May 2023

SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR AUTOMATICALLY SCRAPING CATEGORICAL DATA FROM A PLURALITY OF WEBSITES

PATRICK RYAN FLANAGAN
VISA

PIERS WILLIAM BUTLER CLOUGH
VISA

.

“SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR AUTOMATICALLY SCRAPING CATEGORICAL DATA FROM A PLURALITY OF WEBSITES”

VISA

PATRICK RYAN FLANAGAN
PIERS WILLIAM BUTLER CLOUGH

TECHNICAL FIELD

[0001] This disclosure relates generally to scraping data from websites and, in some non-limiting embodiments or aspects, to systems, methods, and computer program products for automatically scraping categorical data from a plurality of websites.

BACKGROUND

[0002] The vast and distributed nature of the internet makes it such that it can be difficult for users to find a comprehensive compilation of categorical data in a single location. Instead, users must navigate the myriad internet website in order to try and find the data of interest. This makes web searches for a comprehensive compilation of categorical data burdensome for users.

SUMMARY

[0003] Accordingly, it is an object of the present disclosure to provide systems, methods, and computer program products for automatically scraping categorical data from a plurality of websites that overcome some or all of the deficiencies identified above.

[0004] According to non-limiting embodiments or aspects, provided is a computer-implemented method for automatically scraping categorical data from a plurality of websites including: determining a product category; identifying a first website including data associated with the product category; automatically scraping, with at least one processor, the first website to compile first product data associated with the product category; generating, with at least one processor, a plurality of web queries based on the compiled first product data; executing, with at least one processor, the plurality of web queries to identify a plurality of websites; automatically scraping, with at least one processor, at least a portion of the plurality of websites to compile supplier data associated with suppliers in the product category; and storing, with at least one processor, at least a portion of the compiled supplier data in a database.

[0005] In some non-limiting embodiments or aspects, the method may include: generating and exposing, with at least one processor, an application programming interface (API) enabling user querying of the database; receiving, by the API and from a user device, a first query associated

with the product category; querying, by the API, the database based on the first query to generate a first query result; and transmitting, by the API, the first query result to cause the first query result to be displayed on the user device. The first website may include a website of a distributor of products associated with the product category. The method may include analyzing, with a machine learning model, the compiled first product data to generate the plurality of web queries. The method may include: determining, with at least one processor, a subset of websites associated with the product category from the plurality of websites based on a plurality of rules; and where the automatically scraping the at least a portion of the plurality of websites includes automatically scraping the subset of websites to compile the supplier data. The plurality of web queries may be executed using a search engine. The automatic scraping of the first website and/or the at least a portion of the plurality of websites may be performed using a web crawler.

[0006] In some non-limiting embodiments or aspects, the method may include: determining a second product category; identifying a second website including data associated with the second product category; automatically scraping, with at least one processor, the second website to compile second product data associated with the second product category; generating, with at least one processor, a second plurality of web queries based on the compiled second product data; executing, with at least one processor, the second plurality of web queries to identify a second plurality of websites; automatically scraping, with at least one processor, at least a portion of the second plurality of websites to compile second supplier data associated with suppliers in the second product category; and storing, with at least one processor, at least a portion of the compiled second supplier data in the database.

[0007] According to non-limiting embodiments or aspects, provided is a system including: at least one processor; and at least one non-transitory computer-readable medium storing instructions that, when executed by the at least one processor, cause the at least one processor to perform any of the methods described herein.

[0008] According to non-limiting embodiments or aspects, provided is a computer program product including at least one non-transitory computer-readable medium including program instructions that, when executed by at least one processor, cause the at least one processor to perform any of the methods described herein.

[0009]

In an embodiment of the disclosure, a computer-implemented method for automatically scraping categorical data from a plurality of websites, comprises: determining a product category; identifying a first website comprising data associated with the product category; automatically scraping, with at least one processor, the first website to compile first product data associated with the product category; generating, with at least one processor, a plurality of web queries based on the compiled first product data; executing, with at least one processor, the plurality of web queries to identify a plurality of websites; automatically scraping, with at least one processor, at least a portion of the plurality of websites to compile supplier data associated with suppliers in the product category; and storing, with at least one processor, at least a portion of the compiled supplier data in a database.

[0010] In an embodiment of the disclosure, the method comprises: generating and exposing, with at least one processor, an application programming interface (API) enabling user querying of the database; receiving, by the API and from a user device, a first query associated with the product category; querying, by the API, the database based on the first query to generate a first query result; and transmitting, by the API, the first query result to cause the first query result to be displayed on the user device.

[0011] In an embodiment of the disclosure, the first website comprises a website of a distributor of products associated with the product category.

[0012] In an embodiment of the disclosure, the method comprises: analyzing, with a machine learning model, the compiled first product data to generate the plurality of web queries.

[0013] In an embodiment of the disclosure, the method comprises: determining, with at least one processor, a subset of websites associated with the product category from the plurality of websites based on a plurality of rules; and wherein the automatically scraping the at least a portion of the plurality of websites comprises automatically scraping the subset of websites to compile the supplier data.

[0014] In an embodiment of the disclosure, the plurality of web queries are executed using a

search engine.

[0015] In an embodiment of the disclosure, the automatic scraping of the first website and/or the at least a portion of the plurality of websites is performed using a web crawler.

[0016] In an embodiment of the disclosure, the method comprises determining a second product category; identifying a second website comprising data associated with the second product category; automatically scraping, with at least one processor, the second website to compile second product data associated with the second product category; generating, with at least one processor, a second plurality of web queries based on the compiled second product data; executing, with at least one processor, the second plurality of web queries to identify a second plurality of websites; automatically scraping, with at least one processor, at least a portion of the second plurality of websites to compile second supplier data associated with suppliers in the second product category; and storing, with at least one processor, at least a portion of the compiled second supplier data in the database.

[0017] In an embodiment of the disclosure, a system is disclosed which comprises: at least one processor; and at least one non-transitory computer-readable medium storing instructions that, when executed by the at least one processor, cause the at least one processor to perform the method of any of the preceding embodiments.

[0018] In an embodiment of the disclosure, a computer program product is disclosed which comprises at least one non-transitory computer-readable medium including program instructions that, when executed by at least one processor, cause the at least one processor to perform the method of the present disclosure.

[0019] These and other features and characteristics of the present disclosure, as well as the methods of operation and functions of the related elements of structures and the combination of parts and economies of manufacture, will become more apparent upon consideration of the following description with reference to the accompanying drawings and appendix, all of which form a part of this specification, wherein like reference numerals designate corresponding parts in the various figures. It is to be expressly understood, however, that the drawings and appendix are for the purpose of illustration and description only and are not intended as a definition of the limits

of the disclosed subject matter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0001] Additional advantages and details of non-limiting embodiments are explained in greater detail below with reference to the exemplary embodiments that are illustrated in the accompanying schematic figures, in which:

[0002] **FIG. 1** is a schematic diagram of a system for automatically scraping categorical data from a plurality of websites according to some non-limiting embodiments or aspects;

[0003] **FIG. 2** is a flow diagram for a method of automatically scraping categorical data from a plurality of websites according to some non-limiting embodiments or aspects;

[0004] **FIG. 3** is a diagram of an exemplary environment in which methods, systems, and/or computer program products, described herein, may be implemented according to some non-limiting embodiments or aspects; and

[0005] **FIG. 4** is a schematic diagram of example components of one or more devices of FIG. 1 and/or FIG. 3 according to some non-limiting embodiments or aspects; and

[0006] Appendix A includes additional details regarding methods, systems, and computer program products for automatically scraping categorical data from a plurality of websites according to some non-limiting embodiments or aspects.

DESCRIPTION OF THE DISCLOSURE

[0007] For purposes of the description hereinafter, the terms “end,” “upper,” “lower,” “right,” “left,” “vertical,” “horizontal,” “top,” “bottom,” “lateral,” “longitudinal,” and derivatives thereof shall relate to the embodiments as they are oriented in the drawing figures. However, it is to be understood that the embodiments may assume various alternative variations and step sequences, except where expressly specified to the contrary. It is also to be understood that the specific devices and processes illustrated in the attached drawings and appendix, and described in the following specification, are simply exemplary embodiments or aspects of the disclosed subject matter.

Hence, specific dimensions and other physical characteristics related to the embodiments or aspects disclosed herein are not to be considered as limiting.

[0008] No aspect, component, element, structure, act, step, function, instruction, and/or the like used herein should be construed as critical or essential unless explicitly described as such. Also, as used herein, the articles "a" and "an" are intended to include one or more items and may be used interchangeably with "one or more" and "at least one." Furthermore, as used herein, the term "set" is intended to include one or more items (e.g., related items, unrelated items, a combination of related and unrelated items, and/or the like) and may be used interchangeably with "one or more" or "at least one." Where only one item is intended, the term "one" or similar language is used. Also, as used herein, the terms "has," "have," "having," or the like are intended to be open-ended terms. Further, the phrase "based on" is intended to mean "based at least partially on" unless explicitly stated otherwise.

[0009] As used herein, the term "acquirer institution" may refer to an entity licensed and/or approved by a transaction service provider to originate transactions (e.g., payment transactions) using a payment device associated with the transaction service provider. The transactions the acquirer institution may originate may include payment transactions (e.g., purchases, original credit transactions (OCTs), account funding transactions (AFTs), and/or the like). In some non-limiting embodiments or aspects, an acquirer institution may be a financial institution, such as a bank. As used herein, the term "acquirer system" may refer to one or more computing devices operated by or on behalf of an acquirer institution, such as a server computer executing one or more software applications.

[0010] As used herein, the term "account identifier" may include one or more primary account numbers (PANs), tokens, or other identifiers associated with a customer account. The term "token" may refer to an identifier that is used as a substitute or replacement identifier for an original account identifier, such as a PAN. Account identifiers may be alphanumeric or any combination of characters and/or symbols. Tokens may be associated with a PAN or other original account identifier in one or more data structures (e.g., one or more databases, and/or the like) such that they may be used to conduct a transaction without directly using the original account identifier. In some examples, an original account identifier, such as a PAN, may be associated with a plurality of

tokens for different individuals or purposes.

[0011] An "application program interface" (API) refers to computer code or other data sorted on a computer-readable medium that may be executed by a processor to facilitate the interaction between software components, such as a client-side front-end and/or server-side back-end for receiving data from the client. An "interface" refers to a generated display, such as one or more graphical user interfaces (GUIs) with which a user may interact, either directly or indirectly (e.g., through a keyboard, mouse, etc.).

[0012] As used herein, the term "communication" may refer to the reception, receipt, transmission, transfer, provision, and/or the like of data (e.g., information, signals, messages, instructions, commands, and/or the like). For one unit (e.g., a device, a system, a component of a device or system, combinations thereof, and/or the like) to be in communication with another unit means that the one unit is able to directly or indirectly receive information from and/or transmit information to the other unit. This may refer to a direct or indirect connection (e.g., a direct communication connection, an indirect communication connection, and/or the like) that is wired and/or wireless in nature. Additionally, two units may be in communication with each other even though the information transmitted may be modified, processed, relayed, and/or routed between the first and second unit. For example, a first unit may be in communication with a second unit even though the first unit passively receives information and does not actively transmit information to the second unit. As another example, a first unit may be in communication with a second unit if at least one intermediary unit processes information received from the first unit and communicates the processed information to the second unit.

[0013] As used herein, the term "computing device" may refer to one or more electronic devices configured to process data. A computing device may, in some examples, include the necessary components to receive, process, and output data, such as a processor, a display, a memory, an input device, a network interface, and/or the like. A computing device may be a mobile device. As an example, a mobile device may include a cellular phone (e.g., a smartphone or standard cellular phone), a portable computer, a wearable device (e.g., watches, glasses, lenses, clothing, and/or the like), a personal digital assistant (PDA), and/or other like devices. A computing device may also be a desktop computer or other form of non-mobile computer.

[0014] As used herein, the terms "electronic wallet" and "electronic wallet application" refer to one or more electronic devices and/or software applications configured to initiate and/or conduct payment transactions. For example, an electronic wallet may include a mobile device executing an electronic wallet application, and may further include server-side software and/or databases for maintaining and providing transaction data to the mobile device. An "electronic wallet provider" may include an entity that provides and/or maintains an electronic wallet for a customer, such as Google Pay®, Android Pay®, Apple Pay®, Samsung Pay®, and/or other like electronic payment systems. In some non-limiting examples, an issuer bank may be an electronic wallet provider.

[0015] As used herein, the term "issuer institution" may refer to one or more entities, such as a bank, that provide accounts to customers for conducting transactions (e.g., payment transactions), such as initiating credit and/or debit payments. For example, an issuer institution may provide an account identifier, such as a PAN, to a customer that uniquely identifies one or more accounts associated with that customer. The account identifier may be embodied on a portable financial device, such as a physical financial instrument, e.g., a payment card, and/or may be electronic and used for electronic payments. The term "issuer system" refers to one or more computer devices operated by or on behalf of an issuer institution, such as a server computer executing one or more software applications. For example, an issuer system may include one or more authorization servers for authorizing a transaction.

[0016] As used herein, the term "merchant" may refer to an individual or entity that provides goods and/or services, or access to goods and/or services, to customers based on a transaction, such as a payment transaction. The term "merchant" or "merchant system" may also refer to one or more computer systems operated by or on behalf of a merchant, such as a server computer executing one or more software applications.

[0017] As used herein, a "point-of-sale (POS) device" may refer to one or more devices, which may be used by a merchant to conduct a transaction (e.g., a payment transaction) and/or process a transaction. For example, a POS device may include one or more client devices. Additionally or alternatively, a POS device may include peripheral devices, card readers, scanning devices (e.g., code scanners), Bluetooth® communication receivers, near-field communication (NFC) receivers,

radio frequency identification (RFID) receivers, and/or other contactless transceivers or receivers, contact-based receivers, payment terminals, and/or the like. As used herein, a "point-of-sale (POS) system" may refer to one or more client devices and/or peripheral devices used by a merchant to conduct a transaction. For example, a POS system may include one or more POS devices and/or other like devices that may be used to conduct a payment transaction. In some non-limiting embodiments or aspects, a POS system (e.g., a merchant POS system) may include one or more server computers programmed or configured to process online payment transactions through webpages, mobile applications, and/or the like.

[0018] As used herein, the terms "client" and "client device" may refer to one or more client-side devices or systems (e.g., remote from a transaction service provider) used to initiate or facilitate a transaction (e.g., a payment transaction). As an example, a "client device" may refer to one or more POS devices used by a merchant, one or more acquirer host computers used by an acquirer, one or more mobile devices used by a user, and/or the like. In some non-limiting embodiments or aspects, a client device may be an electronic device configured to communicate with one or more networks and initiate or facilitate transactions. For example, a client device may include one or more computers, portable computers, laptop computers, tablet computers, mobile devices, cellular phones, wearable devices (e.g., watches, glasses, lenses, clothing, and/or the like), PDAs, and/or the like. Moreover, a "client" may also refer to an entity (e.g., a merchant, an acquirer, and/or the like) that owns, utilizes, and/or operates a client device for initiating transactions (e.g., for initiating transactions with a transaction service provider).

[0019] As used herein, the term "payment device" may refer to a payment card (e.g., a credit or debit card), a gift card, a smartcard, smart media, a payroll card, a healthcare card, a wristband, a machine-readable medium containing account information, a keychain device or fob, an RFID transponder, a retailer discount or loyalty card, a cellular phone, an electronic wallet mobile application, a personal digital assistant (PDA), a pager, a security card, a computing device, an access card, a wireless terminal, a transponder, and/or the like. In some non-limiting embodiments or aspects, the payment device may include volatile or non-volatile memory to store information (e.g., an account identifier, a name of the account holder, and/or the like).

[0020] As used herein, the term "payment gateway" may refer to an entity and/or a payment

processing system operated by or on behalf of such an entity (e.g., a merchant service provider, a payment service provider, a payment facilitator, a payment facilitator that contracts with an acquirer, a payment aggregator, and/or the like), which provides payment services (e.g., transaction service provider payment services, payment processing services, and/or the like) to one or more merchants. The payment services may be associated with the use of payment devices managed by a transaction service provider. As used herein, the term "payment gateway system" may refer to one or more computer systems, computer devices, servers, groups of servers, and/or the like, operated by or on behalf of a payment gateway.

[0021] As used herein, the term "server" may refer to or include one or more computing devices that are operated by or facilitate communication and processing for multiple parties in a network environment, such as the internet, although it will be appreciated that communication may be facilitated over one or more public or private network environments and that various other arrangements are possible. Further, multiple computing devices (e.g., servers, point-of-sale (POS) devices, mobile devices, etc.) directly or indirectly communicating in the network environment may constitute a "system." Reference to "a server" or "a processor," as used herein, may refer to a previously-recited server and/or processor that is recited as performing a previous step or function, a different server and/or processor, and/or a combination of servers and/or processors. For example, as used in the specification, a first server and/or a first processor that is recited as performing a first step or function may refer to the same or different server and/or a processor recited as performing a second step or function.

[0022] As used herein, the term "transaction service provider" may refer to an entity that receives transaction authorization requests from merchants or other entities and provides guarantees of payment, in some cases through an agreement between the transaction service provider and an issuer institution. For example, a transaction service provider may include a payment network such as Visa® or any other entity that processes transactions. The term "transaction processing system" may refer to one or more computer systems operated by or on behalf of a transaction service provider, such as a transaction processing server executing one or more software applications. A transaction processing server may include one or more processors and, in some non-limiting embodiments or aspects, may be operated by or on behalf of a transaction service provider.

[0023] Non-limiting embodiments or aspects of the disclosed subject matter are directed to systems, methods, and computer program products for automatically scraping categorical data from a plurality of websites. For example, non-limiting embodiments or aspects of the disclosed subject matter automatically scrape internet websites to compile data associated with a product category and store the scraped data in a central database. This enables users to efficiently query the central database for data related to the product categories without attempting to find the data themselves by searching the myriad of internet websites over which the data may originally have been distributed. An API may be generated to interface the user device and the central database to facilitate the querying and return of query results.

[0024] Non-limiting embodiments or aspects of the systems, methods, and computer program products automatically scrape the relevant data associated with the product categories by identifying and scraping a first website comprising data associated with the product category. Based on the data scraped from the first website, non-limiting embodiments or aspects generate a plurality of web queries and execute those web queries to identify a plurality of websites potentially relevant to the product category. The web queries may be generated using machine learning techniques to optimize the queries to those most likely to generate the relevant results. At least a portion of these potentially relevant websites are scraped to compile supplier data associated with the product category, and at least a portion of this supplier data is stored in the central database. The portion of the websites to be scraped may be determined using a plurality of rules which save processing resources by causing the system to only scrape the websites most likely to have the irrelevant information, thereby avoiding scraping websites less likely to have the relevant information. These features of the systems, methods, and computer program products enable the relevant data associated with the product category to be compiled and centralized in a queryable manner to allow users to search for data associated with product categories more efficiently.

[0025] Referring to FIG. 1, shown is a system 100 for automatically scraping categorical data from a plurality of websites according to some non-limiting embodiments or aspects. The system 100 may include a scraper system 102, a communication network 104, a central database 106, an API 108, and a user device 110.

[0026] The scraper system 102 may include a computing device, such as a server (e.g., a single server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, the scraper system 102 may include a processor and/or memory. The scraper system 102 may comprise a web crawler and/or a web scraper to crawl and/or scrape internet websites (e.g., web content and data) as described herein. The internet websites may contain structured and/or unstructured data.

[0027] The communication network 104 may refer to any communication network, such as one or more wired and/or wireless networks. For example, the communication network may include a private network (e.g., a private network associated with a transaction service provider), an ad hoc network, an intranet, the internet, and/or the like, and/or a combination of these or other types of networks. The communication network 104 may refer to the internet.

[0028] The central database 106 may include memory for storing data. The central database 106 may communicate with the scraper system 102 to receive data to be stored in the central database 106. The central database 106 may communicate with the API 108 so as to return a query result to the user device 110 in response to a user query, as described herein.

[0029] The API 108 may comprise computer code or other data sorted on a computer-readable medium that may be executed by a processor to facilitate the interaction between software components. The API 108 may facilitate interaction between the user device 110 and the scraper system 102 and/or the central database 106, as described herein.

[0030] The user device 110 may comprise a computing device. The user device 110 may communicate with the API 108 to request that the scraper system 102 scrape a plurality of websites for data specified by a scraper inquiry (e.g., data related to a specified product category) communicated by the user device 110. The user device 110 may communicate with the API 108 to request that the central database 106 return a query result in response to a query communicated by the user device 110. The user device 110 may be a device of a user, such as a consumer having a payment device, and/or the user device 110 may be any of the devices or systems shown and described in FIG. 3.

[0031] The number and arrangement of systems and devices shown in FIG. 1 are provided as

an example. There may be additional systems and/or devices, fewer systems and/or devices, different systems and/or devices, and/or differently arranged systems and/or devices than those shown in FIG. 1. Furthermore, two or more systems or devices shown in FIG. 1 may be implemented within a single system or device, or a single system or device shown in FIG. 1 may be implemented as multiple, distributed systems or devices. Additionally, or alternatively, a set of systems (e.g., one or more systems) or a set of devices (e.g., one or more devices) of system 100 may perform one or more functions described as being performed by another set of systems or another set of devices of system 100.

[0032] Referring now to FIG. 2, shown is a process 200 for automatically scraping categorical data from a plurality of websites according to some non-limiting embodiments or aspects. The steps shown in FIG. 2 are for example purposes only. It will be appreciated that additional, fewer, different, and/or a different order of steps may be used in non-limiting embodiments or aspects. In some non-limiting embodiments or aspects, one or more of the steps of process 200 may be performed (e.g., completely, partially, and/or the like) by the scraper system 102 (e.g., one or more devices of the scraper system 102). In some non-limiting embodiments or aspects, one or more of the steps of process 200 may be performed (e.g., completely, partially, and/or the like) by another system, another device, another group of systems, or another group of devices, separate from or including the scraper system 102, such as the central database 106, the API 108, and/or the user device 110.

[0033] As shown in FIG. 2, at step 202, the process 200 may include determining a product category. As used herein, a “product” refers to a product or a service produced and/or sold by a merchant. Product categories refer to groups of products and/or services related to each other by at least one characteristic. As a non-limiting example, a product category may be “computer and electronic manufacturing”, which may encompass goods or services related to computers or other electronic components. It will be appreciated that product categories may be defined according to any technique for grouping related products or services together. For example, determining product categories may comprise the scraper system 102 generating at least one, such as plurality of, groups, each group comprising related products and/or services.

[0034] As shown in FIG. 2, at step 204, the process 200 may include identifying a first website

comprising data associated with the product category. For example, the scraper system 102 may identify a website of a distributor of products associated with the product category as the first website.

[0035] As shown in FIG. 2, at step 206, the process 200 may include automatically scraping the first website to compile first product data associated with the product category. For example, the scraper system 102 may comprise a web scraper and/or a web crawler to automatically scrape the first website for the first product data. The first product data may include any of the data described in Appendix A, such as: product category data, product stock keeping unit (SKU) data, product name and/or description data, product certification data, product characteristic data, supplier name and/or URL data, search tag data, product/supplier relationship data, or the like. In non-limiting embodiments, scraping may include parsing through publicly viewable website source code to identify content based on one or more rules. For example, the one or more rules may define web content to be scraped as words arranged between HTML tags or XML tags, words associated with web objects, metadata associated with images and/or other objects, and/or other like content of a website, whether visible or not.

[0036] As shown in FIG. 2, at step 208, the process 200 may include generating a plurality of web queries based on the compiled first product data. For example, the scraper system 102 may analyze, with a machine learning model, the compiled first product data to generate the plurality of web queries.

[0037] As shown in FIG. 2, at step 210, the process 200 may include executing the plurality of web queries to identify a plurality of websites. For example, the scraper system 102 may execute the plurality of web queries using a search engine, in order to identify a plurality of websites potentially relevant to the product category.

[0038] As shown in FIG. 2, at step 212, the process 200 may include automatically scraping at least a portion of the plurality of websites to compile supplier data associated with suppliers in the product category. For example, the scraper system 102 may comprise a web scraper and/or a web crawler to automatically scrape at least a portion of the plurality of websites to compile supplier data associated with suppliers in the product category. The scraper system 102 may not

scrape all websites of the plurality of websites potentially relevant to the product category, but may first determine a subset of websites returned in the web queries to scrape, in order to save processing resources. This may be done by the scraper system 102 determining a subset of websites associated with the product category from the plurality of websites based on a plurality of rules. The plurality of rules may be generated by a machine learning algorithm, according to some non-limiting embodiments or aspects. The rules may comprise any rules which reduce the plurality of websites identified by the plurality of web queries to some subset thereof, which subset of websites are identified by the scraper system 102 as more relevant than those websites from the plurality of websites excluded from the subset. The scraper system 102 may automatically scrape the subset of websites to compile the supplier data. Thus, the scraper system 102 may automatically scrape the subset of websites, instead of each of the plurality of websites, to obtain the desired data. The supplier data may include any of the data described in Appendix A, such as: supplier address, supplier phone number, supplier email address, supplier physical address, supplier legal business name, supplier DBA name, supplier business description, supplier product categories, supplier product SKUs, supplier product names, supplier certifications, supplier Google business identifier (or other business identifier), supplier website, supplier liens, supplier agents, supplier standing (e.g., commercial standing, financial standing, tax standing, and the like), supplier entity type, supplier search tags, supplier years in business, or the like.

[0039] As shown in FIG. 2, at step 214, the process 200 may include storing at least a portion of the compiled supplier data in a database. For example, the scraper system 102 may store at least a portion of the compiled supplier data in the central database 106. The compiled supplier data may be stored in the central database 106 using any suitable arrangement. The compiled supplier data may be stored in the central database 106 in association with the product category.

[0040] The above-described process 200 describes the process for automatically scraping categorical data from a plurality of websites, compiling said data, and storing said data in the central database 106 for a single product category. However, the process may be performed for a plurality of different product categories so that the central database 106 contains compiled supplier data for each of the plurality of product categories. The central database 106 may comprise supplier data associated with at least one product category.

[0041] For example, for a second product category, the process 200 may comprise: determining a second product category; identifying a second website comprising data associated with the second product category; automatically scraping the second website to compile second product data associated with the second product category; generating a second plurality of web queries based on the compiled second product data; executing the second plurality of web queries to identify a second plurality of websites; automatically scraping at least a portion of the second plurality of websites to compile second supplier data associated with suppliers in the second product category; and storing at least a portion of the compiled second supplier data in the database.

[0042] In some non-limiting embodiments or aspects, the process 200 may further include operating a queryable database (e.g., the central database 106) to enable a user to query data stored thereon. The API 108 may be generated and exposed to enable user querying of the central database 106. The user may query the central database 106 by the API 108 using the user device 110. The API 108 may receive a first query associated with the product category from the user device 110. The API 108 may query the central database 106 based on the first query to generate a first query result. For example, the API 108 may query the data in the central database 106 and locate data stored thereon associated with the product category associated with the first query (e.g., the at least a portion of the compiled supplier data in the central database 106 associated with the product category) to generate the first query result. The API 108 may transmit the first query result to the user device 110, the first query result including at least a portion of the located data associated with the product category associated with the first query. The transmission of the first query result may cause the first query result (e.g., data included therein) to be displayed on the user device 110. In some non-limiting embodiments or aspects, the first query may identify a product category, and the returned first query result may identify suppliers associated with the product category and/or data associated with those suppliers.

[0043] From the foregoing description, it will be appreciated that the central database 106 may comprise compiled supplier data associated with at least one product category. As such, the central database 106 may constitute a comprehensive supplier directory that includes, in a single location (e.g., the central database 106) relevant supplier data for various product categories. The central database 106 may be a searchable/queryable supplier directory which the user device 110 may

interact with in order to obtain desired information about suppliers in a particular product category. For example, the user device 110 may obtain supplier data associated with a product category of interest.

[0044] However, it will be appreciated that the disclosure is not limited to the generation of a supplier directory, and other applications will be readily apparent from the foregoing description. For example, then central database 106 may store any type of categorical data which may help a user search/query the categorical data in a single location, as opposed to the user navigating the myriad internet website to obtain the same data distributed across the plurality of websites.

[0045] Referring now to FIG. 3, FIG. 3 is a diagram of a non-limiting embodiment or aspect of an exemplary environment 300 in which systems, products, and/or methods, as described herein, may be implemented. As shown in FIG. 3, environment 300 may include transaction service provider system 302, issuer system 304, customer device 306, merchant system 308, acquirer system 310, and communication network 312. In some non-limiting embodiments or aspects, the scraper system 102, the central database 106, and/or the API 108 may be implemented by (e.g., part of) transaction service provider system 302. In some non-limiting embodiments or aspects, the scraper system 102, the central database 106, and/or the API 108 may be implemented by (e.g., part of) another system, another device, another group of systems, or another group of devices, separate from or including transaction service provider system 302, such as issuer system 304, merchant system 308, acquirer system 310, and/or the like. For example, the scraper system 102, the central database 106, and/or the API 108 may be implemented by (e.g., part of) transaction service provider system 302, and/or the user device 110 may be implemented by one of issuer system 304, customer device 306, merchant system 308, acquirer system 310, and/or the like. The user device 110 may be implemented by customer device 306.

[0046] Transaction service provider system 302 may include one or more devices capable of receiving information from and/or communicating information to issuer system 304, customer device 306, merchant system 308, and/or acquirer system 310 via communication network 312. For example, transaction service provider system 302 may include a computing device, such as a server (e.g., a transaction processing server), a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, transaction service provider system 302 may be associated

with a transaction service provider as described herein. In some non-limiting embodiments or aspects, transaction service provider system 302 may be in communication with a data storage device, which may be local or remote to transaction service provider system 302. In some non-limiting embodiments or aspects, transaction service provider system 302 may be capable of receiving information from, storing information in, communicating information to, or searching information stored in the data storage device.

[0047] Issuer system 304 may include one or more devices capable of receiving information and/or communicating information to transaction service provider system 302, customer device 306, merchant system 308, and/or acquirer system 310 via communication network 312. For example, issuer system 304 may include a computing device, such as a server, a group of servers, and/or other like devices. In some non-limiting embodiments or aspects, issuer system 304 may be associated with an issuer institution as described herein. For example, issuer system 304 may be associated with an issuer institution that issued a credit account, debit account, credit card, debit card, and/or the like to a user associated with customer device 306.

[0048] Customer device 306 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 302, issuer system 304, merchant system 308, and/or acquirer system 310 via communication network 312. Additionally, or alternatively, each customer device 306 may include a device capable of receiving information from and/or communicating information to other customer devices 306 via communication network 312, another network (e.g., an ad hoc network, a local network, a private network, a virtual private network, and/or the like), and/or any other suitable communication technique. For example, customer device 306 may include a client device and/or the like. In some non-limiting embodiments or aspects, customer device 306 may or may not be capable of receiving information (e.g., from merchant system 308 or from another customer device 306) via a short-range wireless communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like), and/or communicating information (e.g., to merchant system 308) via a short-range wireless communication connection.

[0049] Merchant system 308 may include one or more devices capable of receiving

information from and/or communicating information to transaction service provider system 302, issuer system 304, customer device 306, and/or acquirer system 310 via communication network 312. Merchant system 308 may also include a device capable of receiving information from customer device 306 via communication network 312, a communication connection (e.g., an NFC communication connection, an RFID communication connection, a Bluetooth® communication connection, a Zigbee® communication connection, and/or the like) with customer device 306, and/or the like, and/or communicating information to customer device 306 via communication network 312, the communication connection, and/or the like. In some non-limiting embodiments or aspects, merchant system 308 may include a computing device, such as a server, a group of servers, a client device, a group of client devices, and/or other like devices. In some non-limiting embodiments or aspects, merchant system 308 may be associated with a merchant as described herein. In some non-limiting embodiments or aspects, merchant system 308 may include one or more client devices. For example, merchant system 308 may include a client device that allows a merchant to communicate information to transaction service provider system 302. In some non-limiting embodiments or aspects, merchant system 308 may include one or more devices, such as computers, computer systems, and/or peripheral devices capable of being used by a merchant to conduct a transaction with a user. For example, merchant system 308 may include a POS device and/or a POS system.

[0050] Acquirer system 310 may include one or more devices capable of receiving information from and/or communicating information to transaction service provider system 302, issuer system 304, customer device 306, and/or merchant system 308 via communication network 312. For example, acquirer system 310 may include a computing device, a server, a group of servers, and/or the like. In some non-limiting embodiments or aspects, acquirer system 310 may be associated with an acquirer as described herein.

[0051] Communication network 312 may include one or more wired and/or wireless networks. For example, communication network 312 may include a cellular network (e.g., a long-term evolution (LTE®) network, a third generation (3G) network, a fourth generation (4G) network, a fifth generation (5G) network, a code division multiple access (CDMA) network, and/or the like), a public land mobile network (PLMN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), a telephone network (e.g., the public switched telephone

network (PSTN)), a private network (e.g., a private network associated with a transaction service provider), an ad hoc network, an intranet, the internet, a fiber optic-based network, a cloud computing network, and/or the like, and/or a combination of these or other types of networks.

[0052] The number and arrangement of systems, devices, and/or networks shown in FIG. 3 are provided as an example. There may be additional systems, devices, and/or networks; fewer systems, devices, and/or networks; different systems, devices, and/or networks; and/or differently arranged systems, devices, and/or networks than those shown in FIG. 3. Furthermore, two or more systems or devices shown in FIG. 3 may be implemented within a single system or device, or a single system or device shown in FIG. 3 may be implemented as multiple, distributed systems or devices. Additionally, or alternatively, a set of systems (e.g., one or more systems) or a set of devices (e.g., one or more devices) of environment 300 may perform one or more functions described as being performed by another set of systems or another set of devices of environment 300.

[0053] Referring now to FIG. 4, shown is a diagram of example components of a device 400 according to non-limiting embodiments or aspects. Device 400 may correspond to at least one of scraper system 102, central database 106, API 108, or user device 110 in FIG. 1 and/or at least one of transaction service provider system 302, issuer system 304, customer device 306, merchant system 308, and/or acquirer system 310 in FIG. 3, as an example. In some non-limiting embodiments or aspects, such systems or devices in FIG. 1 or FIG. 3 may include at least one device 400 and/or at least one component of device 400. The number and arrangement of components shown in FIG. 4 are provided as an example. In some non-limiting embodiments or aspects, device 400 may include additional components, fewer components, different components, or differently arranged components than those shown in FIG. 4. Additionally, or alternatively, a set of components (e.g., one or more components) of device 400 may perform one or more functions described as being performed by another set of components of device 400.

[0054] As shown in FIG. 4, device 400 may include bus 402, processor 404, memory 406, storage component 408, input component 410, output component 412, and communication interface 414. Bus 402 may include a component that permits communication among the components of device 400. In some non-limiting embodiments or aspects, processor 404 may be

implemented in hardware, firmware, or a combination of hardware and software. For example, processor 404 may include a processor (e.g., a central processing unit (CPU), a graphics processing unit (GPU), an accelerated processing unit (APU), etc.), a microprocessor, a digital signal processor (DSP), and/or any processing component (e.g., a field-programmable gate array (FPGA), an application-specific integrated circuit (ASIC), etc.) that can be programmed to perform a function. Memory 406 may include random access memory (RAM), read only memory (ROM), and/or another type of dynamic or static storage device (e.g., flash memory, magnetic memory, optical memory, etc.) that stores information and/or instructions for use by processor 404.

[0055] With continued reference to FIG. 4, storage component 408 may store information and/or software related to the operation and use of device 400. For example, storage component 408 may include a hard disk (e.g., a magnetic disk, an optical disk, a magneto-optic disk, a solid state disk, etc.) and/or another type of computer-readable medium. Input component 410 may include a component that permits device 400 to receive information, such as via user input (e.g., a touch screen display, a keyboard, a keypad, a mouse, a button, a switch, a microphone, etc.). Additionally, or alternatively, input component 410 may include a sensor for sensing information (e.g., a global positioning system (GPS) component, an accelerometer, a gyroscope, an actuator, etc.). Output component 412 may include a component that provides output information from device 400 (e.g., a display, a speaker, one or more light-emitting diodes (LEDs), etc.). Communication interface 414 may include a transceiver-like component (e.g., a transceiver, a separate receiver and transmitter, etc.) that enables device 400 to communicate with other devices, such as via a wired connection, a wireless connection, or a combination of wired and wireless connections. Communication interface 414 may permit device 400 to receive information from another device and/or provide information to another device. For example, communication interface 414 may include an Ethernet interface, an optical interface, a coaxial interface, an infrared interface, a radio frequency (RF) interface, a universal serial bus (USB) interface, a Wi-Fi® interface, a cellular network interface, and/or the like.

[0056] Device 400 may perform one or more processes described herein. Device 400 may perform these processes based on processor 404 executing software instructions stored by a computer-readable medium, such as memory 406 and/or storage component 408. A computer-readable medium may include any non-transitory memory device. A memory device includes

memory space located inside of a single physical storage device or memory space spread across multiple physical storage devices. Software instructions may be read into memory 406 and/or storage component 408 from another computer-readable medium or from another device via communication interface 414. When executed, software instructions stored in memory 406 and/or storage component 408 may cause processor 404 to perform one or more processes described herein. Additionally, or alternatively, hardwired circuitry may be used in place of or in combination with software instructions to perform one or more processes described herein. Thus, embodiments described herein are not limited to any specific combination of hardware circuitry and software. The term "programmed or configured," as used herein, refers to an arrangement of software, hardware circuitry, or any combination thereof on one or more devices.

[0057] Further details regarding non-limiting embodiments or aspects of methods, systems, and computer program products for automatically scraping categorical data from a plurality of websites are disclosed in the Appendix filed herewith, the entire disclosure of which is hereby incorporated by reference in its entirety.

[0058] Although embodiments have been described in detail for the purpose of illustration, it is to be understood that such detail is solely for that purpose and that the disclosure is not limited to the disclosed embodiments or aspects, but, on the contrary, is intended to cover modifications and equivalent arrangements that are within the spirit and scope of the present disclosure. For example, it is to be understood that the present disclosure contemplates that, to the extent possible, one or more features of any embodiment or aspect can be combined with one or more features of any other embodiment or aspect.

SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR AUTOMATICALLY SCRAPING CATEGORICAL DATA FROM A PLURALITY OF WEBSITES

ABSTRACT

Systems, methods, and computer program products are provided for automatically scraping categorical data from a plurality of websites. These include: determining a product category; identifying a first website including data associated with the product category; automatically scraping the first website to compile first product data associated with the product category; generating a plurality of web queries based on the compiled first product data; executing the plurality of web queries to identify a plurality of websites; automatically scraping at least a portion of the plurality of websites to compile supplier data associated with suppliers in the product category; and storing at least a portion of the compiled supplier data in a database.

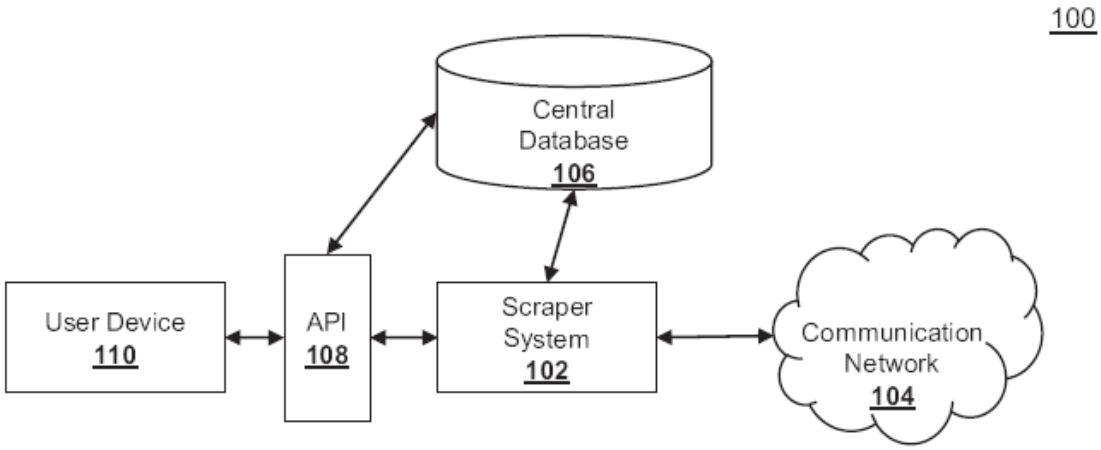
100
Central Database 106
User Device 110
API 108
Scraper System 102
Communication Network 104

FIG. 1

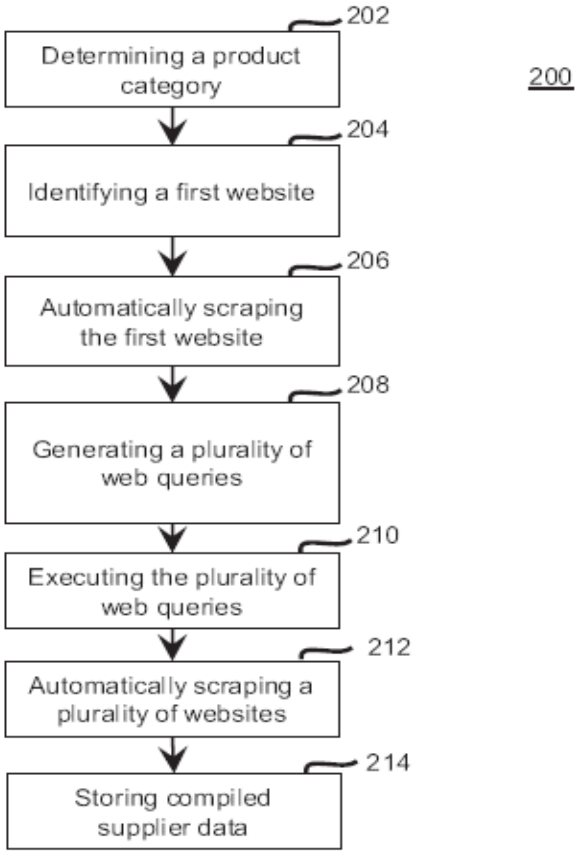
200
202
Determining a product category
204
Identifying a first website
206
Automatically scraping the first website
208
Generating a plurality of web queries
210
Executing the plurality of web queries
212
Automatically scraping a plurality of websites
214
Storing compiled supplier data

FIG. 2

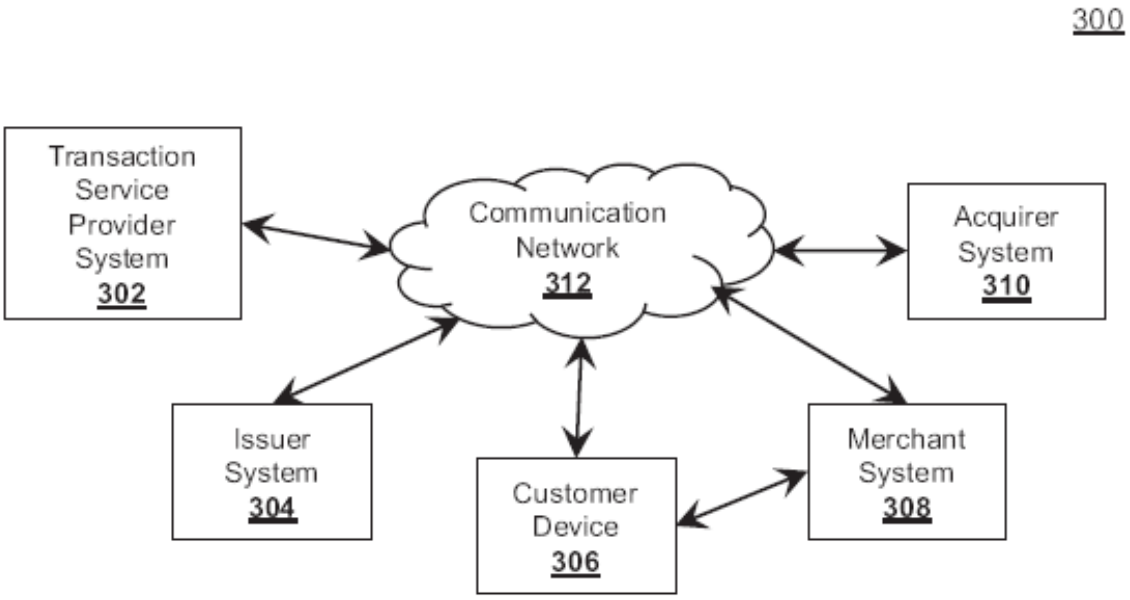
300
Transaction Service Provider System 302
Communication Network 312
Acquirer System 310
Issuer System 304
Customer Device 306
Merchant System 308

FIG. 3

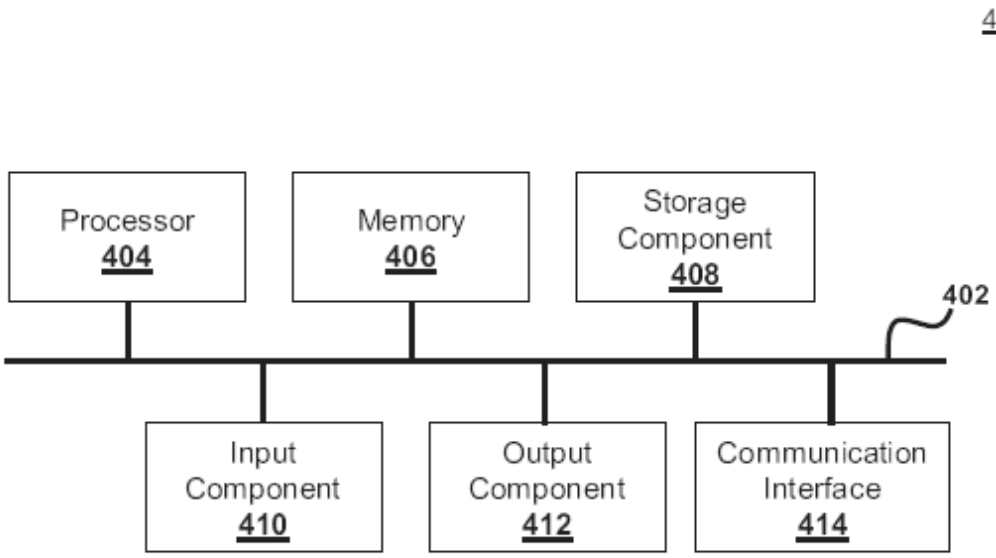
400
Processor 404
Memory 406
Storage Component 408
402
Input Component 410
Output Component 412
Communication Interface 414

FIG. 4