sinh trong quá trình quan sát, như:

a. Sự dính líu của các ngôn ngữ khác;

b. Sự tác động của tính nhất quán của nói chuyện nhóm;

c. Phép ẩn dụ được chia sẻ trong nói chuyện nhóm (được du nhập bởi hoàn cảnh hoặc bởi kinh nghiệm được chia sẻ);

d. Các từ vay mượn từ ngôn ngữ này sang ngôn ngữ khác; và

e. Tính đa thức của người nói (ví dụ, người nói bản địa so với người nói không bản địa, dân bản địa thời đại số so với dân nhập cư thời đại số).

Nghiên cứu truyền thông xã hội là một phần của khảo sát rộng hơn trong dữ liệu không được cấu trúc. Toàn bộ nỗ lực này là một phần của việc phát triển tiếp dung lượng trong việc phát hiện, xử lý và tổng hợp dữ liệu thích hợp cho việc kinh doanh và cho những người trong ngữ cảnh kinh doanh.

Trọng tâm chủ yếu của sáng chế là về dung lượng mà đóng góp vào việc hiểu biết chung đối với toàn bộ nguy cơ và/hoặc toàn bộ cơ hội. Các nhu cầu kế tiếp liên quan đến sự tuân thủ luật định, sự độc lập và đạo đức, và phát hiện hành động phi pháp.

Fig.1 thể hiện sơ đồ khối của hệ thống 100, để nhập, gán thuộc tính, tạo các chiến lược sắp xếp và xuất nguồn dữ liệu qua phép gán thuộc tính thực nghiệm. Hệ thống 100 bao gồm máy tính 105 liên kết với mạng 135.

Mạng 135 là mạng truyền thông dữ liệu. Mạng 135 có thể là mạng cá nhân hoặc

mạng công cộng, và có thể bao gồm bất kỳ hoặc toàn bộ (a) mạng cá nhân, ví dụ, bao phủ một căn phòng, (b) mạng cục bộ, ví dụ, bao phủ một tòa nhà, (c) mạng trường học, ví dụ, bao phủ sân trường, (d) mạng đô thị, ví dụ, bao phủ một thành phố, (e) mạng diện rộng, ví dụ, bao phủ vùng liên kết qua các ranh giới đô thị, khu vực hoặc quốc gia, hoặc (f) Internet. Các thông tin được truyền qua mạng 135 nhờ các tín hiệu điện và tín hiệu quang học.

Máy tính 105 bao gồm bộ xử lý 110, và bộ nhớ 115 liên kết với bộ xử lý 110. Mặc dù máy tính 105 được trình bày ở đây dưới dạng thiết bị độc lập, nó không được giới hạn ở đó, mà thực tế có thể được liên kết với các thiết bị khác (không được thể hiện) trong hệ thống xử lý được bố trí.

Bộ xử lý 110 là thiết bị điện được cấu trúc bằng sơ đồ mạch logic để đáp lại và thi hành các lệnh.

Bộ nhớ 115 là phương tiện lưu trữ hữu hình đọc được bằng máy tính được mã hóa bởi chương trình máy tính. Trong trường hợp này, bộ nhớ 115 lưu trữ dữ liệu và các lệnh, tức là mã chương trình, đọc được và thi hành được bởi bộ xử lý 110 để điều khiển hoạt động của bộ xử lý 110. Bộ nhớ 115 có thể được thi hành trong bộ nhớ truy cập ngẫu nhiên (random access memory: RAM), ổ cứng, bộ nhớ chỉ đọc (read only memory: ROM), hoặc tổ hợp của chúng. Một trong số các thành phần của bộ nhớ 115 là môđun chương trình 120.

Modun chương trình 120 chứa các lệnh để điều khiển bộ xử lý 110 thi hành các quy trình được mô tả trong bản mô tả này. Trong bản mô tả này, mặc dù các tác giả sáng chế mô tả các hoạt động được thực hiện bởi máy tính 105, hoặc bằng phương pháp hoặc quy trình hoặc các quy trình phụ thuộc của nó, các hoạt động này thực tế được thực hiện bởi bộ xử lý 110.

Thuật ngữ "môđun" được dùng trong bản mô tả này để chỉ phép toán hàm mà có thể được mô tả dưới dạng thành phần độc lập hoặc dưới dạng cấu trúc tích hợp của nhiều thành phần phụ thuộc. Do đó, môđun chương trình 120 có thể được thi hành dưới dạng môđun duy nhất hoặc dưới dạng nhiều môđun vận hành phối hợp với nhau. Hơn nữa, mặc dù môđun chương trình 120 được mô tả trong bản mô tả này là được cài đặt trong bộ nhớ 115, và do đó được thi hành bằng phần mềm, nó cần được thi hành trong bất kỳ phần

cứng (ví dụ, sơ đồ mạch điện), phần sụn, phần mềm, hoặc tổ hợp của chúng.

Mặc dù môđun chương trình 120 được chỉ ra là được tải sẵn vào bộ nhớ 115, nó có thể được cấu trúc trên thiết bị lưu trữ 140 để sau đó tải vào bộ nhớ 115.Thiết bị lưu trữ 140 là phương tiện lưu trữ hữu hình đọc được bằng máy tính, lưu trữ môđun chương trình 120 trên đó. Ví dụ về thiết bị lưu trữ 140 bao gồm đĩa cứng, băng từ, bộ nhớ chỉ đọc, phương tiện lưu trữ quang học, ổ cứng hoặc hoặc bộ nhớ gồm nhiều ổ cứng song song, và ổ bus nối tiếp đa năng (universal serial bus: USB). Theo cách khác, thiết bị lưu trữ 140 có thể là bộ truy cập ngẫu nhiên, hoặc loại thiết bị lưu trữ điện tử khác, nằm trên hệ thống lưu trữ từ xa (không được thể hiện) và liên kết với máy tính 105 qua mạng 135.

Hệ thống 100 cũng bao gồm nguồn dữ liệu 150A và nguồn dữ liệu 150B, các nguồn này được gọi chung ở đây là nguồn dữ liệu 150, và liên kết truyền thông với mạng 135. Thực tế, nguồn dữ liệu 150 có thể bao gồm số lượng nguồn dữ liệu bất kỳ, tức là một hoặc nhiều nguồn dữ liệu. Nguồn dữ liệu 150 chứa dữ liệu không được cấu trúc, và có thể bao gồm truyền thông xã hội.

Hệ thống 100 cũng bao gồm thiết bị người dùng 130 vận hành bởi người dùng 101 và liên kết với máy tính 105 qua mạng 135. Thiết bị người dùng 130 bao gồm thiết bị đầu vào, như bàn phím hoặc hệ thống con nhận biết giọng nói, để cho phép người dùng 101 truyền các lựa chọn thông tin và lệnh đến bộ xử lý 110. Thiết bị người dùng 130 cũng bao gồm thiết bị đầu ra như màn hình hoặc máy in, hoặc thiết bị tổng hợp giọng nói. Bộ phận điều khiển con trỏ như chuột, bi xoay, hoặc màn hình cảm ứng, cho phép người dùng 101 thao tác con trỏ trên màn hình để truyền các lựa chọn thông tin và lệnh bổ sung đến bộ xử lý 110.

Bộ xử lý 110 xuất, đến thiết bị người dùng 130, kết quả 122 của thi hành bởi môđun chương trình 120. Theo cách khác, bộ xử lý 110 có thể định hướng đầu ra đến thiết bị lưu trữ 125, ví dụ, cơ sở dữ liệu hoặc bộ nhớ, hoặc thiết bị từ xa (không được thể hiện) qua mạng 135.

Luồng công việc mà trong đó hệ thống 100 có thể sử dụng liên quan đến việc tiếp nhận, phát hiện và xử lý nguồn dữ liệu không được cấu trúc, ví dụ, nguồn dữ liệu 150. Việc tiếp nhận, phát hiện và xử lý này có thể là phần thi hành phục vụ số lượng trường hợp sử dụng bất kỳ, bao gồm, nhưng không chỉ giới hạn ở, tạo ra các quan điểm về tình

cảm tập thể trong truyền thông xã hội, nắm bắt các chuyển dịch trong tình hình thị trường đối với các yêu cầu đặt ra, phát hiện sắc thái dẫn đến phát hiện sự giả mạo hình tướng hoặc hành động phi pháp khác, suy luận các tín hiệu xã hội báo trước sự kiện hoặc hành vi sắp diễn ra, hoặc đơn giản là đánh giá giá trị gia tăng của việc nhập nguồn không được cấu trúc mới vào quy trình tồn tại trước đó.

Fig.2 thể hiện sơ đồ khối dạng hàm của phương pháp 200 được thực hiện bởi hệ thống 100, và cụ thể hơn, bởi bộ xử lý 110 theo môđun chương trình 120. Phương pháp 200 là quy trình chung tiếp nhận dữ liệu, gán thuộc tính nguồn dữ liệu và dữ liệu của chúng ở nhiều cấp độ (nghĩa là, cấp độ ngữ cảnh, cấp độ nguồn và cấp độ nội dung nêu trên) và ra các quyết định để sắp xếp nguồn dữ liệu và dữ liệu, chuyển dữ liệu, ví dụ, các tập con cụ thể của nó, đến một hoặc nhiều hệ thống xuôi dòng, khởi đầu các hàm cung cấp thông tin phản hồi về bộ sắp xếp, và các hàm khởi đầu quá trình phát hiện và thu nhận các nguồn dữ liệu bổ sung. Phương pháp 200 truy nhập và xử lý dữ liệu từ một hoặc nhiều nguồn 150, song để dễ giải thích, dưới đây các tác giả sáng chế sẽ mô tả việc thi hành phương pháp 200 bằng cách sử dụng ví dụ về nguồn dữ liệu duy nhất, tức là nguồn dữ liệu 150A. Phương pháp 200 khởi đầu bằng quy trình 205.

Quy trình 205 truy cập, phân tích và gán thuộc tính nguồn dữ liệu 150A ở nhiều cấp độ, tức là các cấp độ "Ngữ cảnh", "Tệp nguồn" và "Nội dung", như nêu ở trên, và quyết định bộ sắp xếp thích hợp nhất đối với dữ liệu chứa trong nguồn dữ liệu 150A để thu được bộ sắp xếp 212.

Fig.3 là đại diện đồ họa của các cấp độ gán thuộc tính nguồn và mối quan hệ phân cấp của chúng.

Ở cấp độ gán thuộc tính nguồn bất kỳ, và cụ thể là ở cấp độ nội dung, việc gán thuộc tính có thể bao gồm các hàm xử lý nhập nhằng và phân biệt hoạt động trong các chiều được mô tả ở trên, tức là trích xuất thực thể, phân tích tình cảm, xử lý nhập nhằng ngữ nghĩa và chuyển đổi ngôn ngữ. Hơn thế nữa, bằng cách sử dụng các hàm xử lý nhập nhằng và phân biệt này, quy trình 205 sẽ nỗ lực giải quyết các thách thức gán thuộc tính do, ví dụ, đặc tính trùng hợp được mô tả ở trên gây ra, tức là mỉa mai, từ mới, v.v..

Fig.4 thể hiện sơ đồ khối dạng hàm của quy trình 205. Quy trình 205 khởi đầu bằng quy trình 405.

Quy trình 405 tiếp nhận dữ liệu từ nguồn dữ liệu 150A, và các thuộc tính nguồn dữ liệu 150A bằng cách sử dụng các quy tắc và thông tin tham chiếu lưu trữ trong logic thuộc tính 410, do đó tạo ra bảng thuộc tính 403. Các quy tắc và thông tin tham chiếu này là, ví dụ, bộ thuật toán quét dữ liệu để xác định dữ liệu này có phải là văn bản hoặc đa phương tiện hay không.Ví dụ, quy trình 405 phân tích nguồn dữ liệu 150A, và xác định rằng nó là nguồn dữ liệu bên thứ ba, ví dụ, được mua, và ngày tạo là 1-1-2015.

Bảng 1 là đại diện được lấy làm ví dụ của bảng thuộc tính 403, và bao gồm một số thuộc tính và các giá trị của chúng được lấy làm ví dụ.

Bảng 1

(Ví dụ về Bảng thuộc tính 403)

THUỘC TÍNH	GIÁ TRỊ
Kiểu tệp	Văn bản
Định giới	Có
Nguồn (Người tạo ra các đặc tính tệp)	Các tệp dữ liệu ACME
Định dạng	DFC001
Ngày tạo:	1-1-2015
Mã nhận dạng phát hiện web:	- không có mặt -
Mã hóa	UTF-8
Chữ viết được phát hiện	Điều khiển CO và chữ Latin cơ sở

"Kiểu tệp" là thuộc tính cấp độ nguồn và là xác định được tạo ra dưới dạng kết quả của quy trình quét siêu dữ liệu và nội dung của tệp dữ liệu để xác định đặc điểm kiểu dữ liệu của tệp. Các giá trị khác có thể là "ảnh", "video", "nhị phân", "không biết", v.v..

"Định giới" là nhãn có/không đại diện cho kết luận được tạo ra sau khi quét tệp để xác định dữ liệu có được chứa trong các hàng riêng biệt không.

"Nguồn", trong ví dụ, đại diện cho nhà cung cấp tệp; trong trường hợp này được đọc từ siêu dữ liệu "Người tạo ra ("Author" (hoặc "Đặc tính" ("properties")) của tệp dữ liệu.

"Ngày tạo" có thể cũng được đọc từ siêu dữ liệu của tệp.

"Mã nhận dạng phát hiện web" được trình bày dưới dạng ví dụ về thuộc tính

không được tìm thấy, là thẻ chi tiết (explicit) được chèn vào tệp bằng quy trình phát hiện được khởi động bởi hàm 210 (được mô tả ở dưới).

"Mã hóa" cũng được đọc từ siêu dữ liệu tệp và chỉ việc xác định đặc điểm của cách mà tệp được cấu trúc. Các giá trị khác có thể bao gồm "ASCII", "BIG5", "SHIFT-JIS", "EBCDIC", v.v..

"Chữ viết được phát hiện" được cung cấp trong ví dụ để thể hiện thuộc tính không được dẫn xuất từ siêu dữ liệu, mà từ việc quét kho ngữ liệu của chính dữ liệu, để hiểu loại Unicode gì có mặt trong tệp. Giá trị "điều khiển CO và chữ Latin cơ sở" thực tế là bộ dữ liệu Latin chuẩn.

Các kiểu và giá trị thuộc tính thể hiện trong Bảng 1 chỉ là ví dụ, và không nhất thiết đại diện cho các kiểu và giá trị thuộc tính mà hệ thống 100 sẽ đính kèm đối với tệp hoặc dữ liệu cụ thể. Hệ thống 100 có thể được cấu trúc để tạo ra siêu dữ liệu bất kỳ mà có thể là hữu ích.

Quy trình 415 phân tích kho ngữ liệu của nguồn dữ liệu 150A để tạo ra các thuộc tính qua nhiều chiều, bao gồm (nhưng không chỉ giới hạn ở):

Trích xuất thực thể

a) Xử lý nhập nhằng ngữ nghĩa

b) Phân tích tình cảm

c) Trích xuất ngôn ngữ

d) Siêu dữ liệu cơ sở

Quy trình 415 cũng gán thuộc tính và đo sự có mặt và thịnh hành của "đặc tính trùng hợp" trong nguồn dữ liệu 150A, và do đó tạo ra bảng đặc tính trùng hợp 420 liệt kê các đặc tính trùng hợp Ql, Q2, Q3 ... Qn. Một số ví dụ về đặc tính trùng hợp được nêu ở trên.

Bảng 2 là ví dụ về bảng đặc tính trùng hợp 420, và bao gồm một số ví dụ về chỉ số và các giá trị của chúng.

Bảng 2

(Ví dụ về Bảng đặc tính trùng hợp 420)

CHỈ SỐ	GIÁ TRỊ
Độ thịnh hành của từ mới	AX2
Biến ngữ pháp	0,56
Điểm số chấm câu	0
Tình cảm	-0,5
Đặc tính chính tả	THẤP
Điểm số che giấu	0
Tính đồng nhất truyền thông	1,0
Phương sai đoạn	0,01

Trong ví dụ trong Bảng 2, quy mô và phạm vi của các giá trị là độc lập. Một số có thể là con số, các giá trị khác có thể là mã mà cần có các phương pháp phi số học để tạo ra các điểm số khả dụng.

Lưu ý rằng các số đo đặc tính trùng hợp được nêu và minh họa trong bản mô tả này là hoàn toàn độc lập, và thứ hạng không phải là chặt chẽ, ở chỗ hệ thống sẽ có khả năng bổ sung các đặc tính trùng hợp mới nếu chúng được nhận dạng. Ví dụ, trong Bảng 2 ở trên không có đầu vào đối với "dữ liệu đa ngôn ngữ" do các số đo và tác động của các đặc tính trùng hợp này chưa được nhận dạng trong thi hành ví dụ của hệ thống này.

"Độ thịnh hành của từ mới" đại diện cho điểm số tính được từ việc quét đối tượng của nguồn dữ liệu 150A và tạo ra điểm số đo có bao nhiêu từ mới, tức là từ mới và/hoặc khác thường, có mặt trong kho ngữ liệu của nguồn dữ liệu 150A. Trong ví dụ này, "AX2" có thể đại diện cho sự có mặt áp đảo của các từ mới được biết rõ,"ZA9" có thể đại diện cho sự khan hiếm của từ mới, song lại thịnh hành trong tập hợp từ mới rất bất thường hoặc không được nhận biết.

"Biến ngữ pháp" là số đo về tính đồng nhất của kiểu ngữ pháp. Các thuật toán được sử dụng để thiết lập chỉ số này có thể là các phương pháp chuẩn công nghiệp như thuật toán Cocke-Younger-Kasami, hoặc các thuật toán và số đo theo yêu cầu, hoặc thuật toán kết hợp một số số đo. Các số đo con này bản thân chúng có thể được lưu trữ dưới dạng chỉ số trong bảng đặc tính trùng hợp 420, và tiếp theo kết hợp để tạo ra các lối

vào khác trong bảng đặc tính trùng hợp 420.

"Điểm số chấm câu" là số đo về sự có mặt của dấu chấm câu. Trong ví dụ này, dấu chấm câu được phát hiện ít hoặc không đáng kể, do đó giá trị đối với chỉ số này bằng 0.

"Tình cảm" thể hiện liệu "người nói" trong văn bản có truyền đạt tình cảm tích cực về đối tượng hay không (nghĩa là, tán thành, khuyến cáo, chấp thuận, v.v.), tình cảm tiêu cực (nghĩa là, chỉ trích hoặc không tán thành), hoặc tình cảm trung lập (không tích cực cũng không tiêu cực, hoặc có thể mập mờ), số âm chỉ tình cảm tiêu cực (chỉ trích), số 0 chỉ tình cảm trung lập, và số dương chỉ tình cảm tích cực (chấp thuận). Giá trị ví dụ đối với tình cảm ở đây là -0,5, chỉ thứ mà có thể được mô tả là "tình cảm tiêu cực vừa phải".

"Đặc tính chính tả" là số đo về sự thịnh hành của các lỗi chính tả mà không được nhận biết là từ mới. Giá trị "THẤP" ở đây chỉ tỷ lệ lỗi chính tả thấp. Lưu ý rằng "lỗi chính tả" được dùng trong bản mô tả này đơn giản để chỉ sai lệch so với từ vựng đã biết; điểm số "CAO" có thể chỉ, ví dụ, sự thịnh hành cao của các danh từ chính xác không được nhận biết, mà không phải là các lỗi đánh máy hoặc viết chính tả thực sự.

"Điểm số che giấu" là số đo về mức độ rõ ràng rằng các nỗ lực có chủ đích đã được tạo ra để che dấu ý nghĩa - mã hóa văn bản có thể là ví dụ đơn giản về điều này. Giá trị ở đây là bằng 0, chỉ ra không có che giấu nào được phát hiện.

"Tính đồng nhất truyền thông" chỉ ra dữ liệu rõ ràng là kiểu dữ liệu duy nhất (ví dụ, văn bản), hoặc phương tiện hỗn hợp (ví dụ, văn bản có nhúng ảnh hoặc siêu liên kết). Trong ví dụ này, điểm số là 1,0, chỉ ra rằng tệp này là kiểu phương tiện duy nhất. Thông tin này có thể được ghép bởi quy trình 435 (được mô tả ở dưới) với các thuộc tính được dẫn xuất bởi quy trình 405 và được thể hiện trong Bảng 1 để kết luận rằng tệp dữ liệu ví dụ là được bao gồm hoàn toàn bằng văn bản có cấu trúc cột.

"Phương sai đoạn" là điểm số từ 0 đến 1 mô tả tính nhất quán chung về kích thước của các khối riêng biệt của tệp. Trong Bảng 2, điểm số 0,01 chỉ ra rằng các đoạn là rất đồng nhất.Ví dụ này là về tệp rất có cấu trúc, nên đây là giá trị kỳ vọng do các đoạn sẽ đại diện cho các dòng trong tệp. Một tệp đủ thông điệp từ dịch vụ mạng xã hội trực tuyến mà cho phép người dùng gửi và đọc các thông điệp ngắn, ví dụ, 140 ký tự, có thể có điểm số trung bình, do các đoạn sẽ thay đổi song có xu hướng ở quanh 128 ký tự. Đối với dữ liệu

từ dịch vụ mạng xã hội mà cho phép gửi dữ liệu lớn hơn, các đoạn có thể được kỳ vọng có điểm số rất cao do có khả năng thay đổi đáng kể trong kiểu dữ liệu này.

Các chỉ số và giá trị được thể hiện trong Bảng 2 chỉ là ví dụ, và không nhất thiết đại diện cho các giá trị mà hệ thống 100 sẽ gắn kèm đối với tệp hoặc dữ liệu cụ thể.

Như nêu ở trên, quy trình 415 có thể liên quan đến đến nhiều số đo đối với mỗi chỉ số. Ví dụ, một số thuật toán có thể được sử dụng để đo giá trị của chỉ số "Biến ngữ pháp".Ví dụ, một hoặc nhiều số đo thực tế có thể là các chỉ số khác trong bảng đặc tính trùng hợp 420, các số đo khác có thể là hoặc được dẫn xuất bằng cách sử dụng các giá trị trong bảng thuộc tính 403.

Bảng 3, ở dưới, thể hiện ba ví dụ về các số đo thuật toán đối với tình cảm. Các số đo này có thể được kết hợp thành điểm số tình cảm chung trong Bảng 2 ở trên.

Bảng 3

(Danh mục ví dụ về các số đo thuật toán đối với đặc tính trùng hợp tình cảm)

CHỈ SỐ
Trung bình đơn giản của Tình cảm
Trung bình trọng lượng của Tình cảm
Độ lệch chuẩn của Tình cảm

Sau khi hoàn thành quy trình 405 và 415, quy trình 205 tiếp tục đến quy trình 425.

Quy trình 425 là quy trình đo trọng lượng phỏng đoán/tất định, nó tiếp nhận bảng thuộc tính 403 và bảng đặc tính trùng hợp 420, và tính toán các số đo định tính đối với các thuộc tính nêu trong bảng thuộc tính 403 và bảng đặc tính trùng hợp 420, do đó tạo ra bảng chất lượng 432. Các số đo định tính trong bảng chất lượng 432 được tạo ra liên quan đến các tài nguyên trọng số 430, và có thể là các điểm số, hệ số hoặc trọng số đo nguồn dữ liệu 150A qua nhiều chiều.

Bảng 4 là đại diện được lấy làm ví dụ của bảng chất lượng 432. Trong Bảng 4, "trọng lượng" là số đo định tính, và được thu từ các tài nguyên trọng số 430. Quy trình 425 gán trọng lượng cho chỉ số.

Bảng 4

(Ví dụ về Bảng chất lượng 432)

CHỈ SỐ	GIÁ TRỊ	TRỌNG LƯỢNG
Độ thịnh hành của từ mới	AX2	10
Biến ngữ pháp	0,56	50
Điểm số chấm câu	0	1
Tình cảm	-0,5	77
Đặc tính chính tả	THẤP	30
Điểm số che giấu	0	70
Tính đồng nhất truyền thông	1,0	60
Phương sai đoạn	0,01	44
Ngôn ngữ	1	80
Nguồn	SI	55
Tuổi	76	44

Bảng 4 là ví dụ đơn giản. Các phép đo định tính thực có thể liên quan đến các kết hợp khá phức tạp của các yếu tố.

Bảng 4A thể hiện ví dụ về việc sử dụng các yếu tố kết hợp.

Bảng 4A

CHỈ SỐ	GIÁ TRỊ	TRỌNG LƯỢNG
Nguồn	S1	10
Nguồn > Tuổi	S1: 25	76

Trong ví dụ trong Bảng 4A, chỉ số của Nguồn được dò tìm trong bảng khác (không được thể hiện) liệt kê các nguồn dữ liệu đã biết và các trọng lượng lần lượt được gán cho các nguồn dữ liệu này. Trọng lượng đối với nguồn này, được nhận biết dưới dạng nguồn "S1" và được gán bởi quy trình 425 trong trường hợp này, là 10. Tuy nhiên, quy trình 425 có thể tính toán trọng lượng có tính chất phức tạp hơn. Trọng lượng "Nguồn>Tuổi" (dự tính thể hiện rằng nó nằm trong họ trọng lượng "Nguồn") thể hiện rằng có trọng lượng khác vận hành đối với nguồn S1 và áp dụng hệ số cụ thể (tức là 25) trên cơ sở tuổi dữ liệu trong nguồn S1 (nghĩa là, tệp được tạo ra hoặc theo cách khác ngày cụ thể rõ ràng nếu có cách đây bao lâu) để thu được Trọng lượng 76.

Sau khi hoàn thành quy trình 425, quy trình 205 tiếp tục đến quy trình 435.

Quy trình 435 là quy trình hiệu đính/điều chỉnh, nó tiếp nhận bảng chất lượng 432, bảng đặc tính trùng hợp 420 và bảng thuộc tính 403, và sử dụng quy tắc 440 để phân xử việc sắp xếp thích hợp nguồn dữ liệu 150A, và do đó tạo ra bộ sắp xếp 212. Quy tắc 440 có thể có dạng ma trận, bảng tìm kiếm, thẻ ghi điểm, ôtômat hữu hạn không tất định, cây quyết định hoặc tổ hợp bất kỳ của các dạng này hoặc logic ra quyết định khác.

Bộ sắp xếp 212 có thể bao gồm các lệnh hoặc hoặc tư vấn để:

a) Đặt ra quy tắc mà các tệp tương tự với nguồn dữ liệu 150A được nhập vào.

b) Chia nhỏ các tệp từ nguồn dữ liệu 150A và chỉ nhập các phần đáp ứng các tiêu chuẩn nhất định.

c) Nhập toàn bộ tệp từ nguồn dữ liệu 150A, nhưng dán nhãn bằng nhãn chỉ báo mức chất lượng đặc trưng nguồn.

d) Đặt ra quy tắc mà các tệp từ nguồn dữ liệu 150A luôn bị loại bỏ.

e) Thử nhập các tệp từ nguồn dữ liệu 150A nhưng giữ chúng cho đến khi chứng thực thêm, và khởi động phát hiện web đích qua hàm 210.

Cũng lưu ý rằng ví dụ về bảng 432 thể hiện trong Bảng 4 là bảng tham chiếu hai chiều với các giá trị và trọng lượng, song đây chỉ là ví dụ minh họa. Quy trình 435 có thể, qua quy tắc 440, sử dụng các quy trình khác, như tìm kiếm bảng và ôtômat hữu hạn không tất định để thu được bộ sắp xếp 212.

Trở lại Fig.2, phương pháp 200, sau khi hoàn thành quy trình 205, phương pháp 200 tiếp tục đến quy trình 215.

Quy trình 215 tiếp nhận dữ liệu ở dạng nguồn dữ liệu 150A và bộ sắp xếp 212, và thi hành các quy trình để chia nhỏ và lọc dữ liệu nhận được, để thu được dữ liệu đã trích xuất 217. Trong trường hợp này, quy trình 215 sử dụng dữ liệu được tạo ra bởi quy trình 205, tức là bộ sắp xếp 212, để:

Định tính nguồn dữ liệu 150A;

a) Chia nội dung của nguồn dữ liệu 150A thành các tập con có ý nghĩa; và

b) Nhập dữ liệu từ nguồn dữ liệu 150A vào quy trình xuôi dòng 220, là người

tiêu dùng dữ liệu.

Quy trình 220 tiếp nhận dữ liệu đã trích xuất 217, và chuyển dữ liệu đã trích xuất 217 đến quy trình xuôi dòng (không được thể hiện).

Phương pháp 200 cũng thi hành hàm 225 để tạo ra thông tin phản hồi, ví dụ, sự chấp thuận của người dùng, thực nghiệm, ví dụ, thống kê và định lượng, và đưa thông tin phản hồi này trở lại quy trình 205, để cải thiện quy trình 205. Hàm 225 được thông báo bởi (tức là nhận dữ liệu đầu vào từ) bộ sắp xếp 212, bảng chất lượng 432, bảng đặc tính trùng hợp 420 và bảng thuộc tính 403. Hàm 225 được khởi động bằng quá trình xử lý bộ sắp xếp 212 bởi quy trình 215.

Phương pháp 200 cũng thi hành hàm 210 dưới dạng quy trình không đồng bộ và có khả năng liên tục. Hàm 210 khai thác nguồn dữ liệu 150 mới và hiện có, ví dụ, qua phát hiện web tự động, bằng cách sử dụng dữ liệu được tạo ra trong quy trình 205, tức là bộ sắp xếp 212, bảng chất lượng 432, bảng đặc tính trùng hợp 420 và bảng thuộc tính 403. Dữ liệu này có thể là đầu vào cho hàm 210 để khởi động, hướng dẫn hoặc ràng buộc các quy trình tự động phát hiện nguồn dữ liệu. Thông tin này có thể, ví dụ, ở dạng "nhận dạng khoảng trống" (nhận dạng các vùng mà dữ liệu trong kho ngữ liệu nhập vào cho đến nay đã được quan sát là, ví dụ, thiếu, có chất lượng thấp, hoặc mất giá trị do "quá tuổi"), hoặc "thế hệ tương tự" (hướng đích các lớp nguồn dữ liệu trên cơ sở nhận dạng các lớp nguồn dữ liệu giống hoặc tương tự và xác định tính hiệu quả, tính nhất quán hoặc tính chính xác của các lớp này).

Hàm 210 cấu trúc và thi hành các chương trình con, ứng dụng và hàm phát hiện dữ liệu ngoài. Hàm 210 cung cấp các đầu vào cho các quy trình phát hiện dữ liệu này để chúng đóng vai trò tăng cường dữ liệu tiếp nhận được trước đó bởi phương pháp 200. Ví dụ về các đầu vào như vậy là bộ định vị tài nguyên đồng nhất (Uniform Resource Locator: URL) của trang web mà từ đó dữ liệu mong muốn có thể thu được, và danh mục các thuật ngữ tìm kiếm trên cơ sở nội dung của nguồn dữ liệu 150A.

Hệ thống 100 cho phép khai thác tự động, cấu trúc được, lặp lại được và thích ứng được đối với nguồn dữ liệu mới, đặc biệt là dữ liệu không được cấu trúc. Do hệ thống 100 được tự động hoàn toàn theo thời gian chạy, nên nó có thể mở rộng, và do đó cho phép quản lý thu nhận dữ liệu với hiệu quả, tốc độ và độ nhất quán gia tăng đáng kể.

Để minh họa ví dụ về việc thi hành phương pháp 200, các tác giả sáng chế sẽ bắt đầu bằng tệp nguồn EX1, được thể hiện ở dưới trong Bảng 5.

Bảng 5

(Tệp nguồn EX1)

ID	Địa chỉ e-mail	Ngày và thời gian	Thông điệp
funkyDave	dave. smith@gmail. com	2015-01-02 22:23:45:660	Gah! Bạn không chỉ yêu thích nó khi họ để lại một nửa lớp trên khỏi chiếc bánh pizza của bạn?(Gah! Dnt just luv it when they leave half the toppings off ur pizza?)
2PacGlue	Fnky2rtm@yahoo7.com. u	2015-02-02 22:25:05:424	Sẽ thử hương vị Coke mới. KHÔNG PHẢI. (Gonna try the new Coke flavor. NOT.)

Bảng 6 thể hiện bảng thuộc tính 403 đối với tệp nguồn EX1.

Bảng 6

(Ví dụ về bảng thuộc tính 403 đối với tệp nguồn EX1)

THUỘC TÍNH	**GIÁ TRỊ**
Kiểu tệp	Văn bản
Định giới	Có
Nguồn	GNIP
Định dạng	GNIP01
Ngày tạo:	1-7-2015
Mã hóa	UTF-8

Bảng 7 thể hiện bảng đặc tính trùng hợp 420 đối với tệp nguồn EX1.

Bảng 7

(Ví dụ về bảng đặc tính trùng hợp 420 đối với tệp nguồn EX1)

CHỈ SỐ	GIÁ TRỊ
Độ thịnh hành của từ mới	AG7
Biến ngữ pháp	0,88
Điểm số chấm câu	55
Hạng mỉa mai/thật tình	-3
Tình cảm	-0,95
Đặc tính chính tả	CAO
Điểm số che giấu	0
Tính đồng nhất truyền thông	1,0

Trong tập hợp của bảng đặc tính trùng hợp 420 đến đoạn dữ liệu "Sẽ thử hương vị Coke mới. KHÔNG PHẢI." (Gonna try the new Coke flavor. NOT.), quy trình 415 sẽ thực hiện phân tích bao gồm phân tích ngữ nghĩa nội dung theo các dòng trình bày trong Bảng 8.

Bảng 8

(Ví dụ về phân tích được thực hiện để định vị bảng đặc tính trùng hợp 420)

Từ	Phân tích
Đi (Gonna)	Đây có thể là tuyên bố dự định về hành động tương lai. Từ mới "Đi" ("Gonna") có nghĩa đen là "Đi đến) ("Going to") song được sử dụng để chỉ "Tôi nghĩ tôi sẽ"(I think I will).
Thử (try)	
Mạo từ (the)	Đây là đối tượng của câu. Sản phẩm được nhận dạng: Không thể là "Coca-cola mới" ("New Coke"), rất có thể là hương vị không cụ thể mới nào đó? Kết luận: Coca-cola (Coke) có sản phẩm mới.
Mới (new)	
Coca-cola (Coke)	
Hương vị (flavor)	
KHÔNG (NOT)	Từ mới phủ định thông thường - được xác nhận bằng cách trình bày toàn chữ viết hoa. Rất có thể chỉ sự mỉa mai, cũng như đối lập theo nghĩa đen của tuyên bố trước đó.

Phân tích được trình bày trong Bảng 8 là giải cấu trúc "tiếng Anh giản dị" của phân tích thuật toán và thống kê được thực hiện bởi quy trình 415. Phân tích này có thể được sử dụng để định vị Độ thịnh hành của từ mới, do các từ "Đi" ("Gonna") và

"KHÔNG" ("NOT") là các từ mới theo cách chúng được sử dụng, song thực tế chính chúng không phải là các từ mới. Điều này cũng thể hiện tại sao điểm số đối với Độ thịnh hành của từ mới không chỉ là số đơn. Từ mới là cả về các từ mới và các cách dùng mới của các từ cũ. Điểm số chấm câu cũng sẽ bị ảnh hưởng bởi việc dùng dấu chấm câu trong ví dụ, tức là chấm câu và viết hoa được sử dụng một cách nhất quán. Hạng mỉa mai/thật tình là rất liên quan ở đây và bị ảnh hưởng nặng nề bởi việc sử dụng "KHÔNG" ("NOT") đối với phủ định tuyên bố trước đó và chỉ ra sự mỉa mai. Nói chung, dữ liệu này có tính thật tình rất thấp, mặc dù toàn bộ cấu trúc là "thật tình" ở chỗ được dự tính truyền đạt rõ ý định tiêu cực.

Lưu ý rằng cách phân tích trình bày trong Bảng 8 là "tốc ký" được tạo ra nhằm mục đích của ví dụ này. Quy trình 415 sẽ sử dụng nhiều hàm tinh vi để tách các cụm từ, thực hiện phân tích ngữ nghĩa, và bù đặc tính trùng hợp. Cũng lưu ý rằng quy trình 415 thực hiện phân tích và ghi lại các kết quả qua toàn bộ tệp hoặc nguồn dữ liệu.

Bảng 9 thể hiện kết quả Bảng chất lượng 432 với "tỷ lệ phần trăm điểm số" thu được đối với tệp nguồn EX1 thể hiện trong cột ngoài cùng bên phải, cho phép biểu diễn đơn giản việc thi hành quy trình 435 và quy tắc 440. Thực tế, các quy trình và thuật toán tính toán sẽ là cấu trúc được và nói chung phức tạp hơn nhiều so với ví dụ trong Bảng 9.

Bảng 9

(Ví dụ về bảng chất lượng 432 đối với tệp nguồn EX1)

CHỈ SỐ	GIÁ TRỊ	TRỌNG LƯỢNG	ĐIỂM SỐ
Độ thịnh hành của từ mới	AG7	10	34
Biến ngữ pháp	0,88	50	44
Điểm số chấm câu	55	1	55
Hạng mỉa mai/thật tình	-3	77	23
Tình cảm	-0,95	30	33
Đặc tính chính tả	CAO	70	80
Điểm số che giấu	0	60	0
Tính đồng nhất truyền thông	1,0	44	44
Ngôn ngữ	1	80	80
Nguồn	SI	55	23

Tuổi	76	44	50

Bảng 10 thể hiện cách thông dịch "tiếng Anh giản dị" của bộ sắp xếp 212.

Bảng 10

(Ví dụ về bộ sắp xếp 212 đối với tệp nguồn EX1)

1	Dùng làm bản ghi hạt giống gốc	SAI
2	Dùng làm bản ghi chứng thực	ĐÚNG
3	Khớp với cơ sở dữ liệu Business	SAI
4	Khớp với cơ sở dữ liệu Contact	ĐÚNG
5	Chỉ số trung thực ở cái nhìn đầu tiên	0,1
6	Dùng làm hạt giống cho việc phát hiện tự động	SAI
7	Dùng làm hạt giống cho việc điều chỉnh các quy tắc [Phương pháp 100]	ĐÚNG

Lưu ý rằng trong Bảng 10, đầu vào 6 chỉ ra rằng hàm 210 sẽ không được khởi động bởi dữ liệu này (hoặc dữ liệu từ nguồn này trong tương lai), và đầu vào 7 chỉ ra rằng hàm 225 sẽ được khởi động bởi dữ liệu được tạo ra trong phương pháp 100 khi nó xử lý tệp nguồn EX1.

Các kỹ thuật được mô tả trong bản mô tả này được lấy làm ví dụ, và không được hiểu là giới hạn cụ thể bất kỳ đối với sáng chế. Cần phải hiểu rằng các biến đổi, kết hợp và cải biến có thể được chuyên gia trong lĩnh vực này nghĩ ra. Ví dụ, các bước liên quan đến các quy trình được mô tả trong bản mô tả này có thể được thực hiện theo thứ tự bất kỳ, trừ khi có quy định khác hoặc được ra lệnh bởi chính các bước này. Sáng chế được dự tính bao hàm tất cả các cải biến và biến đổi sao cho vẫn nằm trong phạm vi của sáng chế.

Thuật ngữ "bao gồm" được hiểu là chỉ rõ sự có mặt của các đặc điểm, số nguyên, bước hoặc thành phần được nêu ra, nhưng không loại trừ sự có mặt của một hoặc nhiều đặc điểm, số nguyên, bước hoặc thành phần khác hoặc nhóm của chúng. Các mạo từ mang tính không xác định, và do đó, không loại trừ các phương án có nhiều mạo từ này.

YÊU CẦU BẢO HỘ

1. Phương pháp gán các thuộc tính cho nguồn dữ liệu bao gồm các bước sau:

(a) nhận dữ liệu từ nguồn dữ liệu thứ nhất;

(b) thực hiện quy trình phân tích nguồn bao gồm:

gán nguồn dữ liệu thứ nhất này ở mức ngữ cảnh, mức tệp nguồn và mức nội dung, theo các quy tắc, theo đó thu được thuộc tính của nguồn dữ liệu thứ nhất nêu trên;

phân tích dữ liệu nêu trên để xác định đặc trưng của dữ liệu nêu trên trùng hợp với nghĩa của dữ liệu nêu trên, theo đó thu được đặc trưng trùng hợp;

tính toán giá trị đo chất lượng của thuộc tính nêu trên của nguồn dữ liệu thứ nhất, theo đó thu được thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên;

tính toán giá trị đo chất lượng của đặc trưng trùng hợp, theo đó thu được đặc trưng trùng hợp đã thêm trọng số; và

phân tích thuộc tính đã thêm trọng số nêu trên của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số, để tạo ra cách bố trí mà bao gồm các lệnh bố trí;

(c) xử lý dữ liệu nêu trên theo các lệnh bố trí nêu trên, theo đó thu được dữ liệu đã trích xuất;

(d) truyền dữ liệu đã trích xuất nêu trên đến quy trình hạ nguồn;

(e) xác định, dựa trên thuộc tính đã thêm trọng số nêu trên của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số nêu trên, để sử dụng nguồn dữ liệu thứ nhất nêu trên làm gốc của quy trình phát hiện nguồn dữ liệu tự động;

(f) thực thi quy trình phát hiện nguồn dữ liệu tự động nêu trên sử dụng các thuật ngữ tìm kiếm dựa trên nội dung của nguồn dữ liệu thứ nhất nêu trên để phát hiện nguồn dữ liệu thứ hai; và

(g) thực hiện quy trình phân tích nguồn nêu trên trên dữ liệu từ nguồn dữ liệu thứ hai nêu trên.

2. Phương pháp theo điểm 1, trong đó phương pháp này còn bao gồm bước:

tạo ra phản hồi dựa trên cách bố trí nêu trên; và

cải thiện phương pháp nêu trên dựa trên phản hồi nêu trên.

3. Phương pháp theo điểm 1, trong đó việc phân tích nêu trên được tiến hành theo thứ nguyên được chọn từ nhóm bao gồm trích xuất thực thể, định hướng ngữ nghĩa, phân tích quan điểm, trích xuất ngôn ngữ, chuyển đổi ngôn ngữ, và lý lịch dữ liệu cơ sở.

4. Phương pháp theo điểm 1, trong đó đặc trưng trùng hợp nêu trên được chọn từ nhóm bao gồm ngôn ngữ mỉa mai, từ mới, thay đổi ngữ pháp, văn bản ngắt đoạn không thích hợp, chấm câu, dữ liệu đa ngôn ngữ, đánh vần, tối nghĩa, mã hóa, ngữ cảnh, và việc sử dụng kết hợp các phương tiện nêu trên.

5. Phương pháp theo điểm 1, trong đó cách bố trí nêu trên được chọn từ nhóm bao gồm:

(a) thiết lập quy tắc mà các tệp tương tự với nguồn dữ liệu thứ nhất nêu trên được lấy và lưu trữ vào,

(b) chia nhỏ các tệp từ nguồn dữ liệu thứ nhất nêu trên và lấy và lưu trữ chỉ các phần đáp ứng tiêu chuẩn nhất định,

(c) lấy và lưu trữ toàn bộ tệp từ nguồn dữ liệu thứ nhất nêu trên nhưng dữ liệu cờ có chỉ báo mức chất lượng đặc thù cho nguồn,

(d) thiết lập quy tắc mà các tệp từ nguồn dữ liệu thứ nhất nêu trên luôn bị loại bỏ, và

(e) lấy thử và lưu trữ tệp từ nguồn dữ liệu thứ nhất nêu trên, nhưng giữ chúng chờ chứng thực thêm.

6. Hệ thống gán thuộc tính cho nguồn dữ liệu bao gồm:

bộ xử lý; và

bộ nhớ chứa các lệnh có thể đọc được bởi bộ xử lý nêu trên để làm cho bộ xử lý nêu trên thực hiện:

(a) nhận dữ liệu từ nguồn dữ liệu thứ nhất;

(b) thực hiện quy trình phân tích nguồn trong đó các lệnh nêu trên làm cho bộ xử lý nêu trên thực hiện:

gán thuộc tính cho nguồn dữ liệu thứ nhất nêu trên ở mức ngữ cảnh, mức tệp nguồn, và mức nội dung, theo các quy tắc, theo đó thu được thuộc tính của nguồn dữ liệu

thứ nhất nêu trên;

phân tích dữ liệu nêu trên để xác định đặc trưng của dữ liệu nêu trên mà phù hợp với nghĩa của dữ liệu nêu trên, theo đó thu được đặc trưng trùng hợp;

tính toán giá trị đo chất lượng của thuộc tính nêu trên của nguồn dữ liệu thứ nhất nêu trên, theo đó thu được thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên;

tính toán giá trị đo chất lượng của đặc trưng trùng hợp nêu trên, theo đó thu được đặc trưng trùng hợp đã thêm trọng số; và

phân tích thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số nêu trên, tạo ra cách bố trí mà bao gồm các lệnh bố trí;

(c) xử lý dữ liệu nêu trên theo các lệnh bố trí nêu trên, theo đó thu được dữ liệu đã trích xuất;

(d) truyền dữ liệu đã trích xuất nêu trên đến quy trình hạ nguồn;

(e) xác định, dựa trên thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số nêu trên, để sử dụng nguồn dữ liệu thứ nhất nêu trên làm gốc của quy trình phát hiện nguồn dữ liệu tự động;

(f) thực thi quy trình phát hiện nguồn dữ liệu tự động nêu trên sử dụng các thuật ngữ tìm kiếm dựa trên nội dung của nguồn dữ liệu thứ nhất nêu trên để phát hiện nguồn dữ liệu thứ hai; và

(g) thực hiện quy trình phân tích nguồn nêu trên trên nguồn dữ liệu từ nguồn dữ liệu thứ hai nêu trên.

7. Hệ thống theo điểm 6, trong đó các lệnh nêu trên cũng làm cho bộ xử lý nêu trên thực hiện bước:

tạo ra phản hồi dựa trên cách bố trí nêu trên; và

cải thiện phương pháp nêu trên dựa trên phản hồi nêu trên.

8. Hệ thống theo điểm 6, trong đó các lệnh nêu trên làm cho bộ xử lý nêu trên phân tích dữ liệu nêu trên làm cho bộ xử lý nêu trên phân tích dữ liệu nêu trên theo thứ nguyên được chọn từ nhóm bao gồm trích xuất thực thể, định hướng ngữ nghĩa, phân tích quan

điểm, trích xuất ngôn ngữ, chuyển đổi ngôn ngữ, và lý lịch dữ liệu cơ sở.

9. Hệ thống theo điểm 6, trong đó đặc trưng trùng hợp nêu trên được chọn từ nhóm bao gồm ngôn ngữ mỉa mai, từ mới, thay đổi ngữ pháp, văn bản ngắt đoạn không thích hợp, chấm câu, dữ liệu đa ngôn ngữ, đánh vần, tối nghĩa, mã hóa, ngữ cảnh, và việc sử dụng kết hợp các phương tiện này.

10. Hệ thống theo điểm 6, trong đó cách bố trí nêu trên được chọn từ nhóm bao gồm:

(a) đặt quy tắc mà các tệp tương tự với nguồn dữ liệu thứ nhất nêu trên được lấy và lưu trữ vào,

(b) chia nhỏ các tệp từ nguồn dữ liệu thứ nhất nêu trên và lấy và lưu trữ chỉ các phần đáp ứng tiêu chuẩn nhất định,

(c) lấy và lưu trữ toàn bộ tệp từ nguồn dữ liệu thứ nhất nêu trên nhưng dữ liệu cờ có chỉ báo mức chất lượng đặc thù cho nguồn,

(d) thiết lập quy tắc mà các tệp từ nguồn dữ liệu thứ nhất nêu trên luôn bị loại bỏ, và

(e) lấy thử và lưu trữ tệp từ nguồn dữ liệu thứ nhất nêu trên, nhưng giữ chúng chờ chứng thực thêm.

11. Vật ghi đọc được bằng máy tính không chuyển tiếp để gán các thuộc tính cho nguồn dữ liệu, chứa các lệnh mà có thể đọc được bằng bộ xử lý để làm cho bộ xử lý nêu trên thực hiện các bước sau:

(a) nhận dữ liệu từ nguồn dữ liệu thứ nhất;

(b) thực hiện quy trình phân tích nguồn trong đó các lệnh nêu trên làm cho bộ xử lý nêu trên thực hiện các bước:

gán thuộc tính cho nguồn dữ liệu thứ nhất nêu trên ở mức ngữ cảnh, mức tệp nguồn, và mức nội dung, theo các quy tắc, theo đó thu được thuộc tính của nguồn dữ liệu thứ nhất nêu trên;

phân tích dữ liệu nêu trên để xác định đặc trưng của dữ liệu nêu trên mà trùng hợp với nghĩa của dữ liệu nêu trên, theo đó thu được đặc trưng trùng hợp;

tính toán giá trị đo chất lượng của thuộc tính nêu trên của nguồn dữ liệu thứ nhất

nêu trên, theo đó thu được thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên;

tính toán giá trị đo chất lượng của đặc trưng trùng hợp nêu trên, theo đó thu được đặc trưng trùng hợp đã thêm trọng số; và

phân tích thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số nêu trên, tạo ra cách bố trí mà bao gồm các lệnh bố trí;

(c) xử lý dữ liệu nêu trên theo các lệnh bố trí nêu trên, theo đó thu được dữ liệu đã trích xuất;

(d) truyền dữ liệu đã trích xuất nêu trên đến quy trình hạ nguồn;

(e) xác định, dựa trên thuộc tính đã thêm trọng số của nguồn dữ liệu thứ nhất nêu trên và đặc trưng trùng hợp đã thêm trọng số nêu trên, để sử dụng nguồn dữ liệu thứ nhất nêu trên làm gốc của quy trình phát hiện nguồn dữ liệu tự động;

(f) thực thi quy trình phát hiện nguồn dữ liệu tự động nêu trên sử dụng các thuật ngữ tìm kiếm dựa trên nội dung của nguồn dữ liệu thứ nhất nêu trên để phát hiện nguồn dữ liệu thứ hai; và

(g) thực hiện quy trình phân tích nguồn nêu trên trên nguồn dữ liệu từ nguồn dữ liệu thứ hai nêu trên.

12. Vật ghi đọc được bằng máy tính không chuyển tiếp theo điểm 11, trong đó các lệnh nêu trên cũng làm cho bộ xử lý thực hiện các bước:

tạo ra phản hồi dựa trên cách bố trí nêu trên; và

cải thiện phương pháp nêu trên dựa trên phản hồi nêu trên.

13. Vật ghi đọc được bằng máy tính không chuyển tiếp theo điểm 11, trong đó các lệnh nêu trên mà làm cho bộ xử lý nêu trên phân tích dữ liệu nêu trên làm cho bộ xử lý nêu trên phân tích dữ liệu nêu trên theo thứ nguyên được chọn từ nhóm bao gồm trích xuất thực thể, định hướng ngữ nghĩa, phân tích quan điểm, trích xuất ngôn ngữ, chuyển đổi ngôn ngữ, và lý lịch dữ liệu cơ sở.

14. Vật ghi đọc được bằng máy tính không chuyển tiếp theo điểm 11, trong đó đặc trưng trùng hợp nêu trên được chọn từ nhóm bao gồm ngôn ngữ mỉa mai, từ mới, thay đổi ngữ

pháp, văn bản ngắt đoạn không thích hợp, chấm câu, dữ liệu đa ngôn ngữ, đánh vần, tối nghĩa, mã hóa, ngữ cảnh, và việc sử dụng kết hợp các phương tiện này.

15. Vật ghi đọc được bằng máy tính không chuyển tiếp theo điểm 11, trong đó cách bố trí nêu trên được chọn từ nhóm bao gồm:

(a) đặt quy tắc mà các tệp tương tự với nguồn dữ liệu thứ nhất nêu trên được lấy và lưu trữ vào,

(b) chia nhỏ các tệp từ nguồn dữ liệu thứ nhất nêu trên và lấy và lưu trữ chỉ các phần đáp ứng tiêu chuẩn nhất định,

(c) lấy và lưu trữ toàn bộ tệp từ nguồn dữ liệu thứ nhất nêu trên nhưng dữ liệu cờ có chỉ báo mức chất lượng đặc thù cho nguồn,

(d) thiết lập quy tắc mà các tệp từ nguồn dữ liệu thứ nhất nêu trên luôn bị loại bỏ, và

(e) lấy thử và lưu trữ tệp từ nguồn dữ liệu thứ nhất nêu trên, nhưng giữ chúng chờ chứng thực thêm.

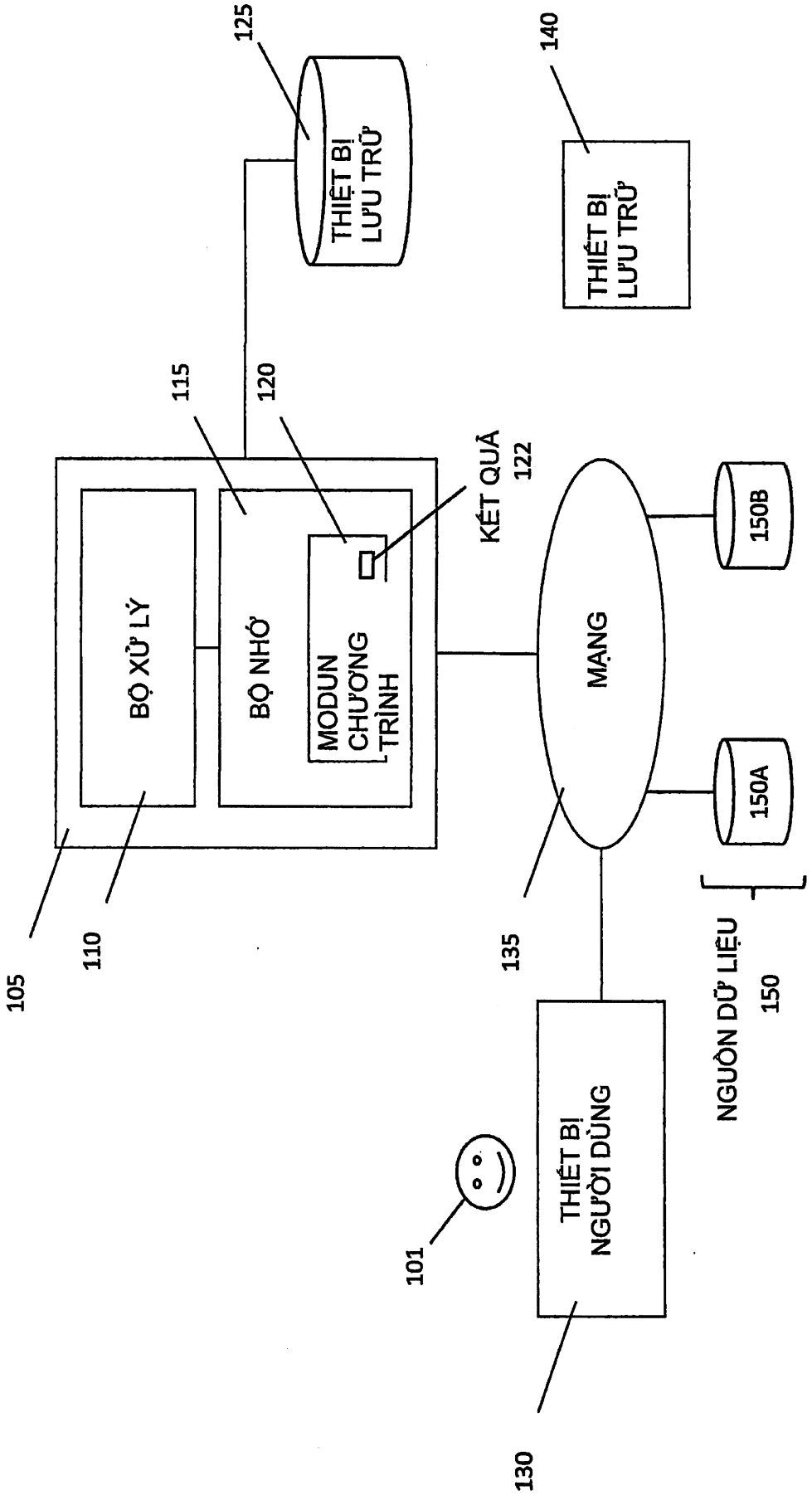

100
105
110
115
120
125
140
122
135
130
101
150
150A
150B
BỘ XỬ LÝ
BỘ NHỚ
MODUN CHƯƠNG TRÌNH
KẾT QUẢ
MẠNG
THIẾT BỊ LƯU TRỮ
THIẾT BỊ LƯU TRỮ
THIẾT BỊ NGƯỜI DÙNG
NGUỒN DỮ LIỆU

FIG. 1

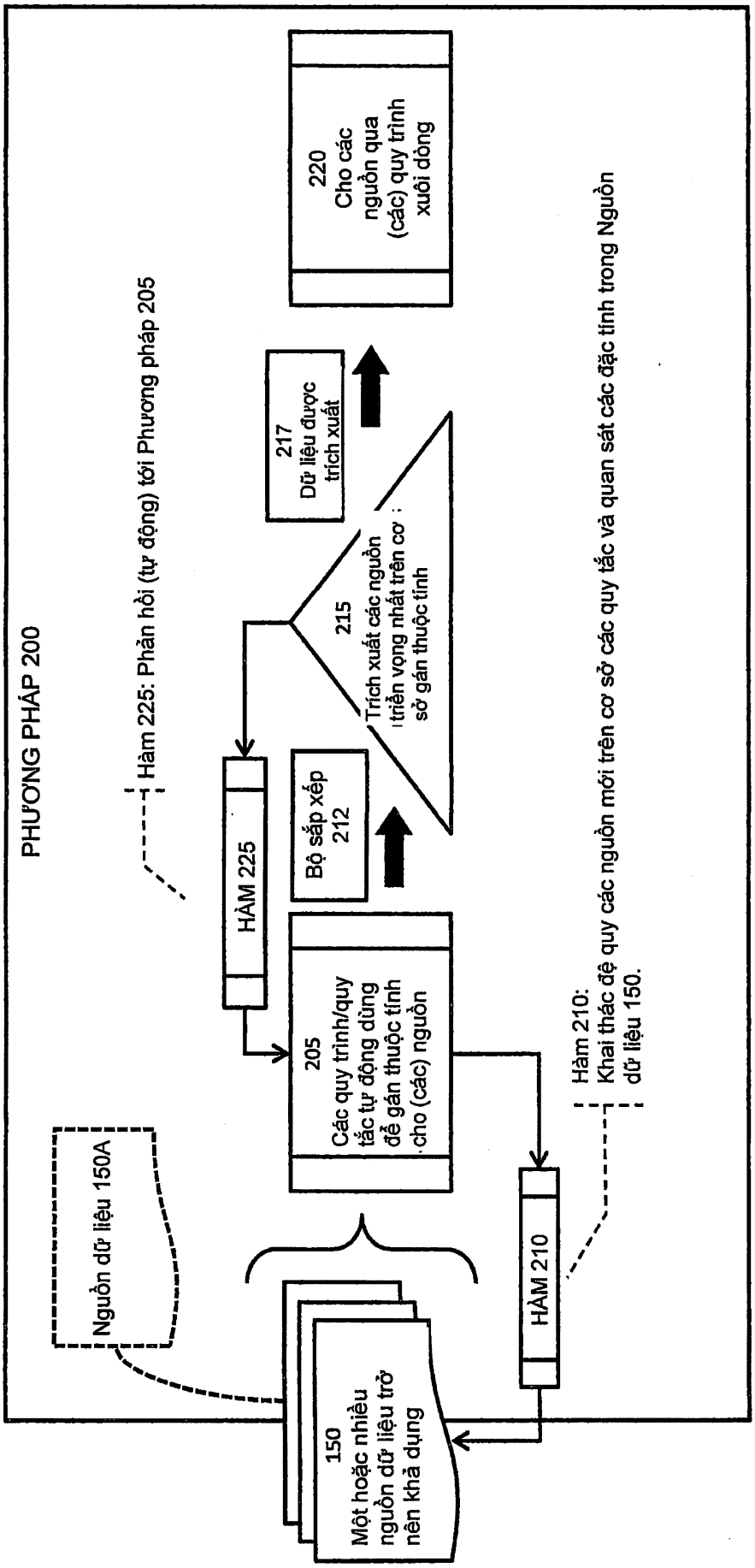

PHƯƠNG PHÁP 200
Hàm 225: Phản hồi (tự động) tới Phương pháp 205
Nguồn dữ liệu 150A
150
Một hoặc nhiều nguồn dữ liệu trở nên khả dụng
205
Các quy trình/quy tắc tự động dùng để gán thuộc tính cho (các) nguồn
HÀM 225
Bộ sắp xếp 212
215
Trích xuất các nguồn triển vọng nhất trên cơ sở gán thuộc tính
217
Dữ liệu được trích xuất
220
Cho các nguồn qua (các) quy trình xuôi dòng
HÀM 210
Hàm 210:
Khai thác đệ quy các nguồn mới trên cơ sở các quy tắc và quan sát các đặc tính trong Nguồn dữ liệu 150.

FIG. 2

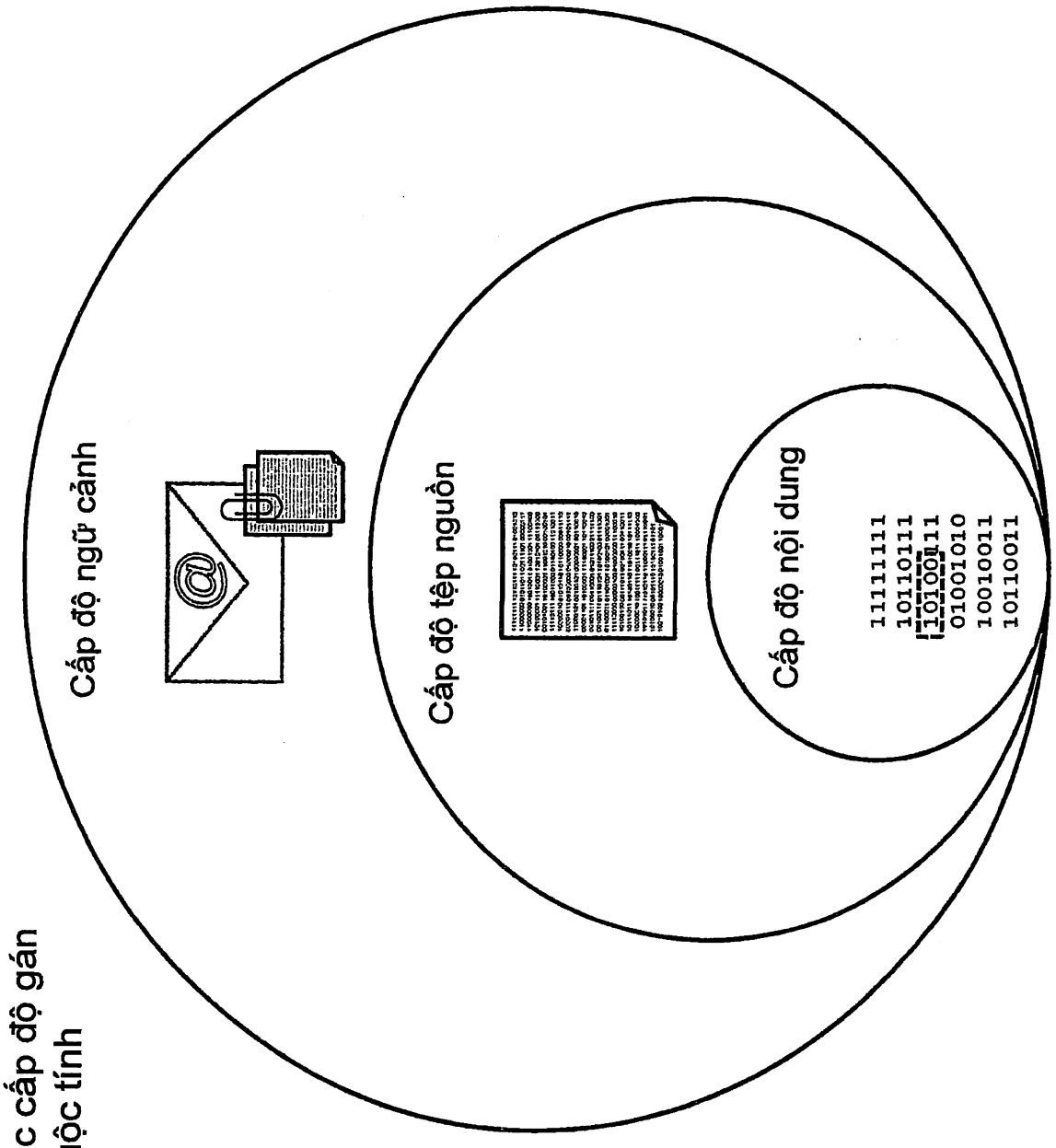

Các cấp độ gán thuộc tính
Cấp độ ngữ cảnh
Cấp độ tệp nguồn
Cấp độ nội dung
11111111
10110111
10100111
01001010
10010011
10110011

FIG. 3

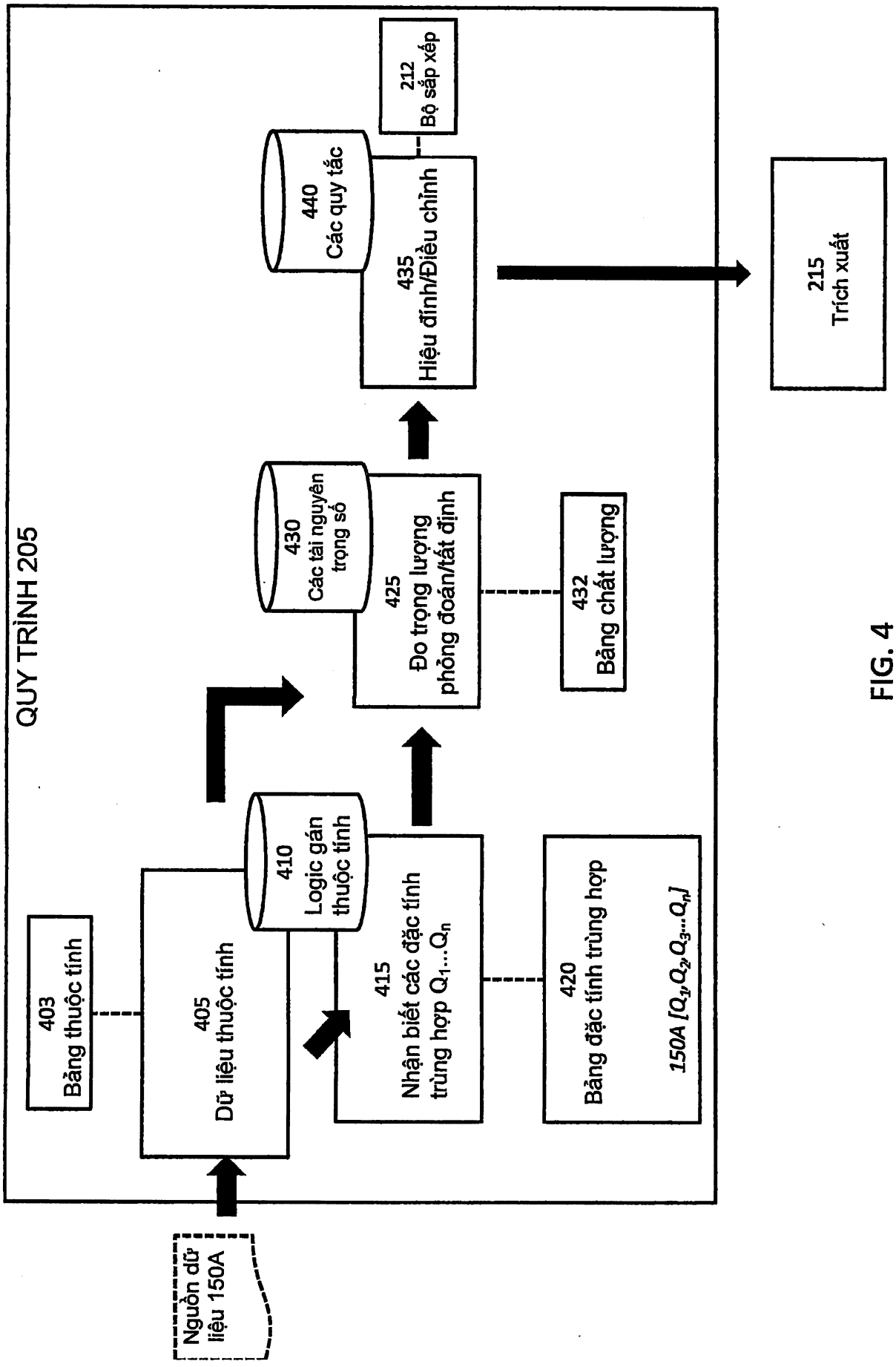

QUY TRÌNH 205
Nguồn dữ liệu 150A
403
Bảng thuộc tính
405
Dữ liệu thuộc tính
410
Logic gán thuộc tính
415
Nhận biết các đặc tính trùng hợp $Q_1...Q_n$
420
Bảng đặc tính trùng hợp
150A $[Q_1,Q_2,Q_3...Q_n]$
425
Đo trọng lượng phỏng đoán/tất định
430
Các tài nguyên trọng số
432
Bảng chất lượng
435
Hiệu đính/Điều chỉnh
440
Các quy tắc
212
Bộ sắp xếp
215
Trích xuất

FIG. 4