

**Methods and systems for determining sets of second attributes associated with
respective first attributes**

Technical Field

[0001] Embodiments generally relate to methods, systems, and computer-readable media for determining sets of first attributes associated with respective second attributes. In some embodiments, each first attribute is or comprises a brand or business name and the second attribute is or comprises an entity identifier, such as an organisation name or organisation number. In some embodiments, a business register, or other data store, is collated, populated and/or updated using the second attributes and determined associated sets of first attributes.

Background

[0002] Any discussion of documents, acts, materials, devices, articles or the like which has been included in the present specification is not to be taken as an admission that any or all of these matters form part of the prior art base or were common general knowledge in the field relevant to the present disclosure as it existed before the priority date of each claim of this application.

Summary

[0003] Embodiments relate to a method comprising: a method comprising: determining a dataset associated with an entity attribute, the dataset comprising character strings for each of a plurality of first attributes; for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a candidate character string of the character string pair to a target character string of the character string pair, and wherein the similarity measure for the character string is based on a lowest value of: (i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and (ii) a minimum number of character edits required to convert the candidate character string into the target character string; determining one or more clusters of the first attributes of the plurality of first attributes according to the similarity measures of the respective first attributes; determining a representative label for each of the one or more clusters; and determining a set of first

attribute labels associated with the entity attribute, the set of first attribute labels comprising the representative labels for each of the one or more clusters.

[0004] The method may further comprise determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string. The entity attribute may be an organisation identifier and/or the plurality of first attribute types may be a plurality of brand names or business names.

[0005] Some embodiments relate to a method comprising: determining a dataset associated with an organisation identifier, the dataset comprising character strings for each of a plurality of brand names; for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a candidate character string of the character string pair to a target character string of the character string pair; determining one or more clusters of the brand names of the plurality of brand names according to the similarity measures of the respective brand names; determining a representative label for each of the one or more clusters; and associating a set of brand labels with the organisation identifier, the set of brand labels comprising the representative labels for each of the one or more clusters.

[0006] The similarity measure may be based on the Levenshtein edit-distance between the character strings of the pair of character strings. The similarity measure may comprise the Levenshtein edit-distance between the character strings of the pair of character strings divided by a length of the longer character string of the pair of character strings. The similarity measure for the character string pair may be based on a lowest value of: (i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and (ii) a minimum number of character edits required to convert the candidate character string into the second character string.

[0007] In some embodiments, the method comprises determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string. In some embodiments, determining the lowest value comprises: initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string; performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target

character string or a substring of the target character string using the initialised shortest path vector; and determining the lowest value as the minimum value of the shortest path vector.

[0008] In some embodiments, the method comprises determining a clustering score for each of the one or more clusters; comparing the clustering scores to a clustering score threshold; and determining validated clusters as only those clusters of the one or more clusters that have a respective clustering score that meets or exceeds the clustering score threshold, wherein determining the representative label for each of the one or more clusters comprises determining the representative label for each of the one or more validated clusters only.

[0009] Determining the one or more clusters may comprise a) determining a cluster dataset associated with the entity identifier, the cluster dataset comprising a plurality of character strings and for each character string, an associated set of similarity measures and a weight, wherein the weight is a fraction of an overall number of unique origins in the cluster dataset that are associated with the respective character string; b) determining each character string of the dataset a centroid, and for each centroid, determining a set of character strings of the cluster dataset that are sufficiently close to the centroid; c) adding the weights of the character strings of the set of character string considered sufficiently similar to the respective centroid to determine a centroid cluster value; d) determining the centroid with the highest centroid cluster value; e) determining that highest centroid cluster value meets a cluster value threshold; and f) determining the centroid with the highest centroid cluster value a valid cluster corresponding to a first attribute of the entity identifier. The method may further comprise g) removing the centroid and the set of character strings belonging to the determined cluster from the cluster dataset, and repeating steps b) to f).

[0010] In some embodiments, the method further comprises populating an entry associated with the entity attribute or organisation identifier in a register based on the determined set of first attribute labels or brand label, the register comprising a plurality of entries and associated first attribute labels or brand labels.

[0011] Some embodiments relate to a method comprising: determining a character string pair, the character string pair comprising a candidate character string and a target character string, wherein both the candidate character string and the target character string represents a first attribute of an entity; determining a similarity measure for the character string pair,

wherein the similarity measure is indicative of the similarity of the candidate character string of the character string pair to the target character string of the character string pair, wherein the similarity measure for the character string is based on a lowest value of: (i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and (ii) a minimum number of character edits required to convert the candidate character string into the target character string; and determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string.

[0012] Determining the lowest value may comprise initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string; performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and determining the lowest value as the minimum value of the shortest path vector. Each character edit may correspond to a delete operation, an insertion operation and/or a replace operation.

[0013] Some embodiments relate to a method comprising: determining a character string pair, the character string pair comprising a candidate character string and a target character string, wherein both the candidate character string and the target character string represents a first attribute of an entity; determining a similarity measure for the character string pair, wherein the similarity measure is indicative of the similarity of the candidate character string of the character string pair to the target character string of the character string pair, wherein determining the similarity measure for the character string comprises: initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string; performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and determining the similarity measure as the minimum value of the shortest path vector.

[0014] Some embodiments relate to a method comprising: determining a dataset associated with an entity attribute, the dataset comprising character strings for each of a plurality of first attributes; for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a

candidate character string of the character string pair to a target character string of the character string pair, and wherein determining the similarity measure comprises: initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string; performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and determining the similarity measure as the minimum value of the shortest path vector; determining one or more clusters of the first attributes of the plurality of first attributes according to the similarity measures of the respective first attributes; determining a representative label for each of the one or more clusters; and determining a set of first attribute labels associated with the entity attribute, the set of first attribute labels comprising the representative labels for each of the one or more clusters.

[0015] Some embodiments relate to a method comprising: a) determining a cluster dataset associated with an entity identifier, the cluster dataset comprising a plurality of character strings and for each character string, an associated set of similarity measures and a weight, wherein the weight is a fraction of an overall number of unique origins in the cluster dataset that are associated with the respective character string; b) determining each character string of the dataset a centroid, and for each centroid, determining a set of character strings of the cluster dataset that are sufficiently close to the centroid; c) adding the weights of the character strings of the set of character string considered sufficiently similar to the respective centroid to determine a centroid cluster value; d) determining the centroid with the highest centroid cluster value; e) determining that highest centroid cluster value meets a cluster value threshold; and f) determining the centroid with the highest centroid cluster value a valid cluster corresponding to a first attribute of the entity identifier. The method may further comprise g) removing the centroid and the set of character strings belonging to the determined cluster from the cluster dataset, and repeating steps b) to f).

[0016] In some embodiments, the associated set of similarity measures are determined according to the method of any one of the described embodiments.

[0017] Some embodiments relate to a system comprising: one or more processors; and memory comprising computer executable instructions, which when executed by the one or

more processors, cause the system to perform the method of any one of the methods described herein.

[0018] Some embodiments relate to a computer-readable storage medium storing instructions that, when executed by a computer, cause the computer to perform the method of any one of the methods described herein.

Brief Description of Drawings

[0019] Various ones of the appended drawings merely illustrate example embodiments of the present disclosure and cannot be considered as limiting its scope.

[0020] Figure 1A is a schematic representation of two entity identifiers and their associated contact names;

[0021] Figure 1B is a schematic representation of two entity identifiers and their associated contact names clustered according to brand, according to some embodiments;

[0022] Figure 1C is an extract of a business register populated with the entity identifiers and related clusters of contact names of Figure 1B;

[0023] Figure 2 is an example block drawings of a system configured to determine sets of second attributes associated with respective first attributes, according to some embodiments;

[0024] Figure 3 is a process flow diagram of a method of determining one or more distinct clusters of first attributes associated with an entity attribute, according to some embodiments;

[0025] Figure 4 is a process flow diagram of a method of determining a similarity measure of a character string pair, according to some embodiments;

[0026] Figure 5 is a process flow diagram of a method of determining one or more clusters using similarity measures, according to some embodiments; and

[0027] Figure 6 is an example t-distributed stochastic neighbor embedding (t-SNE) plot depicting an example clustering of contact names realised according to the method of Figures 3, 4 and 5.

Description of Embodiments

[0028] Embodiments generally relate to methods, systems, and computer-readable media for determining sets of first attributes associated with respective second attributes. In some embodiments, each first attribute is or comprises a brand or business name and the second attribute is or comprises an entity identifier, such as an organisation name or organisation number. In some embodiments, a business register, or other data store, is collated, populated and/or updated using the second attributes and determined associated sets of first attributes.

[0029] There are many organisations that have a plurality of distinct brands or business names associated with them. In such cases, an entity attribute, such as an organisation name or organisation identifier (for example an Australian Business Number (ABN)), may be affiliated with a plurality of brands or business names. For example, the entity name “WOOLWORTHS GROUP LIMITED” may have a unique ABN and be associated with the brands or business names “Woolworths” and “BIGW”. Similarly, the entity name “YYYYY PTY LTD” may have a unique ABN and is associated with the brands or business names “Red Rooster”, “Two Brothers” and “Sushi sushi”. Being able to determine a set of brands or business names associated with a particular organisation can be useful when collating, populating and maintaining a business register, for example.

[0030] The business register may comprise an entity or organisation identifier, such as an organisation name, ABN, or other identifier, which may be unique to the organisation and to the business register. Each entity identifier may be associated with one or more sets of contact names. Each set of contact names may comprise a canonical name, that is, a representative contact name. Each set of contact names may comprise one or more contact aliases or variants. The contact aliases or variants may be names and/or name formats other than the canonical name by which the organisation may be known or referred to. Where an organisation has or is affiliated with multiple brands, the entity identifier may be associated with multiple sets of contact names in the business register. Each set of contact names may relate to a distinct brand or business name of the organisation.

[0031] Such a business register may be used to respond to requests for organisation and/or contact details from a user, such as a user of an accounting platform associated with a particular organisation. For example, such a request may be received from a transaction reconciliation module configured to make reconciliation suggestions to a user to enable the user to reconcile transactions in the accounting platform. In such an example, the request may comprise a candidate attribute in the form of a candidate contact name extracted from a piece of raw data, such as a data line on a financial statement. In response to receiving the query, the business register may be queried to determine a match for the candidate contact name in order to take further actions, such as making a recommendation for a contact name and/or organisation identifier.

[0032] For example, if the candidate contact name was “Woolies”, the business register may return a canonical contact name for the candidate contact name, such as “Woolworths” and/or the organisation name “WOOLWORTHS GROUP LIMITED” or the relevant ABN. Similarly, if the candidate contact name was “Big W”, the business register may return a canonical contact name for the candidate contact name, such as “BIGW” and/or the organisation name “WOOLWORTHS GROUP LIMITED” or the relevant ABN.

[0033] To collate relevant data to populate or update the business register, or indeed any kind of data store, such as a database (relational or non-relational) or a graph data structure, it would be useful to be able to readily or automatically determine or discern one or more distinct business names or brands associated with a particular organisation from the multiple contact names used to refer to those brands. For example, there may exist datasets or data structures of contact names associated with particular organisation identifiers, which may be accessible to the accounting platform. The described embodiments of the present disclosure enable the contact names to be grouped or organised into one or more clusters of contacts names, wherein each cluster relates to, and may be labelled with, a distinct brand or business name.

[0034] For example, consider the schematic 100 of Figure 1A, which depicts an organisation identifier Org X (e.g. WOOLWORTHS GROUP LIMITED) associated with a plurality of contact names including Contact A1 (e.g. Woolies), Contact A2 (e.g. Woolworths), Contact B1 (BIGW), Contact B2 (Big W) and Contact B3 (BW). Similarly, Figure 1A depicts an organisation identifier Org Y (e.g. “YYYYYY PTY LTD”) associated

with a plurality of contact names including Contact C1 (e.g. RedRooster), Contact C2 (e.g. RedR), Contact D1 (Sushi sushi), Contact D2 (e.g. Sushisushi), Contact D3 (e.g. sushi), Contact E1 (e.g. 2Bros), Contact E2 (e.g. twobrothers). Figure 1B depicts the contact names of both organisations, Org X and Org Y organised or clustered into clusters of contacts names - clusters A and B for Org X, and clusters C, D, and E for Org Y. Each cluster has a representative label, which may be a canonical or representative name for the contacts of the cluster. The representative name may correspond to a distinctive brand or business name associated or affiliated with the organisation. Figure 1C depicts an extract of a business register populated with the determined cluster labels for each entity identifier. As illustrated, each entity identifier is associated with a set of contact names. In the case of Org X, it is associated with labels representative of brands of type A and type B. In this case, the label “Contact A” has been allocated to all contact names that are considered sufficiently similar to one another, and for example sufficiently different from other names associated with Org X, such as Contacts B1, B2 and B3. “Contact B” has been allocated to all contact names that are considered sufficiently similar to one another, and for example sufficiently different from other names associated with Org X, such as Contacts A1 and A2. Similarly, Org Y has been associated with three distinct brands, Contact C, Contact D and Contact E.

[0035] To generate cluster(s) from a dataset of contact names associated with an entity identifier, character strings of the contact names are determined, and a similarity measure for each pair of character strings is determined. Each character string pair may comprise a candidate character string and a target character string.

[0036] In some embodiments, the similarity measure is based on the Levenshtein edit-distance between the character strings of the pair of character strings.

[0037] A substring of a string is a sequence of contiguous characters forming a proper subset of the characters of the string. A proper subset of a set X is a subset of set X that is not equal to set X; if Y is a proper subset of X, then all elements of set Y are in set X, but set X includes at least one element that is not in set Y. In other words, the number of characters in a substring is less than the number of characters in the full string. In some embodiments, the similarity measure is based on a minimum number of character edits required to convert or transform the candidate character string of the pair into a substring of the target character string, or into the same character string as that of the target character string (i.e., the full

string), whichever is the smaller of the two. In other words, if the number of edits to transform the candidate string into a substring of the target string is less than the number of edits to transform the candidate string into the target string, then the similarity measure is based on that. On the other hand, if the number of edits to transform the candidate string into a substring of the target string is greater than the number of edits to transform the candidate string into the target string, then the similarity measure is based on the number of edits to transform it into the target string. This embodiment tends to produce relatively fewer false positives, and is effective in differentiating between two different brands corresponding to the same entity identifier (e.g. Big W and Woolworths), but also at recognising that character string pairs having a relatively large distance between them don't necessarily mean they are different brands. With other embodiments, this can be problematic due to prefixes and/or suffixes, such as geographical locations, "Pty" and "Ltd", etc. For example, this embodiment tends to recognise that "Xero", "Xero Australia" and "Xero Pty Ltd" are all one distinct brand.

[0038] The contact names are clustered according to the determined similarity measures, and a number of clusters identified. Each cluster may be labelled with a representative label indicative of the brand.

[0039] In some embodiments, a clustering score is determined which is indicative of the quality of the clustering, and which may be relied upon to determine whether or not to rely on the clusters. For example, a relatively high score may indicate that intra-cluster distances are small and inter-cluster distances are large, meaning that the clusters are crisp and well separated. The clustering score may be compared with a clustering quality threshold, to determine the validity or reliability of the determined cluster(s), and for example, to determine whether or not to populate the business register, or other data store, with the determined distinct brand(s). This may ensure the integrity and consistency in terms of reliability of the business register.

[0040] Figure 2 is a schematic of a communications system 200 comprising a system 202, such as an accounting system (or a system providing an accounting platform) in communications with one or more computing devices 204 across a communications network 206. Examples of a suitable communications network 206 include a cloud server network, wired or wireless internet connection, BluetoothTM or other near field radio communication, and/or physical media such as USB.

[0041] The system 202 comprises one or more processors 208 and memory 210 storing instructions (e.g. program code) which when executed by the processor(s) 208 causes the system 202 to manage accounting aspects for a business or entity, provide accounting functionality to the one or more computing devices 204 and/or to function according to the described methods. The processor(s) 208 may comprise one or more microprocessors, central processing units (CPUs), application specific instruction set processors (ASIPs), application specific integrated circuits (ASICs) or other processors capable of reading and executing instruction code.

[0042] Memory 210 may comprise one or more volatile or non-volatile memory types. For example, memory 210 may comprise one or more of random access memory (RAM), read-only memory (ROM), electrically erasable programmable read-only memory (EEPROM) or flash memory. Memory 210 is configured to store program code accessible by the processor(s) 208. The program code comprises executable program code modules. In other words, memory 210 is configured to store executable code modules configured to be executable by the processor(s) 208. The executable code modules, when executed by the processor(s) 208 cause the accounting system 202 to perform certain functionality, as described in more detail below.

[0043] The system 202 further comprises a network interface 212 to facilitate communications with components of the communications system 200 across the communications network 206, such as the computing device(s) 204, data stores 214, 215 and/or other servers, such as a financial institute or banking server 216. The network interface 212 may comprise a combination of network interface hardware and network interface software suitable for establishing, maintaining and facilitating communication over a relevant communication channel.

[0044] The computing device(s) 204 comprise one or more processors 218 and memory 220 storing instructions (e.g. program code) which when executed by the processor(s) 218 causes the computing device(s) 204 to cooperate with the system 202 to provide accounting functionality to users of the computing device(s) 204 and/or to function according to the described methods. To that end, and similarly to the system 202, the computing devices 204 comprise a network interface 222 to facilitate communication with the components of the communications network 206. For example, memory 220 may comprise a web browser

application (not shown) to allow a user to engage with the accounting system 202. In some embodiments, memory 220 may comprise a reconciliation application (not shown) associated with the accounting system 202, which when executed by processor(s) 218, enables the computing device 204 to allow a user to reconcile a business's records with banking records associated with one or more accounts of the business via interaction with the user interface 224.

[0045] The communications system 200 further comprises the data store 214 and/or data store 215, which may form part of or be local to the system 202, or may be remote from and accessible to the system 202. The data store 214 may be a relational or non-relational database. The data store 215 may be a graph data structure.

[0046] For example, the graph data structure (or “graph”) may represent a network of entities (or things), for example entities associated with an accounting platform, and the information and relationships known about those entities. The graph may be an enriched information interface, which may be capable of providing a coherent and/or enhanced view of knowledge about the network of entities. The graph may be generated or created based on raw data derived from how the accounting platform is used by organisations to manage their accounting needs.

[0047] The data store 214 and/or data store 215 may be configured to store business records, banking records, accounting documents and/or accounting records associated with entities having user accounts with the system 202, availing of the services and functionality of the system 202, or otherwise associated with the system 202. The data store 214 and/or data store 215 may be configured to store entity information about a plurality of entities or organisations having accounts with or otherwise managed by or manageable using the system 204 or accounting platform provided by the system 202. For example, the entity information may comprise information about particular organisations such as financial or banking information, accounting information, type of organization, practice area, number of people employed by the organisation, applications (“apps”) used or downloaded by the organisation, and which, for example integrate with their user account with the accounting platform 404, contacts of the organisation, such as people or other organisations that do business, or are otherwise relevant or related to the organisation. The data store 214 and/or data store 215 may comprise a business register, which may include entity or organisation identifiers, such

as an organisation name, ABN, or other identifiers, and associated contact names, contact name aliases and/or contact name variants.

[0048] The system 202 may also be arranged to communicate with financial institute server(s) 216 or other third party financial systems (not shown) to receive financial records and/or financial documents associated with transactions for reconciliation. For example, in some embodiments, the system 202 may be arranged to receive bank feeds having line items (or financial data items) associated with transactions to be reconciled by the system 202. The financial institute server(s) 216 may provide an Application Programming Interface (API) to securely extract transaction information in relation to one or more bank accounts of a business. The APIs may be secured using authentication and encryption mechanisms and the extracted transaction data may be referred to as bank feed data. Bank feed data may comprise information regarding one or more transactions, including transaction data, transaction amount, transaction reference, for example.

[0049] According to some described embodiments, memory 210 comprises a clustering module 226 which when executed by the processor 208, causes the system 202 to cluster first attributes (e.g. contact names) associated with an entity identifier (e.g. organisation identifier) into one or more clusters, wherein each cluster is potentially indicative of a representative first attribute, such as a brand or business name, associated with or corresponding to the entity identifier.

[0050] The sensitivity of the clustering module 226 may be selectively adjusted by adjusting clustering parameters, such as a clustering score threshold. Tight clustering parameters may result in fewer clusters being determined, whereas looser clustering parameters may result in a greater number of clusters being determined. For example, by setting the clustering score threshold to be relatively high, fewer clusters will be considered sufficiently reliable or valid.

[0051] In some embodiments, a density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm is used.

[0052] In some embodiments, the clustering module 226 is configured to determine one or more clusters from a cluster dataset of character strings and for each character string, an associated set of similarity measures. Each character string is associated with a weight

indicative of a fraction or percentage of an overall number of unique origins (such as users of an accounting platform) in the cluster dataset that are associated with the respective character string. For example, this may be a percentage of different users of an accounting system represented in the cluster dataset that used/use the character string. The clustering module 226 sums or adds together the weights of each character string considered sufficiently similar (based on a similarity measure) to respective centroids (i.e., targets) to determine respective centroid cluster values. The centroid with the highest centroid cluster value is determined and if its centroid cluster value exceeds a cluster value threshold, it is identified as a cluster corresponding to a first attribute of the entity identifier, such as a particular brand or business name. The centroid and its associated sufficiently similar character strings may be removed from the cluster dataset and further the process may be repeated in an attempt to identify further clusters.

[0053] The clustering module comprises a similarity determination module 228 to determine similarity measures between character strings. The similarity determination module 228 is configured to determine similarity measures for pairs of character strings.

[0054] In some embodiments, the similarity determination module 228 is configured to determine the similarity measure for the character string pair based on Levenshtein edit-distance between the character strings of the pair of character strings. For example, the similarity measure may be determined as the Levenshtein edit-distance between the character strings of the pair of character strings divided by a length of the longer character string of the pair of character strings.

[0055] In some embodiments, the similarity determination module 228 is configured to determine the similarity measure for the character string pair based on a minimum number of character edits to transform a candidate character string of the pair into at least a subset or substring of the target character string of the pair. In other words, the similarity measure is indicative of how much modification is required to turn a pattern (i.e. the candidate character string) into at least a substring of the target text (i.e., the target character string). For example, the similarity measure may be a transformation measure.

[0056] A character edit may be a delete operation, or a replace operation or an insert operation. The transformation from the candidate character string to the target character

string may apply to both character and position of the character within the string. That is, if transforming the candidate character string into a substring of the target string (that is a subset of the characters of the target string), the substring is a sequence of contiguous characters of the target string.

[0057] In some embodiments, to ensure normalisation, the similarity measure may be determined as the number of edits required to transform the candidate character string into the target string or a substring of the target string, divided by the length of the character string of the candidate character string.

[0058] For example, consider the character string pair format {candidate, target}. For the pair {woolworths, woolies}, the minimal number of edits to transform 'woolworths' into 'woolies' (i.e., the candidate string into the target string) would be 5 (e.g., by deleting 'w', 'o' and 'r' and replacing 't' and 'h' with 'i' and 'e' to get the string 'woolies'). Noticeably, a greater number of edits would be required to convert 'woolworths' into a substring of 'woolies' ('wool'), that is, 6 edits (e.g., delete 'w', 'o', 'r', 't', 'h', and 's'), than would be required to convert 'woolworths' into the full string 'woolies'.

[0059] For the pair {woolies, woolworths}, the minimal number of edits to transform 'woolies' into a substring of 'woolworths' would be 3 (e.g., by deleting 'i', 'e' and changing 's' to 'w', or deleting 'i', 'e' and 's'). In this case, a greater number of edits would be required to convert 'woolies' into the target string 'woolworths', that is, 5 (e.g., replace 'i' and 'e' with 'w' and 'o', respectively, and add 'r', 't' and 'h').

[0060] The similarity determination module 228 may be configured to receive, as inputs, pairs of candidate and target strings, and provide, as an output, the minimal number of edits to the candidate string required to transform it into a substring of the target string, or into the target string, whichever is the lower value of the two. In some embodiments, the similarity determination module 228 provides as an output, being a similarity measure, the lower value divided the length of the candidate character string.

[0061] According to some described embodiments, memory 210 may comprise a business register management module 230 configured to engage with and manage a business register or business register data structure (not shown). For example, the business register

management module 230 may be configured to populate and/or update information stored in the business register based on sets of attribute labels determined by the clustering module 226. The business register or business register data structure may be local to the system 202 or may be stored in data store 214 or data store 215. In some embodiments, the business register management module 230 is deployed on a separate system remote to system 202 and configured to cooperate with system 202, in some embodiments, to allow the system 202 to populate and/or update information stored in the business register based on sets of attribute labels determined by the clustering module 226.

[0062] The business register (or business register data structure) may be configured to store aliases relating to contacts, such as other names and/or name formats by which they may be known. This enables related contact names to be searched to determine the actual contact intended by the read request. Once a match is determined, additional queries and requests can be made to the graph structure based on the one or more identified objects in the business register. Alternatively, if no object is found or identified, it can be determined that there is no contact object(s) matching the candidate contact name, and therefore that the contact does not exist within the business register. From there, a request can either be made to write a new contact object to the graph structure, or to expand search parameters for additional queries. Further discussion of such an embodiment is described in Australian provisional patent application 2023900913, in the name of Xero Limited, filed on 31 March 2023, the entire content of which is incorporated herein by reference.

[0063] According to some described embodiments, memory 210 may comprise a graph interface module 232 configured to engage with and manage a graph structure, as may be stored locally or remotely at data store 215. For example, the graph interface module 232 may be configured to populate and/or update information stored in the graph based on sets of attribute labels determined by the clustering module 226. In some embodiments, the graph interface module 232 is deployed on a separate system remote to system 202 and configured to cooperate with system 202, in some embodiments, to allow the system 202 to populate and/or update information stored in the graph structure based on sets of attribute labels determined by the clustering module 226.

[0064] In some embodiments, the memory 210 comprises a transaction reconciliation module 234 for reconciling or facilitating the reconciling of financial transactions. For

example, in some embodiments, a financial statement may be used by the transaction reconciliation module to reconcile financial transactions against accounting documents, such as invoices or credit notes, issued by an organisation. The transaction reconciliation module 234 may be configured to extract candidate information from data lines in a financial statement, which may for example be obtained or received from a band feed of the financial institute servers 216. The candidate information may relate to potential entity attributes, for example, the extracted information from a statement line may be a string value which represents a potential contact name. The transaction reconciliation module 234 may be configured to generate a read request to query a business register (for example, as may be stored in data store 214) or the graph structure (for example, as may be stored in data store 215) for the potential contact name, or representative or canonical contact name. The transaction reconciliation module 234 may be configured to pass the request to the business register module 230 and/or to the graph interface module 232, and to receive, in response, a representative contact name and/or entity or organisation identifier, for example.

[0065] The transaction reconciliation module 234 can use the response to reconcile a financial statement line with an accounting record. For example, where a representative contact name is identified within the business register that matches the potential contact name extracted from the statement line, the transaction reconciliation module checks the representative contact against the accounting record issued by the organisation, and reconciles the transaction between the contact and the organisation. Where the statement line includes financial data, for example, a financial amount, the transaction reconciliation module may also compare the financial amount on the statement line with the financial amount from the issued accounting record as part of the reconciliation process.

[0066] Referring now to Figure 3, there is shown a process flow diagram of a method 300 of determining one or more distinct clusters of first attributes associated with an entity attribute. In some embodiments, the system 202 may be configured to perform the method 300 by having the processor(s) 208 execute computer code, such as clustering module 226 and/or similarity determination module 228, stored in memory 210.

[0067] At 302, the system determines a dataset associated with an entity attribute, such as an organisation or company identifier, for example, an organisation name, or an organisation registration number etc. The dataset comprises character strings for each of a plurality of first

attributes. For example, the first attributes may be brand names or business names affiliated with the first entity attribute. The character strings may have different lengths such that some character strings may be longer than others of the character strings, shorter than others of the character strings, and/or the same as others of the character strings.

[0068] In some embodiments, the dataset may be retrieved or received from a raw data source, which may derive data from the system 202, or accounting platform provided by the system 202, and which is used by users and/or organisations, or be retrieved or received from data store 214 (such as a database) or data store 215 (such as a graph structure).

[0069] Each characters string of the dataset may be associated with one or more unique origins, such as users of the system 202 or accounting platform. Each dataset may comprises character strings derived from an overall number of unique origins. For example, the character string “woollies” might have been used by or associated with 62 out of the 100 unique organisations represented in the dataset. In some embodiments, the dataset is pre-processed or filtered to remove character strings that are not associated with at least a minimum number of unique origins. In some embodiments, there is an upper limit on the number of character strings included in the dataset, for example, to ensure computational efficiency. For example, the system 202 may select the character strings derived from or associated with the highest overall number of unique origins in determining the dataset and ensuring compliance with the upper limit.

[0070] At 304, the system determines a similarity measure for each character string pair of the character strings of the dataset. Each character string pair may comprise a candidate character string and a target character string. The similarity measure is indicative of the similarity of the candidate character string of the character string pair to the target character string of the character string pair. The similarity measure may be indicative of character or text similar between the character strings of the pair of character strings. For example, where the dataset comprises N contact names, an NxN matrix of similarity measures may be determined.

[0071] In some embodiments, the system 202 determines the similarity measure based on the Levenshtein edit-distance between the character strings of the pair of character strings. For example, the similarity measure may be determined as the Levenshtein edit-distance between

the character strings of the pair of character strings divided by a length of the longer character string of the pair of character strings.

[0072] In some embodiments, the system 202 determines the similarity measure according to method 400 described below. For example, the similarity measure may be based on a lowest value of: (i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and (ii) a minimum number of character edits required to convert the candidate character string into the second character string.

[0073] At 306, the system 202, for example, clustering module 226, determines one or more clusters of the first attributes of the plurality of first attributes according to the similarity measures of the respective first attributes. For example, the clustering module 226 may be configured to receive, as an input, a determined NxN matrix of similarity measures for the N contact names of the dataset. The clustering module 226 may be configured to cluster or group the first attributes, such as contact names, into one or more clusters based on the similarity measures. The clustering module 226 may determine a number of distinct clusters. The number of distinct clusters determined may depend on a clustering score of each cluster.

[0074] In some embodiments, the system 202 assesses the validity or reliability of the determined cluster(s) based on the respective clustering score(s). For example, in some embodiments, the clustering module 226 may determine or calculate a clustering score for each cluster identified. The clustering score may be indicative of the quality of the cluster, and which may be relied upon to determine whether or not to rely on the cluster. For example, a relatively high score may indicate that intra-cluster distances are small and inter-cluster distances are large, meaning that the clusters are crisp and well separated. The clustering score may be compared with a clustering quality threshold, to determine the validity or reliability of the determined cluster(s), and for example, to determine whether or not to populate the business register, or other data store, with the determined distinct brand(s). This may ensure that integrity and consistency in terms of reliability of the business register.

[0075] In some embodiments, the system 202 determines one or more clusters according to method 500 described below with reference to Figure 5.

[0076] At 308, the system 202 determines a representative first attribute label for each of the one or more clusters. For example, the first representative label may be a canonical name by which the first attribute may be referred to. For example, for the organisation “WOOLWORTHS GROUP LIMITED” or its relevant ABN, the system 202 may determine first representative labels “Woolworths” and “Big W” as labels for the two distinct clusters of contact names determined.

[0077] At 310, the system 202 determines a set of first attribute labels with the entity attribute, the set of first attribute labels comprising the representative first attribute labels for each of the one or more clusters.

[0078] In some embodiments, the determined sets of first attribute labels for respective entity identifiers may be used by the business register module 230 and/or graph interface module 232 to create or populate new entries and/or update respective entries.

[0079] Referring now to Figure 4, there is shown a process flow diagram of a method 400 of determining a similarity measure of a pair of character strings, according to some embodiments. In some embodiments, the system 202 may be configured to perform the method 400 by having the processor(s) 208 execute computer code, such as clustering module 226 and/or similarity determination module 228, stored in memory 210.

[0080] At 402, the system 202 determines a character pair comprising a candidate character string and a target character string. The pair of character strings may be one of a plurality of pairs of character strings of a dataset of character strings. The dataset, and for example, each of the candidate and target character strings of the pair(s) may be associated with an entity attribute. The candidate and target character strings of the pair(s) may correspond to first attributes, for example, of the entity attribute. As discussed above, the entity attribute may be an organisation identifier and/or the first attribute may be a brand name or a business name.

[0081] The similarity measure for a character string pair is based on a lowest value of: (i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and (ii) a minimum number of character edits required to convert the candidate character string into the target character string.

[0082] To make this determination, the system 202 may execute a modified Levenshtein edit-distance calculation between the character strings of the pairs of character strings, according to some embodiments. As opposed to the standard Levenshtein edit-distance calculation, which seeks to determine minimum the number of edits to transform a candidate string into the target string, the modified version seeks to determine the minimum number of edits required to transform the candidate character string into a substring of the target string, or into the target string, whichever is the lower number of edits. To do this, the system 202, and for example, the similarity determination module 228, considers a distance or number of edits from an empty string to any other string to be zero. This contrasts with the standard Levenshtein edit-distance calculation where the number of edits from an empty string to any other string be the number of characters/prefixes in the target string. The system 202 initialises to zero, or zeros, a vector of the shortest path (e.g. $res(N+1)$, where N is the length or number of characters) of the target string $y[1], y[2], \dots y[N]$,) from an initial or first considered prefix or character of the candidate character string $(x[1], x[2], \dots x[n]$, where n is the length (or number of characters) in the candidate character string) to each prefix or character of the target character string.

[0083] At 404, the system 202 initialises with zeros a shortest path vector of the shortest path from a prefix or character of the candidate character string to each prefix or character of the target character string.

[0084] At 406, the system 202 performs an edit distance calculation (such as a Levenshtein edit-distance calculation) to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the zero initialised shortest path vector.

[0085] For example, for each character of the candidate character string, (that is, for each value of i , from i to $n+1$), the system 202 updates the shortest path vector (res (vector of length $N+1$)), such that it contains the shortest path or least number of edits from the last considered prefix of candidate character string $[:i]$ to each prefix of target character string $[:j]$ ($j=1\dots N$) such that candidate character string $[i]$ is aligned with target character string $[j]$.

[0086] For example, a variable “current” may define or hold a value of the shortest distance from candidate character string $[:i]$ to target character string $[:j]$ and will be stored in the

corresponding res value of the shortest path vector on the (j+1)st iteration. The variable “current” may be initialised as zero. A variable “prev” may be the same as current, but lagged by an iteration over j. In other words, after iteration i, j, it will contain the shortest distance from candidate character string [:i] to target character string [:j-1]. The variable “prev” is initialised with i at each outer iteration or loop (i.e., the iteration through all prefixes for the candidate character string). This is because the distance to an empty string is the length of the candidate character string.

[0087] If the system 202 determines that the candidate character string is the same as the target character string, then no edits are required, and the minimum value of the shortest path vector or the similarity measure is determined as being the value stored in the shortest path vector for the previous *j*th iteration, $res[j-1]$, which would be zero.

[0088] Otherwise, for each iteration through the prefixes/characters of the candidate character string (the outer loop/iteration) and the prefixes/characters of the target character string (the inner loop/iteration), the system 202:

- a. aligns candidate character string [i-1] with target character string [j], deletes candidate character string [i] and determines the shortest path value to be $res[j]+1$;
- b. aligns candidate character string [i-1] with target character string [j-1], deletes candidate character string [i] and determines the shortest path value to be $res[j-1]+1$;
- c. aligns candidate character string [i] with target character string [j-1], deletes candidate character string [i] and determines the shortest path value to be $current+1$;

and determines the value for “current” variable as being minimum of the shortest path values determined at a, b and c. The system then updates the variable “prev” and updates the shortest path vector with the value of the “current” variable. The system 202 therefore performs nN iterations.

[0089] At 408, the system 202 determines the similarity measure for the character string pair as the minimum value of the shortest path vector. The minimum value of the shortest path vector corresponds with the minimum number of character edits to transform the candidate character string into the target character string, or into a substring of the target character

string, whichever is the lower value. For example, in some embodiments, the similarity measure is the determined minimum number of character edits. In some embodiments, the system 202 determines the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string.

[0090] Where the lowest value is the minimum number of edits to transform the candidate string into the target string, the method 500 need not keep track of the actual substring which is closest to the target string as it is only concerned with determining the lowest number of edits required as a similarity measure.

[0091] In some examples, the method 500 of Figure 5 can be encoded in the Python programming language as follows:

```
def edit_substring_dist(pattern: str, text: str) -> float:
    """
    Inputs:
        - pattern: a string we are editing
        - text: a string we are looking for substrings in
    """
    if len(pattern) == 0:
        return 0
    if len(text) == 0:
        return len(pattern)

    n = len(pattern)
    N = len(text)

    res = [0] * (N+1)
    current = 0
    for i in range(1, n+1):
        prev = i
        for j in range(1, N+1):
            if pattern[i-1] == text[j-1]:
```

```

        current = res[j-1]
    else:
        current = min(min(res[j-1]+1, prev+1), res[j]+1)
    res[j-1] = prev
    prev = current
    res[N] = prev
    return min(res) / len(pattern)

```

[0092] Referring now to Figure 5, there is shown a process flow diagram of a method 500 of determining one or more clusters of first attributes associated with an entity identifier, according to some embodiments. In some embodiments, the system 202 may be configured to perform the method 400 by having the processor(s) 208 execute computer code, such as clustering module 226 stored in memory 210.

[0093] At 502, the system 202 determines for an entity identifier, a cluster dataset of character strings and for each character string, an associated set of similarity measures and a weight. For example, the cluster dataset may be based on the dataset used to determine the similarity measures discussed with reference to methods 300 and/or 400 above.

[0094] The character strings may be different spellings of brands or business names affiliated with the entity identifier, for example, that may have been used by users of an accounting platform in reconciling transactions. Each similarity measure is indicative of relative similarity of another character string (i.e. a candidate character string) in the dataset to the associated character string (i.e. a target character string). The weight for a character string may be based on of unique or different origins from which the character string was associated with. For example, the unique or different origins may be unique or different users (such as organisations, individuals or business users) or may be unique or different accounts or users of an accounting platform provided by the system 202. The dataset may comprise character strings retrieved from or associated with an overall number of unique origins. The weight may be a fraction or percentage of the overall number of unique origins that used the respective character string. The weight may be a fraction of an overall number of unique origins in the cluster dataset that are associated with the respective character string.

For example, if the dataset includes character strings used by 100 unique organisations, a specific character string that was used by 62 of those unique organisations would have a weight of 0.62.

[0095] At 504, system 202 determines each character string of the dataset a centroid (i.e., a target character string), and for each centroid, determines a set of character strings of the dataset that are sufficiently close to the centroid. The system 202 may determine if a character strings (i.e. a candidate character string) is sufficiently close to the centroid by determining whether the similarity measure of the character string to the centroid meets a similarity threshold. For example, in some embodiments, the lower the similarity measure, the greater the level of similarity between a candidate character string, or centroid, and the target character string. In such an embodiments, meeting the threshold may involve the similarity measure being less than a similarity threshold. Responsive to the system 202 determining that a candidate character string meets the similarity threshold, the character string is added to the set of character strings of the dataset that are sufficiently close to the centroid.

[0096] At 506, for each centroid, the system 202 sums or adds together the weights of the character strings of the set of character string considered sufficiently similar to the respective centroid to determine a centroid cluster value.

[0097] At 508, the system 202 determines the centroid with the highest centroid cluster value. In some embodiments, the system 202 determines the centroid with the highest centroid cluster value to be a cluster corresponding to a first attribute of the entity identifier, such as a brand or business name. In some embodiments, the system 202 determines whether the cluster value meets a cluster value threshold, and responsive to the cluster value meeting a cluster value threshold, the system 202 determines the centroid a valid or distinct cluster. Otherwise, the system 202 determines the cluster as invalid or insufficiently distinct.

[0098] At 510, the system 202 may remove the centroid and the set of character strings belonging to the determined cluster (that is, those having a similarity measure that meets the similarity threshold) from the cluster dataset, for example, to thereby determine or generate a modified cluster dataset.

[0099] At 512, the system 202 may repeat 504 to 510 using the modified cluster dataset to attempt to determine further clusters.

[00100] Referring now to Figure 6, there is shown an example t-distributed stochastic neighbor embedding (t-SNE) plot depicting an example clustering of contact names realised according to the methods of Figures 3, 4 and 5.

[00101] For a dataset associated with the entity identifier of the ABN for “WOOLWORTHS GROUP LIMITED”, the system 202, in performing methods 300 and 400, determined two clusters with a sufficiently high clustering score. The first cluster, shown in Figure 6 as 602, is a cluster of all contact names of the set {‘big w’, ‘bigw’}. The second cluster, shown in Figure 5 as 604, is a cluster of all contact name of the set {‘woolworths’, ‘woolworths##’, ‘woolworth’s’, ‘woolwhorts’, ‘wooloworths’, ‘woolowrths’, ‘woolsworths’, ‘woolwort’}. Although other clusters were identified, the clustering score associated with those clusters was too low and did not meet the clustering score threshold. Accordingly, they were not identified as valid or reliable clusters.

[00102] It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the above-described embodiments, without departing from the broad general scope of the present disclosure. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.

CLAIMS

1. A method comprising:

determining a dataset associated with an entity attribute, the dataset comprising character strings for each of a plurality of first attributes;

for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a candidate character string of the character string pair to a target character string of the character string pair, and wherein the similarity measure for the character string is based on a lowest value of:

(i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and

(ii) a minimum number of character edits required to convert the candidate character string into the target character string;

determining one or more clusters of the first attributes of the plurality of first attributes according to the similarity measures of the respective first attributes;

determining a representative label for each of the one or more clusters; and

determining a set of first attribute labels associated with the entity attribute, the set of first attribute labels comprising the representative labels for each of the one or more clusters.

2. The method of claim 1, further comprising:

determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string.

3. The method of claim 1 or 2, wherein the entity attribute is an organisation identifier and/or the plurality of first attribute types are a plurality of brand names or business names.

4. A method comprising:

determining a dataset associated with an organisation identifier, the dataset comprising character strings for each of a plurality of brand names;

for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a candidate character string of the character string pair to a target character string of the character string pair;

determining one or more clusters of the brand names of the plurality of brand names according to the similarity measures of the respective brand names;

determining a representative label for each of the one or more clusters; and

associating a set of brand labels with the organisation identifier, the set of brand labels comprising the representative labels for each of the one or more clusters.

5. The method of claim 4, wherein the similarity measure is based on the Levenshtein edit-distance between the character strings of the pair of character strings.

6. The method of claim 4, wherein the similarity measure comprises the Levenshtein edit-distance between the character strings of the pair of character strings divided by a length of the longer character string of the pair of character strings.

7. The method of claim 4, wherein the similarity measure for the character string pair is based on a lowest value of:

(i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and

(ii) a minimum number of character edits required to convert the candidate character string into the second character string.

8. The method of claim 7, further comprising:

determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string.

9. The method of any one of claims 1 to 3, 7 or 8, wherein determining the lowest value comprises:

initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string;

performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and

determining the lowest value as the minimum value of the shortest path vector.

10. The method of any one of the preceding claims, further comprising:

determining a clustering score for each of the one or more clusters;

comparing the clustering scores to a clustering score threshold; and

determining validated clusters as only those clusters of the one or more clusters that have a respective clustering score that meets or exceeds the clustering score threshold;

wherein determining the representative label for each of the one or more clusters comprises determining the representative label for each of the one or more validated clusters only.

11. The method of any one of the preceding claims, wherein determining the one or more clusters comprises:

a) determining a cluster dataset associated with the entity identifier, the cluster dataset comprising a plurality of character strings and for each character string, an associated set of similarity measures and a weight, wherein the weight is a fraction of an overall number of unique origins in the cluster dataset that are associated with the respective character string;

b) determining each character string of the dataset a centroid, and for each centroid, determining a set of character strings of the cluster dataset that are sufficiently close to the centroid;

c) adding the weights of the character strings of the set of character string considered sufficiently similar to the respective centroid to determine a centroid cluster value;

d) determining the centroid with the highest centroid cluster value;

e) determining that highest centroid cluster value meets a cluster value threshold; and

f) determining the centroid with the highest centroid cluster value a valid cluster corresponding to a first attribute of the entity identifier.

12. The method of claim 11, further comprising:

g) removing the centroid and the set of character strings belonging to the determined cluster from the cluster dataset, and repeating steps b) to f).

13. The method of any one of the preceding claims, further comprising:

populating an entry associated with the entity attribute or organisation identifier in a register based on the determined set of first attribute labels or brand label, the register comprising a plurality of entries and associated first attribute labels or brand labels.

14. A method comprising:

determining a character string pair, the character string pair comprising a candidate character string and a target character string, wherein both the candidate character string and the target character string represents a first attribute of an entity;

determining a similarity measure for the character string pair, wherein the similarity measure is indicative of the similarity of the candidate character string of the character string pair to the target character string of the character string pair,

wherein the similarity measure for the character string is based on a lowest value of:

(i) a minimum number of character edits required to convert the candidate character string into a substring of the target character string; and

(ii) a minimum number of character edits required to convert the candidate character string into the target character string; and

determining the similarity measure for the character string pair as being the lowest value divided by the length of the candidate character string.

15. The method of claim 14, wherein determining the lowest value comprises:

initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string;

performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and

determining the lowest value as the minimum value of the shortest path vector.

16. A method comprising:

determining a character string pair, the character string pair comprising a candidate character string and a target character string, wherein both the candidate character string and the target character string represents a first attribute of an entity;

determining a similarity measure for the character string pair, wherein the similarity measure is indicative of the similarity of the candidate character string of the character string pair to the target character string of the character string pair,

wherein determining the similarity measure for the character string comprises:

initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string;

performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and

determining the similarity measure as the minimum value of the shortest path vector.

17. A method comprising:

determining a dataset associated with an entity attribute, the dataset comprising character strings for each of a plurality of first attributes;

for each character string pair of the character strings of the dataset, determining a similarity measure, wherein the similarity measure is indicative of the similarity of a candidate character string of the character string pair to a target character string of the character string pair, and wherein determining the similarity measure comprises:

initialising with zeros a shortest path vector of a shortest path from a character of the candidate character string to each character of the target character string;

performing an edit distance calculation to determine the shortest path vector from the candidate character string to the target character string or a substring of the target character string using the initialised shortest path vector; and

determining the similarity measure as the minimum value of the shortest path vector;

determining one or more clusters of the first attributes of the plurality of first attributes according to the similarity measures of the respective first attributes;

determining a representative label for each of the one or more clusters; and

determining a set of first attribute labels associated with the entity attribute, the set of first attribute labels comprising the representative labels for each of the one or more clusters.

18. The method of any one of claims 1 to 3 or 7 to 17, wherein each character edit corresponds to a delete operation, an insertion operation and/or a replace operation.

19. A method comprising:

a) determining a cluster dataset associated with an entity identifier, the cluster dataset comprising a plurality of character strings and for each character string, an associated set of similarity measures and a weight, wherein the weight is a fraction of an overall number of unique origins in the cluster dataset that are associated with the respective character string;

b) determining each character string of the dataset a centroid, and for each centroid, determining a set of character strings of the cluster dataset that are sufficiently close to the centroid;

c) adding the weights of the character strings of the set of character string considered sufficiently similar to the respective centroid to determine a centroid cluster value;

d) determining the centroid with the highest centroid cluster value;

e) determining that highest centroid cluster value meets a cluster value threshold; and

f) determining the centroid with the highest centroid cluster value a valid cluster corresponding to a first attribute of the entity identifier.

20. The method of claim 19, further comprising:

g) removing the centroid and the set of character strings belonging to the determined cluster from the cluster dataset, and repeating steps b) to f).

21. The method of claim 19 or 20, wherein the associated set of similarity measures are determined according to the method of any one of claims 1 to 18.

22. A system comprising:

one or more processors; and

memory comprising computer executable instructions, which when executed by the one or more processors, cause the system to perform the method of any one of claims 1 to 21.

23. A computer-readable storage medium storing instructions that, when executed by a computer, cause the computer to perform the method of any one of claims 1 to 21.