CATEGORY PREDICTION METHOD AND APPARATUS, STORAGE MEDIUM, AND COMPUTER SYSTEM

TECHNICAL FIELD

[0001] One or more embodiments of this specification relate to the artificial intelligence field, and in particular, to a category prediction method and apparatus, a storage medium, and a computer system.

BACKGROUND

[0002] In Internet services, a search function plays an important role in aspects such as information obtaining, a research search, and problem solving. For example, on e-commerce websites, a search function is an important portal for users to find and purchase products or services. The user enters, in a search bar, content that the user wants to search. A system returns a project list matching search content in a database of the system.

[0003] When a quantity of projects in a database of a system is relatively small, for example, in an early stage of developing a new market, when the user searches for some content on a website, a corresponding result may be unable to be found in the database of the system in an existing search method, and then an empty list is returned, resulting in an empty search. The empty search reduces use willingness of the user, and further affects a user activity degree and a user retention rate of the website. Therefore, a better method is needed to match search content of a user with a corresponding project in a database.

SUMMARY

[0004] Embodiments of this specification provide a category prediction method and apparatus, a storage medium, and a computer system. Category prediction is performed on an input search term, and a corresponding project is found based on an obtained category prediction result and is used as a search result, to reduce an empty search rate.

[0005] This specification provides a category prediction method. The method includes:

obtaining an empty search term set, where the empty search term is a search term whose corresponding search result is empty; inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term; constructing a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and training a category prediction model based on the category database, where the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

[0006] This specification provides a computer system, including a storage, a processor, and a computer program that is stored in the storage and that is capable of running on the processor. When the processor executes the program, the following steps are implemented: obtaining an empty search term set, where the empty search term is a search term whose corresponding search result is empty; inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term; constructing a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and training a category prediction model based on the category database, where the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

[0007] This specification provides a category prediction apparatus. The apparatus includes: an obtaining module, configured to obtain an empty search term set, where the empty search term is a search term whose corresponding search result is empty; a prediction module, configured to input each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term; a construction module, configured to construct a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and a training module, configured to train a category prediction model based on the category database, where the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

[0008] This specification provides a computer-readable non-volatile storage medium. The storage medium stores a computer program, and when the computer program is executed by a processor, the category prediction method is implemented.

[0009] At least one of the following beneficial effects can be achieved by using the above-mentioned at least one technical solution in the embodiments of this specification: Category prediction is performed by using the large language model, to expand the category database, thereby training the category prediction model. According to an offline and real-time solution in which the category database and the category prediction model are combined, both efficiency and accuracy of category prediction can be considered. Further, an empty search rate in a search process is reduced, and a user activity degree and a user retention rate are improved. In addition, manual labeling is not needed in a training process, to reduce model training costs.

BRIEF DESCRIPTION OF DRAWINGS

[0010] The accompanying drawings described herein are used to provide a further understanding of this specification and constitute a part of this specification. The example embodiments of this specification and descriptions thereof are used to explain this specification, and do not constitute an improper limitation on this specification. In the accompanying drawings:

[0011] FIG. 1 is a schematic diagram of an implementation scenario of a category prediction method in this specification;

[0012] FIG. 2 is a schematic flowchart of a category prediction method in this specification;

[0013] FIG. 3 is a schematic diagram of a category prediction apparatus in this specification; and

[0014] FIG. 4 is a schematic structural diagram of a computer system according to an embodiment of this specification.

DESCRIPTION OF EMBODIMENTS

[0015] To make the objectives, technical solutions, and advantages of this specification clearer, the following clearly and comprehensively describes the technical solutions of this specification with reference to specific embodiments and corresponding accompanying drawings of this specification. Clearly, the described embodiments are merely some but not all of embodiments of this specification. All other embodiments obtained by a person of ordinary skill in the art based on the embodiment of this specification without creative efforts shall fall within the protection scope of this specification.

[0016] As described above, an empty search problem that an empty result list is returned when a user searches affects a user activity degree and a user retention rate of a website. For example, the website can be a takeaway website. The user enters a "search term A", the search term A is, for example, a name of a restaurant, but cooperators of the takeaway website do not have the restaurant. If an empty list is returned directly in this case, an empty search occurs. When the user finds that a desired result cannot be found, the user may reduce willingness to continue to use the website. In this case, if the "search term A" is analyzed, it can be found that the "search term A" is a restaurant, further, for example, a restaurant that mainly manages fried chicken and hamburger, and a search engine of the takeaway website can search the cooperators of the website for a restaurant that sells fried chicken and hamburger, and return the restaurant to the user as a current search result. In this way, an intention of the user is analyzed based on a search term, and related content is returned, to not only reduce an empty search rate, but also enrich content in a search result list. A process of predicting, based on the search term entered by the user, a project category or a category that the user may want to find is referred to as category prediction.

[0017] It can be learned from the above-mentioned analysis that category prediction is a very important function in a search system. Category prediction can improve quality and correlation of search results, and help the user to quickly locate a project range of interest, to improve search experience of the user.

[0018] In view of this, an embodiment of this specification provides a category prediction method. FIG. 1 is a schematic diagram of an implementation scenario of a category prediction method in this specification. As shown in FIG. 1, category prediction includes a total of three parts, as shown in (a), (b), and (c). (a) in FIG. 1 shows a process of constructing a category database and training a category prediction model. (b) in FIG. 1 shows a process of performing category prediction by using the category database and the category prediction model. (c) in FIG. 1 shows a process of matching a search result corresponding to a search term based on a category prediction result.

[0019] As shown in (a) in FIG. 1, an empty search term set is first obtained. The empty search term set includes a plurality of empty search terms, and the empty search term is a search term whose search result is empty. Then, each empty search term is input into a large language model to perform category prediction, to obtain a category prediction result corresponding to each empty search term. Then, the category database is constructed based on each empty search term and the

category prediction result corresponding to the empty search term. Content stored in the category database can be, for example, a pair including empty search term-category prediction result. Then, the empty search term in the category database is used as a sample, and the category prediction result is used as a label to train the category prediction model.

[0020] After the category prediction model is obtained through training, as shown in (b) in FIG. 1, a search term entered by a user is obtained, and the search term is searched for in the category database. When the search succeeds, a search result is determined as a target category corresponding to the search term; or when the search fails, the search term is input into the category prediction model to perform category prediction, and a prediction result of the category prediction model is determined as a target category corresponding to the search term.

[0021] After the target category corresponding to the search term is obtained, as shown in (c) in FIG. 1, the target category is input into a query parsing service engine, so that the query parsing service engine searches a preset project database for several projects with top-ranked degrees of matching with the target category, uses the several projects as a search result corresponding to the search term, and returns the search result to the user.

[0022] With reference to specific embodiments, the following describes specific implementation steps of the above-mentioned category prediction.

[0023] FIG. 2 is a schematic flowchart of a category prediction method in this specification. The method can be performed by any platform, server, or device cluster that has a computing and processing capability. As shown in FIG. 2, the method includes at least the following steps: Step 202: Obtain an empty search term set, where the empty search term is a search term whose corresponding search result is empty. Step 204: Input each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term. Step 206: Construct a category database based on each empty search term and the first category prediction result corresponding to the empty search term. Step 208: Train a category prediction model based on the category database, where the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

[0024] The following describes a specific execution process of the above-mentioned steps.

[0025] First, in step 202, the empty search term set is obtained. The empty search term is a search term whose corresponding search result is empty.

[0026] The empty search term is a search term for which an empty search result list is returned when a search function is invoked on a platform that performs the category prediction method described in this specification to search for the search term.

[0027] Then, in step 204, each empty search term is input into the large language model to perform category prediction, to obtain the first category prediction result of each empty search term.

[0028] Any publicly or privately trained large language model (LLM) can be used. For example, ChatGPT, Gemini, Claude, etc. can be used. This is not limited herein.

[0029] A corresponding prompt is assembled, and the prompt is input into the large language model, to use the large language model to perform category prediction.

[0030] In an embodiment, step 204 specifically includes: inputting a target empty search term in all the empty search terms into a first prompt template, to obtain a target prompt, where the target prompt further includes a plurality of preset categories, to indicate the large language model to perform category prediction on the target empty search term based on the plurality of preset categories; and inputting the target prompt into the large language model, to obtain a first category prediction result corresponding to the target empty search term.

[0031] In a more specific embodiment, the target prompt further includes description information of each preset category and a hierarchical relationship between all the preset categories.

[0032] The first prompt template can include a statement used to indicate the large language model to perform category prediction, a first placeholder, and a second placeholder. The first placeholder needs to be filled with a hierarchical relationship between a plurality of preset categories and each category, and optionally, may be further filled with description information of each category. The second placeholder needs to be filled with an empty search term. Optionally, the first prompt template can further include a third placeholder, and the third placeholder needs to be filled with a country or a region in which the platform that performs the category prediction method described in this specification is located.

[0033] For example, the first prompt template can be as follows:

"you are expert in the ecommerce, and experienced in the search, now I have several queries, that the customers didn't get any result, so I want to judge what industry categories they intend to search. I will give you the definition of industry categories, the definition is below: {placeholder1}. And please help to judge the intention of category from these queries below: {placeholder2}. Notice:

1. the region is Indonesia, so you can limit the analysis in Indonesia.

2. you can just base on the data of your training.

3. you need to try your best to decide more categories.

4. you should format the results of all queries like such structure(MAP JSON), the placeholder need to replace by the id: {"query":"firstCategoryId&secondCategoryId, firstCategoryId&secondCategoryId "}".

{placeholder1} can be the first placeholder, and needs to be filled with the hierarchical relationship between a plurality of preset categories and each category. For example, {placeholder1} can be filled with the following content:

"[{firstCategory":{"id":"1","name":"O2O","secondCatgories":[{"id":"2","name":"FoodDelivery"},{"id":"3","name":"OnlineSupermarket"},{"id":"4","name":"RideHailing"},{"id":"5","name":"Others"}]}}]".

{placeholder2} can be the second placeholder, and needs to be filled with one or more empty search terms.

[0034] After the empty search term and information about the preset category are input into the first prompt template, the target prompt can be obtained. The first category prediction result of each empty search term can be obtained by entering the target prompt into the large language model.

[0035] For example, a first category prediction result of a specific empty search term can be as follows:

"{"token1": "1&2, 1&5"}".

[0036] The large language model is used to perform category prediction, to obtain a category prediction result of each empty search term. However, because a parameter quantity of the large language model is too large, a reasoning speed of the large language model is relatively slow. In addition, because the large language model is designed for a general purpose, when the large language model is directly used to perform category prediction, category prediction can be implemented by assembling a relatively complex prompt similar to the above-mentioned process.

[0037] To overcome the above-mentioned two problems, in this embodiment of this specification, based on an idea of knowledge distillation, an answer of the large language model is used to train a dedicated category prediction model with a relatively small parameter quantity, to improve a reasoning speed and reduce use difficulty.

[0038] Next, back to FIG. 2, in step 206, the category database is constructed based on each empty search term and the first category prediction result corresponding to the empty search term.

[0039] The category database can include an empty search term-first category prediction result pair including each empty search term and the corresponding first category prediction result; or can be a database entry in which the empty search term can be used as a primary key and the first category prediction result is a field. A plurality of types of database software can be used to construct the category database. For example, HBASE can be used. This is not limited herein.

[0040] Then, in step 208, the category prediction model is trained based on the category database. The category prediction model and the category database are configured to predict the category corresponding to the search term entered by the user.

[0041] Specifically, the category prediction model is trained by using each empty search term in the category database as a training sample and using the first category prediction result corresponding to each empty search term as a label.

[0042] In an embodiment, any empty search term is input into the category prediction model, to obtain a first prediction score for measuring a degree of matching between the empty search term and each preset category; a second prediction score for measuring a degree of matching between a label of the empty search term and each preset category is determined based on one or more categories included in a first category prediction result corresponding to the empty search term; and a parameter value in the category prediction model is adjusted based on a loss value determined based on a difference between the first prediction score and the second prediction score.

[0043] The prediction result output by the category prediction model can be the first prediction score for measuring the degree of matching between the empty search term and each preset category. A higher degree of matching between an empty search term and a preset category indicates a higher prediction score corresponding to the preset category.

[0044] The label includes one or more categories output by the large language model for the empty search term, and does not directly include a prediction score. Therefore, a specific mapping method needs to be used to map the label onto a prediction score. For example, a prediction score corresponding to each category included in the label can be set to 1, and a prediction score corresponding to each category not included in the label can be set to 0. For another example, when there is a sequence between categories in the label, prediction scores of the categories in the label can be set in descending order in the sequence, and prediction scores corresponding to the categories not included in the label are set to 0. For example, a prediction score of a 1st category in the label is set to 1, or a prediction score of a 2nd category is set to 0.8.

[0045] A plurality of existing models can be used as the category prediction model. For example, a combination of a bidirectional encoder representations from transformers (BERT) model and a multi-layer neural network multilayer perceptron (MLP) can be used as the category prediction model.

[0046] In a process of training the category prediction model, a plurality of loss functions can be used to calculate a loss value of the category prediction model between a prediction score of each preset category and a prediction score of each preset category in the label, and a value of each parameter in the category prediction model is adjusted in a gradient descent method in back propagation. For example, a cross-entropy loss function, for example, sigmoid_cross_entropy, based on a sigmoid function can be used. This is not limited herein.

[0047] In the method described in step 202 to step 208, a category prediction result of the large language model for each empty search term is collected and stored in the category database. In addition, knowledge distillation is performed based on data in the category database, and the dedicated category prediction model with a small parameter quantity is trained. The category database and the category prediction model are combined, to perform efficient and accurate category prediction on the search term entered by the user.

[0048] In some possible implementations, after the category prediction model is obtained through training, the method further includes step 210 to step 214, to perform category prediction on the search term entered by the user.

[0049] First, in step 210, the search term entered by the user is obtained.

[0050] The search term entered by the user can be a search term entered by the user in a search box on an interface such as a website or an APP, or can be a search term entered by the user when a search function is invoked in a dialog and chat interface.

[0051] After the search term is received, in some embodiments, the search term can be preprocessed first, to unify formats of the search term and the above-mentioned empty search term. For example, for a Chinese part of the search term, a traditional Chinese character can be converted into a simplified Chinese character, and a sentence is segmented. For an English part of the search term, an upper case letter can be converted into a lower case letter.

[0052] Then, in step 212, the search term is searched for in the category database, and if the search succeeds, it is determined that a found second category prediction result is a target category corresponding to the search term.

[0053] That the search term is searched for in the category database can be searching for the empty search term-first category prediction result pair including the search term, and using the first category prediction result included in the found pair as the second category prediction result; or searching a database entry in which the search term is used as a primary key and using a first category prediction result field in the found entry as the second category prediction result.

[0054] When the search succeeds in step 212, the search result can be used as a category prediction result of the search term, or can be referred to as the target category. When the search fails in step 212, in step 214, the search term is input into the category prediction model, and it is determined that a third category prediction result output by the category prediction model is a target category corresponding to the search term.

[0055] The third category prediction result output by the category prediction model can include one category, or can include a plurality of categories. For example, each category for which the category prediction model outputs a prediction score greater than a preset threshold for each preset category can be used as the third category prediction result.

[0056] In the method described in steps 210 to 214, for a search term existing in the category database, a saved category prediction result is directly output; for a search term that cannot be found in the category database, the category prediction model is invoked to perform category prediction. The category database and the category prediction model are combined, to perform efficient and accurate category prediction on the search term entered by the user.

[0057] In some possible implementations, after the target category corresponding to the search term is determined in step 210 to step 214, the method further includes:

[0058] In step 216, a search result corresponding to the search term is determined based on the target category.

[0059] A project with a high degree of matching with the target category can be found from a project database of the platform based on the target category corresponding to the search term, and is used as the search result corresponding to the search term.

[0060] Specifically, the target category is input into a query parsing service engine, so that the query parsing service engine searches a preset project database for several projects with top-ranked degrees of matching with the target category; and the several projects are determined as the search result corresponding to the search term.

[0061] In step 202 to step 216, a project with a higher correlation can be found for the search

term of the user, and an empty search rate is reduced. In addition, because manual labeling is not required, the method shown in this specification can be automatically completed, thereby reducing labor costs.

[0062] Because the search of the search term fails in step 212 is actually a sample that does not participate in training of the category prediction model. Therefore, in some possible implementations, after the target category corresponding to the search term is determined in step 210 to step 214, the method further includes:

[0063] In step 218, several target search terms unsuccessfully found in the category database are obtained.

[0064] The category prediction model is trained by using the category database. If a search term is not included in the category database, the search term is not used as a training sample to participate in training of the category prediction model. Therefore, after the search function goes online, such a search term can be continuously collected, and the category database and the category prediction model can be updated periodically.

[0065] In step 220, each target search term is input into the large language model to perform category prediction, to obtain a fourth category prediction result of each target search term.

[0066] For a specific implementation process of step 220, refer to step 204. Details are not described herein again.

[0067] In step 224, each target search term and the fourth category prediction result corresponding to the target search term are added to the category database, to obtain an updated category database.

[0068] When each target search term and the fourth category prediction result corresponding to the target search term are added to the category database, a saving form the same as that of existing data in the category database can be used.

[0069] In step 226, the category prediction model is updated based on the updated category database.

[0070] For a specific implementation process of step 226, refer to step 208. Details are not described herein again.

[0071] By using the above-mentioned steps, after the search function of the category prediction method shown in this embodiment of this specification goes online, the category database and the category prediction model can be updated periodically, so that an overall effect

of category prediction is continuously improved.

[0072] After the category prediction model is obtained through training in step 202 to step 208, accuracy of the category prediction model can be further evaluated. In some possible implementations, after the fourth category prediction result of each target search term is obtained in step 220, the method further includes:

[0073] In step 222, prediction accuracy of the category prediction model is evaluated based on a similarity between a third category prediction result and the fourth category prediction result that correspond to each target search term.

[0074] The prediction accuracy of the category prediction model can be positively correlated with the similarity between the third category prediction result and the fourth category prediction result. That is, a higher similarity indicates higher accuracy of the category prediction model.

[0075] The similarity between the third category prediction result and the fourth category prediction result can be measured in a plurality of methods. For example, a Euclidean distance and a cosine similarity can be used. This is not limited herein.

[0076] The foregoing describes the category prediction method provided in the embodiments of this specification. Based on the same idea, this specification further provides a corresponding category prediction apparatus. As shown in FIG. 3, the category prediction apparatus includes: an obtaining module 302, configured to obtain an empty search term set, where the empty search term is a search term whose corresponding search result is empty; a prediction module 304, configured to input each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term; a construction module 306, configured to construct a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and a training module 308, configured to train a category prediction model based on the category database, where the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

[0077] For ease of description, the above-mentioned apparatus is described by dividing the apparatus into various modules based on functions. Certainly, during implementation of this specification, functions of all modules can be implemented in one or more pieces of software and/or hardware.

[0078] The specification also provides a computer-readable non-transitory storage medium,

wherein the storage medium stores a computer program, which, when executed by a processor, can be used to perform one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Herein, the computer-readable non-transitory storage medium or media may include one or more semiconductor-based or other integrated circuits (ICs) (such, as for example, field-programmable gate arrays (FPGAs) or application-specific ICs (ASICs)), hard disk drives (HDDs), hybrid hard drives (HHDs), optical discs, optical disc drives (ODDs), magneto-optical discs, magneto-optical drives, floppy diskettes, floppy disk drives (FDDs), magnetic tapes, solid-state drives (SSDs), RAM-drives, SECURE DIGITAL cards or drives, any other suitable computer-readable non-transitory storage media, or any suitable combination of two or more of these, where appropriate. A computer-readable non-transitory storage medium may be volatile, non-volatile, or a combination of volatile and non-volatile, where appropriate.

[0079] The specification also provides a computer system. FIG. 4 illustrates an example computer system **1000**. In particular embodiments, one or more computer systems **1000** perform one or more steps of one or more methods described or illustrated herein. In particular embodiments, one or more computer systems **1000** provide functionality described or illustrated herein. In particular embodiments, software running on one or more computer systems **1000** performs one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Particular embodiments include one or more portions of one or more computer systems **1000**. Herein, reference to a computer system may encompass a computing device, and vice versa, where appropriate. Moreover, reference to a computer system may encompass one or more computer systems, where appropriate.

[0080] This disclosure contemplates any suitable number of computer systems **1000**. This disclosure contemplates computer system **1000** taking any suitable physical form. As example and not by way of limitation, computer system **1000** may be an embedded computer system, a system-on-chip (SOC), a single-board computer system (SBC) (such as, for example, a computer-on-module (COM) or system-on-module (SOM)), a desktop computer system, a laptop or notebook computer system, an interactive kiosk, a mainframe, a mesh of computer systems, a mobile telephone, a personal digital assistant (PDA), a server, a tablet computer system, or a combination of two or more of these. Where appropriate, computer system **1000** may include one or more computer systems **1000**; be unitary or distributed; span multiple locations; span multiple machines;

span multiple data centers; or reside in a cloud, which may include one or more cloud components in one or more networks. Where appropriate, one or more computer systems **1000** may perform without substantial spatial or temporal limitation one or more steps of one or more methods described or illustrated herein. As an example and not by way of limitation, one or more computer systems **1000** may perform in real time or in batch mode one or more steps of one or more methods described or illustrated herein. One or more computer systems **1000** may perform at different times or at different locations one or more steps of one or more methods described or illustrated herein, where appropriate.

[0081] In particular embodiments, computer system **1000** includes a processor **1002**, memory **1004**, storage **1006**, an input/output (I/O) interface **1008**, a communication interface **1010**, and a bus **1012**. Although this disclosure describes and illustrates a particular computer system having a particular number of particular components in a particular arrangement, this disclosure contemplates any suitable computer system having any suitable number of any suitable components in any suitable arrangement.

[0082] In particular embodiments, processor **1002** includes hardware for executing instructions, such as those making up a computer program. As an example and not by way of limitation, to execute instructions, processor **1002** may retrieve (or fetch) the instructions from an internal register, an internal cache, memory **1004**, or storage **1006**; decode and execute them; and then write one or more results to an internal register, an internal cache, memory **1004**, or storage **1006**. In particular embodiments, processor **1002** may include one or more internal caches for data, instructions, or addresses. This disclosure contemplates processor **1002** including any suitable number of any suitable internal caches, where appropriate. As an example and not by way of limitation, processor **1002** may include one or more instruction caches, one or more data caches, and one or more translation lookaside buffers (TLB s). Instructions in the instruction caches may be copies of instructions in memory **1004** or storage **1006**, and the instruction caches may speed up retrieval of those instructions by processor **1002**. Data in the data caches may be copies of data of **80** in memory **1004** or storage **1006** for instructions executing at processor **1002** to operate on; the results of previous instructions executed at processor **1002** for access by subsequent instructions executing at processor **1002** or for writing to memory **1004** or storage **1006**; or other suitable data. The data caches may speed up read or write operations by processor **1002**. The TLBs may speed up virtual-address translation for processor **1002**. In particular embodiments, processor

1002 may include one or more internal registers for data, instructions, or addresses. This disclosure contemplates processor **1002** including any suitable number of any suitable internal registers, where appropriate. Where appropriate, processor **1002** may include one or more arithmetic logic units (ALUs); be a multi-core processor; or include one or more processors **1002**. Although this disclosure describes and illustrates a particular processor, this disclosure contemplates any suitable processor.

[0083] In particular embodiments, memory **1004** includes main memory for storing instructions for processor **1002** to execute or data for processor **1002** to operate on. As an example and not by way of limitation, computer system **1000** may load instructions from storage **1006** or another source (such as, for example, another computer system **1000**) to memory **1004**. Processor **1002** may then load the instructions from memory **1004** to an internal register or internal cache. To execute the instructions, processor **1002** may retrieve the instructions from the internal register or internal cache and decode them. During or after execution of the instructions, processor **1002** may write one or more results (which may be intermediate or final results) to the internal register or internal cache. Processor **1002** may then write one or more of those results to memory **1004**. In particular embodiments, processor **1002** executes only instructions in one or more internal registers or internal caches or in memory **1004** (as opposed to storage **1006** or elsewhere) and operates only on data in one or more internal registers or internal caches or in memory **1004** (as opposed to storage **1006** or elsewhere). One or more memory buses (which may each include an address bus and a data bus) may couple processor **1002** to memory **1004**. Bus **1012** may include one or more memory buses, as described below. In particular embodiments, one or more memory management units (MMUs) reside between processor **1002** and memory **1004** and facilitate accesses to memory **1004** requested by processor **1002**. In particular embodiments, memory **1004** includes random access memory (RAM). This RAM may be volatile memory, where appropriate. Where appropriate, this RAM may be dynamic RAM (DRAM) or static RAM (SRAM). Moreover, where appropriate, this RAM may be single-ported or multi-ported RAM. This disclosure contemplates any suitable RAM. Memory **1004** may include one or more memories **1004**, where appropriate. Although this disclosure describes and illustrates particular memory, this disclosure contemplates any suitable memory.

[0084] In particular embodiments, storage **1006** includes mass storage for data or instructions. As an example and not by way of limitation, storage **1006** may include a hard disk drive (HDD),

a floppy disk drive, flash memory, an optical disc, a magneto-optical disc, magnetic tape, or a Universal Serial Bus (USB) drive or a combination of two or more of these. Storage **1006** may include removable or non-removable (or fixed) media, where appropriate. Storage **1006** may be internal or external to computer system **1000**, where appropriate. In particular embodiments, storage **1006** is non-volatile, solid-state memory. In particular embodiments, storage **1006** includes read-only memory (ROM). Where appropriate, this ROM may be mask-programmed ROM, programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), electrically alterable ROM (EAROM), or flash memory or a combination of two or more of these. This disclosure contemplates mass storage **1006** taking any suitable physical form. Storage **1006** may include one or more storage control units facilitating communication between processor **1002** and storage **1006**, where appropriate. Where appropriate, storage **1006** may include one or more storages **1006**. Although this disclosure describes and illustrates particular storage, this disclosure contemplates any suitable storage.

[0085] In particular embodiments, I/O interface **1008** includes hardware, software, or both, providing one or more interfaces for communication between computer system **1000** and one or more I/O devices. Computer system **1000** may include one or more of these I/O devices, where appropriate. One or more of these I/O devices may enable communication between a person and computer system **1000**. As an example and not by way of limitation, an I/O device may include a keyboard, keypad, microphone, monitor, mouse, printer, scanner, speaker, still camera, stylus, tablet, touch screen, trackball, video camera, another suitable I/O device or a combination of two or more of these. An I/O device may include one or more sensors. This disclosure contemplates any suitable I/O devices and any suitable I/O interfaces **1008** for them. Where appropriate, I/O interface **1008** may include one or more device or software drivers enabling processor **1002** to drive one or more of these I/O devices. I/O interface **1008** may include one or more I/O interfaces **1008**, where appropriate. Although this disclosure describes and illustrates a particular I/O interface, this disclosure contemplates any suitable I/O interface.

[0086] In particular embodiments, communication interface **1010** includes hardware, software, or both providing one or more interfaces for communication (such as, for example, packet-based communication) between computer system **1000** and one or more other computer systems **1000** or one or more networks. As an example and not by way of limitation, communication interface **1010** may include a network interface controller (NIC) or network adapter for communicating with an

Ethernet or other wire-based network or a wireless NIC (WNIC) or wireless adapter for communicating with a wireless network, such as a WI-FI network. This disclosure contemplates any suitable network and any suitable communication interface **1010** for it. As an example and not by way of limitation, computer system **1000** may communicate with an ad hoc network, a personal area network (PAN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), or one or more portions of the Internet or a combination of two or more of these. One or more portions of one or more of these networks may be wired or wireless. As an example, computer system **1000** may communicate with a wireless PAN (WPAN) (such as, for example, a BLUETOOTH WPAN), a WI-FI network, a WI-MAX network, a cellular telephone network (such as, for example, a Global System for Mobile Communications (GSM) network), or other suitable wireless network or a combination of two or more of these. Computer system **1000** may include any suitable communication interface **1010** for any of these networks, where appropriate. Communication interface **1010** may include one or more communication interfaces **1010**, where appropriate. Although this disclosure describes and illustrates a particular communication interface, this disclosure contemplates any suitable communication interface.

[0087] In particular embodiments, bus **1012** includes hardware, software, or both coupling components of computer system **1000** to each other. As an example and not by way of limitation, bus **1012** may include an Accelerated Graphics Port (AGP) or other graphics bus, an Enhanced Industry Standard Architecture (EISA) bus, a front-side bus (FSB), a HYPERTRANSPORT (HT) interconnect, an Industry Standard Architecture (ISA) bus, an INFINIBAND interconnect, a low-pin-count (LPC) bus, a memory bus, a Micro Channel Architecture (MCA) bus, a Peripheral Component Interconnect (PCI) bus, a PCI-Express (PCIe) bus, a serial advanced technology attachment (SATA) bus, a Video Electronics Standards Association local (VLB) bus, or another suitable bus or a combination of two or more of these. Bus **1012** may include one or more buses **1012**, where appropriate. Although this disclosure describes and illustrates a particular bus, this disclosure contemplates any suitable bus or interconnect.

[0088] It should be further noted that, the terms "include", "comprise", or any other variant thereof are intended to cover a non-exclusive inclusion, so that a process, a method, a product or a device that includes a list of elements not only includes those elements but also includes other elements which are not expressly listed, or further includes elements inherent to such process, method, product, or device. Without more constraints, an element preceded by "includes a …"

does not preclude the existence of additional identical elements in the process, method, product or device that includes the element.

[0089] The embodiments of this specification are described in a progressive way. For same or similar parts in the embodiments, references can be made to each other. Each embodiment focuses on a difference from another embodiment. Particularly, some system embodiments are briefly described because they are basically similar to some method embodiments. For related parts, references can be made to related descriptions in some method embodiments.

[0090] The above-mentioned descriptions are merely embodiments of this specification, and are not intended to limit this specification. A person skilled in the art can make various changes and changes to this specification. Any modification, equivalent replacement, improvement, etc. made without departing from the spirit and principle of this specification shall fall within the scope of the claims in this application.

CLAIMS

What is claimed is:

1. A category prediction method, wherein the method comprises:

obtaining an empty search term set, wherein the empty search term is a search term whose corresponding search result is empty;

inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term;

constructing a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and

training a category prediction model based on the category database, wherein the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

2. The method according to claim 1, wherein the method further comprises:

obtaining the search term entered by the user;

searching the category database for the search term; and

determining that a found second category prediction result is a target category corresponding to the search term if the search succeeds; or

if the search fails, inputting the search term into the category prediction model, and determining that a third category prediction result output by the category prediction model is a target category corresponding to the search term.

3. The method according to claim 2, wherein the method further comprises:

determining a search result corresponding to the search term based on the target category.

4. The method according to claim 2, wherein the method further comprises:

obtaining several target search terms unsuccessfully found in the category database;

inputting each target search term into the large language model to perform category prediction, to obtain a fourth category prediction result of each target search term;

adding each target search term and the fourth category prediction result corresponding to the target search term to the category database, to obtain an updated category database; and

updating the category prediction model based on the updated category database.

5. The method according to claim 4, wherein the method further comprises:

evaluating prediction accuracy of the category prediction model based on a similarity between a third category prediction result and the fourth category prediction result that correspond to each target search term.

6. The method according to claim 1, wherein the inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term comprises:

inputting a target empty search term in all the empty search terms into a first prompt template, to obtain a target prompt, wherein the target prompt further comprises a plurality of preset categories, to indicate the large language model to perform category prediction on the target empty search term based on the plurality of preset categories; and

inputting the target prompt into the large language model, to obtain a first category prediction result corresponding to the target empty search term.

7. The method according to claim 6, wherein the target prompt further comprises description information of each preset category and a hierarchical relationship between all the preset categories.

8. The method according to claim 1, wherein the training a category prediction model based on the category database comprises:

training the category prediction model by using each empty search term in the category database as a training sample and using the first category prediction result corresponding to each empty search term as a label.

9. The method according to claim 8, wherein the training the category prediction model comprises:

inputting any empty search term into the category prediction model, to obtain a first prediction score for measuring a degree of matching between the empty search term and each preset category;

determining, based on one or more categories comprised in a first category prediction result corresponding to the empty search term, a second prediction score for measuring a degree of matching between a label of the empty search term and each preset category; and

adjusting a parameter value in the category prediction model based on a loss value determined based on a difference between the first prediction score and the second prediction score.

10. The method according to claim 3, wherein the determining a search result corresponding to the search term based on the target category comprises:

inputting the target category into a query parsing service engine, so that the query parsing service engine searches a preset project database for several projects with top-ranked degrees of matching with the target category; and

determining the several projects as the search result corresponding to the search term.

11. A computer system, comprising a storage, a processor, and a computer program that is stored in the storage and that is capable of running on the processor, wherein when the processor executes the program, the following steps are implemented:

obtaining an empty search term set, wherein the empty search term is a search term whose corresponding search result is empty;

inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term;

constructing a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and

training a category prediction model based on the category database, wherein the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

12. The computer system according to claim 11, wherein when the processor executes the program, the following steps are further implemented:

obtaining the search term entered by the user;

searching the category database for the search term; and

determining that a found second category prediction result is a target category corresponding to the search term if the search succeeds; or

when the search fails, inputting the search term into the category prediction model, and determining that a third category prediction result output by the category prediction model is a target category corresponding to the search term.

13. The computer system according to claim 12, wherein when the processor executes the program, the following steps are further implemented:

determining a search result corresponding to the search term based on the target category.

14. The computer system according to claim 12, wherein when the processor executes the program, the following steps are further implemented:

obtaining several target search terms unsuccessfully found in the category database;

inputting each target search term into the large language model to perform category prediction, to obtain a fourth category prediction result of each target search term;

adding each target search term and the fourth category prediction result corresponding to the target search term to the category database, to obtain an updated category database; and

updating the category prediction model based on the updated category database.

15. The computer system according to claim 14, wherein when the processor executes the program, the following steps are further implemented:

evaluating prediction accuracy of the category prediction model based on a similarity between a third category prediction result and the fourth category prediction result that correspond to each target search term.

16. The computer system according to claim 11, wherein the inputting each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term comprises:

inputting a target empty search term in all the empty search terms into a first prompt template, to obtain a target prompt, wherein the target prompt further comprises a plurality of preset categories, to indicate the large language model to perform category prediction on the target empty search term based on the plurality of preset categories; and

inputting the target prompt into the large language model, to obtain a first category prediction result corresponding to the target empty search term.

17. The computer system according to claim 16, wherein the target prompt further comprises description information of each preset category and a hierarchical relationship between all the preset categories.

18. The computer system according to claim 11, wherein the training a category prediction model based on the category database comprises:

training the category prediction model by using each empty search term in the category database as a training sample and using the first category prediction result corresponding to each empty search term as a label.

19. The computer system according to claim 18, wherein the training the category prediction model comprises:

inputting any empty search term into the category prediction model, to obtain a first prediction score for measuring a degree of matching between the empty search term and each

preset category;

determining, based on one or more categories comprised in a first category prediction result corresponding to the empty search term, a second prediction score for measuring a degree of matching between a label of the empty search term and each preset category; and

adjusting a parameter value in the category prediction model based on a loss value determined based on a difference between the first prediction score and the second prediction score.

20. The computer system according to claim 13, wherein the determining a search result corresponding to the search term based on the target category comprises:

inputting the target category into a query parsing service engine, so that the query parsing service engine searches a preset project database for several projects with top-ranked degrees of matching with the target category; and

determining the several projects as the search result corresponding to the search term.

21. A category prediction apparatus, wherein the apparatus comprises:

an obtaining module, configured to obtain an empty search term set, wherein the empty search term is a search term whose corresponding search result is empty;

a prediction module, configured to input each empty search term into a large language model to perform category prediction, to obtain a first category prediction result of each empty search term;

a construction module, configured to construct a category database based on each empty search term and the first category prediction result corresponding to the empty search term; and

a training module, configured to train a category prediction model based on the category database, wherein the category prediction model and the category database are configured to predict a category corresponding to a search term entered by a user.

22. A computer-readable non-volatile storage medium, wherein the storage medium stores a computer program, and when the computer program is executed by a processor, the method according to any one of claims 1 to 10 is implemented.