

SAMPLE CLEANING METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present disclosure claims the priority to the Chinese Patent Application CN202410332295.6 filed on March 22, 2024, and entitled “SAMPLE CLEANING METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM”, which is incorporated herein by reference in entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the technical field of artificial intelligence, and particularly to a sample cleaning method and apparatus, an electronic device, and a storage medium.

BACKGROUND ART

[0003] The quality of text labeling in a corpus has a significant impact on result of a text model, and high-quality corpus data can help text model obtain a better service result after training. Therefore, a text classification task typically requires accurate and unique labeling results of texts in the corpus. However, in practical service, due to wide sources and rapid growth of texts in the corpus, there may be inaccurate labeling and inconsistent labeling results for the same text over time, and such labeling results will adversely affect performance of the model after training. Therefore, it is necessary to clean the corpus so as to elevate data quality of the corpus, thereby improving a service effect of the model after training.

[0004] In the prior art, the corpus is typically cleaned by the following methods. In method 1, all labeled texts are re-checked for quality, by which method, mislabeled samples can be found, but it is equivalent to performing a second round of labeling, thus incurring high labor and time investment costs. In method 2, labeled texts are divided into batches, and each batch of data is randomly sampled to check labeling quality. This method avoids performing a full-coverage quality check on the texts, but if a pass rate of each batch of data is relatively low, it is equivalent to full-coverage re-labeling, thus leading to cost investment comparable to that in Method 1. In method 3, it is assumed that a low-quality text is prone to mislabeling, and a classification model for identifying text quality is used to remove low-quality texts in the corpus. This method can render a corpus with few labeling errors, but cannot determine whether a mislabeled high-quality text is mislabeled; as low-quality texts have a certain service

value, removing them does not meet service requirement, and additionally training a classification model for determining whether the text quality is high or low is also costly.

SUMMARY

[0005] The present disclosure provides a sample cleaning method and apparatus, an electronic device and a storage medium, which can quickly locate mislabeled samples in a corpus, with high accuracy of mislabeled samples determined, thereby effectively reducing time and labor cost investment.

[0006] In the first aspect, the present disclosure provides a sample cleaning method, including:

[0007] performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories;

[0008] determining tag categories contained in the sample set of each cluster category, and dividing the sample set of each of the cluster categories into a plurality of sample subsets based on the tag categories; and

[0009] determining, for the sample set of each of the cluster categories, a mislabeled target sample subset from the plurality of sample subsets, thus implementing a sample cleaning task for the corpus.

[0010] Further, the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories includes: determining multiple clustering features and weights corresponding to respective clustering features; performing semantic clustering on the samples based on the multiple clustering features and the weights corresponding to the respective clustering features so as to obtain the sample sets of the multiple cluster categories, where the clustering features include at least one of text edit distance, keyword semantic similarity and text semantic similarity; and the sample sets each include a plurality of samples and label tags corresponding to respective samples.

[0011] Further, for the sample set of a current cluster category, determining the mislabeled target sample subset from the plurality of sample subsets includes: determining number of samples in each sample subset; if the number of samples in the plurality of sample subsets is the same, determining that there is no mislabeled target sample subset in the sample set of the current cluster category; if the number of samples in each of the plurality of sample subsets is different, determining the maximum number of samples and the minimum number of samples based on sample data of the plurality of sample subsets; and calculating whether each sample subset is a mislabeled target sample subset successively based on the number of samples in

each sample subset, the maximum number of samples and the minimum number of samples, so as to obtain the mislabeled target sample subset in the sample set of the current cluster category.

[0012] Further, the step of calculating whether each sample subset is a mislabeled target sample subset successively based on the number of samples in each sample subset, the maximum number of samples and the minimum number of samples includes: for the current sample subset, calculating a mislabeling probability of the current sample subset based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples; if the mislabeling probability is greater than a first preset value, determining the current sample subset as the mislabeled target sample subset; and traversing the plurality of sample subsets, and obtaining the mislabeling probability of each sample subset, thus determining, for each sample subset, whether it is the mislabeled target sample subset.

[0013] Further, the mislabeling probability of the current sample subset is calculated based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples through the following formula:

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

[0014] In the formula, S represents the mislabeling probability, i represents an index number of the current sample subset among the plurality of sample subsets, m_i represents the number of samples in the current sample subset, m1 represents the maximum number of samples, and m2 represents the minimum number of samples.

[0015] Further, the step of determining a mislabeled target sample subset from the plurality of sample subsets includes: determining the number of samples in each sample subset; and if the number of samples is smaller than a second preset value, determining the sample subset corresponding to the number of samples as the mislabeled target sample subset.

[0016] Further, prior to the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories, the method further includes: performing format conversion on the samples in the corpus based on a predefined format to obtain processed samples.

[0017] In the second aspect, the present disclosure provides a sample cleaning apparatus, including:

[0018] a sample clustering module, configured to perform semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories;

[0019] a sample division module, configured to determine tag categories contained in the sample set of each cluster category, and divide the sample set of each cluster category into a plurality of sample subsets based on the tag categories; and

[0020] a sample cleaning module, configured to determine a mislabeled target sample subset from the plurality of sample subsets for the sample set of each of the cluster categories, thus implementing a sample cleaning task for the corpus.

[0021] In the third aspect, the present disclosure provides an electronic device, including: at least one processor and a memory communicating with the at least one processor, where the memory stores a computer program executable by the at least one processor, and the computer program is executed by the at least one processor to enable the at least one processor to implement the sample cleaning method according to any embodiment of the present disclosure.

[0022] In the fourth aspect, the present disclosure provides a computer-readable storage medium, storing computer instructions, where the computer instructions, when executed by a processor, implement the sample cleaning method according to any embodiment of the present disclosure.

[0023] In the fifth aspect, embodiments of the present disclosure provide a computer program product, including computer program instructions, where the computer program instructions, when read and run by a processor, implement the sample cleaning method according to any embodiment of the present disclosure.

[0024] In order to overcome the defects of the prior art in Background Art, embodiments of the present disclosure provide a sample cleaning method, and the following beneficial effects can be brought about by implementing the method. The present disclosure first clusters the samples in the corpus into a plurality of sample sets, then divides each sample set into corresponding sample subsets according to the tag categories contained therein, and finally analyses each sample subset to determine whether each sample subset is mislabeled. The present disclosure enables rapid localization of mislabeled samples in the corpus, with high accuracy of mislabeled samples determined, thereby effectively reducing time and labor cost investments.

[0025] It should be noted that the above computer instructions can be fully or partially stored on a computer-readable storage medium. In the above, the computer-readable storage medium can be packaged together with or separately from a processor of the sample cleaning apparatus, which is not limited in the present application.

[0026] For description of the second aspect, the third aspect and the fourth aspect in the present disclosure, reference can be made to detailed description of the first aspect. Moreover,

for the beneficial effects described in the second aspect, the third aspect and the fourth aspect, reference can be made to the beneficial effect analysis of the first aspect, which is not repeated herein.

[0027] It should be understood that the contents described in this section are neither intended to identify key or critical features of embodiments of the present disclosure, nor to limit the scope of the present disclosure. Other features of the present disclosure will be easy to understand from the following description.

[0028] It can be understood that, before using the technical solutions disclosed in various embodiments of the present disclosure, in compliance with applicable laws and regulations, users shall be informed through appropriate means regarding types, scope of use, and applicable scenarios, etc. of personal information involved in the present disclosure, and users' explicit authorization shall be obtained.

BRIEF DESCRIPTION OF DRAWINGS

[0029] In order to illustrate technical solutions of embodiments of the present disclosure more clearly, drawings required for use in the description of the embodiments will be described briefly below. Obviously, the drawings in the following description are merely for some embodiments of the present disclosure, and those of ordinary skill in the art still could obtain other drawings in light of these drawings without using any inventive efforts.

[0030] FIG. 1 is a first schematic flowchart of a sample cleaning method provided by embodiments of the present disclosure;

[0031] FIG. 2 is a second schematic flowchart of the sample cleaning method provided by embodiments of the present disclosure;

[0032] FIG. 3 is a structural schematic view of a sample cleaning apparatus provided by embodiments of the present disclosure; and

[0033] FIG. 4 is a block diagram of an electronic device for implementing the sample cleaning method in embodiments of the present disclosure.

DETAILED DESCRIPTION OF EMBODIMENTS

[0034] In order to make objectives, technical solutions and advantages of the embodiments of the present disclosure clearer, the technical solutions in the embodiments of the present disclosure will be described clearly and completely below in conjunction with drawings in the embodiments of the present disclosure, and apparently, only some but not all embodiments of the present disclosure are described. All of other embodiments obtained by those ordinarily

skilled in the art based on the embodiments in the present disclosure without using any inventive efforts shall fall within the scope of protection of the present disclosure.

[0035] It should be noted that the terms “first”, “second”, “target”, “original”, etc. in the description, the claims, and the drawings of the present disclosure are used for distinguishing similar objects, rather than for describing a specific order or sequence. It should be understood that the terms used in this way are exchangeable under appropriate circumstances, so that the embodiments of the present disclosure described herein can be implemented in an order other than those illustrated or described herein. Besides, terms “comprise”, “have” and any variants thereof are intended to cover non-exclusive inclusions, for example, a process, method, system, product or apparatus including a series of steps or units is not necessarily limited to those steps or units expressly listed, but may include other steps or units not expressly listed or inherent to this process, method, product or apparatus.

[0036] FIG. 1 is a first schematic flowchart of a sample cleaning method provided by embodiments of the present disclosure. The present embodiment is applicable to cleaning samples in a corpus. The sample cleaning method provided by the present embodiment can be executed by the sample cleaning apparatus provided by embodiments of the present disclosure, and the apparatus can be implemented by means of software and/or hardware, and is integrated in an electronic device for executing the present method.

[0037] With reference to FIG. 1, the method in the present embodiment includes, but is not limited to, the following steps:

[0038] S110, performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories.

[0039] The corpus herein is a corpus having undergone a first round of labeling, and the sample sets each include a plurality of samples and corresponding labeling tags of respective samples.

[0040] Further, as there may be samples in different formats in the corpus, prior to the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories, the method further includes: converting the samples in the corpus to a unified format based on a predefined format to obtain processed samples, thereby preparing the corpus into samples to be clustered, the samples being presented in a text-tag form.

[0041] Specifically, the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories includes: determining multiple clustering features and weights corresponding to respective clustering features; performing the semantic clustering on the samples based on the multiple clustering features and the weights

corresponding to respective clustering features so as to obtain the sample sets of the multiple cluster categories. The clustering features herein include at least one of text edit distance, keyword semantic similarity and text semantic similarity; and the sample sets each include a plurality of samples and label tags corresponding to respective samples. According to a preferred embodiment of the present disclosure, the clustering features may include, for example, the text edit distance, the keyword semantic similarity and the text semantic similarity.

[0042] In an embodiment of a content security service scenario, the weight of the text edit distance can be set to be 0.3, the weight of the keyword semantic similarity can be set to be 0.3, and the weight of the text semantic similarity can be set to be 0.4. The clustering features and the corresponding weights in the present embodiment can be adjusted according to different service scenarios. Compared with directly using the text semantics to perform the clustering, this clustering method has higher accuracy, and is suitable for content security service scenarios.

[0043] Optionally, scores of the text edit distance and the semantic similarity can use, for example, a k-Nearest Neighbor (kNN) or other clustering methods to perform the semantic clustering on the samples in the corpus. A more accurate clustering method can further narrow a range of samples to be checked for quality in the corpus, thereby reducing cost investments of quality check.

[0044] Exemplarily, it is assumed that there are ten thousand samples in a corpus, each sample consists of a text and a corresponding tag, and it is assumed that tags are “a, b, c, d, e, and f” respectively. The text edit distance, the keyword semantic similarity and the text semantic similarity of the texts are assigned different weights, and sample sets of 100 cluster categories are generated through the semantic clustering, numbered 0 to 99 respectively.

[0045] S120, determining tag categories contained in the sample set of each cluster category, and dividing the sample set of each cluster category into a plurality of sample subsets based on the tag categories.

[0046] In the above, the number of sample subsets is the same as the number of tag categories contained in each of the sample sets, and one tag category is corresponding to one sample subset.

[0047] Exemplarily, it is assumed that there are 500 samples in total and five tag categories, which are “a, b, c, d, and e” respectively, in a sample set numbered 0. Then, the 500 samples are divided into five sample subsets according to the tag categories.

[0048] S130, determining a mislabeled target sample subset from the plurality of sample subsets for the sample set of each cluster category, thus implementing a sample cleaning task for the corpus.

[0049] In embodiments of the present disclosure, with regard to a sample set of a current cluster category, a mislabeled target sample subset is determined from a plurality of sample subsets in the sample set of the current cluster category; and the sample sets of all cluster categories are traversed, and the mislabeled target sample subsets in respective sample sets can be obtained, so that the sample cleaning task for the corpus can be implemented. The target sample subsets herein are samples to be checked for quality in the corpus.

[0050] Specifically, the step of determining a mislabeled target sample subset from the plurality of sample subsets includes: determining the number of samples in each sample subset; and if the number of samples is smaller than a second preset value, determining the sample subset corresponding to the number of samples as the mislabeled target sample subset.

[0051] In embodiments of the present disclosure, it is generally recognized in the art that texts with similar semantics probably have an identical tag. Therefore, the samples in the corpus can be clustered, thereby counting distribution of the number of samples of each tag in each cluster.

[0052] It is generally recognized in the art that texts corresponding to a label with a relatively small number of samples within a cluster are probably mislabeled texts, and require quality check. The second preset value can be set according to actual application requirements.

[0053] Exemplarily, it is assumed that statistics reveal that there are 500 samples and five tag categories, which are “a, b, c, d, and e” respectively, in the sample set numbered 0, in which there are 300 samples in a sample subset with the tag a, 100 samples in a sample subset with the tag b, 80 samples in a sample subset with the tag c, 10 samples in a sample subset with the tag d, and 10 samples in a sample subset with the tag e. In this case, within this cluster category, the samples corresponding to the tag d and the tag e are probably mislabeled samples and should be checked for quality. Based on the same method, by traversing the sample sets of the cluster categories from 0 to 99, potential mislabeled samples can be screened out.

[0054] In the technical solution provided by the present embodiment, the sample sets of multiple cluster categories are obtained by performing the semantic clustering on the samples in the corpus; the tag categories contained in the sample set of each cluster category are determined, and the sample set of each cluster category is divided into a plurality of sample subsets based on the tag categories; for the sample set of each cluster category, the mislabeled target sample subset is determined from the plurality of sample subsets, thus implementing the sample cleaning task for the corpus. The present disclosure first clusters the samples in the corpus into a plurality of sample sets, then divides each sample set into corresponding sample subsets according to the tag categories contained therein, and finally analyses each sample

subset to determine whether each sample subset is mislabeled. The present disclosure enables rapid localization of mislabeled samples in the corpus, with high accuracy of mislabeled samples determined, thereby effectively reducing time and labor cost investments.

[0055] The sample cleaning method provided by embodiments of the present disclosure is further described below. FIG. 2 is a second schematic flowchart of the sample cleaning method provided by an embodiment of the present disclosure. The present embodiment is optimization on the basis of the above various embodiments, with specific optimization as follows: for the sample set of the current cluster category, the present embodiment illustrates in detail a process of determining the mislabeled target sample subset from the plurality of sample subsets.

[0056] With reference to FIG. 2, the method in the present embodiment includes, but is not limited to, following steps:

[0057] S210, determining the number of samples in each sample subset.

[0058] In the present embodiment, the semantic clustering is performed on the samples in the corpus to obtain the sample sets of multiple cluster categories, the tag categories contained in the sample set of each cluster category are determined, and the sample set of each cluster category is divided into a plurality of sample subsets based on the tag categories. For the sample set of each cluster category, the number of samples in each sample subset is determined, and whether the number of samples in each sample subset is the same is determined.

[0059] S220, if the number of samples in the plurality of sample subsets is the same, determining that there is no mislabeled target sample subset in the sample set of the current cluster category.

[0060] In embodiments of the present disclosure, when the number of samples in a plurality of sample subsets is the same, it is indicated that a mislabeling probability of corresponding sample set is 0, demonstrating that there is no mislabeled sample in the sample set.

[0061] S230, if the number of samples in the plurality of sample subsets is different, determining the maximum number of samples and the minimum number of samples based on sample data of the plurality of sample subsets.

[0062] In embodiments of the present disclosure, when the number of samples in the plurality of sample subsets is different, the sample subset of which tag category has the maximum number of samples is determined, so as to obtain the maximum number of samples; and the sample subset of which tag category has the minimum number of samples is determined, so as to obtain the minimum number of samples.

[0063] S240, for the current sample subset, calculating a mislabeling probability of the current sample subset based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples.

[0064] Specifically, the mislabeling probability of the current sample subset is calculated based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples through the following formula:

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

[0065] In the formula, S represents the mislabeling probability (equivalent to tag error rate score), being a normalization score of the number of tags in the sample set of each cluster category, i represents an index number of a current sample subset among the plurality of sample subsets, m_i represents the number of samples in the current sample subset, m1 represents the maximum number of samples, and m2 represents the minimum number of samples.

[0066] The larger the value of S corresponding to a certain sample subset is, the higher the probability that the sample subset is mislabeled (i.e. the sample subset is the target sample subset), and vice versa. The mislabeling probability of a certain sample subset is associated with the number of tag categories within that sample subset.

[0067] S250, if the mislabeling probability is greater than a first preset value, determining the current sample subset as the mislabeled target sample subset.

[0068] The first preset value can be set according to different practical application scenarios. According to a preferred embodiment of the present disclosure, in a service security scenario, the first preset value can be set to be, for example, 0.8. Experiments prove that a range of samples to be checked for quality, as determined by a conventional method, accounts for 50% to 100% of a total number of samples in the corpus. In the service security scenario, the sample subsets with the mislabeling probability greater than 0.8 are checked for quality, and finally by checking only a small part of the samples (about 10% of the total number of the samples in the corpus, saving the labor cost by more than 50% compared with the conventional method), the corpus can be finally enabled to achieve the same sample accuracy rate as the conventional method (the accuracy rate is generally higher than 99.5%), thus significantly reducing the cost investments for cleaning the corpus.

[0069] S260, traversing the plurality of sample subsets, and obtaining the mislabeling probability of each sample subset, thus obtaining that there is a mislabeled target sample subset in the sample set of the current cluster category.

[0070] In the embodiment corresponding to FIG. 1, when the number of samples in a certain sample subset is smaller than the second preset value, this sample subset is considered as being mislabeled. Because the number of tag categories in the sample set of each cluster category is different, it is difficult to uniformly set a numerical value of the second preset value for the sample set of each cluster category, and then whether the sample subset is mislabeled can also be determined by the method for calculating the mislabeling probability of the sample subset in the present embodiment.

[0071] In the technical solution provided by the present embodiment, the number of samples in each sample subset is determined; if the number of samples in the plurality of sample subsets is the same, it is determined that there is no mislabeled target sample subset in the sample set of the current cluster category; if the number of samples in the plurality of sample subsets is different, the maximum number of samples and the minimum number of samples are determined based on sample data of the plurality of sample subsets; for the current sample subset, the mislabeling probability of the current sample subset is calculated based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples; if the mislabeling probability is greater than the first preset value, the current sample subset is determined as the mislabeled target sample subset; and the plurality of sample subsets are traversed, and the mislabeling probability of each sample subset is obtained, thus obtaining that there is a mislabeled target sample subset in the sample set of the current cluster category. The present disclosure determines whether each sample subset is mislabeled by the method of calculating the mislabeling probability of each sample subset, and can reduce the number of target sample subsets (i.e. samples to be checked for quality) compared with the conventional sample cleaning method. The present disclosure further enables rapid localization of mislabeled samples in the corpus, with high accuracy of mislabeled samples determined, thereby effectively reducing time and labor cost investments.

[0072] FIG. 3 is a structural schematic view of a sample cleaning apparatus provided embodiments of the present disclosure. As shown in FIG. 3, the apparatus 300 may include:

[0073] a sample clustering module 310, configured to perform semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories;

[0074] a sample division module 320, configured to determine tag categories contained in the sample set of each cluster category, and divide the sample set of each cluster category into a plurality of sample subsets based on the tag categories; and

[0075] a sample cleaning module 330, configured to determine a mislabeled target sample subset from the plurality of sample subsets for the sample set of each cluster category, thus implementing a sample cleaning task for the corpus.

[0076] Further, the above sample clustering module 310 can be specifically configured to determine multiple clustering features and weights corresponding to respective clustering features; perform the semantic clustering on the samples based on the multiple clustering features and the weights corresponding to respective clustering features so as to obtain the sample sets of the multiple cluster categories, where the clustering features include at least one of text edit distance, keyword semantic similarity and text semantic similarity; and the sample sets each include a plurality of samples and label tags corresponding to respective samples.

[0077] Further, the above sample cleaning module 330 can be specifically configured to: determine the number of samples in each sample subset for the sample set of the current cluster category; if the number of samples in the plurality of sample subsets is the same, determine that there is no mislabeled target sample subset in the sample set of the current cluster category; if the number of samples in the plurality of sample subsets is different, determine the maximum number of samples and the minimum number of samples based on sample data of the plurality of sample subsets; calculate whether each sample subset is a mislabeled target sample subset successively based on the number of samples in each sample subset, the maximum number of samples and the minimum number of samples, thus obtaining that there is a mislabeled target sample subset in the sample set of the current cluster category.

[0078] Further, the above sample cleaning module 330 can be further specifically configured to: for the current sample subset, calculate a mislabeling probability of the current sample subset based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples; determine the current sample subset as the mislabeled target sample subset if the mislabeling probability is greater than a first preset value; and traverse the plurality of sample subsets, and obtain the mislabeling probability of each sample subset, thus obtaining whether each sample subset is the mislabeled target sample subset

[0079] Further, the above sample cleaning module 330 can be further specifically configured to: calculate the mislabeling probability of the current sample subset based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples through the following formula:

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

[0080] In the formula, S represents the mislabeling probability, i represents an index number of a current sample subset among the plurality of sample subsets, m_i represents the number of samples in the current sample subset, m_1 represents the maximum number of samples, and m_2 represents the minimum number of samples.

[0081] Further, the above sample cleaning module 330 can be specifically configured to: determine the number of samples in each sample subset; and if the number of samples is smaller than a second preset value, determine that the sample subset corresponding to the number of samples as the mislabeled target sample subset.

[0082] Further, the above sample cleaning apparatus further may include: a sample processing module.

[0083] The sample processing module is configured to, prior to performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories, perform format conversion on the samples in the corpus based on a predefined format to obtain processed samples.

[0084] The sample cleaning apparatus provided by the present embodiment is applicable to the sample cleaning method provided by any of the above embodiments, and has corresponding functions and beneficial effects.

[0085] FIG. 4 is a block diagram of an electronic device for implementing the sample cleaning method in embodiments of the present disclosure. An electronic device 10 is intended to represent various forms of digital computers, such as laptop computers, desktop computers, workstations, personal digital assistants, servers, blade servers, mainframe computers and other suitable computers. The electronic device may further represent various forms of mobile apparatus, such as personal digital processors, cellular phones, smart phones, wearable device (e.g., helmets, glasses, and watches) and other similar computing apparatus. Components shown herein, their connections and relationships and their functions are merely by way of example, and are not intended to limit implementation of the present disclosure described and/or claimed herein.

[0086] As shown in FIG. 4, the electronic device 10 includes at least one processor 11 and a memory communicating with the at least one processor 11, such as a read only memory (ROM) 12, and random access memory (RAM) 13, where a computer program executable by the at least one processor is stored in the memory, the at least one processor 11 may perform various appropriate actions and processes according to a computer program stored in the read only memory (ROM) 12 or a computer program loaded from a storage unit 18 into the random

access memory (RAM) 13. Various programs and data necessary for operations of the electronic device 10 can also be stored in the RAM 13. The processor 11, the ROM 12 and the RAM 13 are connected to each other via a bus 14. An input/output (I/O) interface 15 is also connected to the bus 14.

[0087] A plurality of components in the electronic device 10 are connected to the I/O interface 15, including: an input unit 16, such as a keyboard and mouse; an output unit 17, such as various types of displays and speakers; a storage unit 18, such as a magnetic disk and a compact disk; and a communication unit 19 such as a network card, a modem, and a wireless communication transceiver. The communication unit 19 allows the electronic device 10 to exchange information/data with other devices through a computer network such as the Internet and/or various telecommunication networks.

[0088] The processor 11 may be various general-purpose and/or special-purpose processing components having processing and computing capabilities. Some examples of the processor 11 include, but are not limited to, a central processing unit (CPU), a graphics processing unit (GPU), various special-purpose artificial intelligence (AI) computing chips, various processors running machine learning model algorithms, digital signal processors (DSPs) and any appropriate processors, controllers, microcontrollers, etc. The processor 11 executes various methods and processes described above, such as the sample cleaning method.

[0089] In some embodiments, the sample cleaning method may be implemented as a computer program tangibly embodied on a computer-readable storage medium, such as the storage unit 18. In some embodiments, the computer program may be partially or fully loaded into and/or installed on the electronic device 10 via the ROM 12 and/or the communication unit 19. When the computer program is loaded into the RAM 13 and executed by the processor 11, one or more steps of the sample cleaning method described in the above can be performed. Alternatively, in other embodiments, the processor 11 can be configured to perform the sample cleaning method by any other suitable means (for example, by means of firmware).

[0090] Various implementations of systems and techniques described herein above may be implemented in digital electronic circuitry, integrated circuitry, field programmable gate array (FPGA), application specific integrated circuit (ASIC), application specific standard product (ASSP), system on a chip (SOC), complex programmable logic device (CPLD), computer hardware, firmware, software and/or their combinations. These various embodiments may include: implementation in one or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable processor, where the at least one programmable processor may be special-purpose or general-purpose

programmable processor, can receive data and instructions from a storage system, at least one input apparatus and at least one output apparatus, and transmit data and instructions to the storage system, the at least one input apparatus and the at least one output apparatus.

[0091] Computer programs for implementing the method of the present disclosure can be written in one or any combination of more programming languages. These computer programs can be provided to a processor of a general-purpose computer, a special-purpose computer or other programmable data processing apparatus, such that the computer programs, when executed by the processor, implement functions/operations specified in the flowchart and/or block diagram. The computer programs can be executed fully on a machine, partly on the machine, as a stand-alone software package partly on the machine and partly on a remote computer or fully on the remote computer or server.

[0092] In the context of the present disclosure, the computer-readable storage medium may be a tangible medium that can contain or store a computer program for use by or in connection with an instruction execution system, apparatus or device. The computer-readable storage medium may include, but is not limited to, an electronic, magnetic, optical, electromagnetic, infrared or semiconductor system, apparatus or device, or any suitable combination of the above contents. Alternatively, the computer-readable storage medium may be a machine-readable signal medium. More specific examples of the machine-readable storage medium include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or flash memory), an optical fiber, a portable compact disc read-only memory (CD ROM), an optical storage device, a magnetic storage device or any suitable combination of the above.

[0093] In order to provide for interaction with a user, systems and techniques described herein can be implemented on an electronic device having: a display device (for example, a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user; and a keyboard and a pointing device (for example, a mouse or a trackball) by which the user can provide input to the electronic device. Other kinds of devices can also be used to provide for interaction with the user; for example, feedback provided to the user can be any form of sensory feedback (for example, visual feedback, auditory feedback or tactile feedback); and input from the user can be received in any form (including acoustic, speech, or tactile input).

[0094] The systems and techniques described herein can be implemented in a computing system (for example, as a data server) that includes a background component or a computing system (for example, an application server) that includes an middleware component or a

computing system (for example, a user computer having a graphical user interface or a web browser, the user can interact with embodiments of the systems and techniques described herein through the graphical user interface or the web browser) including a front-end component or a computing system that includes any combination of such background component, middleware component or front-end component. The components of the system can be interconnected by digital data communication (for example, communication network) of any form or medium. Examples of communication network include a local area network (LAN), a wide area network (WAN), a block chain network and the Internet.

[0095] A computing system may include a client and a server. The client and server are generally remote from each other and typically interact through a communication network. A relationship between the client and the server is generated by programs running on respective computers, and there is the client-server relationship between these programs. The server may be a cloud server, also referred to as a cloud computing server or a cloud host, and is a host product in a cloud computing service system, so as to solve the defects of great management difficulty and poor service expansibility in traditional physical hosts and VPS services.

[0096] Note that the above are only for preferred embodiments and technical principles used in the present disclosure. Those skilled in the art would understand that the present disclosure is not limited to the specific embodiments described herein, and various obvious changes, readjustments and substitutions could be made by those skilled in the art without departing from the scope of protection of the present disclosure. For example, those skilled in the art could rearrange, add or delete steps using various forms of flows shown above; the steps described in the present disclosure may be executed in parallel, sequentially or in different orders, as long as desired results of the technical solutions of the present disclosure can be achieved, which is not limited herein.

[0097] The above embodiments do not constitute limitation to the scope of protection of the technical solution. Those skilled in the art should understand that various modifications, combinations, sub-combinations and replacements could be made according to design requirements and other factors. Any modifications, equivalent substitutions, improvements and so on made within the spirit and principle of the present disclosure should be covered within the scope of protection of the present disclosure.

CLAIMS

What is claimed is:

1. A sample cleaning method, comprising steps of:
performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories;
determining tag categories contained in the sample set of each cluster category, and dividing the sample set of each of the cluster categories into a plurality of sample subsets based on the tag categories; and
determining, for the sample set of each of the cluster categories, a mislabeled target sample subset from the plurality of sample subsets, thus implementing a sample cleaning task for the corpus.
2. The sample cleaning method according to claim 1, wherein the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories comprises:
determining multiple clustering features and weights corresponding to respective clustering features;
performing semantic clustering on the samples based on the multiple clustering features and the weights corresponding to the respective clustering features so as to obtain the sample sets of the multiple cluster categories,
wherein the clustering features comprise at least one of text edit distance, keyword semantic similarity and text semantic similarity; and the sample sets each comprise a plurality of samples and label tags corresponding to respective samples.
3. The sample cleaning method according to claim 1, wherein for the sample set of a current cluster category, the step of determining a mislabeled target sample subset from the plurality of sample subsets comprises steps of:
determining number of samples in each sample subset;
if the number of samples in each of the plurality of sample subsets is the same, determining that there is no mislabeled target sample subset in the sample set of the current cluster category;
if the number of samples in each of the plurality of sample subsets is different, determining a maximum number of samples and a minimum number of samples based on sample data of the plurality of sample subsets; and

calculating whether each sample subset is a mislabeled target sample subset successively based on the number of samples in each sample subset, the maximum number of samples and the minimum number of samples, so as to obtain the mislabeled target sample subset in the sample set of the current cluster category.

4. The sample cleaning method according to claim 3, wherein the step of calculating whether each sample subset is a mislabeled target sample subset successively based on the number of samples in each sample subset, the maximum number of samples and the minimum number of samples comprises:

for the current sample subset, calculating a mislabeling probability of the current sample subset based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples;

if the mislabeling probability is greater than a first preset value, determining the current sample subset as the mislabeled target sample subset; and

traversing the plurality of sample subsets, and obtaining the mislabeling probability of each sample subset, thus determining, for each sample subset, whether it is the mislabeled target sample subset.

5. The sample cleaning method according to claim 4, wherein the mislabeling probability of the current sample subset is calculated based on the number of samples in the current sample subset, the maximum number of samples and the minimum number of samples through the following formula:

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

in the formula, S represents the mislabeling probability, i represents an index number of the current sample subset among the plurality of sample subsets, m_i represents the number of samples in the current sample subset, m1 represents the maximum number of samples, and m2 represents the minimum number of samples.

6. The sample cleaning method according to claim 1, wherein the step of determining a mislabeled target sample subset from the plurality of sample subsets comprises:

determining the number of samples in each sample subset; and

if the number of samples is smaller than a second preset value, determining the sample subset corresponding to the number of samples as the mislabeled target sample subset.

7. The sample cleaning method according to claim 1, prior to the step of performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories, further comprising:

performing format conversion on the samples in the corpus based on a predefined format to obtain processed samples.

8. A sample cleaning apparatus, comprising:

a sample clustering module, configured to perform semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories;

a sample division module, configured to determine tag categories contained in the sample set of each cluster category, and divide the sample set of each cluster category into a plurality of sample subsets based on the tag categories; and

a sample cleaning module, configured to determine a mislabeled target sample subset from the plurality of sample subsets for the sample set of each of the cluster categories, thus implementing a sample cleaning task for the corpus.

9. An electronic device, comprising: at least one processor and a memory communicating with the at least one processor, wherein the memory stores a computer program executable by the at least one processor, and the computer program is executed by the at least one processor to enable the at least one processor to implement the sample cleaning method according to any one of claims 1 to 7.

10. A computer-readable storage medium, storing computer instructions, wherein the computer instructions, when executed by a processor, implement the sample cleaning method according to any one of claims 1 to 7.

11. A computer program product, comprising computer program instructions, wherein the computer program instructions, when read and run by a processor, implementing the sample cleaning method according to any one of claims 1 to 7.