DATA PROCESSING METHOD AND SYSTEM

TECHNICAL FIELD

[0001] This specification relates to the field of data processing technologies, and in particular, to a data processing method and system.

BACKGROUND

[0002] With rapid development of software engineering and information technology, as an important tool to express step logic, an algorithm structure, or a system workflow, a flowchart is widely used in fields such as software development, project management, and academic research. How to extract professional knowledge in the flowchart becomes particularly important. In a related technology, a logical relationship in the flowchart is extracted by manually establishing a rule. In such a manner, costs are high and maintenance is difficult. When the flowchart is updated, data needs to be manually updated.

[0003] Therefore, an automatic data processing method and system for a flowchart need to be provided.

[0004] Content of the background part is merely information learned of by the inventor, and neither means that the information has entered a public domain before an application date of this specification, nor means that the information can become a conventional technology of this specification.

SUMMARY

[0005] This specification provides a data processing method and system, which can be applied to a flowchart recognition scenario.

[0006] According to a first aspect, this specification provides a data processing method, including: obtaining arrow data of an arrow in a to-be-processed flowchart; obtaining character data of characters in the to-be-processed flowchart; and generating first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data,

where the first structured data comprises a connection relationship between the characters.

[0007] According to a second aspect, this specification further provides a data processing system, including at least one storage medium and at least one processor. The at least one storage medium stores at least one instruction set, to perform a data processing method. The at least one processor is communicatively connected to the at least one storage medium, and when running, the at least one processor reads the at least one instruction set, and performs the following operations based on an indication of the at least one instruction set: obtaining arrow data of an arrow in a to-be-processed flowchart; obtaining character data of characters in the to-be-processed flowchart; and generating first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data, where the first structured data comprises a connection relationship between the characters.

[0008] According to a third aspect, this specification further provides a computer-readable nonvolatile storage medium. The computer-readable nonvolatile storage medium stores at least one instruction set, and when the at least one instruction set is executed by at least one processor, the data processing method according to the first aspect is implemented.

[0009] It can be learned from the above-mentioned technical solutions that, according to the data processing method and the system provided in this specification, the arrow data of the arrow and the character data of the characters in the to-be-processed flowchart can be automatically extracted, and the connection relationship between the characters is determined based on the arrow data and the character data, to generate the first structured data. According to the data processing method and the system provided in this specification, the characters and the arrow in the flowchart can be recognized, to establish the connection relationship between the characters based on the arrow in the flowchart, thereby automatically performing knowledge extraction on the flowchart, to obtain a logical relationship in the flowchart. The first structured data obtained in this specification can be used for calculation of a subsequent algorithm or program, so that the subsequent algorithm or program can perform further data processing based on the logical relationship in the flowchart.

[0010] Other functions of the data processing method and system provided in this specification are partially listed in the following descriptions. Creative aspects of the data processing method and system provided in this specification can be fully explained by practice or using the methods, apparatuses, and combinations described in the following detailed examples.

BRIEF DESCRIPTION OF DRAWINGS

[0011] To describe the technical solutions in the embodiments of this specification more clearly, the following briefly describes the accompanying drawings needed for describing the embodiments. Clearly, the accompanying drawings in the following descriptions show merely some embodiments of this specification, and a person of ordinary skill in the art can still derive other drawings from these accompanying drawings without creative efforts.

[0012] FIG. 1 is a schematic diagram of an application scenario of data processing according to an embodiment of this specification;

[0013] FIG. 2 is a schematic diagram of a hardware structure of a computing device according to some embodiments of this specification;

[0014] FIG. 3 is a flowchart of a data processing method according to an embodiment of this specification; and

[0015] FIG. 4 is a flowchart of a data processing method according to another embodiment of this specification.

DESCRIPTION OF EMBODIMENTS

[0016] Specific application scenarios and requirements of this specification are provided in the following descriptions, so that a person skilled in the art can manufacture and use content of this specification. For the person skilled in the art, various local modifications to the disclosed embodiments are clear, and the general principles defined herein can be applied to other embodiments and applications without departing from the spirit and scope of this specification. Therefore, this specification is not limited to the shown embodiments, but has a widest scope consistent with that of the claims.

[0017] The terms used herein are only intended to describe a particular example embodiment, but do not impose a limitation. For example, unless otherwise specified in the context, the singular forms "one", "a", and "the" used herein can also include a plural form. When used in this specification, the terms "include", "comprise", and/or "have" indicate/indicates existence of an associated integer, step, operation, element, and/or component, but does not exclude existence of one or more other features, integers, steps, operations, elements, components, and/or groups, or another feature, integer, step, operation, element, component, and/or group can be added to the

system/method.

[0018] In consideration of the following descriptions, these and other features of this specification, operations and functions of related components of the structure, and economy of combination and manufacturing of components can be significantly improved. With reference to the accompanying drawings, all of these form a part of this specification. However, it should be clearly understood that the accompanying drawings are merely used for the purpose of illustration and description, and are not intended to limit the scope of this specification. It should be further understood that the accompanying drawings are not drawn to scale.

[0019] The flowchart used in this specification illustrates operations implemented by a system in some embodiments of this specification. It should be clearly understood that operations in the flowchart can be implemented out of sequence. On the contrary, operations can be implemented reversely or simultaneously. In addition, one or more other operations can be added to the flowchart. One or more operations can be removed from the flowchart.

[0020] In this specification, "X includes at least one of A, B, or C" means that X includes at least A, or X includes at least B, or X includes at least C. In other words, X can include only any one of A, B, and C, or simultaneously include any combination of A, B, and C and other possible content/elements. Any combination of A, B, and C can be A, B, C, AB, AC, BC, or ABC.

[0021] In this specification, unless expressly stated, an association relationship generated between structures can be a direct association relationship or an indirect association relationship. For example, when "A is connected to B" is described, unless it is explicitly stated that A is directly connected to B, it should be understood as that A can be directly connected to B or can be indirectly connected to B. For another example, when "A is above B" is described, unless it is explicitly stated that A is directly above B (A is adjacent to B and A is above B), it should be understood as that A can be directly above B, or A can be indirectly above B (A is separated from B by another element and A is above B), and so on.

[0022] It should be noted that user data obtained in this specification is authorized by the user, and does not relate to user privacy.

[0023] For ease of description, the terms that appear below in this specification are first explained.

[0024] A large language model is usually used in the artificial intelligence field of natural language processing (NLP), and is especially a large machine learning model with a large quantity

of parameters and computing resources. The large language model is named for having a large quantity of parameters and a complex network structure. The large language model has a powerful feature representation capability and feature understanding capability, and can better capture a pattern and a rule in data when processing a complex task. The large language model is designed and trained to better understand and generate a natural language.

[0025] Optical character recognition (OCR) is a process of performing analysis and recognition processing on an image file of a text material to obtain text and layout information. In this process, various types of images such as handwritten and printed images are converted into a machine-coded text. A core of an OCR technology is to recognize text information in the image and convert the text information into a text format that can be edited and processed by a computer.

[0026] In a related technology, a logical relationship in a flowchart is usually extracted by manually establishing a rule. In such a method, an expert and an engineering person need to perform a large amount of communication and coding tasks, design rules for various situations and patterns, recognize and extract various information in a document based on the rules, and perform rewriting and normalization for database storage. In such a manner, a workload is heavy, costs are high, and maintenance is difficult. Once a flowchart is modified or updated, the rule needs to be manually re-established. However, in a machine-based learning manner, a large quantity of professional persons need to label data to perform model training. Consequently, costs are high, and maintenance is difficult.

[0027] A data processing method provided in this specification can be applied to a flowchart recognition scenario. After arrow data of an arrow and character data of characters in a to-be-processed flowchart are obtained, a connection relationship between the characters can be determined based on the arrow data and the character data, and first structured data is generated based on the connection relationship between the characters. The first structured data obtained in this way not only includes the character data included in the flowchart, but also includes the connection relationship between the characters, to implement automatic extraction of a logical relationship in the flowchart. In the above-mentioned process, the first structured data can be automatically extracted without manual participation. In this way, extraction efficiency is high. When the flowchart is updated, only an updated flowchart needs to be automatically extracted. In this way, maintenance is simple.

[0028] The data processing method and system provided in this specification have a wide

application scenario. The data processing method and system can be applied to various technical support or operation and maintenance scenarios. For example, the data processing method provided in this specification can be applied to an operation and maintenance scenario, for example, a scenario in which a running status of an internal or external system needs to be detected, an alarm scenario in which a health degree of an application scenario needs to be configured, or a scenario in which a large amount of troubleshooting and positioning need to be performed on a consulting work order. A troubleshooting scenario is used as an example for description. Different troubleshooting scenarios may correspond to different troubleshooting logic. The troubleshooting logic is usually used by a professional person to form a large quantity of knowledge base documents based on historical troubleshooting experience. The document describes troubleshooting logic corresponding to each scenario and standard steps in a troubleshooting process by using a natural language, a flowchart, etc. As a workload increases, a troubleshooting workload increases. To automate the troubleshooting process, knowledge extraction needs to be performed on the flowchart in the document, to extract a logical relationship in the flowchart and help a subsequent troubleshooting system performs automatic troubleshooting based on the logical relationship in the flowchart. The method and the system provided in this specification can automatically sort the flowchart in the knowledge base document, to obtain unified and standardized first structured data, so that a subsequent computer algorithm or program further performs data processing or application on the first structured data. For example, an operation and maintenance system can perform fault troubleshooting based on the first structured data, to implement automation of the troubleshooting process. A specific structural requirement for the first structured data is determined based on a standardization requirement for the data structure in a current application scenario. Different application scenarios may correspond to different preset standardized data structures. In the above-mentioned process of generating the first structured data, manual participation does not need to be performed, and the character data in the flowchart and the connection relationship between the characters can be automatically recognized and extracted, to generate the first structured data, and improve efficiency of performing structural processing on the flowchart.

[0029] In some other embodiments, the method provided in this specification can be further applied to a learning scenario of a large language model. For example, in a learning process of the large language model, some flowcharts can be input into the large language model as knowledge,

so that the large language model learns of a logical relationship between steps in the flowchart. To enable the large language model to better understand the input flowchart and improve a learning effect of the large language model, an engineer or an expert usually needs to sort a logical relationship/a connection relationship between steps in each flowchart, to generate a unified preset structured data format (for example, a preset code format or a preset tree format). Then, structured data corresponding to the sorted flowchart is input to the large language model, so that the large language model performs learning based on the structured data. In this way, a manner of generating the structured data corresponding to the flowchart depends on manpower because the flowchart is structured. Therefore, in the above-mentioned manner, not only manpower costs and time costs are greatly consumed, but also corresponding structured data needs to be manually regenerated for each flowchart. The manner has low applicability. According to the data processing method provided in this specification, after the to-be-processed flowchart is obtained, the arrow data and the character data in the to-be-processed flowchart can be directly and automatically extracted, and first structured data of a preset structure is generated based on the extracted data and a structural requirement for the preset structure. The first structured data corresponding to the to-be-processed flowchart is automatically extracted and generated. The first structured data can clearly and completely display procedure steps (character data) corresponding to the to-be-processed flowchart and an association relationship between the procedure steps. The data processing method provided in this specification is automated processing, manual participation does not need to be performed, and efficiency of performing structured sorting on the to-be-processed flowchart is high. Even if there are currently a plurality of to-be-processed flowcharts, first structured data corresponding to each to-be-processed flowchart can be quickly obtained, to input the first structured data to the large language model for learning, thereby improving learning efficiency of the large language model when the large language model learns of the flowchart.

[0030] It should be noted that the above-mentioned related application scenario is merely one of a plurality of use scenarios provided in this specification. A person skilled in the art should understand that, when the data processing method and system provided in this specification are applied to another use scenario, an implementation and a technical effect thereof are similar. The data processing method provided in this specification can be applied to any scenario in which structured data of a flowchart needs to be sort. A specific form of the structured data is not limited here, and can be flexibly adjusted based on a scenario.

[0031] FIG. 1 is a schematic diagram of an application scenario of data processing according to an embodiment of this specification.

[0032] As shown in FIG. 1, an application scenario 100 can include a data processing system 11. The data processing system 11 can obtain arrow data of an arrow and character data of characters in a to-be-processed flowchart, determine a connection relationship between the characters based on the arrow data and the character data, and generate first structured data of a preset structure based on the connection relationship between the characters. The preset structure can be a preset code structure, a preset hierarchical structure, a preset tree structure, etc. Setting of the specific preset structure can be flexibly adjusted according to a requirement of an application scenario, and the preset structure only needs to be a preset structure that can completely and clearly display a logical structure corresponding to the to-be-processed flowchart, and is not limited to that provided in the above-mentioned embodiments.

[0033] The to-be-processed flowchart can be a historical flowchart, can be obtained from a local database, can be downloaded from a webpage, can be obtained from a database at a position such as a cloud or a server, or can be drawn and uploaded in real time by a user. A specific source of the to-be-processed flowchart can be flexibly adjusted according to a user requirement, and is not limited here.

[0034] The data processing system 11 can be an electronic device that has a specific computing capability. The data processing system 11 can perform the data processing method provided in this specification. The data processing system 11 can store data or instructions for performing the data processing method described in this specification, and can execute or be configured to execute the data or the instructions. The data processing system 11 can include a hardware device with a data information processing function and a necessary program required for driving the hardware device to work.

[0035] The data processing system 11 can be a single computing device, or can be a cluster system including a plurality of computing devices. Data or instructions that are stored in the data processing system 11 and that are used to perform the data processing method described in this specification can be in any form of system architecture, for example, a layered architecture, an event-driven architecture, a microcore architecture, a microservice architecture, or a cloud architecture.

[0036] A system that provides real-time data processing is used as an example for description.

The above-mentioned system architecture can include a terminal and a server. The terminal can be deployed with a system client, and the server can be deployed with a system serving end. The terminal can include a mobile device, a tablet computer, a laptop computer, or similar content, or any combination thereof. In some embodiments, the mobile device can include a smart home device, a smart mobile device, a virtual reality device, an augmented reality device or a similar device, or any combination thereof. In some embodiments, the smart home apparatus can include a smart television, a desktop computer, or any combination. In some embodiments, the smart mobile device can include a smartphone, a personal digital assistant, a game device, a navigation device, or any combination thereof.

[0037] In some embodiments, the data processing system 11 can serve as a terminal in the above-mentioned system architecture. For example, the data processing system 11 can receive a to-be-processed flowchart input by the user; perform the data processing method provided in this specification, to obtain first structured data corresponding to the to-be-processed flowchart; and then perform a subsequent operation based on the first structured data, for example, output the first structured data to the user.

[0038] In some embodiments, the data processing system 11 can serve as a server in the above-mentioned system architecture. For example, the terminal receives the to-be-processed flowchart input by the user, and sends the to-be-processed flowchart to the data processing system 11. The data processing system 11 performs the data processing method provided in this specification to process the to-be-processed flowchart, to obtain the first structured data corresponding to the to-be-processed flowchart. Further, the data processing system 11 can feed back the first structured data to the terminal.

[0039] It should be understood that the application scenario shown in FIG. 1 is only one of a plurality of application scenarios provided in this specification. A quantity of data processing systems 11 shown in FIG. 1 is merely an example. Based on an implementation requirement, there can be any quantity.

[0040] FIG. 2 is a schematic diagram of a hardware structure of a computing device 200 according to some embodiments of this specification. The computing device 200 can be a computing device in a data processing system 11. As shown in FIG. 2, the computing device 200 includes at least one storage medium 230 and at least one processor 220. In some embodiments, the computing device 200 can further include an internal communication bus 210. In some

embodiments, the computing device 200 can further include a communication port 250. In some embodiments, the computing device 200 can further include an I/O component 260.

[0041] The internal communication bus 210 can be connected to different system components, including a storage medium 230 and a processor 220. The I/O component 260 supports an input/output between the computing device 200 and another component.

[0042] The storage medium 230 can include a data storage apparatus. The data storage apparatus may be a non-transitory storage medium, or may be a transitory storage medium. For example, the data storage apparatus can include one or more of a disk 232, a read-only storage medium (ROM) 234, or a random access storage medium (RAM) 236. The storage medium 230 further includes at least one instruction set stored in the data storage apparatus. The instruction set is computer program code. The computer program code can include a program, a routine, an object, a component, a data structure, a process, a module, etc. for performing the data processing method provided in this specification.

[0043] The communication port 250 is used for data communication between the computing device 200 and the outside. For example, the computing device 200 can be connected to a network through the communication port 250. The at least one processor 220 is communicatively connected to the at least one storage medium 230 through the internal communication bus 210. The at least one processor 220 is configured to execute the at least one instruction set. When the system 130 runs, the at least one processor 220 reads the at least one instruction set and performs, based on an indication of the at least one instruction set, the data processing method provided in this specification.

[0044] The processor 220 can perform all steps included in the data processing method. The processor 220 can be in a form of one or more processors. The processor 220 may send an execution instruction. The processor 220 can include one or more hardware processors, for example, a microcontroller, a microprocessor, a reduced instruction set computer (RISC), an application-specific integrated circuit (ASIC), an application-specific instruction set processor (ASIP), a central processing unit (CPU), a graphics processing unit (GPU), a physical processing unit (PPU), a microcontroller unit, a digital signal processor (DSP), a field programmable gate array (FPGA), an advanced RISC machine (ARM), a programmable logic device (PLD), any circuit or processor that can perform one or more functions, or any combination thereof.

[0045] For only the purpose of description, only one processor 220 is described in the

computing device 200 in this specification. However, it should be noted that the computing device 200 in this specification can further include a plurality of processors. Therefore, the operations and/or method steps disclosed in this specification can be performed by one processor or can be performed jointly by a plurality of processors, as described in this specification. For example, if the processor 220 in the computing device 200 in this specification performs step A and step B, it should be understood as that step A and step B can be jointly or separately performed by two different processors 220 (for example, a first processor performs step A, and a second processor performs step B, or a first processor and a second processor jointly perform step A and step B).

[0046] FIG. 3 is a flowchart of a data processing method according to an embodiment of this specification. The data processing method P300 is applied to a data processing system 11. As described above, a user can submit a to-be-processed flowchart by using a terminal. The data processing system 11 can receive a to-be-processed flowchart entered by the user, and perform the data processing method P300 provided in this specification. As described above, a computing device 200 can perform the data processing method P300 in this specification. Specifically, a processor 220 in the computing device 200 can read an instruction set stored in a local storage medium of the computing device 20s0, and then perform the data processing method P300 in this specification based on a stipulation of the instruction set. As shown in FIG. 3, the method P300 provided in this specification can include S310 to S350.

[0047] S310: Obtain arrow data of an arrow in the to-be-processed flowchart.

[0048] The to-be-processed flowchart can be sent by the user by using the above-mentioned terminal. Alternatively, the to-be-processed flowchart can be directly extracted by the terminal from a local database.

[0049] In some embodiments, the to-be-processed flowchart can be a troubleshooting flowchart in an operation and maintenance scenario. Alternatively, the to-be-processed flowchart can be a flowchart in an information technology development or software development scenario. Alternatively, the to-be-processed flowchart can be a flowchart in a project management scenario. Alternatively, the to-be-processed flowchart can be a flowchart in a manufacturing scenario such as a process flowchart. Alternatively, the to-be-processed flowchart can be a flowchart in an experimental scenario. Alternatively, the to-be-processed flowchart can be a flowchart in an engineering scenario, etc. A type that may be included in a specific flowchart can be determined based on a specific application scenario, and is not limited to that provided in the above-mentioned

embodiments.

[0050] In the flowchart, an arrow is usually used to indicate a flow direction between procedure steps and indicate a sequence of the steps. Therefore, obtaining the arrow data of the arrow in the to-be-processed flowchart is particularly important for subsequently determining a connection relationship between characters.

[0051] In this embodiment of this specification, the arrow data can include an arrow position and an arrow direction. The arrow position can be a position of the arrow in the to-be-processed flowchart. Specifically, the data processing system 11 can establish a reference coordinate system based on the to-be-processed flowchart. The arrow position can be a coordinate position of the arrow relative to the reference coordinate system. A base coordinate system is a coordinate system in the to-be-processed flowchart. The arrow direction is a direction in which the arrow points to the top of the arrow from the end (a wider part) of the arrow. The arrow direction can be a direction of the arrow relative to the to-be-processed flowchart. The data processing system 11 can use, as the arrow direction, the direction pointing to the top of the arrow from the end (the wider part) of the arrow.

[0052] A manner in which the data processing system 11 recognizes the arrow data can be: recognizing the to-be-processed flowchart based on a preset convolution kernel set, and recognizing the arrow in the to-be-processed flowchart based on each preset convolution kernel in the preset convolution kernel set, to determine the arrow data of the arrow in the to-be-processed flowchart.

[0053] In some embodiments, the preset convolution kernel set can include a plurality of different convolution kernels, values of some pixels in one of the plurality of different convolution kernels are a preset value, and a shape comprising the some pixels is an arrow shape. The preset value is different from a value of another pixel. The plurality of different convolution kernels have different sizes and/or some different pixels. Each convolution kernel corresponds to, for example, an arrow of a shape/size, and different convolution kernels are set to match arrows of different shapes/sizes in the to-be-processed flowchart in a recognition process, to improve recognition accuracy. In a process of recognizing the arrow in the to-be-processed flowchart, pixel-by-pixel traversal can be performed on all pixels in the to-be-processed flowchart based on the plurality of different convolution kernels in the convolution kernel set, and the arrow data included in the to-be-processed flowchart is determined based on a traversal convolution result.

[0054] In some possible embodiments, the preset convolution kernel set includes only convolution kernels of different sizes or includes arrows of different shapes, and no additional convolution kernel is designed for arrows at different angles. Therefore, in this embodiment of this specification, when an arrow angle of each arrow in the to-be-processed flowchart is recognized, the to-be-processed flowchart needs to be rotated to obtain a plurality of to-be-processed flowcharts at different angles, and an arrow angle that corresponds to an arrow and that is recognized at a current rotation angle is determined based on a rotation angle of the to-be-processed flowchart. For example, 45° can be used as a preset angle interval, and the to-be-processed flowchart can be separately rotated by 45°, 90°, 135°, 180°, 225°, 270°, 315°, and 360°. For example, 30° can be used as a preset angle interval, and the to-be-processed flowchart can be separately rotated by 30°, 60°, 90°, 120°, 150°, 180°, 210°, 240°, 270°, 300°, 330°, and 360°. In some embodiments, another angle can be used as an interval, to rotate the to-be-processed flowchart by, for example, 15° or 10°. The to-be-processed flowcharts at the plurality of different angles are recognized based on a preset convolution kernel set, to determine an arrow position of an arrow in a to-be-processed flowchart at each angle. Then, the arrow position and the arrow direction of the arrow in the to-be-processed flowchart is determined based on the arrow position of the arrow in the to-be-processed flowchart at each angle and a rotation angle of the to-be-processed flowchart at each angle. The arrow direction is related to the rotation angle of the to-be-processed flowchart at the corresponding angle. For example, the arrow direction of the arrow in the to-be-processed flowchart can be the rotation angle of the to-be-processed flowchart.

[0055] It should be understood that the above-mentioned embodiment is merely an example for description, and a specific to-be-processed flowchart can be rotated by more or fewer angles. The to-be-processed flowchart can be rotated by more angles, so that a recognition result of the recognized arrow angle is more precise and more accurate. The to-be-processed flowchart can be rotated by fewer angles. Because the to-be-processed flowchart can be rotated by fewer angles, in other words, a quantity of times that the to-be-processed flowcharts at different angles are matched with the preset convolution kernel for a smaller quantity of times, a speed of recognizing the arrow angle of the arrow in the to-be-processed flowchart can be faster, and recognition efficiency is high. Therefore, an angle by which the to-be-processed flowchart can be rotated can be preset according to an actual requirement of the user. For example, 15° is used as a preset angle interval, and the angle by which the to-be-processed flowchart can be rotated is split into 16 angles. In other

words, the to-be-processed flowchart can be rotated by 15°, 30°, 45°, ..., 330°, 345°, or 360°. The above-mentioned embodiments are merely an example for description, and any angle can be used as a preset angle interval. For example, the preset angle interval can also be set to 30°, 50°, or 60°. A specific setting situation of the angle by which the to-be-processed flowchart can be rotated can be flexibly adjusted according to a user requirement, and is not limited to that provided in the above-mentioned embodiments.

[0056] To improve recognition efficiency and reduce calculation costs, before recognizing the to-be-processed flowchart based on the preset convolution kernel set, a size of the to-be-processed flowchart further needs to be adjusted. Convolutional kernels of different sizes correspond to to-be-processed flowcharts of different preset sizes. An example in which the preset convolution kernel set includes a 9*9 convolution kernel is used for description. A preset size of a to-be-processed flowchart corresponding to the 9*9 convolution kernel is 361*361. If the size of the to-be-processed flowchart is inconsistent with the above-mentioned preset size, the to-be-processed flowchart needs to be scaled up/down, and scaled up/down to a preset size corresponding to a current convolution kernel. A manner of scaling up/down the to-be-processed flowchart can be: performing pixel filling on the to-be-processed flowchart in a difference method when an image size of the to-be-processed flowchart is less than the preset size, to adjust the to-be-processed flowchart to the preset size. When the image size of the to-be-processed flowchart is greater than the preset size, an average value can be obtained for two adjacent lines in the to-be-processed flowchart, to scale up/down the to-be-processed flowchart, until the image size of the to-be-processed flowchart is equal to the preset size. In other words, for each convolution kernel in the preset convolution kernel set, before the to-be-processed flowchart is recognized, size standardization needs to be performed on the to-be-processed flowchart, and the size of the to-be-processed flowchart is adjusted to a preset size corresponding to each convolution kernel. Subsequently, the to-be-processed flowchart after size standardization is recognized based on the convolutional kernel.

[0057] An example in which the arrow data of the arrow in the to-be-processed flowchart is recognized by using a convolution kernel A in the convolution kernel set is used for description. For example, the convolution kernel A may be in the following form:

0	0	0	0	1	0	0	0	0
0	0	0	3	1	3	0	0	0

0	0	1	3	3	3	1	0	0
0	1	3	3	3	3	3	1	0
1	3	3	3	3	3	3	3	1
0	0	0	1	3	1	0	0	0
0	0	0	1	3	1	0	0	0
0	0	0	1	3	1	0	0	0
0	0	0	1	3	1	0	0	0

[0058] The convolution kernel A is a 9*9 convolution kernel, and a weight distribution of the convolution kernel A is similar to an arrow shape. The weight distribution of the convolution kernel is designed, to represent arrows of different shapes/sizes by using the convolution kernel. In this way, in a process in which the convolution kernel slides in the to-be-processed flowchart, a feature of the arrow shape in the to-be-processed flowchart is more easily recognized. In a process in which the convolution kernel A performs pixel-by-pixel recognition on the to-be-processed flowchart through translation, because the convolution kernel A has a shape similar to an arrow, in a process in which the to-be-processed flowchart of the convolution kernel A is traversed, when a shape similar to that of the arrow in the to-be-processed flowchart is encountered, a convolutional result is significantly higher than a convolutional result in another case. A position with a significantly high convolutional result is determined as an arrow position with an arrow. To be specific, in this embodiment of this specification, arrow information that is in the to-be-processed flowchart and that corresponds to an arrow shape/size similar to the convolution kernel A can be accurately extracted based on a traversal result of the convolution kernel A.

[0059] In a process in which the convolution kernel A performs pixel-by-pixel traversal on the to-be-processed image, if an arrow is recognized in the to-be-processed flowchart, position coordinates of the convolution kernel in an image at this time are recorded, and the position coordinates are used as recognized arrow coordinates. In addition, a current rotation angle corresponding to the to-be-processed image is recorded when an arrow is recognized. Subsequently, an arrow angle corresponding to the recognized arrow can be determined based on the current rotation angle of the to-be-processed image. In the above-mentioned manner, recognition of arrow positions and arrow angles of different arrows in the to-be-processed flowchart can be implemented, to recognize arrows at various positions in various directions.

[0060] In some embodiments, the arrow data can further include data such as an arrow length, an arrow width, an arrow shape, an arrow connection line, etc.

[0061] S330: Obtain character data of characters in the to-be-processed flowchart.

[0062] In this embodiment of this specification, the character data of the characters in the to-be-processed flowchart can be obtained based on a character recognition algorithm. The character data includes character content and a character position. For example, a manner of obtaining the character data is: ruling out a non-character image element in the to-be-processed flowchart based on the character recognition algorithm, and extracting only all characters in the to-be-processed flowchart. Then, the extracted characters are filtered, and only a character including valid information is reserved. The character including the valid information includes but is not limited to a step description character, an operation guide character, or another procedure-related parameter character (for example, an input parameter character or an output parameter character) in the flowchart. In addition, character position coordinates of each character content can be further recognized based on the character recognition algorithm. For example, a central position of each character is used as a coordinate position corresponding to the character. Certainly, an upper left/right corner position of each character can also be used as a coordinate position of the character. A specific manner of determining a character coordinate position of each character can be flexibly adjusted according to a user requirement, and is not limited here.

[0063] S350: Generate first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data.

[0064] The first structured data includes the connection relationship between characters, and is a series of first structured data that is formed after all characters are serially connected based on the connection relationship between the characters and that have an upstream-downstream relationship. The connection relationship between the characters can be, for example, the upstream-downstream relationship, a child-node relationship or a parent-node relationship, a parameter input or a parameter output relationship, a sequential relationship, a cyclic relationship, or a parallel relationship. A manner of serially connecting the characters can be, for example, serially connecting the characters based on the connection relationship in a preset format.

[0065] For example, the preset format can be a code format. Specific format content is as follows: "Start node": {"node name": "Start troubleshooting", "Parent node": "Empty", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": ["End node"]}, "End node":

{"Node name": "End troubleshooting", "Parent node": "Start node", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": []}}. Content included in "XXXX" needs to be filled in based on a recognition result of the to-be-processed flowchart. It should be understood that the above-mentioned embodiments are merely an example for description. A specific data format of the first structured data can be adjusted according to an actual requirement of a current application scenario, and is not limited to that provided in the above-mentioned embodiments, and only a connection relationship between a plurality of characters needs to be represented.

[0066] In some embodiments, a manner of generating the first structured data corresponding to the to-be-processed flowchart can be: matching the arrow position and the arrow direction that are of the arrow with the character position of the character, to determine a connection relationship between the character content. For example, a first character is determined from the characters in the to-be-processed flowchart based on an arrow position of each arrow and a character position of each character in the to-be-processed flowchart. A first arrow in the arrow comprised in the to-be-processed flowchart points to a character position of the first character. Then, a second character corresponding to the first character is determined based on an arrow direction of the first arrow pointing to the first character. The second character is a character whose character position is on a reverse extension line in the arrow direction of the first arrow and that is closest to the character position of the first character. After the first character and the second character corresponding to the first character are determined, character content of the first character is used as downstream content of character content of the second character corresponding to the first character, and a connection relationship between the character content of the first character and the character content of the second character corresponding to the first character is established.

[0067] In this embodiment of this specification, when the first character is determined, for example, the character position of the character in the to-be-processed flowchart is matched with the arrow position of the arrow in the to-be-processed flowchart, and a character with an arrow in a preset range of the character position is used as the first character.

[0068] In some embodiments, the preset range is related to, for example, the image size of the to-be-processed flowchart, and image sizes of different to-be-processed flowcharts may correspond to different preset ranges. An example in which the preset range is a circular preset range is used for description. For example, the preset range is determined based on a search radius, and the search radius is related to the image size of the to-be-processed flowchart. A larger image

size of the to-be-processed flowchart indicates a larger search radius, that is, a larger preset range. On the contrary, a smaller image size of the to-be-processed image indicates a smaller search radius, that is, a smaller preset range.

[0069] In this embodiment of this specification, the search radius can be determined based on the pixel size of the to-be-processed image, and the search radius can be, for example, a product of a preset proportion a and the pixel size. In other words, the preset range is determined based on a center of a circle and the search radius by using each character position as a center/the center of a circle and using a*pixel size of the to-be-processed flowchart as the search radius. A preset range around each character position in the to-be-processed flowchart is searched, to determine whether an arrow exists within the preset range around each character position. A character with an arrow at a preset position around a character is determined as the first character.

[0070] In some other embodiments, the search range can be, for example, a rectangular range, an arc range, or any search range of a preset size. A specific manner of determining the search range and a range format of the search range can be flexibly adjusted according to a user requirement, and are not limited to those provided in the above-mentioned embodiments.

[0071] A manner of determining the reverse extension line in the arrow direction of the first arrow can be, for example, determining the arrow direction of the first arrow based on the rotation angle of the to-be-processed flowchart. For example, if the rotation angle of the to-be-processed flowchart is rotated by 315° in a clockwise manner, it indicates that the reverse extension line of the first arrow is a line that is inclined to a lower left direction by 45°, and the direction of the first arrow is inclined to an upper right direction by 45°. If the rotation angle of the to-be-processed flowchart is rotated by 135° in a clockwise manner, it indicates that the reverse extension line of the first arrow is a line that is inclined to an upper right direction by 45°, and the direction of the first arrow is inclined to a lower left direction by 45°. If the rotation angle of the to-be-processed flowchart is rotated by 45° in a clockwise manner, it indicates that the reverse extension line of the first arrow is a line that is inclined to a lower right direction by 45°, and the direction of the first arrow is inclined to an upper left direction by 45°. After the arrow position and the arrow direction of the first arrow are determined, an extension line is constructed in a reverse direction of the arrow direction starting from the arrow position of the first arrow. In the extension direction of the reverse extension line of the first arrow, it is determined that a character that intersects with the reverse extension line and that is closest to the arrow position of the first arrow is the second character. In

this case, because the arrow direction points to the first character from the second character, it indicates that when each procedure step in the to-be-processed flowchart is to be performed, a procedure step corresponding to the second character needs to be first performed, and then a procedure step corresponding to the first character needs to be performed. In other words, the connection relationship between the first character and the second character can be determined based on the arrow direction: The first character is a downstream character of the second character.

[0072] In some embodiments, the arrow data further includes an arrow length. A manner of generating the first structured data corresponding to the to-be-processed flowchart can be as follows: After an arrow position and an arrow direction of each arrow in the to-be-processed flowchart are recognized, a start position of a line segment can be further determined based on a line segment recognition method (for example, a longest straight line distance detected through Hough transform), by using the arrow position as an end position of the line segment, and based on an end position of the line segment, an arrow direction, and a line segment recognition result. A distance between the start position of the line segment and the end position of the line segment is the arrow length. Then, matching is performed based on the arrow position, the arrow direction, the arrow length, and the arrow position of the character, to determine the connection relationship between the character content.

[0073] For example, after the arrow position, the arrow direction, and the arrow length are determined, the arrow position is used as the end position of the arrow, and the start position of the line segment is used as the start position of the arrow. Then, the character position of the first character corresponding to the arrow is determined based on the position coordinates of the end position of the arrow. A distance between the character position of the first character and the end position of the arrow is shortest, and a difference between the character position of the first character and the end position of the arrow, for example, falls within the preset range. The character position of the second character corresponding to the arrow is determined based on the position coordinates of the start position of the arrow. A distance between the character position of the second character and the start position of the arrow is shortest, and a difference between the character position of the second character and the end position of the arrow, for example, falls within the preset range. After the first character and the second character that correspond to the arrow are determined, the association relationship between the first character and the second character is established: The character content of the first character is used as downstream content

of the character content of the second character corresponding to the first character, and the connection relationship between the character content of the first character and the character content of the second character corresponding to the first character is established. It should be understood that how to specifically determine the connection relationship between the character content based on the arrow data and the character data can be flexibly adjusted according to a user requirement, and is not limited to that provided in the above-mentioned embodiments.

[0074] According to the data processing method provided in this specification, the arrow data of the arrow and the character data of the character in the to-be-processed flowchart are obtained, and the connection relationship between the characters is determined based on the arrow data and the character data, to generate the first structured data based on the determined connection relationship. The first structured data obtained in this way not only includes the character data included in the flowchart, but also includes the connection relationship between the characters. The first structured data obtained in this specification can be used for calculation of a subsequent algorithm or program, so that the subsequent algorithm or program can perform further data processing based on the logical relationship in the flowchart.

[0075] In some embodiments, to ensure data integrity, in addition to performing structured processing on the to-be-processed flowchart, structured processing may also need to be performed on some to-be-processed documents related to the to-be-processed flowchart. The to-be-processed document may include document content and the to-be-processed flowchart. Alternatively, the to-be-processed document may include document content. In addition, the to-be-processed document and the to-be-processed flowchart further have a preset association relationship, and are complementary to each other. Therefore, in the data processing method provided in this specification, in addition to recognizing the first structured data in the to-be-processed flowchart, document content in the to-be-processed document can be recognized, to obtain second structured data corresponding to the document content. The first structured data is related to the second structured data.

[0076] A manner of obtaining the second structured data corresponding to the document content can be: invoking the first large language model to recognize the document content in the to-be-processed document, to generate the second structured data corresponding to the document content. The to-be-processed document is related to the to-be-processed flowchart, and the second structured data comprises a connection relationship between logical entities in the to-be-processed

document.

[0077] The logical entity can be any character element that has a logical significance in a specific context, for example, a decision point in a procedure, a state in an algorithm, or a node in a data stream. The logical entity does not necessarily directly correspond to a physical object, but is important for understanding and implementing logic of steps in the to-be-processed document. In addition, content included in the logical entity is not fixed, and the logical entity may change based on a specific application scenario, a system architecture, and a running environment. This means that in different application scenarios, for example, when the to-be-processed document belongs to a different field, a logical entity that needs to be concerned may be different.

[0078] For example, an example in which the second structured data is coding data is used for description. In a process of generating the second structured data, for example, a first prompt language can be used to assist the first large language model in generating the second structured data based on the document content of the to-be-processed document. The preset first prompt language may include one or more pieces of content of a target description, understanding requirement understanding, knowledge base use, an output format, an output template, an input example, and an answer requirement. For a data form that is of the second structured data and that is expected to be obtained, for example, in some embodiments, the preset first prompt language can be designed as follows:

"Please help us extract information such as a troubleshooting step in an input document. Understand problem logic and then make determining.\n"+

"\n"+

"You can use the following information as knowledge to supplement determining:\n"+

"{knowledgebase}\n"+

"\n"+

"You need to use the extracted information in the following JSON format as an answer. When the information includes a plurality of values, a sequence number is added after a JSON KEY value for distinguishing:\n"+

"{JSON template}\n"+

"The input information is as follows:\n"+

"{INPUT_DOCUMENT}\n"+

"Now give an answer. Do not give an explanation or reason:\n"+

"---";"

JSON template {"start node": {"Node name": "Start troubleshooting", "Parent node":

"Empty", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": ["End node"]},

"End node": {"Node name": "End troubleshooting", "Parent node": "Start node", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": []}}

[0079] The preset first prompt language provided in the above-mentioned embodiments includes content of seven parts:

[0080] Target description: "Help us extract information such as a troubleshooting step in an input document." The above-mentioned content indicates that a current task of the first major language model is to analyze a given to-be-processed document and to recognize and extract a step of troubleshooting or problem resolving.

[0081] Requirement understanding: "Understand problem logic and then make determining." The above-mentioned content describes that before the first large language model extracts the document content in the to-be-processed document, logic and a context in the document need to be understood first, to ensure that a step is accurately recognized.

[0082] Knowledge base use: "You can use the following information as knowledge to supplement determining: {knowledgebase}". The above-mentioned content indicates that the first large language model can refer to additional knowledge base information to assist in understanding and judgment. {knowledgebase} here is actual available knowledge base content corresponding to a current application scenario, and is used to define a structured model of a troubleshooting step, to instruct the first large language model to perform a troubleshooting step for the to-be-processed document. The troubleshooting step is a logical step chain with clear semantics and indication. The troubleshooting step includes a start node and an end node. In addition to these two nodes, each node has a parent node and a child node, and both the parent node and the child node are not unique. In addition, each node also needs to input a parameter and output a parameter. An output of the parent node needs to be a part of an input parameter of the child node.

[0083] Output format: "You need to use the extracted information in the following JSON format as an answer". The above-mentioned content is used to indicate that an output of the large language model needs to follow a specified JSON format. This helps sort the document content in the to-be-processed document into structured data for subsequent processing or analysis.

[0084] Input example: "The input information is as follows: {INPUT_DOCUMENT}". The above-mentioned content is to insert an actual document text of the to-be-processed document for

analysis by the first large language model.

[0085] Answer requirement: "Now give an answer. Do not give an explanation or reason." The above-mentioned content indicates that the first large language model only needs to provide an answer formatted according to a requirement, and does not need to explain why such an answer is made.

[0086] JSON template: The provided JSON template shows an expected answer format (namely, a data format of the second structured data) in a current scenario. Each node needs to include information such as a node name, a parent node, a parameter input, a parameter output, and a child node. If the information includes a plurality of values, a sequence number needs to be added after a JSON key to distinguish the plurality of values.

[0087] It should be understood that the above-mentioned embodiments are merely an example for description. Specifically, content included in the first prompt language can be flexibly adjusted based on an actual requirement of the current scenario, and is not limited to that provided in the above-mentioned embodiments. The preset first prompt language can include more or less content, and a language format of the prompt language can also be adjusted according to a user requirement. The content can include only a template corresponding to an output format and an output format, or include only a template corresponding to the target description and the output format, and is not limited to that provided in the above-mentioned embodiments.

[0088] In this embodiment of this specification, after the first structured data and the second structured data are generated, the first structured data and the second structured data can be directly separately stored; or the first structured data and the second structured data can be separately sent to a preset client.

[0089] In some other possible embodiments, after the first structured data and the second structured data are generated, to ensure comprehensiveness and simplicity of a data structure, the first structured data and the second structured data can be further integrated, to obtain and output integrated target structured data corresponding to the to-be-processed document and the to-be-processed flowchart. A manner of determining the target structured data can be: invoking a second large language model to integrate the second structured data and the first structured data, to generate the target structured data corresponding to the to-be-processed document and the to-be-processed flowchart. An integration requirement and a standardization requirement of the target structured data can be input to the second large language model in a manner 1 of a preset second

prompt language. For example, the preset second prompt language can be: "Integrate the following input content, remove repeated information, and generate data of a code structure from integrated data". It should be understood that the above-mentioned embodiments are merely an example for description. Specifically, content of the preset second prompt language is not limited to that provided in the above-mentioned embodiments, provided that a requirement of the user for the target structured data is included, and can be flexibly adjusted according to a user requirement.

[0090] For example, an example in which a preset structure is the code structure is still used for description. The preset second prompt language input to the second large language model can be, for example, as follows:

[0091] "You are an expert in the technical support field. Please help us fuse information such as troubleshooting steps in two JSON formats, and output a standardized complete step troubleshooting map in the JSON format. Understand problem logic and then make determining"+

"\n"+

"You can use the following information as knowledge to supplement determining:\n"+

"{knowledgebase}\n"+

"\n"+

"You need to use the extracted information in the following JSON format as an answer. When the information includes a plurality of values, a sequence number is added after a JSON KEY value for distinguishing:\n"+

"{JSON template}\n"+

"The input information is as follows:\n"+

"{INPUT_OCR}\n"+

"{INPUT_DOCUMENT}\n"+

"Now give an answer. Do not give an explanation or reason:\n"+

"---";"

JSON template {"start node": {"Node name": "Start troubleshooting", "Parent node": "Empty", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": ["End node"]},

"End node": {"Node name": "End troubleshooting", "Parent node": "Start node", "Parameter input": "XXXX", "Parameter output": "XXXX", "Child node": []}}.

[0092] The preset second prompt language provided in the above-mentioned embodiments includes content of eight parts:

[0093] Role definition: "You are an expert in the technical support field". The above-

mentioned content is used to clarify a professional background of the second large language model, and suggests that the second large language model needs to have knowledge and experience in a related field when resolving the above-mentioned problems.

[0094] Task description: "Please help us fuse information such as troubleshooting steps in two JSON formats, and output a standardized complete step troubleshooting map in the JSON format." The above-mentioned content describes that a current task of the second large language model is to combine the troubleshooting step information in the two sources (the first structured data and the second structured data), eliminate redundancy, and create a comprehensive and standardized troubleshooting step map.

[0095] Understanding requirements: "Understand problem logic and then make determining". The above-mentioned content is used to describe that nature and logic of a problem need to be deeply understood before the second large language model starts to execute a task, to ensure that a processing method is correct.

[0096] Knowledge base use: "You can use the following information as knowledge to supplement determining: {knowledgebase}". The above content provides additional knowledge base information, to help an answer better understand a context and a rule of the troubleshooting step.

[0097] Output format: "You need to use the extracted information in the following JSON format as an answer ...". The above-mentioned content specifies a specific format (namely, a specific format of the target structured data) of an output answer of the second large language model, including fields such as a node name, a parent node, a parameter input, a parameter output, and a child node, and an output needs to be of a JSON structure.

[0098] Input information: "The input information is as follows: {INPUT_OCR}\n{INPUT_DOCUMENT}". The above-mentioned content indicates two input files: first structured data obtained through optical character recognition (OCR) and second structured data directly read from a document, the two inputs are presented in the JSON format.

[0099] Knowledge base guide: The part "Knowledgebase" provides a detailed explanation of the input information, and emphasizes that the two inputs (the first structured data and the second structured data) may include repeated information or may be complementary to each other. Integration and de-duplication need to be performed, and unique information is combined. The knowledge base guide task is used to describe a data integration task. Designed data integration

logic is as follows: INPUT_OCR (first structured data) and INPUT_DOCUMENT (second structured data) each are a troubleshooting logic chain in the JSON format, and include an upstream-downstream relationship between respective troubleshooting steps and parameter output/input information. Information included in INPUT_OCR (first structured data) and INPUT_DOCUMENT (second structured data) may overlap each other or may be complementary to each other. The repeated information needs to be removed, and only one of INPUT_OCR (first structured data) and INPUT_DOCUMENT (second structured data) needs to be retained. In addition, unique information in INPUT_OCR (first structured data) and INPUT_DOCUMENT (second structured data) is combined into new target structured data to form a complete chain.

[0100] Answer requirement: "Now give an answer. Do not give an explanation or reason". The above-mentioned content requires the second large language model to directly provide an answer when answering a question, that is, directly answer the integrated target structured data based on the first structured data and the second structured data, without any additional explanation or reason.

[0101] A manner of integrating the second structured data and the first structured data can be as follows: The first structured data and the second structured data include an upstream-downstream relationship between the character content/the logical entities in respective troubleshooting steps. Information included in the first structured data and the second structured data may overlap each other, or may be different from each other (complementary to each other). To ensure data integrity and simplicity, repeated information in the first structured data and the second structured data can be removed as redundant information, and only one of the first structured data and the second structured data is reserved. In addition, unique information in the first structured data and the second structured data is combined and sorted, and it is determined that the structured data obtained after de-duplication and sorting is the target structured data. The target structured data includes a complete procedure chain. Complete steps of the entire procedure chain can be clearly and completely viewed based on the target structured data.

[0102] In this embodiment of this specification, the first large language model and the second large language model can be the same large language model, or can be two large language models that are independent of each other. A specific setting manner of the first large language model and the second large language model can be flexibly adjusted according to a user requirement, and is not limited to that provided in the above-mentioned embodiments.

[0103] FIG. 4 is a flowchart of a data processing method P300 according to another embodiment of this specification. As shown in FIG. 4, a complete procedure of the method provided in this specification can be the following procedure:

[0104] After to-be-processed data (a to-be-processed document and a to-be-processed procedure related to the to-be-processed document) is obtained, steps S310 to 350 are performed for the to-be-processed flowchart, to extract character information and arrow information in the to-be-processed flowchart, and generate, based on the character information and the arrow information, first structured data corresponding to the to-be-processed flowchart. In addition, for the to-be-processed document, document content of the to-be-processed document can be recognized based on a first major language model, to extract a logical entity in the to-be-processed document and a connection relationship between logical entities. Second structured data corresponding to the to-be-processed document is generated based on the connection relationship between the extracted logical entities corresponding to the to-be-processed document. The first structured data and the second structured data are integrated based on a second major language model, to generate the target structured data corresponding to the to-be-processed document and the to-be-processed flowchart. The target structured data includes complete character content in the to-be-processed document and the to-be-processed flowchart, and a connection relationship between character content (the logical entities). Finally, the generated target structured data is output for subsequent viewing or use.

[0105] In an embodiment of this specification, the target structured data can be, for example, target coding data. In the conventional technology, an expert and an engineering person usually need to perform coding work based on a view result summary by viewing a related document or a related flowchart of a large quantity of application scenarios, to generate coding rules (target coding data) corresponding to various cases or patterns of the application scenarios. In the above-mentioned coding rule generation manner, for the related document or the related flowchart in each application scenario, the expert or the engineering person needs to perform summarization and coding work. In this way, efficiency is low and large labor costs are consumed. Each new application scenario or each new related document/flowchart needs to be manually sorted again, and a corresponding coding rule needs to be generated.

[0106] For example, when a running status or a health degree of a current application scenario is troubleshot, a large amount of log data or work orders need to be detected by depending on a

coding program. If the method provided in the conventional technology is used, a troubleshooting rule of the current application scenario needs to be manually viewed, and corresponding troubleshooting coding data is written based on the troubleshooting rule. However, according to the method provided in this specification, the coding work does not need to be performed manually, and character data or a logical entity in the document or the flowchart is automatically extracted and sorted based on the related document or the related flowchart corresponding to the troubleshooting rule of the current scenario, to generate the target structured data and form normalized coding data. In other words, according to the data processing method provided in this specification, structured coding data can be directly generated based on the to-be-processed flowchart and the to-be-processed document, and the generated coding data can be directly used for subsequent intelligent work order troubleshooting; or in a log data troubleshooting scenario, efficiency and accuracy of generating the structured coding data are high, and manual participation does not need to be performed.

[0107] In some other embodiments, the target structured data can be data of a document type. According to the data processing method provided in this specification, structural sorting is performed on the to-be-processed flowchart and the to-be-processed document, to remove redundant data and complement and perfect missing information to generate clear and complete target structured data. The target structured data can be presented in a form of a document for subsequent viewing or learning by a person. It should be understood that the above-mentioned embodiments are merely an example for description. A specific structural form of the target structured data can be determined based on prompt information. For example, the prompt information can be input to the second large language model together with the first structured data and the second structured data, to prompt the second large language model to generate corresponding target structured data based on a data form of structured data specified in the prompt information.

[0108] In conclusion, in the data processing method P300 and the system provided in this specification, the arrow data and the character data in the to-be-processed flowchart can be recognized, and a connection relationship between character content is determined based on the arrow data and the character data. Then, the first structured data is generated based on the connection relationship between character content. In addition, to ensure data integrity and smoothness, in the method provided in this specification, the to-be-processed document related to

the to-be-processed flowchart is further obtained, a logical entity relationship in the to-be-processed document is recognized, and the second structured data corresponding to the to-be-processed document is determined based on the recognized logical entity relationship. Then, the first structured data and the second structured data are integrated to remove a repeated part in the first structured data and the second structured data, and the integrated data is sorted to generate the target structured data corresponding to the to-be-processed document and the to-be-processed flowchart. The target structured data includes a complete and smooth to-be-processed document and all procedure steps in the related to-be-processed flowchart. Complete logical steps corresponding to the current application scenario can be clearly and completely reflected based on the target structured data.

[0109] Another aspect of this specification provides a computer-readable non-transitory storage medium, storing at least one instruction set, including executable instructions for performing a data processing method. When the at least one instruction set is executed by a processor, the at least one instruction set instructs the processor to implement the steps of the data processing method P300 in this specification. In some possible implementations, aspects of this specification can further be implemented in a form of a program product, including program code. When the program product runs in a system, the program code is used to enable the system to perform the steps of the method P300 described in this specification. The program product configured to implement the foregoing method can be a portable compact disc read-only memory (CD-ROM) that includes program code, and can run in the system. However, the program product in this specification is not limited thereto. In this specification, a readable storage medium can be any tangible medium that includes or stores a program, and the program can be used by or in combination with an instruction execution system. The program product can be any combination of one or more readable media. The readable medium can be a readable signal medium or a readable storage medium. The readable storage medium can be, for example, an electrical system, apparatus, or device, a magnetic system, apparatus, or device, an optical system, apparatus, or device, an electromagnetic system, apparatus, or device, an infrared system, apparatus, or device, or a semiconductor system, apparatus, or device, or any combination thereof. More specific examples of the readable storage medium include: an electrical connection including one or more conducting wires, a portable disk, 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 component, a magnetic storage component, or any proper combination thereof. The computer-readable storage medium can include a data signal propagated in a baseband or as part of a carrier, and carries readable program code. The propagated data signal can be in a plurality of forms, and includes but is not limited to an electromagnetic signal, an optical signal, or any proper combination thereof. The readable storage medium can further be any readable medium other than the readable storage medium, and the readable medium can be used to send, propagate, or transmit a program used by or in combination with an instruction execution system, apparatus, or component. The program code included in the readable storage medium can be transmitted in any proper medium, including but not limited to a wireless medium, a wired medium, optical cable, an RF medium, or any proper combination thereof. Program code for performing an operation of this specification can be written in any combination of one or more program design languages. The program design language includes an object-oriented program design language, for example, Java or C++, and further includes a conventional procedural program design language, for example, "C" language, or a similar program design language.

[0110] In conclusion, according to the data processing method P300 and the system provided in this specification, a to-be-processed flowchart can be recognized, and a connection relationship between characters in the to-be-processed flowchart is sorted based on recognized arrow data and recognized character data, to obtain first structured data corresponding to the to-be-processed flowchart, thereby automatically extracting a logical relationship in the to-be-processed flowchart. Therefore, a subsequent algorithm or program can perform further data processing based on the obtained first structured data, to use a logical relationship in the to-be-processed flowchart. In addition, the method P300 provided in this specification and the system can automatically recognize and sort the to-be-processed flowchart and the related to-be-processed document, to generate complete and smooth target structured data corresponding to the to-be-processed document and the related to-be-processed flowchart. According to the method P300 and the system provided in this specification, structured data is automatically generated, data processing efficiency is improved, and applicability is strong. The method P300 and the system can be extended to different application scenarios, and corresponding procedure rules in different application scenarios are automatically sorted and integrated. In addition, maintenance of the method P300 and the system provided in this specification is simple. When the flowchart is

modified and updated, a logical relationship can be extracted again from a new flowchart.

[0111] Specific embodiments of this specification are described above. Other embodiments fall within the scope of the appended claims. In some cases, the actions or steps described in the claims can be performed in an order different from that in the embodiments, and the desired results can still be achieved. In addition, the process depicted in the accompanying drawings does not necessarily need a particular order or consecutive order to achieve the desired results. In some implementations, multitasking and parallel processing are possible or may be advantageous.

[0112] In conclusion, after reading the detailed disclosure, a person skilled in the art can understand that the detailed disclosure can be presented only in an example manner, and can impose no limitation. Although it is not explicitly stated here, a person skilled in the art can understand that the requirements of this specification include various proper changes, improvements, and modifications to the embodiments. These changes, improvements, and modifications are intended to be presented in this specification, and fall within the spirit and scope of the example embodiments of this specification.

[0113] In addition, some terms in this specification are used to describe the embodiments of this specification. For example, "one embodiment", "an embodiment", and/or "some embodiments" mean that particular features, structures, or characteristics described with reference to the embodiments can be included in at least one embodiment of this specification. Therefore, it can be emphasized and understood that two or more of "an embodiment", "one embodiment", or "alternative embodiments" referred in parts of this specification do not necessarily indicate the same embodiment. In addition, specific features, structures, or features can be properly combined in one or more embodiments of this specification.

[0114] It should be understood that in the above-mentioned descriptions of the embodiments of this specification, to help understand a feature, for the purpose of simplifying this specification, various features are combined in a single embodiment, accompanying drawing, or description thereof in this specification. However, this does not mean that the combination of these features is necessary. It is entirely possible for a person skilled in the art to mark some devices and understand the devices as separate embodiments when reading this specification. That is, the embodiments in this specification can also be understood as an integration of a plurality of secondary embodiments. Content of each secondary embodiment is also true when features are less than all features in one of the above-mentioned disclosed embodiments.

[0115] Each patent, patent application, publication of a patent application, and other materials, for example, articles, books, instructions, publications, documents, and products, that are cited in this specification and that are other than any document that is inconsistent or conflict with this document or that imposes limiting impact on the widest scope of the claims can be incorporated here by reference and used for all purposes now or later associated with this document. In addition, the term in this document shall prevail when there is any inconsistency or conflict between the description, definition and/or use of a related term in any material and the description, definition and/or use of a related term in this document.

[0116] Finally, it should be understood that the principles of the implementation solutions of this specification are described in the implementation solutions of the application disclosed in this specification. Other modified embodiments also fall within the scope of this specification. Therefore, the embodiments disclosed in this specification are merely examples and not limitations. A person skilled in the art can use an alternative configuration according to the embodiments of this specification to implement the application in this specification. Therefore, the embodiments of this specification are not limited to the embodiments accurately described in the application.

CLAIMS

What is claimed is:

1. A data processing method, wherein the method comprises:

obtaining arrow data of an arrow in a to-be-processed flowchart;

obtaining character data of characters in the to-be-processed flowchart; and

generating first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data, wherein the first structured data comprises a connection relationship between the characters.

2. The data processing method according to claim 1, wherein the arrow data comprises an arrow position and an arrow direction, and the character data comprises character content and a character position; and

the generating first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data comprises:

matching the arrow position and the arrow direction that are of the arrow with the character position of the character, to determine a connection relationship between the character content.

3. The data processing method according to claim 2, wherein the obtaining arrow data of an arrow in a to-be-processed flowchart comprises:

rotating the to-be-processed flowchart, to obtain to-be-processed flowcharts at a plurality of different angles;

recognizing the to-be-processed flowcharts at the plurality of different angles based on a preset convolution kernel set, to determine an arrow position of an arrow in a to-be-processed flowchart at each angle; and

determining the arrow position and the arrow direction of the arrow in the to-be-processed flowchart based on the arrow position of the arrow in the to-be-processed flowchart at each angle and a rotation angle of the to-be-processed flowchart at each angle, wherein the arrow direction is related to the rotation angle of the to-be-processed flowchart at the corresponding angle.

4. The data processing method according to claim 3, wherein the preset convolution kernel set comprises a plurality of different convolution kernels, values of some pixels in one of the plurality of different convolution kernels are a preset value, and a shape comprising the some pixels is an arrow shape.

5. The data processing method according to claim 4, wherein the plurality of different convolution kernels have different sizes and/or some different pixels.

6. The data processing method according to claim 2, wherein the matching the arrow position and the arrow direction that are of the arrow with the character position of the character, to determine a connection relationship between the character content comprises:

determining a first character from the characters in the to-be-processed flowchart based on the arrow position of the arrow and the character position of the character, wherein a first arrow in the arrow comprised in the to-be-processed flowchart points to a character position of the first character;

determining a corresponding second character based on an arrow direction of the first arrow, wherein the second character is a character whose character position is on a reverse extension line in the arrow direction of the first arrow and that is closest to the first character; and

using character content of the first character as downstream content of character content of the second character corresponding to the first character, and establishing a connection relationship between the character content of the first character and the character content of the second character corresponding to the first character.

7. The data processing method according to claim 6, wherein the determining a first character from the characters in the to-be-processed flowchart comprises:

matching the character position of the character in the to-be-processed flowchart with the arrow position of the arrow in the to-be-processed flowchart, and using a character with an arrow in a preset range of the character position as the first character.

8. The data processing method according to claim 1, further comprising:

invoking a first large language model to recognize document content in a to-be-processed document, to generate second structured data corresponding to the document content, wherein the to-be-processed document is related to the to-be-processed flowchart, and the second structured data comprises a connection relationship between logical entities in the to-be-processed document.

9. The data processing method according to claim 8, further comprising:

invoking a second large language model to integrate the second structured data and the first structured data, to generate target structured data corresponding to the to-be-processed document and the to-be-processed flowchart.

10. The data processing method according to claim 1, wherein the obtaining character data of

characters in the to-be-processed flowchart comprises:

obtaining the character data of the characters in the to-be-processed flowchart based on a character recognition algorithm.

11. A data processing system, comprising:

at least one storage medium, storing at least one instruction set, to perform a data processing method; and

at least one processor, communicatively connected to the at least one storage medium, wherein when running, the at least one processor reads the at least one instruction set, and performs the following operations based on an indication of the at least one instruction set:

obtaining arrow data of an arrow in a to-be-processed flowchart;

obtaining character data of characters in the to-be-processed flowchart; and

generating first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data, wherein the first structured data comprises a connection relationship between the characters.

12. The data processing system according to claim 11, wherein the arrow data comprises an arrow position and an arrow direction, and the character data comprises character content and a character position; and

to generate the first structured data corresponding to the to-be-processed flowchart based on the arrow data and the character data, the at least one processor performs the following operations:

matching the arrow position and the arrow direction that are of the arrow with the character position of the character, to determine a connection relationship between the character content.

13. The data processing system according to claim 12, wherein to obtain the arrow data of the arrow in the to-be-processed flowchart, the at least one processor performs the following operations:

rotating the to-be-processed flowchart, to obtain to-be-processed flowcharts at a plurality of different angles;

recognizing the to-be-processed flowcharts at the plurality of different angles based on a preset convolution kernel set, to determine an arrow position of an arrow in a to-be-processed flowchart at each angle; and

determining the arrow position and the arrow direction of the arrow in the to-be-processed flowchart based on the arrow position of the arrow in the to-be-processed flowchart at each angle

and a rotation angle of the to-be-processed flowchart at each angle, wherein the arrow direction is related to the rotation angle of the to-be-processed flowchart at the corresponding angle.

14. The data processing system according to claim 13, wherein the preset convolution kernel set comprises a plurality of different convolution kernels, values of some pixels in one of the plurality of different convolution kernels are a preset value, and a shape comprising the some pixels is an arrow shape.

15. The data processing system according to claim 14, wherein the plurality of different convolution kernels have different sizes and/or some different pixels.

16. The data processing system according to claim 12, wherein to match the arrow position and the arrow direction that are of the arrow with the character position of the character, to determine the connection relationship between the character content, the at least one processor performs the following operations:

determining a first character from the characters in the to-be-processed flowchart based on the arrow position of the arrow and the character position of the character, wherein a first arrow in the arrow comprised in the to-be-processed flowchart points to a character position of the first character;

determining a corresponding second character based on an arrow direction of the first arrow, wherein the second character is a character whose character position is on a reverse extension line in the arrow direction of the first arrow and that is closest to the first character; and

using character content of the first character as downstream content of character content of the second character corresponding to the first character, and establishing a connection relationship between the character content of the first character and the character content of the second character corresponding to the first character.

17. The data processing method according to claim 16, wherein to determine the first character from the characters in the to-be-processed flowchart, the at least one processor performs the following operations:

matching the character position of the character in the to-be-processed flowchart with the arrow position of the arrow in the to-be-processed flowchart, and using a character with an arrow in a preset range of the character position as the first character.

18. The data processing system according to claim 11, wherein the at least one processor performs the following operation:

invoking a first large language model to recognize document content in a to-be-processed document, to generate second structured data corresponding to the document content, wherein the to-be-processed document is related to the to-be-processed flowchart, and the second structured data comprises a connection relationship between logical entities in the to-be-processed document.

19. The data processing system according to claim 18, wherein the at least one processor performs the following operation:

invoking a second large language model to integrate the second structured data and the first structured data, to generate target structured data corresponding to the to-be-processed document and the to-be-processed flowchart.

20. A computer-readable nonvolatile storage medium, wherein the computer-readable nonvolatile storage medium stores at least one instruction set, and when the at least one instruction set is executed by at least one processor, the data processing method according to any one of claims 1 to 10 is implemented.