(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) **1-0021371**

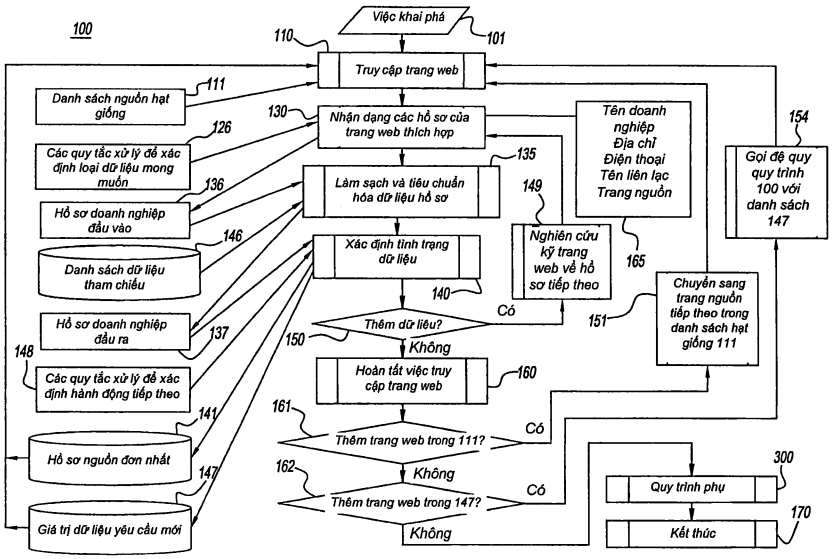
(51)[7] **G06F 17/30**

(13) **B**

(21) 1-2015-01214
(22) 07.10.2013
(86) PCT/US2013/063737 07.10.2013
(87) WO2014/058805A1 17.04.2014
(30) 61/711,673 09.10.2012 US
(45) 25.07.2019 376
(43) 27.07.2015 328
(73) The Dun & Bradstreet Corporation (US)
103 JFK Parkway Short Hills, New Jersey 07078 United States of America
(72) Anthony J. Scriffignano (US), Michael Klein (US), Hoang Q. Thang (US), Vindra Rampaul (US), Robin Davies (US), Anjali Reddi (US)
(74) Công ty TNHH Quốc tế D & N (D&N INTERNATIONAL CO.,LTD.)

(54) HỆ THỐNG VÀ PHƯƠNG PHÁP TÌM KIẾM DỮ LIỆU PHÙ HỢP VỚI YÊU CẦU VÀ VẬT GHI LƯU TRỮ BẤT BIẾN ĐỌC ĐƯỢC BỞI MÁY TÍNH

(57) Sáng chế đề cập đến hệ thống và phương pháp được sử dụng để phát hiện dữ liệu phù hợp với yêu cầu trong đó nhiều nguồn dữ liệu, có thể là các trang web hoặc các nguồn dữ liệu khác, được kiểm tra về dữ liệu phù hợp với yêu cầu. Quy trình và phương pháp này được thực hiện theo cách đệ quy số lượng không xác định các bước lặp, nhờ sử dụng dữ liệu và siêu dữ liệu từ nhiều nguồn để chứng thực dữ liệu phát hiện được và siêu dữ liệu từ các nguồn khác, cho đến khi không còn dữ liệu hoặc các nguồn phù hợp được tìm thấy, hoặc các quy tắc phân xử hoặc ngoại lệ được đáp ứng. Dữ liệu và siêu dữ liệu phát hiện được được tổ chức và hợp nhất, phân xử để đánh giá độ tin cậy, tổng hợp, và nhóm thành các hồ sơ tổng hợp nhờ sử dụng các quy luật ưu tiên và nguồn gốc để xác định các nguồn dữ liệu đáng tin cậy nhất cũng như các điều khoản sử dụng cho mỗi nguồn. Dữ liệu, siêu dữ liệu, và thông tin về mỗi lần tìm kiếm được giữ lại và có thể được sử dụng cho các mục đích tiếp theo, như các tìm kiếm tiếp theo hoặc các hoạt động ngược dòng khác. Ngoài ra, sáng chế còn đề cập đến vật ghi lưu trữ bất biến đọc được bởi máy tính.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến hệ thống và phương pháp tìm kiếm dữ liệu phù hợp với yêu cầu, và vật ghi lưu trữ bất biến đọc được bởi máy tính lưu trữ các lệnh của chương trình máy tính.

Tình trạng kỹ thuật của sáng chế

Các phương pháp được mô tả trong phần này là các phương pháp mà có thể được theo đuổi, nhưng không nhất thiết là các phương pháp mà trước đó đã được hình thành hoặc theo đuổi. Do đó, trừ khi có quy định khác, các phương pháp được mô tả trong phần này có thể không phải là tình trạng kỹ thuật đã biết với các yêu cầu bảo hộ của đơn này và không được thừa nhận là tình trạng kỹ thuật bởi việc đưa vào trong phần này.

Việc truy cập dữ liệu hiệu quả, thông qua việc tìm kiếm, so khớp, và các tính năng phân tích khác, là quan trọng để phát hiện và phân xử nhận dạng và các thông tin liên quan về các doanh nghiệp và các loại thực thể khác. Điều cốt yếu cho mục đích này là khả năng thực hiện truy cập hiệu quả, tìm kiếm và đối chiếu thông tin từ một hoặc nhiều nguồn dữ liệu. Ngoài ra, rất cần có khả năng đánh giá và xác định quy trình và các nguồn mà dữ liệu như vậy được truy cập từ đó cũng như chính dữ liệu. Điều này bao gồm việc phân tích dữ liệu và nguồn dữ liệu mà từ đó các phản hồi có tác dụng có thể được tạo ra và sau đó sẵn sàng để sử dụng để đưa ra quyết định liên quan đến quy trình, dữ liệu, siêu dữ liệu về dữ liệu được phát hiện, các nguồn mà từ đó dữ liệu được phát hiện, siêu dữ liệu về các nguồn đó và phản hồi có tác dụng từ toàn bộ quy trình.

Có các sản phẩm và tính năng phát hiện bán trên thị trường mà xử lý yêu cầu về một thực thể hoặc một nhóm thực thể thông qua tính năng xử lý khối hoặc các tính năng thao tác, ví dụ, khi nó được người dùng nhập theo các phương pháp khác nhau như (a) một người nhập dữ liệu vào trường yêu cầu bằng cách gõ hoặc "trích xuất" dữ liệu từ các nguồn khác nhau, (b) máy tính tạo ra giá trị yêu cầu, hoặc (c) hệ thống tương

tác trực tiếp với một hệ thống khác và sau đó truy vấn các trang web hoặc các nguồn dữ liệu khác theo các mục nhập chứa các thuộc tính yêu cầu này hoặc thông tin suy ra liên quan đến các thuộc tính này. Trong các trường hợp khác, theo các phương pháp phát hiện web truyền thống, công nghệ có thể thu thập dữ liệu, mà có thể là dạng tự do hoặc thuộc bản thể luận cố định (hoặc cấu trúc logic).

Các sản phẩm và tính năng phát hiện hiện nay thông thường bị giới hạn ở một số khuôn khổ, bao gồm cách mà dữ liệu yêu cầu được phân tích và định hướng để xác định các thuộc tính mà có thể được sử dụng để nhận biết dữ liệu từ các nguồn dữ liệu sẵn có, cách mà theo đó các nguồn dữ liệu này được truy cập, cách mà theo đó dữ liệu từ nguồn truy cập được sử dụng để bắt đầu hoặc hỗ trợ các yêu cầu phân tích tiếp theo, thông tin được cung cấp liên quan đến quy trình phân tích và các đặc tính của dữ liệu truy cập như chất lượng, tính đầy đủ, và độ trễ, và cách mà các thông tin này được sử dụng làm một phần của quy trình quản lý bao gồm việc phát hiện, tuyển lựa, phân xử, và tổ chức và hợp nhất, và tuân thủ các điều khoản sử dụng và các ràng buộc hiện hành.

Các sản phẩm và tính năng phát hiện hiện nay thông thường cung cấp dữ liệu trực tiếp đến hệ thống và người dùng cuối mà đã đưa ra yêu cầu mà không lưu trữ các thông tin về sự thành công của quy trình phân tích và các kết quả của nó để sử dụng sau này. Ngoài ra, các sản phẩm và tính năng phát hiện hiện hành này thông thường không duy trì siêu dữ liệu về các dữ liệu và nguồn dữ liệu được phát hiện. Hơn nữa, các sản phẩm và tính năng phát hiện hiện hành này không sử dụng dữ liệu và siêu dữ liệu được truy cập từ một lần tìm kiếm dưới dạng đầu vào cho tìm kiếm khác qua quy trình học tập đệ quy.

Các sản phẩm và tính năng phát hiện hiện hành này có khả năng giới hạn đối với việc tự động sử dụng việc học hỏi dựa trên kinh nghiệm về độ xác thực, nguồn gốc, và nội dung của dữ liệu và nguồn dữ liệu của mỗi kinh nghiệm, để tạo ra các ý kiến đánh giá mà ảnh hưởng đến khả năng truy cập và sử dụng các nguồn này và dữ liệu của chúng trong tương lai, hoặc các thay đổi về đặc tính có hiệu quả hoặc các khía cạnh chất lượng của các nguồn này.

Bản chất kỹ thuật của sáng chế

Phương án được bộc lộ đề xuất quy trình được thực hiện theo cách đệ quy nhằm sử dụng các kết quả của một yêu cầu hoặc tập hợp các thuộc tính khách quan để bắt đầu các yêu cầu tiếp theo từ cùng nguồn dữ liệu cũng như từ các nguồn dữ liệu khác, sao cho dữ liệu phát hiện được trở thành dữ liệu yêu cầu.

Theo phương án được bộc lộ, thông tin dựa trên kinh nghiệm về độ xác thực, nguồn gốc, và nội dung của dữ liệu và nguồn dữ liệu của mỗi kinh nghiệm được tạo ra, tổ chức và hợp nhất, tổng hợp, và tự động gộp để tạo ra các ý kiến đánh giá mà ảnh hưởng đến khả năng truy cập và sử dụng các nguồn này và dữ liệu của chúng trong tương lai.

Theo phương án được bộc lộ, phản hồi đối với mỗi kinh nghiệm được tạo ra để cho phép người dùng cuối, mà có thể là một người, hệ thống máy tính, hoặc đối tượng nhận khác hoặc quy trình phía sau, sử dụng các quy luật thương mại để điều khiển việc sử dụng và dùng quy trình và các kết quả cuối để đưa ra các quyết định đối với việc sử dụng thông tin theo cách thống nhất và lặp lại được, và theo cách mở rộng nếu cùng các quy luật thương mại đó được áp dụng cho các công nghệ, sản phẩm hoặc giải pháp khác nhau.

Theo phương án được bộc lộ, các quy trình có thể được thực hiện mà không có các hạn chế về phạm vi, vị trí địa lý, ngôn ngữ, hoặc hệ thống ghi. Kết quả này đạt được thông qua khả năng bất khả tri ngôn ngữ cho phép sử dụng loại hoặc số lượng bất kỳ nhãn ngôn ngữ web chuẩn và không bị hạn chế bởi việc mã hóa và logic đặc trưng cho ngữ cảnh, vị trí địa lý, ngôn ngữ hoặc hệ thống ghi.

Sáng chế đề cập đến hệ thống tự động và phương pháp thực hiện quy trình phát hiện đệ quy tự động mà không đòi hỏi sự can thiệp của con người để nhận biết, tuyển lựa, tổ chức và hợp nhất, phân xử và tổng hợp dữ liệu, như, ví dụ, nhận dạng hoặc trạng thái của doanh nghiệp, và thông tin về siêu dữ liệu liên quan mà được truy cập từ nhiều nguồn.

Sáng chế đề cập đến hệ thống tìm kiếm dữ liệu phù hợp với yêu cầu, bao gồm: thiết bị lưu trữ chứa danh sách các nguồn hạt giống mà xác định mục tiêu tìm kiếm ban

đầu cho yêu cầu đó; công cụ tìm kiếm để tìm kiếm các mục tiêu tìm kiếm ban đầu cho dữ liệu phù hợp với yêu cầu dựa trên dữ liệu và siêu dữ liệu được nhận biết từ các mục tiêu tìm kiếm ban đầu này và các mục tiêu tìm kiếm bổ sung trước; thiết bị lưu trữ này lưu trữ, từ các mục tiêu tìm kiếm ban đầu và bổ sung, dữ liệu và siêu dữ liệu này; và trong đó công cụ tìm kiếm tìm kiếm các mục tiêu tìm kiếm bổ sung đối với ít nhất một dữ liệu bổ sung phù hợp với yêu cầu, và dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm bổ sung khác.

Công cụ tìm kiếm này tìm kiếm các mục tiêu tìm kiếm bổ sung, và các mục tiêu tìm kiếm khác được tìm thấy. Dữ liệu và siêu dữ liệu liên quan bất kỳ xác định các mục tiêu tìm kiếm khác được công cụ tìm kiếm lưu trữ trong thiết bị lưu trữ này để sử dụng trong việc truy cập các đích khác. Các mục tiêu tìm kiếm bổ sung được tìm kiếm, và các mục tiêu tìm kiếm khác được tìm thấy, cho đến khi việc tìm kiếm các mục tiêu tìm kiếm bổ sung không tạo ra mục tiêu tìm kiếm nào để tìm kiếm hoặc cho đến khi các quy tắc phân xử chấp nhận được hoặc các quy tắc ngoại lệ đã được đáp ứng.

Các mục tiêu tìm kiếm khác được tìm thấy thông qua quá trình toàn diện và đệ quy nhờ đó các mục tiêu tìm kiếm được xác định dựa trên dữ liệu và siêu dữ liệu từ các kết quả của lần tìm kiếm và các mục tiêu tìm kiếm trước.

Công cụ tìm kiếm được tạo cấu hình để tìm kiếm các trang web hoặc các nguồn khác, và danh sách các nguồn hạt giống là danh sách các trang web hoặc các nguồn khác.

Bộ xử lý được tạo cấu hình để làm sạch dữ liệu có được từ mỗi mục tiêu tìm kiếm. Việc làm sạch có thể bao gồm ít nhất một trong số các bước phân tích cú pháp dữ liệu, loại bỏ các giá trị lỗi hoặc không phù hợp cho dữ liệu, và loại bỏ các mã thông báo định trước từ dữ liệu.

Bộ xử lý có thể được tạo cấu hình để thực hiện việc hợp thức hóa dữ liệu trên dữ liệu cho trước bằng cách so sánh dữ liệu cho trước từ các mục tiêu tìm kiếm đã được tìm kiếm, và chọn lọc, khi cần, dữ liệu từ nguồn được xem là đáng tin cậy và hữu ích nhất, dựa trên tập hợp các quy tắc ưu tiên và sử dụng. Bộ xử lý cũng có thể được tạo cấu hình để tổ chức và hợp nhất, phân xử, tổng hợp, và nhóm các dữ liệu liên quan từ

các mục tiêu tìm kiếm khác nhau để tạo ra các bộ dữ liệu được phân nhóm. Bộ xử lý có thể tạo ra bộ dữ liệu tập hợp từ các bộ dữ liệu được phân nhóm.

Sáng chế còn đề cập đến phương pháp tìm kiếm dữ liệu phù hợp với yêu cầu, bao gồm kiểm tra dữ liệu tìm thấy trong tập hợp các mục tiêu tìm kiếm ban đầu; lưu trữ trong thiết bị lưu trữ, từ các mục tiêu tìm kiếm ban đầu, (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (b) ít nhất một trong số dữ liệu và siêu dữ liệu mà xác định các mục tiêu tìm kiếm bổ sung cần tìm kiếm; tìm kiếm các mục tiêu tìm kiếm bổ sung cho (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (c) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác cần tìm kiếm; và lưu trữ trong thiết bị lưu trữ này, từ các mục tiêu tìm kiếm khác (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu và (c) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác nữa.

Khi các mục tiêu tìm kiếm bổ sung được tìm kiếm, nếu các mục tiêu khác nữa được tìm thấy, phương pháp còn bao gồm bước sử dụng (c) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác để truy cập vào các mục tiêu khác nữa; và lưu trữ trong thiết bị lưu trữ này, từ các mục tiêu tìm kiếm khác (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (d) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác cần tìm kiếm.

Phương pháp này được dừng lại khi việc tìm kiếm các mục tiêu tìm kiếm khác không tạo ra mục tiêu tìm kiếm bổ sung nào để tìm kiếm.

Theo phương pháp này, công cụ tìm kiếm có thể được cấu hình để tìm kiếm các trang web hoặc các nguồn khác. Danh sách các nguồn hạt giống là danh sách các trang web hoặc các nguồn khác.

Dữ liệu có được từ mục tiêu tìm kiếm có thể được làm sạch. Việc làm sạch dữ liệu có thể được thực hiện bằng các bước bao gồm phân tích cú pháp dữ liệu, loại bỏ các giá trị lỗi cho dữ liệu, và loại bỏ các mã thông báo định trước từ dữ liệu.

Việc hợp thức hóa dữ liệu có thể được thực hiện bằng cách so sánh dữ liệu từ nhiều nguồn khác nhau đã được tìm kiếm, và chọn lọc, khi cần, dữ liệu từ nguồn được xem là đáng tin cậy nhất, dựa trên tập hợp các quy tắc ưu tiên.

Phương pháp này có thể còn bao gồm việc tổ chức và hợp nhất, phân xử, tổng hợp, và phân nhóm các dữ liệu liên quan từ các nguồn khác nhau để tạo ra các bộ dữ liệu được phân nhóm. Các bộ dữ liệu đa nguồn kết hợp có thể được tạo ra từ các tập hợp bộ dữ liệu phân nhóm. Phương pháp này có thể còn bao gồm, ví dụ, nhưng không giới hạn, việc thực hiện ít nhất một việc được chọn từ nhóm bao gồm ghi vào cơ sở dữ liệu, lưu trữ trong cơ sở dữ liệu, tạo thông báo và công bố kết quả đã được tìm thấy bằng cách tìm kiếm dữ liệu phù hợp với yêu cầu.

Phương pháp này có thể còn bao gồm việc áp dụng phương pháp phân tích trong số ít nhất một phương pháp được chọn từ nhóm bao gồm các quy tắc, thuật toán, nghiệm suy, và các hàm phân tích khác để đưa ra quyết định đối với dữ liệu, và quyết định về việc tiếp tục hay dừng phương pháp này.

Sáng chế còn đề cập đến vật ghi lưu trữ bất biến đọc được bởi máy tính lưu trữ các lệnh của chương trình máy tính, mà khi được chạy bởi hệ thống máy tính, sẽ dẫn đến việc thực hiện các bước kiểm tra dữ liệu tìm thấy trong tập hợp các mục tiêu tìm kiếm ban đầu; lưu trữ trong thiết bị lưu trữ, từ các mục tiêu tìm kiếm ban đầu, (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (b) ít nhất một trong số dữ liệu và siêu dữ liệu mà xác định các mục tiêu tìm kiếm bổ sung cần tìm kiếm; tìm kiếm các mục tiêu tìm kiếm bổ sung đối với (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (c) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác cần tìm kiếm; và lưu trữ trong thiết bị lưu trữ này, từ các mục tiêu tìm kiếm khác (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu và (c) dữ liệu xác định các mục tiêu tìm kiếm khác nữa.

Vật ghi lưu trữ bất biến đọc được bởi máy tính có thể lưu trữ các lệnh bổ sung của chương trình máy tính, mà khi được chạy bởi hệ thống máy tính, sẽ dẫn đến việc thực hiện các bước sử dụng (c) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác để truy cập vào các mục tiêu khác nữa; và lưu trữ trong thiết bị lưu trữ này, từ các mục tiêu tìm kiếm khác (a) ít nhất một trong số dữ liệu và siêu dữ liệu phù hợp với yêu cầu, và (d) ít nhất một trong số dữ liệu và siêu dữ liệu xác định các mục tiêu tìm kiếm khác cần tìm kiếm.

Vật ghi lưu trữ bất biến đọc được bởi máy tính có thể lưu trữ các lệnh bổ sung của chương trình máy tính, mà khi được chạy bởi hệ thống máy tính, khi các mục tiêu tìm kiếm bổ sung được tìm kiếm, và dữ liệu xác định các mục tiêu tìm kiếm khác được tìm thấy, để thực hiện lặp lại phương pháp này cho đến khi việc tìm kiếm các mục tiêu tìm kiếm khác không tạo ra mục tiêu tìm kiếm nào để tìm kiếm, hoặc các quy tắc phân xử chấp nhận được hoặc các quy tắc ngoại lệ đã được đáp ứng.

Mô tả vắn tắt các hình vẽ

FIG.1 là sơ đồ khối của phương pháp tìm kiếm và tập hợp dữ liệu từ tập hợp nguồn ban đầu.

FIG.2 là sơ đồ khối của phương pháp phân xử và hợp nhất bản ghi.

FIG.3 là hình minh họa hệ thống máy tính sử dụng để thực hiện phương án được bộc lộ.

Thành phần hoặc đặc điểm mà chung ở nhiều hơn một hình vẽ sẽ được chỉ ra bằng cùng số tham chiếu ở mỗi hình vẽ.

Mô tả chi tiết sáng chế

Định nghĩa

Phép đệ quy xác định quy trình phát hiện đa ngôn ngữ hoặc bất khả tri ngôn ngữ nhiều bước nhờ đó kết quả của các bước trước trở thành đầu vào của các bước sau, nghĩa là mô hình hoạt động đầu cuối đến đầu cuối và các kết quả là không dự đoán được vào lúc bắt đầu nhưng thay vào đó được xác định bằng dữ liệu phát hiện, quy trình thực hiện và các kết quả trung gian. Hoạt động này có thể bao gồm việc sử dụng dữ liệu đầu vào ở trạng thái nguyên bản như được lấy từ bước trước đó hoặc có thể được biến đổi do việc làm sạch hoặc các dạng biến đổi dữ liệu khác.

Phép phân tích xác định quy trình nhận biết dữ liệu theo giá trị yêu cầu. Yêu cầu có thể được giải quyết nhờ sử dụng các phương pháp khác nhau, ví dụ, thông qua các tính năng so khớp, tìm kiếm hoặc truy tìm, đối với nhiều loại tham chiếu khác.

Việc phát hiện là quy trình để nhận biết và tuyển lựa thông tin được lấy ra từ các nguồn dữ liệu phù hợp với yêu cầu.

Tính linh hoạt chỉ ra rằng các quy trình có thể được tự động thích ứng dựa trên việc sử dụng và thông tin về việc các quy trình được thực hiện như thế nào hoặc các siêu dữ liệu khác, và được biến đổi dễ dàng để hỗ trợ cho các mục tiêu khác nhau.

Thuật ngữ không xác định ngụ ý rằng phương pháp linh hoạt được xác định một cách năng động dựa trên thông tin về dữ liệu, siêu dữ liệu, và các quy trình trước mà truy cập vào dữ liệu và siêu dữ liệu đó, và không thể ước tính được đại số trước liên quan đến thời gian hoặc kết quả thực hiện.

Học qua trải nghiệm đề cập đến việc sử dụng thông tin về tính xác thực, nguồn gốc, và nội dung của dữ liệu, các nguồn dữ liệu, và siêu dữ liệu đối với dữ liệu và các nguồn dữ liệu của mỗi trải nghiệm vì một loạt các lý do, ví dụ, nhằm tạo ra các ý kiến đánh giá mà ảnh hưởng đến khả năng truy cập và sử dụng các nguồn này và dữ liệu của chúng trong tương lai.

Phản hồi là thông tin về quy trình và trải nghiệm phát hiện, cũng như về kết quả của quy trình đó. Phản hồi này có thể được sử dụng để cho phép người dùng đầu cuối, mà có thể là một người, hệ thống máy tính, hoặc thiết bị nhận khác, sử dụng các quy tắc hoạt động để điều khiển việc sử dụng và dùng quá trình này và các kết quả cuối cùng để đưa ra quyết định. Phản hồi mà phản ánh chất lượng suy luận là dữ liệu có thể được sử dụng bởi người dùng cuối để xác định mức độ mà các kết quả đáp ứng tiêu chuẩn dựa trên chất lượng của người dùng đầu cuối.

Các chỉ dẫn khác là dữ liệu mà có thể được sử dụng làm một phần của quá trình phân tích hoặc thông tin mà được lấy ra dưới dạng kết quả của quy trình đó mà liên quan đến nhận dạng hoặc các khía cạnh khác của đối tượng của chúng. Các chỉ dẫn có thể bao gồm dữ liệu mà đã được biết hoặc phát hiện trước trong quá trình thực hiện quy trình phát hiện này, mà sau đó có thể được sử dụng vì các lý do sau, như, ví dụ, tổ chức và hợp nhất, phân xử, và tổng hợp nhằm thực hiện các trải nghiệm phát hiện trong tương lai.

Các quy trình tổ chức và hợp nhất và phân xử đánh giá định tính mỗi trải nghiệm phát hiện và các kết quả, và xác định các hành động tiếp theo dựa trên các đánh giá đó. Quy trình này bao gồm giữ lại và đánh giá nguồn gốc và tính xác thực của mỗi nguồn mà được truy cập trong quá trình thực hiện phương pháp này theo phương án được mô tả, nhằm đưa ra thông tin mà sẽ được sử dụng trong các lần thực hiện tiếp theo như các lợi ích về chất lượng của mỗi nguồn trong toàn bộ và mỗi loại dữ liệu truy cập được từ nguồn đó. Thuật ngữ tổ chức và hợp nhất thường nói đến việc tổ chức và hợp nhất, lưu trữ, và bảo quản dữ liệu. Phân xử thường chỉ việc thực hiện sự xác định về chất lượng của dữ liệu và cách thức có thể sử dụng nó một cách có hiệu quả nhất.

Thuật ngữ môđun được sử dụng ở đây để biểu thị hoạt động chức năng mà có thể được thể hiện dưới dạng thành phần độc lập hoặc dưới dạng cấu hình tích hợp của nhiều thành phần phụ. Do đó, môđun chương trình, như được mô tả dưới đây, có thể được thể hiện dưới dạng một môđun đơn nhất hoặc dưới dạng nhiều môđun mà hoạt động kết hợp với nhau.

Công cụ tìm kiếm là một ví dụ về tính năng phân tích, và thông thường là sự kết hợp của phần cứng, và phần mềm điều khiển hoạt động của phần cứng, cho các mục tiêu tìm kiếm và thu thập dữ liệu liên quan đến hoặc được phát hiện là kết quả của yêu cầu.

Thuật ngữ tìm kiếm được sử dụng ở đây chỉ phương pháp bất kỳ để nhận biết thông tin có thể được quan tâm.

Thuật ngữ ứng dụng cụ thể như được sử dụng theo sáng chế nghĩa là các tập hợp quy tắc và quy trình hoạt động khác nhau mà được sử dụng trong việc thực hiện phương pháp này và trong việc thực thi hệ thống có thể được xác định để đáp ứng các mục tiêu cụ thể được bộc lộ ở đây.

Thuật ngữ công bố nghĩa là kết quả đầu cuối của quy trình được thực hiện bởi sáng chế có thể có sẵn cho lần sử dụng tiếp theo, ví dụ, được ghi vào kho dữ liệu, được sử dụng làm đầu vào của hệ thống hoặc ứng dụng, hoặc được ghi vào báo cáo.

Thiết bị lưu trữ dữ liệu có thể là thiết bị bất kỳ được sử dụng trong công nghệ máy tính để lưu trữ dữ liệu, thường trên cơ sở bất biến, như được mô tả thêm dưới đây. Các

phần khác nhau của thiết bị lưu trữ dữ liệu, được sử dụng để lưu trữ các loại dữ liệu khác nhau, có thể nằm trên cùng thiết bị vật lý hoặc trên thiết bị vật lý khác nhau.

FIG.1 và FIG.2 là các sơ đồ khối chức năng mô tả quy trình ví dụ để tập hợp và biến đổi dữ liệu từ mạng lưới toàn cầu (còn được gọi là Internet) thành dữ liệu doanh nghiệp có thể công bố nhờ sử dụng quy trình đệ quy mà bao gồm việc truy cập vào các trang web đã biết và phát hiện được, và lặp lại quy trình này nhiều lần để tìm thêm các trang web khác nhờ sử dụng dữ liệu mà trước đó đã biết hoặc được phát hiện thông qua quy trình này, để nhận biết, tuyển lựa, phân xử, tổ chức và hợp nhất, tổng hợp, và giữ dữ liệu trong một hoặc nhiều kho chứa dữ liệu, lưu trữ theo cách để cho phép xuất ra dữ liệu trong các báo cáo, truy cập hoặc hiển thị nhờ sử dụng các phương tiện trực quan hoặc truyền thông khác, hoặc tạo điều kiện cho việc tích hợp dữ liệu phát hiện vào các ứng dụng phần mềm để sử dụng sau này. Phương án được bộc lộ này là thiết bị, bao gồm các thành phần là các công cụ tìm kiếm và các quy trình liên quan để xử lý phần tử dữ liệu thực thể thương mại, nhưng có thể được sử dụng cho các loại dữ liệu khác, như, ví dụ, dữ liệu y khoa. Các kho dữ liệu có thể bao gồm dữ liệu cho các thực thể mới cũng như các dạng biến thể đối với các thực thể mà trước đó đã có trong kho dữ liệu, mà là một phần của phương án được bộc lộ này.

Trên FIG.1 một phần của quy trình, 100, mà chủ yếu là phần phát hiện dữ liệu của quy trình, tìm kiếm và tập hợp dữ liệu từ tập hợp nguồn hữu hạn ban đầu mà được xác định bằng các địa chỉ trang web hoặc định vị tài nguyên thống nhất (uniform resource locators - URLs), bao gồm cả danh sách nguồn xác định trước để sử dụng cũng như các nguồn khác được nhận biết trong quá trình thực hiện quy trình 100. Theo phương án này, ngay khi tất cả các địa chỉ web ban đầu được được tìm kiếm, quy trình 100 sẽ kiểm tra sự có mặt của các địa chỉ web được phát hiện mới khác, nhưng thứ tự tìm kiếm các địa chỉ web có thể khác theo phương án khác. Quy trình phát hiện này có tầm quan trọng với phương pháp không xác định của phương án được bộc lộ, ở chỗ, quy trình này sẽ tiếp tục truy cập vào các trang web mới và thu thập dữ liệu cho đến khi không phát hiện hoặc thu thập thông tin khác mà được quan tâm, dựa trên các tiêu chuẩn định trước. Đối với mỗi tập hợp mới của các địa chỉ web phát hiện, quy trình 100, nhờ tác dụng đòn bẩy của tập hợp quy tắc, thuật toán, suy nghiệm, hoặc các phương pháp phân tích khác, đưa ra quyết định xem nó sẽ tự yêu cầu tiếp tục chu kỳ phát hiện và thu thập

tiếp theo hay dừng các hoạt động này. Sự tự yêu cầu này có tầm quan trọng với bản chất đệ quy của phương án được bộc lộ và là một đặc tính xác định mà có thể phân biệt với việc tìm kiếm và khai phá web truyền thống.

Ví dụ về thuật toán mà có thể được sử dụng là việc tạo ra dạng viết tắt dưới dạng giá trị thay thế cho tên thương mại bằng cách tách và sau đó ghép các ký tự đầu tiên của mỗi từ riêng rẽ. Ví dụ, theo thuật toán này, Công ty máy tính quốc tế - International Business Machines sẽ trở thành IBM, hoặc với một sự thay đổi nhỏ trong thuật toán, Tập đoàn công nghệ máy tính quốc tế - International Business Machines Corporation cũng sẽ trở thành IBM.

Ví dụ về suy nghiệm mà có thể được sử dụng là để nhận biết phân đoạn nhân khẩu học địa lý dựa trên các loại công nghiệp. Ví dụ, nhiều công ty công nghệ ở Mountain View, CA có thể đại diện cho xu hướng về một tỷ lệ phần trăm cao những người lao động có độ tuổi dưới bốn mươi và xu hướng về một tỷ lệ phần trăm cao các cá nhân có bằng thứ hai hoặc bằng thạc sỹ cao.

Ở bước 101, việc nhận việc phát hiện khởi động quy trình 100. Ở bước 110, trang web được truy cập dựa trên danh sách nguồn hạt giống 111. Mỗi hồ sơ trong danh sách nguồn hạt giống 111 bao gồm địa chỉ trang web (URL). Nó có thể bao gồm URL bất kỳ theo quy tắc định dạng chuẩn như, nhưng không chỉ giới hạn ở, các địa chỉ được bắt đầu bằng "www".

Quy trình 100 tiến đến bước 130, ở đây các hồ sơ trang web thích hợp được xác định nhờ sử dụng danh sách các quy tắc xử lý 126, mà xác định loại dữ liệu mong muốn, dưới dạng giá trị đầu vào. Danh sách các quy tắc xử lý 126 xác định các yếu tố dữ liệu được tìm kiếm trong danh sách nguồn hạt giống 111.

Bảng dưới đây đưa ra một số ví dụ về các quy tắc định nghĩa yếu tố dữ liệu trong danh sách quy tắc xử lý 126.

Yếu tố dữ liệu	Quy tắc
Số điện thoại	Tất cả các chuỗi số
Số điện thoại	dài 10 bit

Số điện thoại	Tìm thấy dấu nối sau con số thứ 3 và thứ 6
Địa chỉ	Chuỗi văn bản chứa các từ khóa "phố", "đại lộ", "đường" hoặc "ngách", v.v.
Địa chỉ	Không thể chỉ chứa các con số
Mã bưu điện	dài 5 đến 9 con số
Mã bưu điện	Tất cả các chuỗi số

Ở bước 130, nhờ sử dụng các quy tắc được định nghĩa trong danh sách các quy tắc xử lý 126, dữ liệu tìm thấy trên trang mã nguồn được nhận biết bởi URL trong danh sách nguồn hạt giống 111, mà được cung cấp thông qua bước 110, được đọc có hệ thống. Đối với mỗi phần dữ liệu đọc được từ trang web, bước 130 thực hiện việc xác định xem dữ liệu có phù hợp với quy tắc bất kỳ trong số các quy tắc được định nghĩa trong danh sách các quy tắc xử lý 126. Nếu phần dữ liệu phù hợp với các quy tắc về các yếu tố cho trước, thì nó được lưu trữ trong hồ sơ doanh nghiệp đầu vào 136 ở yếu tố phù hợp được định nghĩa bởi danh sách yếu tố dữ liệu 165. Ví dụ, chuỗi trị số 999-999-9999 đáp ứng tiêu chí về số điện thoại của Mỹ dựa trên quy tắc mẫu nêu trên. Giá trị này sẽ được ghi vào trường số điện thoại trong hồ sơ doanh nghiệp đầu vào 136.

Thông qua sử dụng các quy tắc chuyên dụng (ví dụ, danh sách các quy tắc xử lý 126), quy trình 100 có khả năng duy trì việc thực hiện không xác định và linh hoạt mà đáp ứng các nhu cầu thay đổi các mục đích sử dụng đầu cuối của quy trình này. Nhờ tác dụng đòn bẩy của các quy tắc này, quy trình 100 có thể phát hiện loại thực thể cụ thể (ví dụ, các thực thể thương mại) mà thỏa mãn tập hợp các đặc tính mà có thể được định nghĩa theo cách đặc biệt đối với mỗi bước thực thi riêng lẻ của quy trình 100.

Hồ sơ doanh nghiệp đầu vào 136 là cấu trúc dữ liệu được sử dụng để lưu trữ dữ liệu tìm thấy và thu thập ở bước 130. Mỗi hồ sơ gồm các yếu tố dữ liệu được định nghĩa trong danh sách yếu tố dữ liệu 165. Danh sách yếu tố dữ liệu 165 về cơ bản là kho chứa các thành phần chính của một thực thể, như trong trường hợp hồ sơ doanh nghiệp là tên, địa chỉ và số điện thoại.

Ở bước 135, mỗi yếu tố của dữ liệu hồ sơ thu được được phân tách thành các thành phần phụ, làm sạch, chuẩn hóa và tiêu chuẩn hóa, nhờ sử dụng các phương pháp đã biết rộng rãi trong lĩnh vực tìm kiếm, nhờ sử dụng hồ sơ doanh nghiệp đầu vào 136 làm giá trị đầu vào. Nhờ sử dụng dữ liệu tham chiếu trong danh sách dữ liệu tham chiếu 146, các giá trị ở mỗi yếu tố dữ liệu trong hồ sơ doanh nghiệp đầu vào 136 được phân tách. Khi mỗi yếu tố được phân tách, tập hợp các thành phần đặc trưng cho mỗi thực thể, như thực thể thương mại, được xác định dựa trên danh sách yếu tố dữ liệu 165. Các thành phần này cung cấp cái nhìn sâu sắc đối với các đặc điểm của doanh nghiệp; ví dụ về thực thể thương mại có thể bao gồm cơ cấu doanh nghiệp và vị trí địa lý (địa chỉ). Ví dụ, thành phần địa chỉ được tiêu chuẩn hóa dựa trên dữ liệu tham chiếu trong danh sách dữ liệu tham chiếu 146. Dữ liệu còn được làm sạch các giá trị không mong muốn, ví dụ khoảng trống, chấm câu thừa, hoặc các đặc điểm khác hoặc các tập hợp đặc điểm thường được gọi là "nhiễu"; các giá trị yếu tố thu được là các thông tin thương mại sử dụng được cho ứng dụng. Dữ liệu được biến đổi được ghi vào hồ sơ doanh nghiệp đầu ra 137. Hồ sơ doanh nghiệp đầu ra 137 này có cùng cấu trúc như hồ sơ doanh nghiệp đầu vào 136 như được định nghĩa bởi danh sách yếu tố dữ liệu 165.

Dưới đây là ví dụ về một dạng dữ liệu tham chiếu mà có thể được tìm thấy, nhưng không bao gồm tất cả các loại dữ liệu tham chiếu có thể có, trong danh sách 146.

Yếu tố	Mã thông báo	Hành động	Giá trị đầu ra	Giá trị đầu vào	Đầu ra chuẩn hóa
Tên thương mại	*	Loại bỏ ra khỏi giá trị			
Tên thương mại	&	Loại bỏ ra khỏi giá trị			
Địa chỉ	"chính"		"chính"		
Địa chỉ	"đường"		"đường"		
Tên liên hệ	"Ông"	Loại bỏ ra			

		khỏi giá trị			
Địa chỉ				"Thành phố bất kỳ, NJ"	"Thành phố bất kỳ, NJ 10000"
Địa chỉ				1600 Đại lộ Pennsylvania, Washington, DC	1600 Đại lộ Pennsylvania, Washington, D.C. 20500-0003

Hồ sơ doanh nghiệp đầu vào mẫu trong danh sách hồ sơ doanh nghiệp đầu vào 136 là:

Tên thương mại	Địa chỉ	Điện thoại	Tên liên hệ	Ngành kinh doanh	Trang nguồn
ABC *&Company	123 đường chính, Thành phố bất kỳ, NJ	1(999)1234567	Ông John Smith	!sản xuất	www.CompanyListing.com

Hồ sơ doanh nghiệp đầu ra trong danh sách hồ sơ doanh nghiệp đầu ra 137, với dữ liệu được phân tích cú pháp, làm sạch, chuẩn hóa và tiêu chuẩn hóa là:

Tên thương mại	Địa chỉ	Điện thoại	Tên liên hệ	Ngành kinh doanh	Trang nguồn
ABC Company	123 đường	19991234567	John Smith	Sản xuất	www.CompanyListing.com

	chính, Thành phố bất kỳ, NJ 10000				

Danh sách dữ liệu tham chiếu 146 là tập hợp các dữ liệu tham chiếu được sử dụng trong bước 135. Danh sách dữ liệu tham chiếu 146 chứa chuỗi mã thông báo và dữ liệu tham chiếu địa lý. Trong giai đoạn phân tách của bước 135, các mã thông báo này được sử dụng để nhận biết yếu tố dữ liệu then chốt trong dữ liệu được thu thập. Ví dụ về mã thông báo là "Đường". Sự có mặt của "Đường" sẽ chỉ ra rằng yếu tố dữ liệu là địa chỉ.

Các mã thông báo trong danh sách dữ liệu tham chiếu 146 còn chứa các giá trị không mong muốn. Quy trình làm sạch tìm kiếm và loại bỏ các mã thông báo này ra khỏi các giá trị trong danh sách yếu tố dữ liệu 165. Ví dụ về các mã thông báo không mong muốn là các từ tục tĩu.

Dữ liệu tham chiếu về địa lý trong danh sách dữ liệu tham chiếu 146 được sử dụng trong quá trình tiêu chuẩn hóa của bước 135, ví dụ, để biến đổi dữ liệu địa chỉ phù hợp theo chuẩn bưu chính địa phương. Dữ liệu đã được chuẩn hóa tạo ra sự thể hiện phù hợp hơn của dữ liệu địa chỉ.

Trong ví dụ dưới đây, giá trị không được tiêu chuẩn hóa ban đầu không có mã bưu điện và có địa chỉ đường phố không hoàn chỉnh. Quy trình tiêu chuẩn hóa ở bước 135 so sánh địa chỉ gốc với dữ liệu tham chiếu lưu trong danh sách dữ liệu tham chiếu 146 và xuất ra danh sách hồ sơ doanh nghiệp đầu ra với địa chỉ hoàn chỉnh 137.

	Đường	Thành phố	Mã bưu điện	Bang
Không được tiêu chuẩn hóa	1600 Đại lộ Pennsylvania	Washington		DC
Tiêu chuẩn hóa	1600 Pennsylvania	Washington	20500-0003	DC

	Đại lộ NW			

Ở bước 140 (sau khi dữ liệu trong hồ sơ doanh nghiệp đầu vào 136 được làm sạch, tiêu chuẩn hóa, và ghi vào hồ sơ doanh nghiệp đầu ra 137 như trên), trạng thái đối với hồ sơ doanh nghiệp đầu ra 137 được xác định, trong đó xác định các hoạt động tiếp theo. Hồ sơ doanh nghiệp đầu ra 137 và danh sách các quy tắc xử lý các hoạt động tiếp theo 148 là các đầu vào được sử dụng ở bước 140.

Danh sách các quy tắc xử lý 148 bao gồm (A) tập hợp các quy tắc xử lý sử dụng để xác định tập hợp các hành động tiếp theo để thực hiện trên hồ sơ doanh nghiệp đầu ra 137 cũng như (B) bước logic tiếp theo cần thực hiện trong quy trình 100. Ví dụ, một loại quy tắc trong danh sách các quy tắc xử lý các hành động tiếp theo 148 là tập hợp các quy tắc hợp thức hóa hồ sơ. Danh sách các quy tắc xử lý 148 cũng có thể bao gồm thuật toán, suy nghiệm, hoặc các phương pháp phân tích khác, như được mô tả ở trên.

Các quy tắc và tiêu chí hợp thức hóa thực tế không bị giới hạn nhưng dành riêng đối với mỗi lần thực hiện phương án cụ thể. Ví dụ về quy tắc có khả năng thực hiện là: "Quy tắc hợp thức hóa 1: Tên thương mại, đường, thành phố, mã bưu điện, và số điện thoại phải cùng nơi."

Nếu hồ sơ doanh nghiệp đầu ra 137 vượt quá tập hợp các quy tắc hợp thức hóa này của danh sách các quy tắc xử lý các hành động tiếp theo 148, nó được ghi vào danh sách các hồ sơ nguồn duy nhất141.

Danh sách các hồ sơ nguồn duy nhất 141 lưu trữ tất cả các dạng hợp lệ của hồ sơ doanh nghiệp đầu ra 137. Nguồn duy nhất ngụ ý rằng hồ sơ dữ liệu riêng biệt được giữ lại đối với mỗi hồ sơ được truy cập và lựa chọn từ các nguồn URL. Tại thời điểm này trong quy trình 100, mỗi hồ sơ doanh nghiệp đầu ra 137 là hồ sơ nguồn duy nhất được sao chép vào danh sách các giá trị dữ liệu yêu cầu mới 147. Do đó, mỗi hồ sơ doanh nghiệp đầu ra trong danh sách hồ sơ doanh nghiệp đầu ra 137 và các giá trị dữ liệu yêu cầu mới trong danh sách các giá trị dữ liệu yêu cầu mới 147 chứa dữ liệu chỉ từ một trang web. Trong danh sách các hồ sơ nguồn duy nhất 141, có thể có nhiều hồ sơ đại diện cho cùng doanh nghiệp, nhưng các nguồn dữ liệu này có thể khác nhau. Ví dụ mẫu về danh sách các hồ sơ nguồn duy nhất 141 là như sau:

Hồ sơ số	Tên thực thể	Địa chỉ	Thành phố	Mã bưu điện	Số điện thoại	Ngành kinh doanh	Trang nguồn
1	ABC Company	123 Đường chính	Thành phố bất kỳ	10000	999-888-1111	Hiệu bánh	abccompany.com
2	Abc Company		Thành phố bất kỳ		999-888-1000	Hiệu bánh/bánh mỳ	Yellowpages.com
3	ABC Company	Đường chính		10000		Hiệu bánh	Thành phố bất kỳBiz.com
4	MyCorp	222 First Đại lộ	New York Thành phố	21212	614-111-1010		Yellowpages.com
5	John Doe, Inc.	Suite A, 100 4th Đường	Washington	11111	888-555-0000	Dịch vụ pháp lý	JohnDoe.com

Trong ví dụ nêu trên, hồ sơ 1 có thông tin về doanh nghiệp chỉ từ một trang web, www.companylisting.com. Mặc dù hồ sơ 2 có thể đề cập đến cùng doanh nghiệp, nhưng thông tin của nó được thu thập từ www.abccompany.com. Mỗi hồ sơ chỉ có một trang web là nguồn dữ liệu của nó.

Trong suốt quá trình thực hiện quy trình 100 qua các bước 130, 135, và 140, các nguồn URLs mới liên quan đến hồ sơ của doanh nghiệp có thể được tìm thấy; chúng được ghi vào danh sách các giá trị dữ liệu yêu cầu mới 147. Chỉ có các URLs chưa được truy cập được ghi vào danh sách các giá trị dữ liệu yêu cầu mới 147. Danh sách

các giá trị dữ liệu yêu cầu mới 147 là giống về cấu trúc với danh sách nguồn hạt giống 111 và chứa các trang web để tiếp tục kiểm tra trong quy trình 100.

Ví dụ:

- Trang nguồn www.companylisting.com là trang gốc trong danh sách nguồn hạt giống 111.
- Trong quá trình đọc dữ liệu trên www.companylisting.com, doanh nghiệp "ABC Company" được tìm thấy cùng với trang web liên quan của nó www.ABC-Company.com.
- www.ABC-Company.com được ghi vào danh sách các giá trị dữ liệu yêu cầu mới 147.

Ở bước 140, dựa trên việc đánh giá dữ liệu thu được và phân tích đến thời điểm này trong quy trình 100, các bước tiếp theo trong quy trình 100 được xác định. Các tiêu chí sử dụng để xác định sự có mặt của các dữ liệu bổ sung là chuyên dụng, và được xác định như mô tả trong các điều kiện và bước dưới đây. Một số ví dụ bao gồm, nhưng không chỉ giới hạn ở, sự có mặt của các mục nhập tên doanh nghiệp bổ sung và/hoặc các liên kết đến các trang web bổ sung mà có thể có các dữ liệu thích hợp bổ sung. Các điều kiện và tiêu chuẩn chuyên dụng này được lưu trữ trong danh sách các quy tắc xử lý các hành động tiếp theo 148.

Danh sách các quy tắc xử lý 148 gồm có tập hợp các quy tắc xử lý sử dụng để xác định tập hợp các hành động tiếp theo để thực hiện trên hồ sơ 137 cũng như bước logic tiếp theo cần thực hiện trong quy trình. Hai loại quy tắc mẫu trong danh sách các quy tắc xử lý 148 là tập hợp các quy tắc hợp thức hóa hồ sơ hoặc quy trình để đánh giá các điều khoản sử dụng của trang web. Phương pháp 140 hợp thức dữ liệu trong hồ sơ 137 nhờ sử dụng các quy tắc này để xác nhận sự có mặt của của hồ sơ doanh nghiệp hoàn chỉnh. Số lượng các quy tắc và tiêu chuẩn hợp thức hóa là chuyên dụng.

Danh sách các quy tắc xử lý 148 có thể còn có các quy tắc để giữ quy trình 100 liên tục mà không bị dừng. Quy trình 100 có thể bị dừng, ví dụ, sau khi một khoảng thời gian cố định từ khi bắt đầu quy trình đã trôi qua, hoặc sau khi số lượng đã định của hoạt động CPU đã được thực hiện.

Có nhiều điều kiện tiềm năng mà sẽ xác định các hoạt động tiếp theo. Trong ví dụ này, hai điều kiện được mô tả. Điều kiện 1 đề cập đến tình huống có thể có nhiều dữ liệu thích hợp hơn cần truy cập từ URL ban đầu hoặc URL bổ sung được phát hiện trong quá trình thực hiện quy trình 100. Điều kiện 2 đề cập đến tình huống không có thêm dữ liệu thích hợp cần tìm thấy trên các URL ban đầu và bổ sung.

Đối với điều kiện 1, ở bước 150, có các dữ liệu bổ sung cần đọc từ trang web hiện hành. Quy trình 100 lặp lại bước 130 theo cách của bước 149, mà trang web được kiểm tra được nghiên cứu kỹ lưỡng cho hồ sơ tiếp theo. Vòng lặp này sẽ lặp lại cho đến khi không có dữ liệu bổ sung cần đọc từ trang web đang được kiểm tra.

Tiếp theo điều kiện 1, vòng lặp qua bước 149 đến bước 130 là một phần có tính chất mở, toàn diện và lặp đi lặp lại của phương án được bộc lộ. Việc lặp đi lặp lại này cho phép quy trình 100 là không xác định về bản chất, do đó không đòi hỏi biết trước về trang web được kiểm tra hoặc số lượng các thực thể kinh doanh có mặt trước khi phát hiện ra web.

Tiếp theo điều kiện 1, ở mỗi lần thực hiện vòng lặp qua bước 149 đến bước 130, danh sách các hồ sơ nguồn duy nhất 141 và danh sách các giá trị dữ liệu yêu cầu mới 147 được duy trì trong các phương tiện lưu trữ dữ liệu mà giữ tất cả các thông tin thu thập được trong mỗi lần thực hiện trước của vòng lặp qua bước 149 đến bước 130. Danh sách các hồ sơ nguồn duy nhất 141 có các hồ sơ doanh nghiệp bổ sung đính kèm vào nó trong giai đoạn phát hiện như được mô tả ở trên, đối với FIG.1. Danh sách các giá trị dữ liệu yêu cầu mới 147 có các URL bổ sung đính kèm vào nó trong giai đoạn phát hiện dữ liệu. Dữ liệu trong danh sách các hồ sơ nguồn duy nhất 141 và danh sách các giá trị dữ liệu yêu cầu mới 147 được sử dụng làm đầu vào của vòng lặp từ bước 149 đến bước 130, cho đến khi toàn bộ quy trình 100 được hoàn thành và không có dữ liệu khác có thể được đọc từ trang web này, như được mô tả trong điều kiện 2 dưới đây.

Đối với điều kiện 2, quy trình 100 thực hiện từ bước 140 đến bước 150. Nếu xác định được ở bước 150 rằng tất cả dữ liệu đã được đọc từ trang web hiện đang được kiểm tra, quy trình 100 sau đó diễn ra từ bước 150 đến bước 160.

Tiếp tục điều kiện 2, ở bước 160, việc truy cập vào trang web hiện đang được kiểm tra được hoàn thành, và kết nối đến trang web đang được kiểm tra được đóng lại. Ngay khi hoàn thành bước 160, quy trình 100 tiến đến bước 161, nơi thực hiện việc xác định xem tất cả các trang web trong nguồn hạt giống 111 có được kiểm tra chưa. Nếu tất cả các trang web trong danh sách nguồn hạt giống 111 chưa được kiểm tra, thì quy trình 100 tiến đến bước 151, trong đó kết nối được mở đến trang web tiếp theo trong danh sách nguồn hạt giống 111. Quy trình 100 quay trở lại bước 110.

Tại thời điểm này có nhiều điều kiện phụ tiềm năng. Trong ví dụ này, mô tả hai điều kiện phụ tiềm năng (dưới đây được gọi là Điều kiện 2A và Điều kiện 2B) mà xác định các hoạt động tiếp theo. Việc tiến triển tiếp theo của các bước được xác định dựa trên một trong số các điều kiện này.

Đối với điều kiện 2A, nếu tất cả các trang nguồn ban đầu trong danh sách nguồn hạt giống 111 đã được truy cập, thì quy trình 100 tiến từ bước 161 đến bước 162. Ở bước 162, lúc đầu, thực hiện việc xác định xem có thêm dữ liệu yêu cầu mới nào (ví dụ, các URL) vào danh sách các giá trị dữ liệu yêu cầu mới 147 không. Quy trình 100 tiến đến bước 154, nếu dữ liệu yêu cầu mới được thêm vào danh sách các giá trị dữ liệu yêu cầu mới 147, mà yêu cầu quy trình 100 một lần nữa từ bước 110 để truy cập vào trang web khác nhờ sử dụng dữ liệu trong danh sách các giá trị dữ liệu yêu cầu mới 147 dưới dạng đầu vào. Danh sách các giá trị dữ liệu yêu cầu mới 147 về cơ bản giả định vai trò mà trước đây là vai trò của danh sách nguồn hạt giống 111, trong yêu cầu này của bước 110. Vòng lặp quay lại bước 110 là một phần khác có tính chất lặp đi lặp lại của phương án được bộc lộ. Việc lặp lại vòng lặp này cho phép quy trình trở nên có tính chất không xác định sao cho không cần biết trước về trang web hoặc số lượng các thực thể có mặt trước khi phát hiện ra web. Vòng lặp tìm kiếm được thực hiện, danh sách các hồ sơ nguồn đơn nhất 141 và danh sách các giá trị dữ liệu yêu cầu mới 147 được duy trì trên các phần của các phương tiện lưu trữ dữ liệu mà lưu trữ tất cả các thông tin thu thập được trong mỗi tiến trình thông qua các vòng lặp qua bước 154. Danh sách các hồ sơ nguồn đơn nhất 141 có các hồ sơ doanh nghiệp bổ sung đính kèm vào nó trong giai đoạn phát hiện như được mô tả trên đây. Danh sách các giá trị dữ liệu yêu cầu mới 147 có các URL bổ sung đính kèm vào nó trong giai đoạn phát hiện này như được mô tả trên đây. Dữ liệu tiếp tục tồn tại trong danh sách các hồ sơ nguồn đơn nhất 141 và

danh sách các giá trị dữ liệu yêu cầu mới 147. Dữ liệu trong danh sách các hồ sơ nguồn đơn nhất 141 và các giá trị dữ liệu yêu cầu mới 147 được sử dụng làm đầu vào của quy trình 100 ở bước 110, cho đến khi, ở bước 162, xác định được rằng không có giá trị dữ liệu yêu cầu mới được thêm vào danh sách các giá trị dữ liệu yêu cầu mới 147. Điều này được tiếp tục theo cách đệ quy và toàn diện cho đến khi không có thêm URL nào cần xử lý, như được mô tả bởi điều kiện 2B dưới đây.

Đối với điều kiện 2B, tất cả các trang nguồn ban đầu trong danh sách nguồn hạt giống 111 đã được kiểm tra và tất cả các dữ liệu yêu cầu mới, nếu có, từ danh sách các giá trị dữ liệu yêu cầu mới 147 cũng được kiểm tra. Quy trình 100 tiến đến quy trình con 300.

Quy trình con 300 được mô tả dưới đây, có tham khảo FIG.2. Sau khi hoàn thành quy trình con 300, quy trình 100 tiến đến bước 170, nơi quy trình 100 kết thúc.

FIG.2 là sơ đồ khối của quy trình con 300, mô tả chi tiết phương pháp phân tích và đánh giá dữ liệu phát hiện được trong phần phát hiện dữ liệu của quy trình 100 nêu trên (FIG.1), để biến đổi nó thành dữ liệu công bố được. Quy trình con 300 bắt đầu bằng tập hợp các hồ sơ dữ liệu có khả năng có liên quan nhưng không tương quan như được phát hiện trong quy trình 100. Quy trình con 300 tạo các khóa duy nhất cho mỗi phần dữ liệu và sau đó xác định mối tương quan giữa các yếu tố dữ liệu không tương quan. Qua quy trình đánh giá dựa trên các quy tắc, quy trình con 300 thiết lập mối quan hệ giữa các yếu tố dữ liệu, và nhóm các thực thể thành các nhóm dựa trên các tiêu chí định trước. Sau đó các nhóm này được tổng hợp thành một loạt các thực thể đơn duy nhất, mỗi thực thể thể hiện "cái nhìn tốt nhất" (như được xác định bởi các quy tắc kinh doanh chuyên dụng) của các giá trị yếu tố của nhóm đó.

Ở bước 301, quy trình con 300 được bắt đầu bằng một hồ sơ từ danh sách hồ sơ nguồn đơn nhất 341 được đọc dưới dạng giá trị đầu vào. Danh sách hồ sơ nguồn đơn nhất 341 là cơ sở dữ liệu của các hồ sơ có nguồn duy nhất mà được thu thập ngay khi thực hiện quy trình 100. Các hồ sơ trong danh sách này có dạng của các hồ sơ trong hồ sơ doanh nghiệp đầu vào 136 (FIG.1), có dạng, ví dụ, như sau:

Tên	Địa chỉ	Thành phố	Mã	Số điện	Ngành	Trang nguồn

thương mại			bưu điện	thoại	kinh doanh	
ABC Company	123 Đường chính	Thành phố bất kỳ	10000	999-888-1111	Hiệu bánh	abccompany.com
Abc Company		Thành phố bất kỳ		999-888-1000	Hiệu bánh/bánh mỳ	Yellowpages.com
ABC Company	Đường chính		10000		Hiệu bánh	anytownBiz.com
MyCorp	222 First Đại lộ	New York Thành phố	21212	614-111-1010		Yellowpages.com
John Doe, Inc.	Phòng A, 100 đường 4^{th}	Washington	11111	888-555-0000	Dịch vụ pháp lý	JohnDoe.com

Ở bước 302, các yếu tố dữ liệu 360, dưới dạng danh sách yếu tố dữ liệu 165 (FIG.2), từ hồ sơ doanh nghiệp trong danh sách hồ sơ doanh nghiệp 346 (danh sách các hồ sơ doanh nghiệp đơn), được đọc. Khóa so khớp được tạo ra cho mỗi trường trong yếu tố dữ liệu 360 trong danh sách hồ sơ doanh nghiệp 346. Khóa so khớp là ký hiệu nhận dạng mà chỉ nhận biết duy nhất giá trị theo ngữ cảnh của yếu tố dữ liệu. Ví dụ, "ABC Company" và "Abc Company" là các chuỗi byte không giống nhau. Tuy nhiên, theo ngữ cảnh, chúng là cùng một tên. Việc đánh giá này có thể cũng được thực hiện bằng cách kiểm tra từ đồng nghĩa, kiểu tên thay thế, hoặc các biến đổi cho phép khác. Khoá so khớp được tạo ra cho hai giá trị này có thể là giống nhau. Bảng sau thể hiện các ví dụ về các hồ sơ có khóa so khớp giống nhau và các hồ sơ có các khóa so khớp duy nhất như được chỉ ra ở cột "khóa so khớp tên".

Tên thương mại	Khóa so khớp tên
ABC Company	X
Abc Company	X
ABC Company	X
MyCorp	A
John Doe, Inc.	B

Ở bước 302, các khóa so khớp, cùng với dữ liệu gốc trong danh sách hồ sơ doanh nghiệp 346, được ghi vào danh sách khóa so khớp 342.

Ở bước 303, thực hiện việc xác định xem liệu có thêm hồ sơ nào trong danh sách hồ sơ doanh nghiệp 346 không. Nếu có thêm các hồ sơ, thì vòng lặp được thực hiện qua bước 304, và khóa so khớp được tạo ra cho hồ sơ tiếp theo. Khi thực hiện vòng lặp qua bước 304, hồ sơ tiếp theo cần đọc được đọc ở bước 301. Khi tất cả các hồ sơ trong danh sách hồ sơ nguồn duy nhất 341 đã được đọc và các khóa so khớp được tạo ra cho chúng, quy trình con 300 thoát ra khỏi vòng lặp ở bước 303, và tiến đến bước 305.

Ở bước 305, với danh sách khóa so khớp 342 làm đầu vào, các hồ sơ khớp được nhóm thành các nhóm dựa trên các khóa so khớp của chúng. Mỗi nhóm được gán một số nhận biết nhóm, được gọi là ID của nhóm. Mỗi hồ sơ được gán ID nhóm được ghi vào danh sách các hồ sơ được tạo nhóm 343 cùng với ID nhóm của nó.

Tên thương mại	Khóa so khớp	ID nhóm
ABC Company	X	Y

Abc Company	X	Y
John Doe, Inc	A	<không có ID nhóm>
MyCorp	B	<không có ID nhóm>

Trong các ví dụ nêu trên, "ABC Company" và "Abc Company" có cùng khóa so khớp do đó chúng đều được gán ID nhóm giống nhau. Cả "John Doe, Inc" hoặc "MyCorp" đều không khớp với hồ sơ khác bất kỳ. Do đó chúng không được gán ID nhóm.

Danh sách các hồ sơ được tạo nhóm 343 chứa các nhóm được tạo ra ở bước 305. Mỗi hồ sơ trong danh sách các hồ sơ được tạo nhóm 343 chứa một hồ sơ doanh nghiệp duy nhất và ID nhóm gắn với nó.

Ví dụ về hai nhóm trong danh sách các hồ sơ được phân nhóm 343:

Tên thương mại	Địa chỉ	Thành phố	Mã bưu điện	Số điện thoại	Ngành kinh doanh	Trang nguồn	ID nhóm
ABC Company	123 Đường chính	Thành phố bất kỳ	10000	999-888-1111	Hiệu bánh	abccompany.com	Y
Abc Company		Thành phố bất kỳ		999-888-1000	Hiệu bánh/bánh mỳ	Yellowpages.com	Y
ABC Company	Đường chính		10000		Hiệu bánh	Thành phố bất kỳ Biz.com	Y
SomeCorp	222 First Đại lộ	Washington	21212	888-555-0000		Yellowpages.com	Z

SomeCORP	Phòng A, 100 đường 4th	Washington		888-555-0000	Dịch vụ pháp lý	LegalSvcLists.com	Z

Quy trình con 300 tiếp theo tiến từ bước 305 đến bước 306, tại đây danh sách các hồ sơ được tạo nhóm 343 và các quy tắc ưu tiên trong danh sách các quy tắc ưu tiên 344 được nhận dưới dạng đầu vào.

Ở bước 306, tập hợp các hồ sơ được tạo nhóm được hợp nhất vào một hồ sơ tổng hợp duy nhất nhờ tác dụng đòn bẩy của các quy tắc ưu tiên trong danh sách các quy tắc ưu tiên 344. Quy trình hợp nhất này chọn giá trị mong muốn nhất cho mỗi yếu tố dữ liệu thu thập được. Quy trình hợp nhất này của bước 306 có thể thực hiện được dựa trên các phát hiện nguồn và thực thể được thực hiện trong phần phát hiện dữ liệu (FIG.1) của quy trình 100. Khi mỗi thực thể được phát hiện, trang nguồn thông tin, bao gồm siêu dữ liệu về trang đó, được giữ lại. Với mỗi bộ siêu dữ liệu mới được tìm thấy, một phần thông tin mới được phát hiện về các nguồn dữ liệu có sẵn. Siêu dữ liệu này, kết hợp với các quy tắc ưu tiên trong danh sách các quy tắc ưu tiên 344, được sử dụng để thực hiện việc đánh giá về chất lượng của dữ liệu thu thập từ nhiều nguồn khác nhau và để xác định dạng hiển thị tốt nhất của dữ liệu trong hồ sơ được tạo nhóm.

Ví dụ về siêu dữ liệu chung về các trang nguồn là dạng đuôi HTML được sử dụng để mô tả vị trí cho các công cụ tìm kiếm. Ví dụ về các đuôi HTML này để nhận biết nguồn hồ sơ có thể là "thư mục kinh doanh của doanh nghiệp", "các trang vàng", hoặc "danh bạ điện thoại trực tuyến".

Ngoài ra, các quy tắc trong danh sách các quy tắc ưu tiên 344 định rõ các tiêu chí về việc trang nào có thể được sử dụng để chứng thực tính xác thực của dữ liệu phát hiện. Dựa trên ứng dụng kinh doanh cụ thể, dựa trên danh sách các quy tắc ưu tiên 344, một số trang có thể có đủ tin cậy để xác nhận tính xác thực của một trang khác. Ví dụ, trang web của công ty điện thoại có thể được xem là nguồn có căn cứ đối với số điện thoại và có thể xác nhận tính xác thực của số điện thoại tìm thấy trên trang web phi viễn thông.

Dưới đây là một ví dụ về cách mà các quy tắc ưu tiên trong danh sách các quy tắc ưu tiên 344 có thể được sử dụng.

Ví dụ: Quy tắc 1: Trang web của công ty viễn thông đã biết sẽ được xem là căn cứ chính cho số điện thoại.

Tên thương mại	Địa chỉ	Thành phố	Mã bưu điện	Số điện thoại	Ngành kinh doanh	Trang nguồn
ABC Company	123Đường chính	Thành phố bất kỳ	10000	999-888-1111	Hiệu bánh	searchengine.com
Abc Company		Thành phố bất kỳ		999-888-1000	Hiệu bánh/bánh mỳ	Yellowpages.com

Dựa vào "Quy tắc 1" và các hồ sơ mẫu nêu trên, số điện thoại từ yellowpages.com sẽ được xem là giá trị đáng tin cậy nhất. Hồ sơ hợp nhất thu được có thể là:

Tên thương mại	Địa chỉ	Thành phố	Mã bưu điện	Số điện thoại	Ngành kinh doanh	Trang nguồn
ABC Company	123 Đường chính	Thành phố bất kỳ	10000	999-888-1000	Hiệu bánh	Yellowpages.com

Khi không có các quy tắc ưu tiên bất kỳ cho yếu tố cho trước, việc chọn lọc mặc định có thể được sử dụng trong đó quy trình hợp nhất của bước 306 chọn giá trị từ hồ sơ thứ nhất trong nhóm. Các giá trị này được ghi vào hồ sơ yếu tố dữ liệu kết hợp 365. Lưu ý rằng không nhất thiết phải có sự phân biệt về cách mà các hồ sơ được bố trí trong yếu tố dữ liệu 360 và trong hồ sơ yếu tố dữ liệu kết hợp 365. Hồ sơ yếu tố dữ liệu kết hợp 365 được ghi vào danh sách hồ sơ từ nhiều nguồn 345.

Danh sách hồ sơ từ nhiều nguồn 345 giữ tất cả các hồ sơ từ nhiều nguồn. Các hồ sơ từ nhiều nguồn là các hồ sơ mà từ đó dữ liệu doanh nghiệp được tìm thấy ở ít nhất hai trang nguồn. Ngay khi tất cả các nhóm được đọc và được xử lý ở bước 306, quy trình con 300 chuyển đến bước 307, tại đó quy trình con 300 được hoàn thành và việc kiểm tra quay lại quy trình 100. Sau bước 170 của quy trình 100, các kết quả được tạo ra bởi phương án được bộc lộ có thể được lưu trữ, ghi vào và/hoặc công bố. Ví dụ, các kết quả này có thể được lưu trữ trong cơ sở dữ liệu, được ghi vào cơ sở dữ liệu, được sử dụng để tạo ra báo cáo, hoặc công bố cho ứng dụng gọi mà được gọi là quy trình 100. Các kết quả này có thể được sử dụng cho mục đích bất kỳ hoặc tất cả các mục đích này hoặc các mục đích khác, và phương pháp sử dụng các kết quả này là chuyên dụng và không phụ thuộc vào việc sử dụng dự tính trong tương lai của các kết quả được tạo ra bởi phương án được bộc lộ.

FIG.3 là sơ đồ khối của hệ thống 400, để sử dụng sáng chế. Hệ thống 400 bao gồm máy tính 405 được nối với mạng 420, ví dụ, mạng Internet.

Máy tính 405 bao gồm giao diện người dùng 410, bộ xử lý 415, và bộ nhớ 425. Máy tính 405 có thể được chạy trên máy vi tính thông dụng. Mặc dù máy tính 405 được thể hiện ở đây là thiết bị độc lập, nó không bị giới hạn ở đó, mà thay vào đó có thể được nối với các thiết bị khác (không được thể hiện trên hình vẽ) qua mạng 420.

Bộ xử lý 415 được tạo cấu hình của hệ mạch logic đáp ứng và thực hiện các lệnh.

Bộ nhớ 425 lưu trữ dữ liệu và lệnh để điều khiển hoạt động của bộ xử lý 415. Bộ nhớ 425 có thể được cài đặt trong bộ nhớ truy cập ngẫu nhiên (RAM), ổ đĩa cứng, bộ nhớ chỉ đọc (ROM), hoặc dạng kết hợp của chúng. Một trong các thành phần của bộ nhớ 425 là môđun chương trình 430.

Môđun chương trình 430 chứa lệnh để điều khiển bộ xử lý 415 để thực hiện các phương pháp được mô tả ở đây. Ví dụ, nhờ quá trình thực thi của môđun chương trình 430, bộ xử lý 415 (a) kiểm tra dữ liệu tìm thấy trong tập hợp các mục tiêu tìm kiếm ban đầu; (b) lưu trữ trong thiết bị lưu trữ dữ liệu, từ các mục tiêu tìm kiếm ban đầu, dữ liệu phù hợp với yêu cầu, và dữ liệu mà xác định các mục tiêu tìm kiếm bổ sung cần tìm

kiếm; (c) tìm kiếm các mục tiêu tìm kiếm bổ sung đối với dữ liệu phù hợp với yêu cầu, và dữ liệu xác định các mục tiêu tìm kiếm khác cần tìm kiếm; và (d) lưu trữ trong thiết bị lưu trữ dữ liệu, từ các mục tiêu tìm kiếm khác nữa, dữ liệu phù hợp với yêu cầu và dữ liệu xác định các mục tiêu tìm kiếm khác nữa.

Thuật ngữ "môđun" được sử dụng ở đây để chỉ hoạt động chức năng mà có thể được thể hiện dưới dạng thành phần độc lập hoặc dưới dạng cấu hình tích hợp của nhiều thành phần phụ phối. Do đó, môđun chương trình 430 có thể được lắp đặt dưới dạng một môđun duy nhất hoặc dưới dạng nhiều môđun mà hoạt động kết hợp với nhau. Ngoài ra, mặc dù môđun chương trình 430 được mô tả ở đây là được cài đặt trong bộ nhớ 425, và do đó được cài đặt trong phần mềm, nó có thể được cài đặt trong phần bất kỳ trong các phần cứng (ví dụ, hệ mạch điện), phần sụn, phần mềm hoặc kết hợp của chúng.

Giao diện người dùng 410 bao gồm thiết bị đầu vào, như bàn phím hoặc hệ thống con nhận biết ngôn ngữ, để cho phép người dùng truyền đạt thông tin và các lệnh chọn đến bộ xử lý 415. Giao diện người dùng 410 còn bao gồm thiết bị đầu ra như màn hình hoặc máy in. Dạng điều khiển bằng con trỏ như chuột, bi điều khiển, hoặc cần trò chơi, cho phép người dùng điều khiển con trỏ trên màn hình để truyền đạt các thông tin và các lệnh chọn bổ sung đến bộ xử lý 415.

Bộ xử lý 415 xuất ra, đến giao diện người dùng 410, kết quả thực hiện các phương pháp được mô tả theo sáng chế. Theo cách khác, bộ xử lý 415 có thể chuyển kết quả đầu ra này đến thiết bị xa (không được thể hiện trên hình vẽ) thông qua mạng 420.

Mặc dù môđun chương trình 430 được chỉ ra là đã được tải vào bộ nhớ 425, nhưng nó có thể được cấu hình trên phương tiện lưu trữ 435 cho lần tải sau đó vào bộ nhớ 425. Phương tiện lưu trữ 435 có thể là phương tiện lưu trữ thông thường bất kỳ để lưu trữ môđun chương trình 430 trên đó ở dạng hữu hình. Ví dụ về vật ghi lưu trữ 435 bao gồm đĩa mềm, đĩa compac, băng từ, bộ nhớ chỉ đọc, vật ghi lưu trữ quang, ổ đĩa kết nối nhanh dùng chủng kết nối tuần tự đa dụng (USB), đĩa đa dụng kỹ thuật số, hoặc ổ đĩa nén. Theo cách khác, vật ghi lưu trữ 435 có thể là bộ nhớ truy cập ngẫu nhiên, hoặc các loại

lưu trữ điện khác, được bố trí trong hệ lưu trữ từ xa và kết nối với máy tính 405 thông qua mạng 420.

Do đó, phương án theo sáng chế cải thiện các sản phẩm và tính năng phát hiện hiện có, bao gồm nhưng không chỉ giới hạn ở một hoặc tất cả các loại sau: (1) quy trình xử lý các yêu cầu trước khi truy cập vào mạng lưới toàn cầu và các nguồn dữ liệu khác, (2) quy trình truy cập vào mạng lưới toàn cầu và các nguồn khác theo ngữ cảnh của yêu cầu, (3) quy trình so sánh dữ liệu yêu cầu với dữ liệu trên mạng lưới toàn cầu và các nguồn khác nhằm nhận biết, đánh giá, tổ chức và hợp nhất, phân xử, và chọn lọc các hồ sơ mà chứa dữ liệu được xác định là tương tự với yêu cầu và siêu dữ liệu liên quan đến các dữ liệu đó, (4) quy trình trong đó thông tin về quy trình phát hiện và các nguồn được giữ và tuyển lựa, (5) quy trình mà dữ liệu và siêu dữ liệu từ các nguồn được phát hiện được sử dụng để truy cập vào các nguồn dữ liệu khác thông qua quy trình đệ quy, (6) quy trình mà dữ liệu thu được có thể còn được tổ chức và hợp nhất, phân xử, và tổng hợp, bao gồm đầu tiên là chuyển vào cũng như cập nhật dữ liệu lên cơ sở dữ liệu mới hoặc đã có, và (7) quy trình trong đó các kết quả phát hiện được bao gồm dữ liệu và siêu dữ liệu được tuyển lựa để xác định sự tương đồng tương đối giữa yêu cầu và kết quả phát hiện, bao gồm dữ liệu mà đã được sử dụng để thực hiện việc xác định này, mà sau đó có thể được cung cấp cho người yêu cầu hoặc hệ thống yêu cầu để xác định có sử dụng các kết quả này hay không và sử dụng như thế nào.

Để thỏa mãn các yêu cầu này, sáng chế đề xuất phương pháp bao gồm, nhưng không giới hạn ở, một hoặc tất cả các bước sau: (a) nhận yêu cầu để bắt đầu quy trình phát hiện nhận dạng của doanh nghiệp và các thông tin liên quan, (b) xác định, dựa trên yêu cầu, chiến lược hoặc các chiến lược đệ quy để truy cập vào mạng lưới toàn cầu và các nguồn dữ liệu khác dựa trên dữ liệu cụ thể có trong yêu cầu, (c) truy cập và phân tích dữ liệu theo cách đệ quy trên mạng lưới toàn cầu và các nguồn dữ liệu khác, theo chiến lược này, đối với dữ liệu mà tương tự với yêu cầu, (d) đánh giá độ chính xác của và tuyển lựa dữ liệu và siêu dữ liệu liên quan, và (e) xuất ra kết quả dữ liệu, phản hồi, và các thông tin liên quan khác đến quy trình truy cập và xác định dữ liệu là có giá trị. Sáng chế còn đề xuất hệ thống thực hiện phương pháp này, và phương tiện lưu trữ chứa các lệnh mà điều khiển bộ xử lý để thực hiện phương pháp này.

Các kỹ thuật được mô tả trong bản mô tả này là để làm ví dụ, không nên được hiểu là hàm ý giới hạn cụ thể bất kỳ đối với sáng chế. Cần hiểu rằng các dạng thay thế, kết hợp và cải biến khác nhau có thể được tạo ra bởi những người có hiểu biết trong lĩnh vực này. Ví dụ, các bước đi kèm với các quy trình được mô tả ở đây có thể được thực hiện theo thứ tự bất kỳ, trừ khi được nói rõ hoặc quy định bởi chính các bước này. Sáng chế dự định bao gồm tất cả các dạng thay thế, cải biến và biến thể này.

YÊU CẦU BẢO HỘ

1. Vật ghi lưu trữ bất biến đọc được bằng máy tính lưu trữ các lệnh của chương trình máy tính, mà khi được thực thi bởi hệ thống máy tính, dẫn đến việc thực hiện các bước bao gồm:

a) cung cấp các mục tiêu tìm kiếm ban đầu dựa trên yêu cầu về các nguồn hạt giống;

b) tìm kiếm mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu để thu được dữ liệu và siêu dữ liệu phù hợp với yêu cầu;

c) lặp lại bước (b) cho đến khi tất cả dữ liệu và siêu dữ liệu của mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

d) tìm kiếm mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu để thu dữ liệu và siêu dữ liệu bổ sung phù hợp với yêu cầu;

e) lặp lại bước (d) cho đến khi dữ liệu và siêu dữ liệu bổ sung của mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

f) lặp lại các bước (d) và (e) cho đến khi tất cả các mục tiêu tìm kiếm ban đầu được tìm kiếm và dữ liệu và siêu dữ liệu bổ sung được thu thập và lưu trữ;

g) xử lý dữ liệu và siêu dữ liệu và dữ liệu và siêu dữ liệu bổ sung lần lượt được lưu trữ ở các bước (c) và (e), nhờ đó tạo ra các mục tiêu tìm kiếm bổ sung;

h) lặp lại các bước từ (b) đến (f) đối với các mục tiêu tìm kiếm bổ sung; và

i) lặp lại các bước (g) và (h) cho đến khi không có dữ liệu và siêu dữ liệu bổ sung được tìm thấy; và hợp thức hóa dữ liệu cho trước bằng cách thực hiện các bước gồm:

so sánh dữ liệu của các mục tiêu tìm kiếm đã được tìm kiếm, và

chọn dữ liệu từ nguồn được cho là tin cậy và hữu dụng nhất làm dữ liệu hợp lệ, dựa trên tập hợp các quy tắc ưu tiên và sử dụng.

2. Phương pháp tìm kiếm dữ liệu phù hợp với yêu cầu, bao gồm bước sử dụng máy tính để thực hiện các bước:

a) cung cấp các mục tiêu tìm kiếm ban đầu dựa trên yêu cầu về các nguồn hạt giống;

b) tìm kiếm mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu để thu được dữ liệu và siêu dữ liệu phù hợp với yêu cầu;

c) lặp lại bước (b) cho đến khi tất cả dữ liệu và siêu dữ liệu của mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

d) tìm kiếm mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu để thu dữ liệu và siêu dữ liệu bổ sung phù hợp với yêu cầu;

e) lặp lại bước (d) cho đến khi tất cả dữ liệu và siêu dữ liệu bổ sung của mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

f) lặp lại các bước (d) và (e) cho đến khi tất cả các mục tiêu tìm kiếm ban đầu được tìm kiếm và dữ liệu và siêu dữ liệu bổ sung được thu thập và lưu trữ;

g) xử lý dữ liệu và siêu dữ liệu và dữ liệu và siêu dữ liệu bổ sung được lưu trữ ở các bước (c) và (e), nhờ đó tạo ra các mục tiêu tìm kiếm bổ sung;

h) lặp lại các bước từ (b) đến (f) cho các mục tiêu tìm kiếm bổ sung; và

i) lặp lại các bước (g) và (h) cho đến khi không có dữ liệu và siêu dữ liệu bổ sung được tìm thấy; và

hợp thức hóa dữ liệu xác định bằng cách thực hiện các bước gồm:

so sánh dữ liệu từ các mục tiêu tìm kiếm đã được tìm kiếm, và

chọn dữ liệu từ nguồn được cho là tin cậy nhất và hữu dụng nhất làm dữ liệu hợp lệ, dựa trên tập hợp các quy tắc ưu tiên và sử dụng.

3. Phương pháp theo điểm 2, trong đó máy tính được tạo cấu hình để tìm kiếm các địa chỉ trang mạng hoặc các nguồn khác và danh sách các nguồn hạt giống là danh sách các địa chỉ trang mạng hoặc các nguồn khác.

4. Phương pháp theo điểm 2, trong đó phương pháp này còn bao gồm bước làm sạch dữ liệu thu được từ mục tiêu tìm kiếm.

5. Phương pháp theo điểm 4, trong đó bước làm sạch dữ liệu được thực hiện bằng các bước bao gồm ít nhất một bước loại bỏ các giá trị lỗi đối với dữ liệu và loại bỏ các mã thông báo định trước ra khỏi dữ liệu.

6. Phương pháp theo điểm 2, trong đó phương pháp này còn bao gồm bước tổ chức và hợp nhất, phân xử, tổng hợp và tạo nhóm dữ liệu liên quan từ các nguồn khác nhau để tạo ra hồ sơ dữ liệu được tạo nhóm.

7. Phương pháp theo điểm 6, trong đó phương pháp này còn bao gồm bước tạo các hồ sơ dữ liệu đa nguồn kết hợp từ tập hợp các hồ sơ dữ liệu được tạo nhóm.

8. Phương pháp theo điểm 2, trong đó phương pháp này còn bao gồm bước thực hiện ít nhất một bước được chọn từ nhóm gồm viết vào cơ sở dữ liệu, lưu trữ trong cơ sở dữ liệu, tạo báo cáo và công bố kết quả được tìm thấy nhờ tìm kiếm dữ liệu phù hợp với yêu cầu.

9. Phương pháp theo điểm 2, trong đó phương pháp này còn bao gồm bước áp dụng phương pháp phân tích trong số ít nhất một phương pháp được chọn từ nhóm gồm các quy tắc, thuật toán, suy nghiệm và các hàm phân tích khác đưa ra quyết định liên quan đến dữ liệu và quyết định xem có tiếp tục hay kết thúc phương pháp này.

10. Phương pháp theo điểm 2, trong đó máy tính thực hiện các bước còn bao gồm:

đọc dữ liệu trong mục tiêu tìm kiếm ban đầu và mục tiêu tìm kiếm bổ sung bất kỳ;

so sánh dữ liệu với các quy tắc trong danh sách quy tắc xử lý, lưu trữ dữ liệu phù hợp với các quy tắc này dưới dạng các hồ sơ doanh nghiệp đầu vào;

phân tích các hồ sơ doanh nghiệp đầu vào này nhờ sử dụng dữ liệu trong danh sách dữ liệu tham chiếu để tạo ra các hồ sơ doanh nghiệp đầu ra;

lưu trữ các hồ sơ doanh nghiệp đầu ra gồm nhiều phần dữ liệu.

11. Phương pháp theo điểm 10, trong đó máy tính còn thực hiện các bước bao gồm:

tạo ra các khóa duy nhất cho mỗi phần dữ liệu;

xác định mối tương quan giữa các phần dữ liệu không tương quan trong các hồ sơ doanh nghiệp đầu ra khác nhau;

tạo các khóa so khớp để xác định giá trị theo ngữ cảnh của mỗi phần dữ liệu;

nhóm các phần dữ liệu so khớp thành các nhóm dựa vào các khóa so khớp;

tổng hợp các nhóm này thành dãy các hồ sơ kết hợp theo quy tắc ưu tiên trong danh sách quy tắc ưu tiên.

12. Phương pháp theo điểm 11, trong đó máy tính thực hiện bước còn bao gồm:

sử dụng các quy tắc ưu tiên để xác định tính xác thực của dữ liệu hoặc siêu dữ liệu thu được từ một trang, dựa trên dữ liệu hoặc siêu dữ liệu thu được từ một trang khác.

13. Hệ thống máy tính có bộ xử lý và bộ nhớ, bộ nhớ có các lệnh khiến cho bộ xử lý tìm kiếm dữ liệu phù hợp với yêu cầu, bằng cách thực hiện các bước bao gồm:

a) cung cấp các mục tiêu tìm kiếm ban đầu dựa trên yêu cầu về các nguồn hạt giống;

b) tìm kiếm mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu để thu được dữ liệu và siêu dữ liệu phù hợp với yều cầu;

c) lặp lại bước (b) cho đến khi tất cả dữ liệu và siêu dữ liệu của mục tiêu thứ nhất trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

d) tìm kiếm mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu để thu dữ liệu và siêu dữ liệu bổ sung phù hợp với yêu cầu;

e) lặp lại bước (d) cho đến khi tất cả dữ liệu và siêu dữ liệu bổ sung của mục tiêu tiếp theo trong số các mục tiêu tìm kiếm ban đầu được thu thập và lưu trữ;

f) lặp lại các bước (d) và (e) cho đến khi tất cả các mục tiêu tìm kiếm ban đầu đã được tìm kiếm và dữ liệu và siêu dữ liệu bổ sung được thu thập và lưu trữ;

g) xử lý dữ liệu và siêu dữ liệu và dữ liệu và siêu dữ liệu bổ sung lần lượt được lưu trữ ở các bước (c) và (e), nhờ đó tạo ra các mục tiêu tìm kiếm bổ sung;

h) lặp lại các bước từ b) đến f) đối với các mục tiêu tìm kiếm bổ sung;

i) lặp lại các bước (g) và (h) cho đến khi không có dữ liệu và siêu dữ liệu bổ sung nào được tìm thấy; và

j) hợp thức hóa dữ liệu xác định bằng cách thực hiện các bước gồm:

so sánh dữ liệu của các mục tiêu tìm kiếm mà đã được tìm kiếm, và

chọn dữ liệu từ nguồn được cho là tin cậy và hữu dụng nhất làm dữ liệu hợp lệ, dựa trên tập hợp các quy tắc ưu tiên và sử dụng.

14. Hệ thống theo điểm 13, trong đó máy tính được tạo cấu hình để tìm kiếm các trang web hoặc các nguồn khác và danh sách các nguồn hạt giống là danh sách của trang web hoặc các nguồn khác.

15. Hệ thống theo điểm 13, trong đó máy tính được tạo cấu hình để làm sạch dữ liệu thu được từ mỗi mục tiêu tìm kiếm.

16. Hệ thống theo điểm 15, trong đó máy tính được tạo cấu hình để làm sạch dữ liệu nhờ các bước gồm ít nhất một bước loại bỏ các giá trị lỗi hoặc không thích hợp đối với dữ liệu và loại bỏ các mã thông báo định trước ra khỏi dữ liệu.

17. Hệ thống theo điểm 13, trong đó hệ thống còn bao gồm bộ xử lý được tạo cấu hình để tổ chức và hợp nhất, phân xử, tổng hợp và tạo nhóm dữ liệu liên quan từ các mục tiêu tìm kiếm khác nhau tạo ra các hồ sơ dữ liệu được tạo nhóm.

18. Hệ thống theo điểm 17, trong đó bộ xử lý được tạo cấu hình để tạo ra hồ sơ dữ liệu kết hợp từ các hồ sơ dữ liệu được tạo nhóm.

19. Hệ thống theo điểm 13, trong đó máy tính thực hiện các bước gồm:

đọc dữ liệu trong mục tiêu tìm kiếm ban đầu và mục tiêu tìm kiếm bổ sung bất kỳ;

so sánh dữ liệu với các quy tắc trong danh sách quy tắc xử lý, lưu trữ dữ liệu phù hợp với các quy tắc dưới dạng hồ sơ doanh nghiệp đầu vào;

phân tích các hồ sơ doanh nghiệp đầu vào này nhờ sử dụng dữ liệu trong danh sách dữ liệu tham chiếu để tạo ra các hồ sơ doanh nghiệp đầu ra;

lưu trữ các hồ sơ doanh nghiệp đầu ra gồm các phần dữ liệu.

20. Hệ thống theo điểm 19, trong đó máy tính thực hiện các bước bao gồm:

tạo các khóa duy nhất cho mỗi phần dữ liệu;

xác định mối tương quan giữa các phần dữ liệu không tương quan trong các hồ sơ doanh nghiệp đầu ra khác nhau;

tạo ra khóa so khớp để nhận dạng giá trị ngữ cảnh của mỗi phần dữ liệu;

nhóm các phần dữ liệu so khớp thành các nhóm dựa trên các khóa so khớp;

tổng hợp các nhóm này thành dãy các hồ sơ kết hợp theo quy tắc ưu tiên trong danh sách quy tắc ưu tiên.

21. Hệ thống theo điểm 20, trong đó máy tính này còn thực hiện bước: sử dụng các quy tắc ưu tiên để xác định tính xác thực của dữ liệu hoặc siêu dữ liệu thu được từ một trang, dựa trên dữ liệu hoặc siêu dữ liệu từ một trang khác.

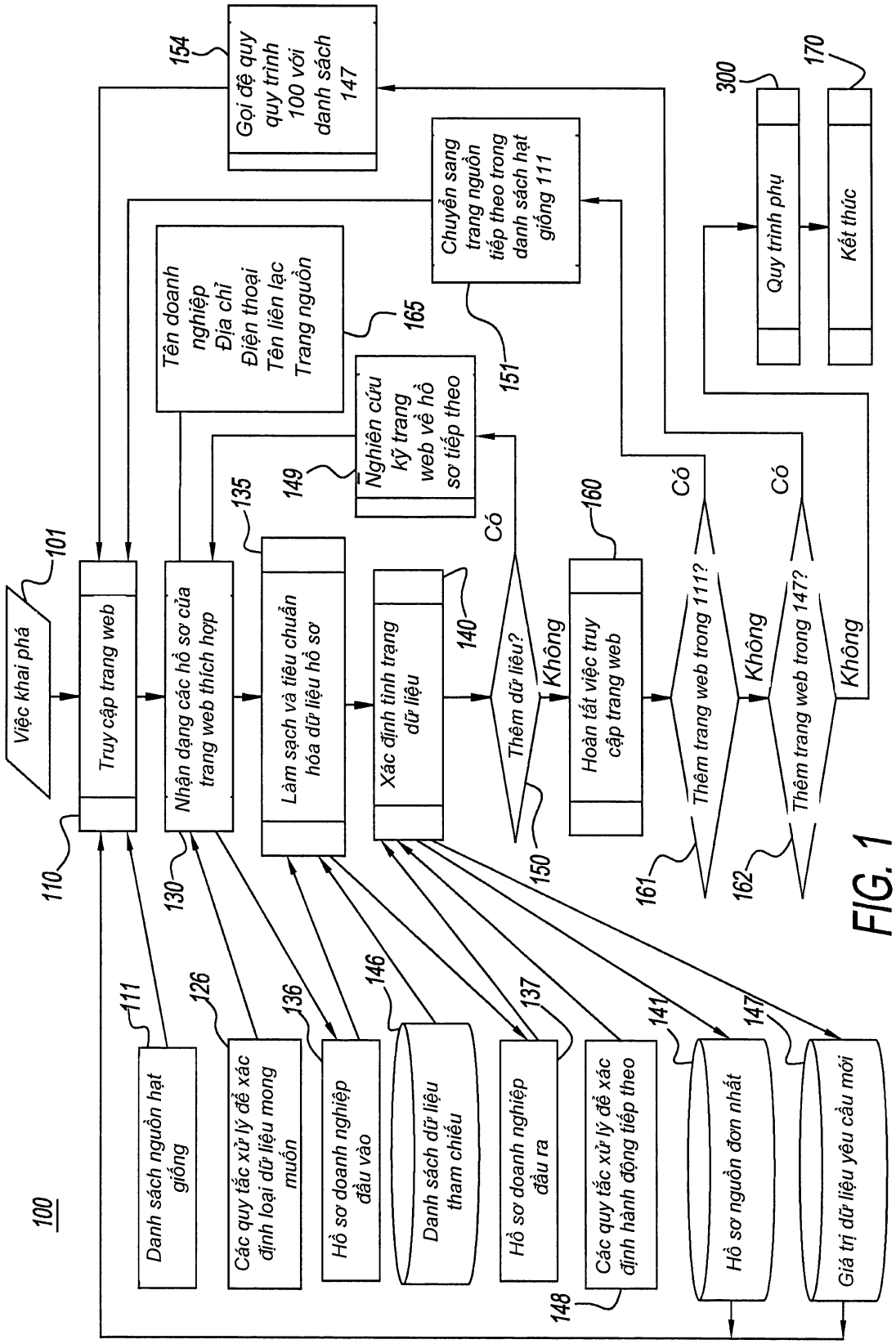

100
Việc khai phá
101
Truy cập trang web
110
Nhận dạng các hồ sơ của trang web thích hợp
130
Làm sạch và tiêu chuẩn hóa dữ liệu hồ sơ
135
Xác định tình trạng dữ liệu
140
Thêm dữ liệu?
150
Có
Không
Hoàn tất việc truy cập trang web
160
Thêm trang web trong 111?
161
Có
Không
Thêm trang web trong 147?
162
Có
Không
Gọi đệ quy quy trình 100 với danh sách 147
154
Chuyển sang trang nguồn tiếp theo trong danh sách hạt giống 111
151
Tên doanh nghiệp
Địa chỉ
Điện thoại
Tên liên lạc
Trang nguồn
165
Nghiên cứu kỹ trang web về hồ sơ tiếp theo
149
Quy trình phụ
300
Kết thúc
170
Danh sách nguồn hạt giống
111
Các quy tắc xử lý để xác định loại dữ liệu mong muốn
126
Hồ sơ doanh nghiệp đầu vào
136
Danh sách dữ liệu tham chiếu
146
Hồ sơ doanh nghiệp đầu ra
137
Các quy tắc xử lý để xác định hành động tiếp theo
148
Hồ sơ nguồn đơn nhất
141
Giá trị dữ liệu yêu cầu mới
147

FIG. 1

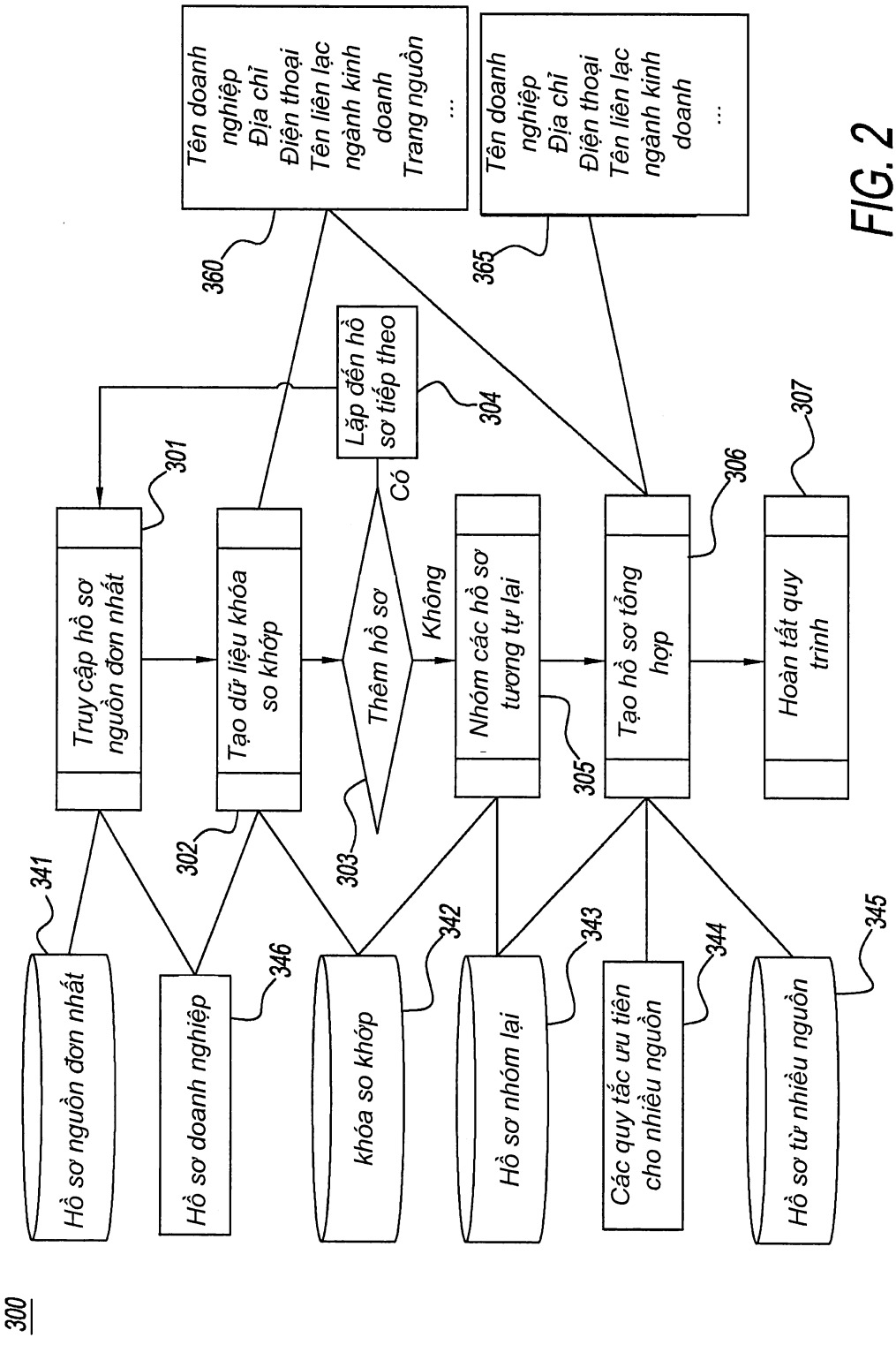

300
Hồ sơ nguồn đơn nhất
Hồ sơ doanh nghiệp
khóa so khớp
Hồ sơ nhóm lại
Các quy tắc ưu tiên cho nhiều nguồn
Hồ sơ từ nhiều nguồn
341
346
342
343
344
345
Truy cập hồ sơ nguồn đơn nhất
Tạo dữ liệu khóa so khớp
Thêm hồ sơ
Có
Không
Lặp đến hồ sơ tiếp theo
Nhóm các hồ sơ tương tự lại
Tạo hồ sơ tổng hợp
Hoàn tất quy trình
301
302
303
304
305
306
307
Tên doanh nghiệp
Địa chỉ
Điện thoại
Tên liên lạc
ngành kinh doanh
Trang nguồn
...
360
Tên doanh nghiệp
Địa chỉ
Điện thoại
Tên liên lạc
ngành kinh doanh
...
365

FIG. 2

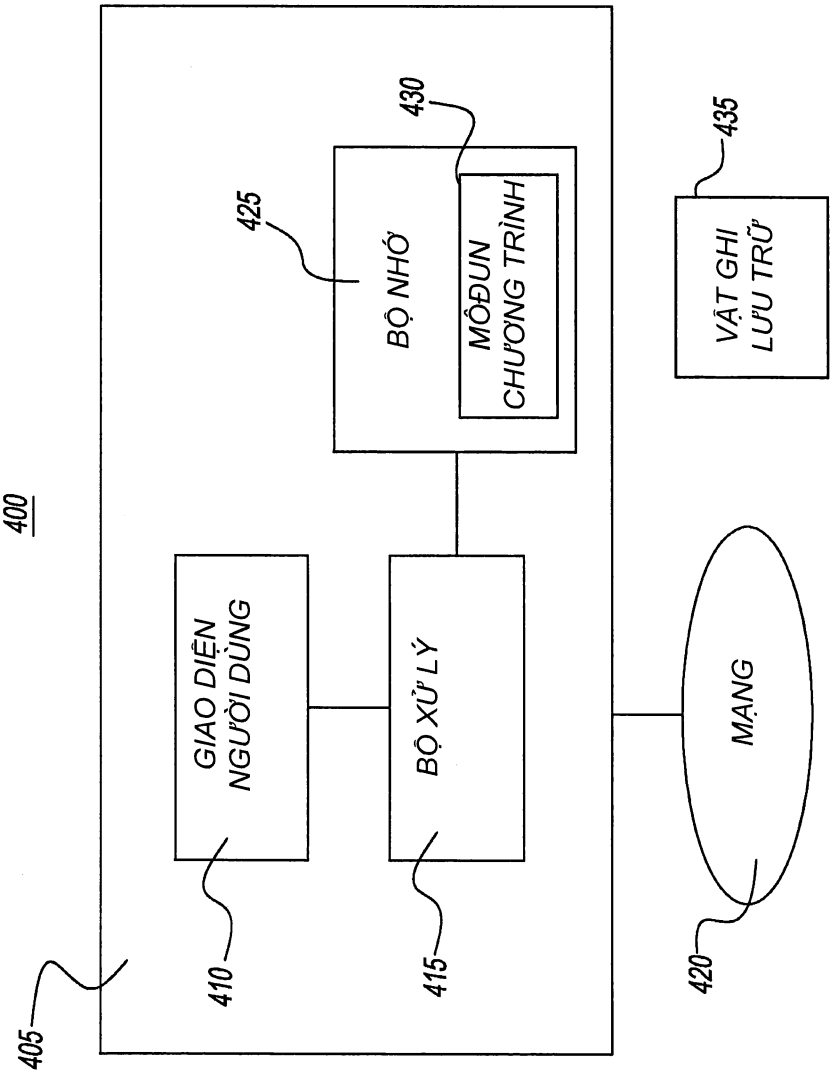

400
405
410
GIAO DIỆN NGƯỜI DÙNG
415
BỘ XỬ LÝ
420
MẠNG
425
BỘ NHỚ
430
MÔĐUN CHƯƠNG TRÌNH
435
VẬT GHI LƯU TRỮ

FIG. 3