



(51) International Patent Classification:
G06F 15/00 (2006.01)

(21) International Application Number:
PCT/CN2018/125993

(22) International Filing Date:
31 December 2018 (31.12.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
201811647685.3 29 December 2018 (29.12.2018) CN

(71) Applicant: **BEIJING DIDI INFINITY TECHNOLOGY AND DEVELOPMENT CO., LTD.** [CN/CN]; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN).

(72) Inventors: **YU, Peng**; Room 201, 2/F, Building 3, No. 29 Dongbeiwang South Road, Haidian District, Beijing 100193 (CN). **ZHENG, Wanji**; Room 201, 2/F, Building 3, No. 29 Dongbeiwang South Road, Haidian District, Beijing 100193 (CN). **ZHAO, Ji**; Room 201, 2/F, Building 3, No. 29 Dongbeiwang South Road, Haidian District, Beijing 100193 (CN). **CHEN, Huan**; Room 201, 2/F, Building 3, No. 29 Dongbeiwang South Road, Haidian District, Beijing

100193 (CN). **SONG, Qi**; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN). **MA, Li**; Room 201, 2/F, Building 3, No. 29 Dongbeiwang South Road, Haidian District, Beijing 100193 (CN).

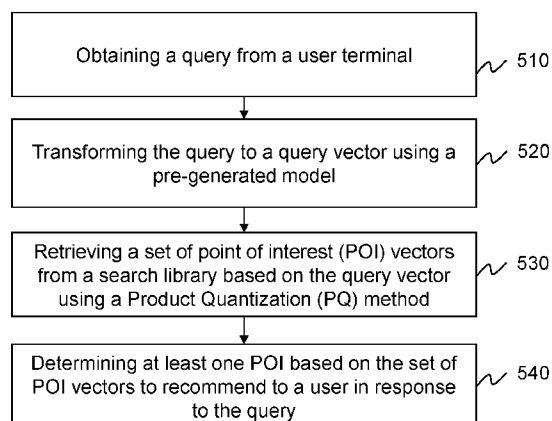
(74) Agent: **METIS IP (CHENGDU) LLC**; Room 808, 8th Floor, Building B7, Block D, Tianfu Jingrong Center, No. 99 West Section of Hupan Road, Tianfu New District, Chengdu, Sichuan 610213 (CN).

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,

(54) Title: ARTIFICIAL INTELLIGENT SYSTEMS AND METHODS FOR SEMANTIC-BASED SEARCH

500



(57) Abstract: A system and a method for semantic-based search. The system may perform the method to: obtaining a query from a user terminal(510); transforming the query to a query vector using a pre-generated model(520); retrieving a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization(PQ) method(530); and determining at least one POI based on the set of POI vectors to recommend to a user in response to the query(540).

FIG. 5

EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

ARTIFICIAL INTELLIGENT SYSTEMS AND METHODS FOR SEMANTIC-BASED SEARCH

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority of Chinese Application No. 201811647685.3, filed on December 29, 2018, the contents of which are incorporated herein in its entirety by reference.

TECHNICAL FIELD

[0002] The present disclosure generally relates to artificial intelligent systems and methods for semantic-based search, and more particularly, to artificial intelligent systems and methods for semantic-based search based on deep learning.

BACKGROUND

[0003] During an online to offline service, a user may input a query relating to a location into a search box of the online to offline service application implemented on a user terminal device when the user wants to retrieve a point of interest (POI). The online to offline system may recommend at least one location to the user based on the semantic of the inputted query. The at least one recommended location may help the user quickly find out the POI that the user wants to retrieve. In existing methods for semantic search, the online to offline system may parse the query, and traverse a plurality of recall policies to retrieve at least one recommended location for the user. However, the problems of the existing methods are that search strategy is complex causing an inefficient search and high search cost. Therefore, it is desirable to provide an artificial intelligence systems and methods for semantic-based search directly based on the query inputted by the user, and more particularly for retrieving POIs according to a deep learning algorithm.

SUMMARY

[0004] An aspect of the present disclosure introduces a system for semantic-based

search may include at least one storage medium including a set of instructions. The system may include a data exchange port communicatively connected to a network. The system may also include at least one processor in communication with the storage medium. When executing the set of instructions, the at least one processor may perform the following operations. The at least one processor may obtain a query from a user terminal via the data exchange port. The at least one processor may transform the query to a query vector using a pre-generated model. The at least one processor may retrieve a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method. The at least one processor may determine at least one POI based on the set of POI vectors to recommend to a user in response to the query.

[0005] In some embodiments, the pre-generated model may be trained based on correlations between query semantics and POIs by performing: obtaining a preliminary model; obtaining a plurality of historical query-POI pairs, each of the plurality of historical query-POI pairs including a historical query and a corresponding historical POI; for each of the plurality of historical query-POI pairs, parsing at least one of the historical query or the corresponding historical POI into one or more segments, and generating one or more text denotations for the historical query-POI pair based on the parsing; and training the preliminary model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method to obtain the pre-generated model.

[0006] In some embodiments, the pre-generated model may include a deep POI semantic model (DPSM) model.

[0007] In some embodiments, to retrieve the set of POI vectors from a search library based on the query vector using a Product Quantization (PQ) method, the at least one processor may obtain an index to a plurality of POIs in a POI database; obtain at least one target location from the query; and determine the set of POI vectors from the search library based on the index, the query vector and the at least one target location. The set of POI vectors may be relevant to the query.

[0008] In some embodiments, to obtain an index to a plurality of POIs in a POI

database, the at least one processor may obtain the plurality of POIs from the POI database; transform the plurality of POIs to the plurality of POI vectors using the pre-generated model; and generate the index to the plurality of POIs based on the plurality of POI vectors using graphics processing unit (GPU).

[0009] In some embodiments, to determine the at least one POI based on the set of POI vectors to recommend to a user in response to the query, the at least one processor may generate an identifier (ID) list based on the set of POI vectors. The ID list may include a set of IDs each directing to a candidate POI. The at least one processor may determine a set of similarity scores for a set of candidate POIs with respect to the set of POI vectors using a trained quality model, each similarity score indicating a relevance level of a corresponding POI to the query. The at least one processor may rank the set of similarity scores of the set of candidate POIs. The at least one processor may determine the at least one POI from the set of candidate POIs based on the ranking of the set of similarity scores.

[0010] In some embodiments, the trained quality model is a Gradient Boosted Decision Tree (GBDT) model.

[0011] In some embodiments, to rank the set of similarity scores of the set of candidate POIs, for each of the set of candidate POIs, the at least one processor may determine whether the corresponding similarity score is greater than a score threshold; obtain one or more relevant candidate POIs with similarity scores greater than the score threshold; and rank the one or more relevant candidate POIs based on a trained ranking model.

[0012] In some embodiments, the trained ranking model may be obtained by a learning to rank method.

[0013] According to another aspect of the present disclosure, an artificial intelligent method for semantic-based search may include: obtaining a query from a user terminal via the data exchange port; transforming the query to a query vector using a pre-generated model; retrieving a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and determining at least one POI based on the set of POI vectors to recommend to a

user in response to the query.

[0014] According to still another aspect of the present disclosure, a non-transitory computer readable medium, comprising at least one set of instructions compatible for semantic-based search, wherein when executed by at least one processor of an electrical device, the at least one set of instructions directs the at least one processor to perform a method, the method comprising: obtaining a query from a user terminal; transforming the query to a query vector using a pre-generated model; retrieving a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and determining at least one POI based on the set of POI vectors to recommend to a user in response to the query.

[0015] According to still another aspect of the present disclosure, a system for semantic-based search may include: a query obtaining module configured to obtain a query from a user terminal; a query vector transforming module configured to transform the query to a query vector using a pre-generated model; a POI vector set obtaining module configured to retrieve a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and a recommended POI determining module configured to determine at least one POI based on the set of POI vectors to recommend to a user in response to the query.

[0016] Additional features will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art upon examination of the following and the accompanying drawings or may be learned by production or operation of the examples. The features of the present disclosure may be realized and attained by practice or use of various aspects of the methodologies, instrumentalities and combinations set forth in the detailed examples discussed below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The present disclosure is further described in terms of exemplary embodiments. These exemplary embodiments are described in detail with

reference to the drawings. These embodiments are non-limiting exemplary embodiments, in which like reference numerals represent similar structures throughout the several views of the drawings, and wherein:

[0018] FIG. 1 is a schematic diagram illustrating an exemplary artificial intelligent system according to some embodiments of the present disclosure;

[0019] FIG. 2 is a schematic diagram illustrating exemplary hardware and/or software components of a computing device according to some embodiments of the present disclosure;

[0020] FIG. 3 is a schematic diagram illustrating exemplary hardware and/or software components of a mobile device according to some embodiments of the present disclosure;

[0021] FIG. 4 is a block diagram illustrating an exemplary processing engine according to some embodiments of the present disclosure;

[0022] FIG. 5 is a flowchart illustrating an exemplary process for determining at least one point of interest (POI) according to some embodiments of the present disclosure;

[0023] FIG. 6 is a flowchart illustrating an exemplary process for obtaining a pre-generated model according to some embodiments of the present disclosure;

[0024] FIG. 7 is a flowchart illustrating an exemplary process for determining a set of PIO vectors according to some embodiments of the present disclosure;

[0025] FIG. 8 is a flowchart illustrating an exemplary process for generating an index to a plurality of POIs according to some embodiments of the present disclosure;

[0026] FIG. 9 is a flowchart illustrating an exemplary process for determining at least one POI according to some embodiments of the present disclosure;

[0027] FIG. 10 is a flowchart illustrating an exemplary process for ranking one or more relevant candidate POIs according to some embodiments of the present disclosure; and

[0028] FIG. 11 is a schematic diagram illustrating an exemplary process for determining at least one POI according to some embodiments of the present

disclosure.

DETAILED DESCRIPTION

[0029] The following description is presented to enable any person skilled in the art to make and use the present disclosure, and is provided in the context of a particular application and its requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present disclosure. Thus, the present disclosure is not limited to the embodiments shown but is to be accorded the widest scope consistent with the claims.

[0030] The terminology used herein is for the purpose of describing particular example embodiments only and is not intended to be limiting. As used herein, the singular forms “a,” “an,” and “the” may be intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises,” “comprising,” “includes,” and/or “including” when used in this disclosure, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0031] These and other features, and characteristics of the present disclosure, as well as the methods of operations and functions of the related elements of structure and the combination of parts and economies of manufacture, may become more apparent upon consideration of the following description with reference to the accompanying drawing(s), all of which form part of this specification. It is to be expressly understood, however, that the drawing(s) are for the purpose of illustration and description only and are not intended to limit the scope of the present disclosure. It is understood that the drawings are not to scale.

[0032] The flowcharts used in the present disclosure illustrate operations that systems implement according to some embodiments of the present disclosure. It is

to be expressly understood, the operations of the flowcharts may be implemented not in order. Conversely, the operations may be implemented in inverted order, or simultaneously. Moreover, one or more other operations may be added to the flowcharts. One or more operations may be removed from the flowcharts.

[0033] An aspect of the present disclosure relates to artificial intelligent systems and methods for semantic-based search. To this end, the artificial intelligent systems and methods may transform a query that a user inputs to a query vector using a deep POI semantic model (DPSM) model. The artificial intelligent systems and methods may retrieve a plurality of similar POI vectors that are relevant to the transformed query vector using a Product Quantization (PQ) method. Then the artificial intelligent systems and methods may select at least one most relevant POI from the plurality of similar POI vectors to recommend to the user in response to the query. The artificial intelligent systems and methods may use a trained Gradient Boosted Decision Tree (GBDT) model to predict similarity score that indicate a relevance level of a POI to the query in the plurality of similar POI vectors. In this way, the artificial intelligent systems and methods may determine at least one most relevant POI and recommend the at least one most relevant POI to the user.

[0034] FIG. 1 is a schematic diagram of an exemplary artificial intelligent (AI) system 100 according to some embodiments of the present disclosure. For example, the AI system 100 may be an online to offline service platform for providing services such as taxi hailing, chauffeur services, delivery vehicles, carpool, bus service, driver hiring, shuttle services, online navigation services, good delivery services, etc. The AI system 100 may include a server 110, a network 120, a user terminal 130, and a storage 140. The server 110 may include a processing engine 112.

[0035] The server 110 may be configured to process information and/or data relating to a query input by a user. For example, the server 110 may transform the query into a query vector. As another example, the server 110 may retrieve a set of POI vectors from a search library based on the query vector. As still another example, the server 110 may determine at least one POI to recommend to the user

in response to the query. In some embodiments, the server 110 may be a single server, or a server group. The server group may be centralized, or distributed (e.g., server 110 may be a distributed system). In some embodiments, the server 110 may be local or remote. For example, the server 110 may access information and/or data stored in the user terminal 130, and/or the storage 140 via the network 120. As another example, the server 110 may connect the user terminal 130, and/or the storage 140 to access stored information and/or data. In some embodiments, the server 110 may be implemented on a cloud platform. Merely by way of example, the cloud platform may be a private cloud, a public cloud, a hybrid cloud, a community cloud, a distributed cloud, an inter-cloud, a multi-cloud, or the like, or any combination thereof. In some embodiments, the server 110 may be implemented on a computing device 200 having one or more components illustrated in FIG. 2 in the present disclosure.

[0036] In some embodiments, the server 110 may include a processing engine 112. The processing engine 112 may process information and/or data relating to the query input by the user to perform one or more functions described in the present disclosure. For example, the processing engine 112 may transform the query into a query vector. As another example, the processing engine 112 may retrieve a set of POI vectors from a search library based on the query vector. As still another example, the processing engine 112 may determine at least one POI to recommend to the user in response to the query. In some embodiments, the processing engine 112 may include one or more processing engines (e.g., single-core processing engine(s) or multi-core processor(s)). Merely by way of example, the processing engine 112 may be one or more hardware processors, such as a central processing unit (CPU), an application-specific integrated circuit (ASIC), an application-specific instruction-set processor (ASIP), a graphics processing unit (GPU), a physics processing unit (PPU), a digital signal processor (DSP), a field programmable gate array (FPGA), a programmable logic device (PLD), a controller, a microcontroller unit, a reduced instruction-set computer (RISC), a microprocessor, or the like, or any combination thereof.

[0037] The network 120 may facilitate exchange of information and/or data. In some embodiments, one or more components of the AI system 100 (e.g., the server 110, the user terminal 130, and the storage 140) may transmit information and/or data to other component(s) in the AI system 100 via the network 120. For example, the server 110 may obtain the query from the user terminal 130 via the network 120. As another example, the server 110 may send at least one recommended POI to the user terminal 130 to render the user terminal 130 to present the at least one recommended POI. In some embodiments, the network 120 may be any type of wired or wireless network, or combination thereof. Merely by way of example, the network 120 may be a cable network, a wireline network, an optical fiber network, a tele communications network, an intranet, an Internet, a local area network (LAN), a wide area network (WAN), a wireless local area network (WLAN), a metropolitan area network (MAN), a wide area network (WAN), a public telephone switched network (PSTN), a Bluetooth network, a ZigBee network, a near field communication (NFC) network, or the like, or any combination thereof. In some embodiments, the network 120 may include one or more network access points. For example, the network 120 may include wired or wireless network access points such as base stations and/or internet exchange points 120-1, 120-2, ... , through which one or more components of the AI system 100 may be connected to the network 120 to exchange data and/or information between them.

[0038] The user terminal 130 may be any electronic device used by a user of the online to offline service. In some embodiments, the user terminal 130 may be a mobile device 130-1, a tablet computer 130-2, a laptop computer 130-3, a desktop computer 130-4, or the like, or any combination thereof. In some embodiments, the mobile device 130-1 may be a wearable device, a smart mobile device, a virtual reality device, an augmented reality device, or the like, or any combination thereof. In some embodiments, the wearable device may be a smart bracelet, a smart footgear, a smart glass, a smart helmet, a smart watch, a smart clothing, a smart backpack, a smart accessory, or the like, or any combination thereof. In some embodiments, the smart mobile device may be a smartphone, a personal digital

assistance (PDA), a gaming device, a navigation device, a point of sale (POS) device, or the like, or any combination thereof. In some embodiments, the virtual reality device and/or the augmented reality device may be a virtual reality helmet, a virtual reality glass, a virtual reality patch, an augmented reality helmet, an augmented reality glass, an augmented reality patch, or the like, or any combination thereof. For example, the virtual reality device and/or the augmented reality device may be a Google Glass™, a RiftCon™, a Fragments™, a Gear VR™, etc. In some embodiments, the desktop computer 130-4 may be an onboard computer, an onboard television, etc.

[0039] In some embodiments, the user terminal 130 may be a device with positioning technology for locating the position of the user and/or the user terminal 130. The positioning technology used in the present disclosure may be a global positioning system (GPS), a global navigation satellite system (GLONASS), a compass navigation system (COMPASS), a Galileo positioning system, a quasi-zenith satellite system (QZSS), a wireless fidelity (WiFi) positioning technology, or the like, or any combination thereof. One or more of the above positioning technologies may be used interchangeably in the present disclosure.

[0040] In some embodiments, the user terminal 130 may further include at least one network port. The at least one network port may be configured to send information to and/or receive information from one or more components in the AI system 100 (e.g., the server 110, the storage 140) via the network 120. In some embodiments, the user terminal 130 may be implemented on a computing device 200 having one or more components illustrated in FIG. 2, or a mobile device 300 having one or more components illustrated in FIG. 3 in the present disclosure.

[0041] The storage 140 may store data and/or instructions. For example, the storage 140 may store a pre-generated model, a trained quality model, and/or a trained ranking model. As another example, the storage 140 may store a search library, a POI database, and/or a plurality of historical query-POI pairs. As still another example, the storage 140 may store data and/or instructions that the server 110 may execute or use to perform exemplary methods described in the present

disclosure. In some embodiments, the storage 140 may be a mass storage, a removable storage, a volatile read-and-write memory, a read-only memory (ROM), or the like, or any combination thereof. Exemplary mass storage may include a magnetic disk, an optical disk, a solid-state drive, etc. Exemplary removable storage may include a flash drive, a floppy disk, an optical disk, a memory card, a zip disk, a magnetic tape, etc. Exemplary volatile read-and-write memory may include a random-access memory (RAM). Exemplary RAM may include a dynamic RAM (DRAM), a double data rate synchronous dynamic RAM (DDR SDRAM), a static RAM (SRAM), a thyristor RAM (T-RAM), and a zero-capacitor RAM (Z-RAM), etc. Exemplary ROM may include a mask ROM (MROM), a programmable ROM (PROM), an erasable programmable ROM (EPROM), an electrically erasable programmable ROM (EEPROM), a compact disk ROM (CD-ROM), and a digital versatile disk ROM, etc. In some embodiments, the storage 140 may be implemented on a cloud platform. Merely by way of example, the cloud platform may be a private cloud, a public cloud, a hybrid cloud, a community cloud, a distributed cloud, an inter-cloud, a multi-cloud, or the like, or any combination thereof.

[0042] In some embodiments, the storage 140 may include at least one network port to communicate with other devices in the AI system 100. For example, the storage 140 may be connected to the network 120 to communicate with one or more components of the AI system 100 (e.g., the server 110, the user terminal 130) via the at least one network port. One or more components in the AI system 100 may access the data or instructions stored in the storage 140 via the network 120. In some embodiments, the storage 140 may be directly connected to or communicate with one or more components in the AI system 100 (e.g., the server 110, the user terminal 130). In some embodiments, the storage 140 may be part of the server 110.

[0043] In some embodiments, one or more components of the AI system 100 (e.g., the server 110, the user terminal 130, and the storage 140) may communicate with each other in form of electronic and/or electromagnetic signals, through wired and/or

wireless communication. In some embodiments, the AI system 100 may further include at least one data exchange port. The at least one exchange port may be configured to receive information and/or send information relating to determining the recommended locations (e.g., in form of electronic signals and/or electromagnetic signals) between any electronic devices in the AI system 100. In some embodiments, the at least one data exchange port may be one or more of an antenna, a network interface, a network port, or the like, or any combination thereof. For example, the at least one data exchange port may be a network port connected to the server 110 to send information thereto and/or receive information transmitted therefrom.

[0044] FIG. 2 is a schematic diagram illustrating exemplary hardware and software components of a computing device 200 on which the server 110, and/or the user terminal 130 may be implemented according to some embodiments of the present disclosure. For example, the processing engine 112 may be implemented on the computing device 200 and configured to perform functions of the processing engine 112 disclosed in this disclosure.

[0045] The computing device 200 may be used to implement an AI system 100 for the present disclosure. The computing device 200 may be used to implement any component of AI system 100 that perform one or more functions disclosed in the present disclosure. For example, the processing engine 112 may be implemented on the computing device 200, via its hardware, software program, firmware, or a combination thereof. Although only one such computer is shown, for convenience, the computer functions relating to the online to offline service as described herein may be implemented in a distributed fashion on a number of similar platforms, to distribute the processing load.

[0046] The computing device 200, for example, may include COM ports 250 connected to and from a network connected thereto to facilitate data communications. The COM port 250 may be any network port or data exchange port to facilitate data communications. The computing device 200 may also include a processor (e.g., the processor 220), in the form of one or more processors (e.g.,

logic circuits), for executing program instructions. For example, the processor may include interface circuits and processing circuits therein. The interface circuits may be configured to receive electronic signals from a bus 210, wherein the electronic signals encode structured data and/or instructions for the processing circuits to process. The processing circuits may conduct logic calculations, and then determine a conclusion, a result, and/or an instruction encoded as electronic signals. The processing circuits may also generate electronic signals including the conclusion or the result (e.g., the at least one recommended POI) and a triggering code. In some embodiments, the trigger code may be in a format recognizable by an operation system (or an application installed therein) of an electronic device (e.g., the user terminal 130) in the AI system 100. For example, the trigger code may be an instruction, a code, a mark, a symbol, or the like, or any combination thereof, that can activate certain functions and/or operations of a mobile phone or let the mobile phone execute a predetermined program(s). In some embodiments, the trigger code may be configured to render the operation system (or the application) of the electronic device to generate a presentation of the conclusion or the result (e.g., a prediction result) on an interface of the electronic device. Then the interface circuits may send out the electronic signals from the processing circuits via the bus 210.

[0047] The exemplary computing device may include the internal communication bus 210, program storage and data storage of different forms including, for example, a disk 270, and a read only memory (ROM) 230, or a random access memory (RAM) 240, for various data files to be processed and/or transmitted by the computing device. The exemplary computing device may also include program instructions stored in the ROM 230, RAM 240, and/or other type of non-transitory storage medium to be executed by the processor 220. The methods and/or processes of the present disclosure may be implemented as the program instructions. The exemplary computing device may also include operation systems stored in the ROM 230, RAM 240, and/or other type of non-transitory storage medium to be executed by the processor 220. The program instructions may be compatible with the operation systems for providing the online to offline service. The computing device

200 also includes an I/O component 260, supporting input/output between the computer and other components. The computing device 200 may also receive programming and data via network communications.

[0048] Merely for illustration, only one processor is illustrated in FIG. 2. Multiple processors are also contemplated; thus, operations and/or method steps performed by one processor as described in the present disclosure may also be jointly or separately performed by the multiple processors. For example, if in the present disclosure the processor of the computing device 200 executes both step A and step B, it should be understood that step A and step B may also be performed by two different processors jointly or separately in the computing device 200 (e.g., the first processor executes step A and the second processor executes step B, or the first and second processors jointly execute steps A and B).

[0049] FIG. 3 is a schematic diagram illustrating exemplary hardware and/or software components of an exemplary mobile device 300 on which the user terminal 130 may be implemented according to some embodiments of the present disclosure.

[0050] As illustrated in FIG. 3, the mobile device 300 may include a communication platform 310, a display 320, a graphic processing unit (GPU) 330, a central processing unit (CPU) 340, an I/O 350, a memory 360, and a storage 390. The CPU may include interface circuits and processing circuits similar to the processor 220. In some embodiments, any other suitable component, including but not limited to a system bus or a controller (not shown), may also be included in the mobile device 300. In some embodiments, a mobile operating system 370 (e.g., iOS™, Android™, Windows Phone™, etc.) and one or more applications 380 may be loaded into the memory 360 from the storage 390 in order to be executed by the CPU 340. The applications 380 may include a browser or any other suitable mobile apps for receiving and rendering information relating to the at least one recommended POI. User interactions with the information stream may be achieved via the I/O devices 350 and provided to the processing engine 112 and/or other components of the system 100 via the network 120.

[0051] To implement various modules, units, and their functionalities described in

the present disclosure, computer hardware platforms may be used as the hardware platform(s) for one or more of the elements described herein (e.g., the AI system 100, and/or other components of the AI system 100 described with respect to FIGs. 1-11). The hardware elements, operating systems and programming languages of such computers are conventional in nature, and it is presumed that those skilled in the art are adequately familiar therewith to adapt those technologies to determine at least one POI to recommend to a user in response to a query as described herein. A computer with user interface elements may be used to implement a personal computer (PC) or other type of work station or terminal device, although a computer may also act as a server if appropriately programmed. It is believed that those skilled in the art are familiar with the structure, programming and general operation of such computer equipment and as a result the drawings should be self-explanatory.

[0052] One of ordinary skill in the art would understand that when an element of the AI system 100 performs, the element may perform through electrical signals and/or electromagnetic signals. For example, when a server 110 processes a task, such as determine at least one POI to recommend to the user in response to the query, the server 110 may operate logic circuits in its processor to process such task. When the server 110 completes determine the at least one POI, the processor of the server 110 may generate electrical signals encoding the the at least one POI to recommend to the user. The processor of the server 110 may then send the electrical signals to at least one data exchange port of a target system associated with the server 110. The server 110 communicates with the target system via a wired network, the at least one data exchange port may be physically connected to a cable, which may further transmit the electrical signals to an input port (e.g., an information exchange port) of the user terminal 130. If the server 110 communicates with the target system via a wireless network, the at least one data exchange port of the target system may be one or more antennas, which may convert the electrical signals to electromagnetic signals. Within an electronic device, such as the user terminal 130, and/or the server 110, when a processor thereof processes an instruction, sends out an instruction, and/or performs an action,

the instruction and/or action is conducted via electrical signals. For example, when the processor retrieves or saves data from a storage medium (e.g., the storage 140), it may send out electrical signals to a read/write device of the storage medium, which may read or write structured data in the storage medium. The structured data may be transmitted to the processor in the form of electrical signals via a bus of the electronic device. Here, an electrical signal may be one electrical signal, a series of electrical signals, and/or a plurality of discrete electrical signals.

[0053] FIG. 4 is a block diagram illustrating an exemplary processing engine 112 according to some embodiments of the present disclosure. As illustrated in FIG. 4, the processing engine 112 may include a query obtaining module 410, a query vector transforming module 420, a POI vector set obtaining module 430, a recommended POI determining module 440, and a model training module 450.

[0054] The query obtaining module 410 may be configured to obtain the query from the user terminal 130. The query may be inputted by the user to search a content relating to the online to offline service (e.g., a content relating to a location of the online to offline service). The online to offline service may include but not limited to a navigation service, a car-hailing service, a meal-delivery service.

[0055] The query vector transforming module 420 may be configured to transform the query obtained by the query obtaining module 410 to a query vector using a pre-generated model. In some embodiments, the query vector may be a semantic representation of the query. For example, the query vector may be a multi-dimensional vector of numerical features that represent the query, for example a 128-dimensional vector. In some embodiments, the pre-generated model may include a deep POI semantic model (DPSM) model. The DPSM model may be trained online and/offline, and the training process of the DPSM may be found elsewhere in the present disclosure (e.g., FIG. 5 and the description thereof).

[0056] The POI vector set obtaining module 430 may be configured to retrieve a set of point of interest (POI) vectors from a search library using a Product Quantization (PQ) method. In some embodiments, the search library may be an algorithms library for searching similar content, for example, the search library may include a

Facebook open source search library (e.g., Faiss). In some embodiments, the PQ method may be used to search a plurality of POI vectors that are nearest to the query that the user inputs. Alternatively or additionally, the PQ method for retrieving the POI vectors may include a tree-based method (e.g., a K-dimension (KD) Tree method, an R-Tree method, etc.), a hash-based method (e.g., a Locality Sensitive hash (LSH) method, etc.), or the like, or any combination thereof.

[0057] The recommended POI determining module 440 may be configured to determine at least one POI based on the set of POI vectors to recommend to the user in response to the query. The at least one POI may include at least one POI that are determined as the most relevant with the query that the user inputs. In some embodiments, the method and/or process for determining the at least one POI may be found elsewhere in the present disclosure (e.g., FIG. 5 and/or FIG. 9 and the description thereof).

[0058] The model training module 450 may be configured to train a model. For example, the model training module 450 may obtain a preliminary model and a plurality of historical query-POI pairs. For each of the plurality of historical query-POI pairs, the model training module 450 may parse at least one of the historical query or the corresponding historical POI into one or more segments, and generate one or more text denotations for the historical query-POI pair based on the parsing. The model training module 450 may train the preliminary model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method to obtain the pre-generated model. Details regarding to the training of the model may be found elsewhere in the present disclosure (e.g., FIG. 6 and the description thereof).

[0059] The modules in the processing engine 112 may be connected to or communicate with each other via a wired connection or a wireless connection. The wired connection may be a metal cable, an optical cable, a hybrid cable, or the like, or any combination thereof. The wireless connection may be a Local Area Network (LAN), a Wide Area Network (WAN), a Bluetooth, a ZigBee, a Near Field Communication (NFC), or the like, or any combination thereof. Two or more of the

modules may be combined into a single module, and any one of the modules may be divided into two or more units. For example, the model training module 450 may be divided into two or more units for training a deep POI semantic model (DPSM) model, a Gradient Boosted Decision Tree (GBDT) model, and a ranking model, respectively. As another example, the processing engine 112 may include a storage module (not shown) used to store data and/or information relating to the query.

[0060] FIG. 5 is a flowchart illustrating an exemplary process 500 for determining at least one point of interest (POI) according to some embodiments of the present disclosure. The process 500 may be executed by the AI system 100. For example, the process 500 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 500. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 500 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 5 and described below is not intended to be limiting.

[0061] In 510, the processing engine 112 (e.g., the processor 220, the query obtaining module 410) may obtain a query from the user terminal 130.

[0062] In some embodiments, the query may be inputted into the user terminal 130 by the user for searching a content relating to the online to offline service (e.g., a content relating to a location of the online to offline service). In some embodiments, the query may include a textual query, a voice query, an image query, or the like, or any combination thereof. For example, the query may include a word, a term, a sentence, or the like, or any expression associated with a location.

[0063] In some embodiments, the user terminal 130 may send the query to the processing engine 112 via a data exchange port that is communicatively connected to a network 120. The processing engine 112 may obtain the query via the data

exchange port.

[0064] In 520, the processing engine 112 (e.g., the processor 220, the query vector transforming module 420) may transform the query to a query vector using a pre-generated model.

[0065] In some embodiments, the query vector may be a semantic representation of the query. For example, the query vector may be a multi-dimensional vector of numerical features that represent the query. For example, the query vector may be a mathematical expression of a 128-dimensional vector.

[0066] In some embodiments, the pre-generated model may include a trained model used for transforming a query to a query vector. For example, the pre-generated model may include a deep POI semantic model (DPSM) model. The DPSM model may include a query model. The query model may include a plurality of layers. For example, the query model may include a contextual representation layer, a feature extraction layer and a vector representation layer. The processing engine 112 may input the query into the DPSM model. The output of the DPSM model may be the corresponding query after the query passing through the contextual representation layer, the feature extraction layer and the vector representation layer, respectively.

[0067] In some embodiments, the DPSM model may also transform a POI to a POI vector, and/or determine a relevance level (e.g., a similarity score) of a corresponding POI to the query. For example, the DPSM model may further include a POI model. The POI model may include a plurality of contextual representation layers, a plurality of feature extraction layers and a plurality of vector representation layers. For example, the processing engine 112 may input a POI name of a POI into a contextual representation layer, a feature extraction layer and a vector representation layer, respectively to obtain a POI name vector. The processing engine 112 may input a POI address of the POI into a contextual representation layer, a feature extraction layer and a vector representation layer, respectively to obtain a POI address vector. Then the processing engine 112 may input the POI name vector and the POI address vector into a vector representation layer to obtain

a POI vector. The DPSM model may finally output a similarity score for a POI with respect to the query based on the query vector and the POI vector. As used herein, the processing engine 112 may use a middle output (i.e., the query vector, the POI vector) of the pre-generated model. In some embodiments, the process and/or method for training of the DPSM model may be found elsewhere in the present disclosure (e.g., FIG. 6 and the description thereof).

[0068] In some embodiments, the processing engine 112 may input the query into the query model of the DPSM model. The contextual representation layer of the DPSM model may segment the inputted query, and a result of the segmentation may be embedded into a vector representation of a single word (e.g., "go", "location") and/or a phrase (e.g., "go to", "a location"). For example, each of the inputted query may be segmented into a plurality of vector representations according to a semantic sequence of the inputted query, and each vector representation may represent a single word. The plurality of vector representations may be transmitted to the feature extraction layer. In the feature extraction layer, any two adjacent vector representations according to the semantic sequence of the inputted query in the plurality of vector representations may be convolved, i.e., a contextual information may be extracted. The convolved information may be maxpooled to obtain a semantic information of the query. For example, the vector representation corresponding to each maximum convolved information may be selected. Then in the vector representation layer, the semantic information may be spliced by full-connecting to obtain the query vector. In some embodiments, a plurality of vector representations representing phrases may be processed according to a same way as the plurality of vector representations representing single words.

[0069] In 530, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may retrieve a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method.

[0070] In some embodiments, the search library may be an algorithms library for searching similar content. For example, the search library may include a Facebook open source search library (e.g., Faiss) for nearest neighbor searching in high-

dimension.

[0071] In some embodiments, the PQ method may be a product quantization-based method for similarity searching of POIs. For example, the PQ method may be used to search a plurality of POI vectors that are nearest to the query that the user inputs. In some embodiments, the PQ method may include a PQ method, an IVFPQ method, or any modified method based on PQ method, or any combination thereof. Alternatively or additionally, the method and/or process for retrieving the set of POI vectors may include a tree-based method (e.g., a K-dimension (KD) Tree method, an R-Tree method, etc.), a hash-based method (e.g., a Locality Sensitive hash (LSH) method, etc.), or the like, or any combination thereof.

[0072] In some embodiments, the set of POI vectors may include a plurality of POI vectors that are highly correlated with the query. For example, a set of POI vectors may include a predetermined number of POI vectors that are most relevant to the query. The predetermine number may be a default value or dynamically determined by the processing engine 112.

[0073] In some embodiments, the processing engine 112 may obtain an index to a plurality of POIs in a POI database. The index may indicate spatial information (e.g., a nation, a city, a street, etc.) of the plurality of POIs. The processing engine 112 may obtain at least one target location from the query. The target location may indicate a location that the user searches, for example, the target location may include a city, a street, a road, a point, or the like, or any combination thereof. The processing engine 112 may determine the set of POI vectors from the search library based on the index, the query vector and the at least one target location. The set of POI vectors may be relevant to the query. For example, the processing engine 112 may input the index, the query vector, and the at least one target location into the search library (e.g., Faiss). The processing engine 112 may first invoke the corresponding index of a city based on a city ID of the at least one target location, and then search the set of POI vectors that are associated with the corresponding POIs in the city according to the PQ method. In some embodiments, the method and/or process for retrieving the set of POI vectors may be found elsewhere in the

present disclosure (e.g., FIG. 7 and the description thereof).

[0074] In 540, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may determine at least one POI based on the set of POI vectors to recommend to a user in response to the query.

[0075] In some embodiments, the at least one POI may include at least one POI that are determined as the most relevant to the query that the user inputs. The processing engine 112 may recommend the at least one POI to the user in response to the query.

[0076] In some embodiments, the processing engine 112 may generate an identifier (ID) list based on the set of POI vectors. The ID list may include a set of IDs each directing to a candidate POI (e.g., a set of candidate POI-ID pairs). The set of POI vectors may correspond to a set of candidate POIs. The ID of a corresponding candidate POI may be an identifier that represent the candidate POI. The processing engine 112 may determine a set of similarity scores for the set of candidate POIs with respect to the set of POI vectors using a trained quality model. Each similarity score may indicate a relevance level of a corresponding candidate POI to the query. For each the candidate POI-ID pair in the ID list, the trained quality model may predict a similarity score of the candidate location and the query. The processing engine 112 may rank the set of similarity scores of the set of candidate POIs, and select the at least one POI that ranks at top of the ranking from the set of candidate POIs. In some embodiments, the method and/or process for determining the at least one POI may be found elsewhere in the present disclosure (e.g., FIG. 9 and the description thereof).

[0077] In some embodiments, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may recommend the at least one POI to the user. For example, the recommended at least one POI may be displayed on the user interface of the user terminal 130 for the user to view and/or select. Merely by way of example, at least one POI recommended to the user in response to a query may be shown in Table 1.

Query	Guiyou Liyuan	Digital 3 Valley	Sunshine 100	No. 731 Aerospace Hospital	Go to Huguo Temple
Recommended POIs	Guiyou Mansion	Shangdong·Digital Valley	Sunshine 100 International Apartment	Outpatient Department, No. 731 Aerospace Hospital	Huguo Temple Hotel
	Park Lot of Guiyou Mansion	Shangdong·Digital Valley (Southeast Door)	West Door, Sunshine 100 International Apartment	Emergency Department, No. 731 Aerospace Hospital	Huguo Temple Snack Bar
	Tower A of Guiyou Mansion	No. 2 Building, A zone, Shangdong·Digital Valley	Est Door, Sunshine 100 International Apartment	Southwest Door, No. 731 Aerospace Hospital	Huguo Temple Street

Table 1

[0078] When a user input query, the processing engine 112 may determine at least one recommended POI and transmit the at least one recommended POI to the user (or a user terminal thereof). For example, as shown in Table 1, when a user input “Guiyou Liyuan” by the user terminal, the processing engine 112 may display recommended POIs, such as “Guiyou Mansion”, “Park Lot of Guiyou Mansion”, and/or “Tower A of Guiyou Mansion” on the user interface of the user terminal for the user to view and/or select.

[0079] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. For example, operation 510 and operation 520 may be integrated into a single step. However, those variations and modifications do not depart from the scope of the present disclosure.

[0080] FIG. 6 is a flowchart illustrating an exemplary process 600 for obtaining a pre-generated model according to some embodiments of the present disclosure.

The process 600 may be executed by the AI system 100. For example, the process 600 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 600. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 600 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 6 and described below is not intended to be limiting.

[0081] In 610, the processing engine 112 (e.g., the processor 220, the model training module 450) may obtain a preliminary model.

[0082] In some embodiments, the preliminary model may be an initial neural network model to be trained to obtain a pre-generated model. For example, if the pre-generated model is a DPSM model, the preliminary model may be an initial DPSM model including an initial query model and an initial POI model. The initial query model may be an initial neural network including a plurality of layers. For example, the initial query model may include a plurality of contextual representation layers, a plurality of feature extraction layers, a plurality of vector representation layers and a plurality of output layers. Similarly, the initial POI model may be an initial neural network including a plurality of layers. For example, the initial POI model may include a plurality of contextual representation layers, a plurality of feature extraction layers, and a plurality of vector representation layers.

[0083] In 620, the processing engine 112 (e.g., the processor 220, the model training module 450) may obtain a plurality of historical query-POI pairs. In some embodiments, each of the plurality of historical query-POI pairs may include a historical query and a corresponding historical POI.

[0084] In some embodiments, a historical query-POI pair in the plurality of historical query-POI pairs may indicate a matching pair of a historical query that a historical user input and a corresponding historical POI that the user selected in response to

the historical query. For example, a historical user may input a query, such as “Guiyou Liyuan”, and a plurality of POIs may be displayed to the historical user, such as “Guiyou Liyuan Building”, “Park lot of Guiyou Building”, “bus station of Guiyou”, etc. The user may click/select the “Park lot of Guiyou Building” in response to the query. Therefore, the “Guiyou Liyuan” and the “Park lot of Guiyou Building” may form a query-POI pair. In some embodiments, the plurality of historical query-POI pairs may be stored in a user log. The user log may include a user information, such as a user name, a historical query of the user, a selected historical POI by the user, or the like, or any combination thereof. The user log may be stored in the ROM 230, RAM 240, and/or other type of storage medium to be executed by the processing engine 112.

[0085] In 630, for each of the plurality of historical query-POI pairs, the processing engine 112 (e.g., the processor 220, the model training module 450) may parse at least one of the historical query or the corresponding historical POI into one or more segments, and generate one or more text denotations for the historical query-POI pair based on the parsing.

[0086] In some embodiments, for each of the plurality of historical query-POI pairs, the processing engine 112 (the model training module 450, or a segment unit in the model training module 450) may segment the historical query or the corresponding historical POI into one or more segments using different segmentation methods. For example, the processing engine 112 may segment a query or a POI of “Beijing University” into three segments (e.g., “Bei”+“jing”+“University”) or two segments (e.g., “Beijing”+“University”), or one segment (e.g., “Beijing University”).

[0087] In some embodiments, one or more text denotations for the historical query-POI pair may be generated based on the parsing. A text denotation may be a textual representation. For example, the processing engine 112 may respectively generate a text denotation of “Bei/jing/University” based on the three segments of “Bei”+“jing”+“University”, a text denotation of “Beijing/University” based on the two segments of “Beijing”+“University”, and a text denotation of “Beijing University” based on the one segment of “Beijing University”.

[0088] In some embodiments, the processing engine 112 may parse both a historical query and the corresponding historical POI in each historical query-POI pair to obtain one or more text denotations for the historical query-POI pair.

Alternatively, the processing engine 112 may only parse the historical query (or only the corresponding historical POI) in each historical query-POI pair to obtain the one or more text denotations for the historical query-POI pair.

[0089] In 640, the processing engine 112 (e.g., the processor 220, the model training module 450) may train the preliminary model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method to obtain the pre-generated model.

[0090] In some embodiments, the processing engine 112 may input the text denotations of the plurality of historical query-POI pairs into the preliminary model to train the preliminary model. For example, the processing engine 112 may input the one or more text denotations for each historical query-POI pair into the initial DPSM model including the initial query model and the initial POI model, and revise or adjust parameters (e.g., layers, nodes, weights, etc.) of the initial query model and the initial POI model according to the deep learning method. The preliminary model may be trained to obtain the pre-generated model until the preliminary model is convergent in the training process.

[0091] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. In some embodiments, one or more other optional operations (e.g., a storing operation) may be added elsewhere in the exemplary process 600. For example, process 600 may further include storing the pre-generated model after obtaining the pre-generated model.

[0092] FIG. 7 is a flowchart illustrating an exemplary process 700 for determining a set of POI vectors according to some embodiments of the present disclosure. The

process 700 may be executed by the AI system 100. For example, the process 700 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 700. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 700 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 7 and described below is not intended to be limiting.

[0093] In 710, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may obtain an index to a plurality of POIs in a POI database.

[0094] In some embodiments, the index may indicate a spatial information (e.g., a nation, a city, a street, etc.) of the plurality of POIs. For example, a POI of “Beijing University - No. 5, Yiheyuan Road, Haidian District, Beijing” may be included in the plurality of POIs, and the index corresponding to the POI may indicate the spatial information of Beijing or Haidian District, etc. The index may be stored in the storage (e.g., the ROM 230, the RAM 240, and/or other type of storage medium to be executed by the processing engine 112).

[0095] In some embodiments, the POI database may include a data set of a large number of POIs. In some embodiments, the processing engine 112 may transform the plurality of POIs in the POI database to the plurality of POI vectors using the pre-generated model (e.g., DPSM model). For example, the processing engine 112 may input each of the plurality of POIs into the DPSM model to obtain the plurality of POI vectors. In some embodiments, the POI vector may be a mathematical expression of a 128-dimensional vector. The processing engine 112 may use a graphics processing unit (GPU) to generate the index by mapping the plurality of POI vectors to spatial information. The transformation of the POI to POI vectors may be found elsewhere of the present disclosure (e.g., FIG. 8, and the description thereof).

[0096] In some embodiments, the processing engine 112 may obtain the index to the plurality of POIs in the POI database using the GPU. Alternatively or additionally, the index may be pre-generated and stored in a storage device (e.g., the storage 140, the ROM 230 the RAM 240, and/or other type of storage medium to be executed by the processing engine 112), and the processing engine 112 may obtain the index from the storage 140 directly.

[0097] In 720, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may obtain at least one target location from the query.

[0098] In some embodiments, the target location may indicate a target area that the user searches, for example, the target location may include a nation, a city, a street, a road, a point, or the like, or any combination thereof. In some embodiments, the processing engine 112 may extract the target location from GPS information obtained from the user terminal. For example, the processing engine 112 may obtain a target city that the user terminal located based on the GPS information. In some embodiments, the processing engine 112 may extract the target location from a selected location that the user selects on a user interface of the user terminal 130. For example, before the user inputs the query, the user may select the target location on the user interface. In some embodiments, the target location may include a nation ID, a city ID, a street ID, or the like, or any combination thereof.

[0099] In 730, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may determine the set of POI vectors from the search library based on the index, the query vector and the at least one target location. The set of POI vectors are relevant to the query.

[0100] In some embodiments, the processing engine 112 may input the query vector, the index and at least one target location into the search library (e.g., Faiss). For example, the processing engine 112 may map the at least one target location (e.g., the city ID) with the index to obtain a plurality of POI vectors in the city of the at least one target location. The processing engine 112 may then determine a similarity between the query vector and the plurality of POI vectors in the city, and select the set of POI vectors from the plurality of POI vectors in the city. Each POI

vector in the set of POI vectors may have a high similarity with the query vector. In some embodiments, the processing engine 112 may perform the similarity search using the PQ method to determine the similarity between the query vector and the plurality of POI vectors in the city. In some embodiments, the set of POI vectors may include a predetermined number of POI vectors that are most similar (or most relevant) to the query vector. The predetermine number may be a default value or dynamically determined by the processing engine 112.

[0101] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. For example, operation 720 may be performed before operation 710.

[0102] FIG. 8 is a flowchart illustrating an exemplary process 800 for generating an index to a plurality of POIs according to some embodiments of the present disclosure. The process 800 may be executed by the AI system 100. For example, the process 800 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 800. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 800 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 8 and described below is not intended to be limiting.

[0103] In 810, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may obtain the plurality of POIs from the POI database.

[0104] In 820, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may transform the plurality of POIs to the plurality of POI vectors using the pre-generated model.

[0105] In some embodiments, each POI in the POI database may be transformed to a POI vector using the pre-generated model (e.g., DPSM model). For example, the processing engine 112 may input each POI in the POI database into the pre-generated model. The POI may be inputted into the POI model of the DPSM model. The POI model may output a POI vector responding to the inputted POI. As used herein, the processing engine 112 may use a middle output (i.e., the POI vector) of the pre-generated model. Each POI in the POI database may be transformed into a POI vector responding to the POI. Therefore, the plurality of POIs in the POI database may be transformed into the plurality of POI vectors responding to the POIs using the pre-generated model.

[0106] In 830, the processing engine 112 (e.g., the processor 220, the POI vector set obtaining module 430) may generate the index to the plurality of POIs based on the plurality of POI vectors using graphics processing unit (GPU).

[0107] In some embodiments, the processing engine 112 may use a graphics processing unit (GPU) to generate the index by mapping the plurality of POI vectors to spatial information. For example, the GPU may extract spatial information from each of the plurality of POI vectors, and generate the index by converting the extracted spatial information into a multi-dimensional vector. The index may be a multi-dimensional vector converted using a binary, hexadecimal value, or other numerical value of the spatial information. In some embodiments, the generating of the index in the GPU may be implemented offline. The generated index may be stored in the ROM 230, RAM 240, and/or other type of storage medium to be executed by the processing engine 112.

[0108] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. For example, one or more other optional operations (e.g., a storing operation) may be added elsewhere in the exemplary process 800. For example, process 800 may

further include storing the index after generating the index.

[0109] FIG. 9 is a flowchart illustrating an exemplary process 900 for determining at least one POI according to some embodiments of the present disclosure. The process 900 may be executed by the AI system 100. For example, the process 900 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 900. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 900 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 9 and described below is not intended to be limiting.

[0110] In 910, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may generate an identifier (ID) list based on the set of POI vectors. In some embodiments, the ID list may include a set of IDs each directing to a candidate POI.

[0111] In some embodiments, the set of POI vectors may include a predetermined number of POI vectors that are nearest to or most similar (or most relevant) to the query vector. Each POI vector in the set of the POI vectors may respond to a candidate POI. The candidate POI may indicate a possible target location where the user may search via the user terminal 130.

[0112] In some embodiments, the ID list may include a plurality of IDs each directing to a candidate POI. Each ID may be an identifier that indicates a candidate POI. The processing engine 112 may generate an ID for each candidate POI corresponding to each POI vector in the set of the POI vectors, and a plurality of IDs may form an ID list.

[0113] In 920, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may determine a set of similarity scores for a set of candidate POIs with respect to the set of POI vectors using a trained

quality model. Each similarity score may indicate a relevance level of a corresponding POI to the query.

[0114] In some embodiments, the similarity score may be a parameter to indicate a relevance level between a candidate POI and a query inputted by the user. The similarity score may be positive correlation with the relevance level between the candidate POI and the query, that is, the higher of the similarity score, the higher of the relevance level between the POI and the query. And the user may be more likely to retrieve the corresponding candidate POI. For example, if a similarity score is 1, and another similarity score is 3, the relevance level between the corresponding candidate POI and the inputted query with the similarity score 3 may be higher than the relevance level between the corresponding candidate POI and the inputted query with the similarity score 1, and the user may be more likely to obtain the candidate POI with the similarity score 3 than the candidate POI with the similarity score 1.

[0115] In some embodiments, the trained quality model may be used to predict a similarity score between a query and a POI. For example, the trained quality model may include a Gradient Boosted Decision Tree (GBDT) model. In some embodiments, the processing engine 112 (e.g., the processor 220, the model training module 450) may obtain a preliminary model. For example, the preliminary model may be a preliminary GBDT model. A large number of historical queries that a plurality of user inputted in the history and the corresponding historical POIs in a historical POI list that recalled in response to the historical queries may be marked manually to obtain labeled data. The labeled data may include similarity scores marked manually. The similarity score may be generated for the query and each POI in the POI list manually according to a pre-determined standard. The standard may be pre-determined manually. The processing engine 112 may input the labeled data into the preliminary GBDT model to train the preliminary GBDT model. The preliminary model may be trained to obtain the trained quality model until the preliminary model is convergent in the training process.

[0116] In some embodiments, the processing engine 112 may input the query and each candidate POI responding to vector in the set of the POI vectors into the

trained quality model, the trained quality model may output the similarity score between the query and the corresponding candidate POI. Thereby, the processing engine 112 may determine a set of similarity scores for the set of candidate POIs with respect to the set of POI vectors.

[0117] In 930, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may rank the set of similarity scores of the set of candidate POIs.

[0118] In some embodiments, the processing engine 112 may rank the similarity scores in the set of similarity scores according to a predetermined rule. For example, the processing engine 112 may rank the set of similarity scores according to an ascending order, a descending order, or the like, or any combination thereof. As another example, the processing engine 112 may first determine whether each similarity score is greater than a score threshold. The score threshold may be set manually or be determined by the processing engine 112. In response to a determination that the corresponding similarity score is not greater than the score threshold, the processing engine 112 may delete the corresponding POI, and in response to a determination that the corresponding similarity score is greater than the score threshold, the processing engine 112 may input the corresponding POI into a ranking model to rank the corresponding candidate POIs. The detailed description of the ranking process may be found elsewhere in the present disclosure (e.g., FIG. 10, and the description thereof).

[0119] In 940, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may determine the at least one POI from the set of candidate POIs based on the ranking of the set of similarity scores.

[0120] In some embodiments, the processing engine 112 may select a predetermined number of POIs from the candidate POIs that rank at top of the ranking of the set of similarity scores, and recommend the predetermined number of POIs to the user via a data exchange port that is communicatively connected to a network 120. The predetermined number may be a default number of the system or may be a number determined according to different conditions. For example, the

processing engine 112 may select top 5 POIs in ranking of set of similarity scores corresponding to the set of the POI in the descending order. The processing engine 112 may recommend the 5 POIs to the user via the user terminal.

[0121] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. In some embodiments, one or more other optional operations (e.g., a storing operation) may be added elsewhere in the exemplary process 900. For example, process 900 may further include storing the at least one POI after determining the at least one POI.

[0122] FIG. 10 is a flowchart illustrating an exemplary process 1000 for ranking a set of similarity scores of a set of candidate POIs according to some embodiments of the present disclosure. The process 1000 may be executed by the AI system 100. For example, the process 1000 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 1000. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 1000 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 10 and described below is not intended to be limiting.

[0123] In 1010, for each of the set of candidate POIs, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may determine whether the corresponding similarity score is greater than a score threshold.

[0124] In some embodiments, the score threshold may be used to determine whether the relevance level of the candidate POI with the query meet a requirement.

For example, in response to a similarity score of a candidate POI in the set of the candidate POIs is higher than the score threshold, the processing engine 112 may determine that the relevance level of the candidate POI to the query meets the requirement. In response to the similarity score of the candidate POI is less than the score threshold, the processing engine 112 may determine that the relevance level of the candidate POI with the query may not meet the requirement. In some embodiments, the score threshold may be a default value stored in a storage device (e.g., the storage 140, the ROM 230 the RAM 240, and/or other type of storage medium to be executed by the processing engine 112), or may be dynamically determined according to different conditions.

[0125] In 1020, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may obtain one or more relevant candidate POIs with similarity scores greater than the score threshold.

[0126] In some embodiments, the processing engine may select one or more candidate POIs in set of candidate POIs with similarity scores that are greater than the score threshold as relevant candidate POIs. For example, the score threshold may be set as 1 by the processing engine 112, the candidate POIs with the similarity score less than 1 may be deleted by the processing engine 112, and the candidate POIs with the similarity score greater than 1 may be selected by the processing engine 112 as the relevant candidate POIs.

[0127] In 1030, the processing engine 112 (e.g., the processor 220, the recommended POI determining module 440) may rank the one or more relevant candidate POIs based on a trained ranking model.

[0128] In some embodiments, the processing engine 112 may input the one or more relevant candidate POIs into a trained ranking model. The ranking model may rank the one or more relevant candidate POIs and output a ranking order. The trained ranking model may be used to predict a relevance ranking order of a plurality of relevant candidate POIs related to the query. The trained ranking model may be pre-trained and stored in a storage device (e.g., the storage 140, the ROM 230 the RAM 240, and/or other type of storage medium to be executed by the processing

engine 112), or may be online trained by the processing engine 112. For example, the processing engine 112 may train the ranking model according to a Learn to Rank method.

[0129] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure.

[0130] FIG. 11 is a schematic diagram illustrating an exemplary process 1100 for determining at least one POI according to some embodiments of the present disclosure. The process 1100 may be executed by the AI system 100. For example, the process 1100 may be implemented as a set of instructions (e.g., an application) stored in the storage ROM 230 or RAM 240. The processor 220 may execute the set of instructions, and when executing the instructions, it may be configured to perform the process 1100. The operations of the illustrated process presented below are intended to be illustrative. In some embodiments, the process 1100 may be accomplished with one or more additional operations not described and/or without one or more of the operations discussed. Additionally, the order in which the operations of the process as illustrated in FIG. 11 and described below is not intended to be limiting.

[0131] As shown in FIG. 11, in the dashed box, the processing engine 112 may train a DPSM model. The processing engine 112 may obtain a plurality of the history query and a plurality of corresponding history POIs in a user log. The history query and the corresponding history POI may form a historical query-POI pair. The processing engine 112 may parse the historical query-POI pair into one or more segments, and generate one or more text denotations for the historical query-POI pair based on the parsing. The processing engine 112 may train a preliminary DPSM model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method, and obtain the DPSM model. In some embodiments, the training process may be performed by the processing

engine 112 offline.

[0132] When a user wants to take an online to offline service, the user may send a user request by inputting a query into a search box of an application of the online to offline service via the user terminal 130. The processing engine 112 may obtain the query and/or a city ID from the user terminal 130 based on the user request. The processing engine 112 may invoke the trained DPSM model trained in the dashed box using a Tensor Flow serving (TF_SVR), and input the query into the trained DPSM model. The trained DPSM model may transform the query to a query vector. The processing engine 112 may obtain a POI database, and input each POI in the POI database into the trained DPPSM model to obtain a plurality of POI vectors. Then the processing engine 112 may generate an index based on the plurality of POI vectors and the GPU. The processing engine 112 may input the query, the city ID, and the index into the search library (e.g., Faiss), and perform a similarity search to retrieve at least one candidate POI vector that is similar to the query using the PQ method.

[0133] The processing engine 112 may generate an ID for each similar POI vector of the at least one similar POI vector to obtain an ID list. The ID list may include a set of IDs each directing to a candidate POI. The processing engine 112 may test a set of candidate POIs with respect to the set of POI vectors to test a relevance level between each candidate POI and the query. For example, the processing engine 112 may determine a set of similarity scores for a set of candidate POIs with respect to the set of POI vectors using a trained quality model. Each similarity score may indicate a relevance level of a corresponding candidate POI to the query. The processing engine 112 may rank the set of similarity scores of the set of candidate POIs using a trained ranking model, and select at least one POI from the set of candidate POIs based on the ranking of the set of similarity scores. Finally, the processing engine 112 may display the ranking results to the user by the user terminal. The processing engine 112 may recommend the selected at least one POI to the user terminal 130 via the data exchange port, and the user may select a POI from the at least one displayed POI responding to the query.

[0134] Having thus described the basic concepts, it may be rather apparent to those skilled in the art after reading this detailed disclosure that the foregoing detailed disclosure is intended to be presented by way of example only and is not limiting.

Various alterations, improvements, and modifications may occur and are intended to those skilled in the art, though not expressly stated herein. These alterations, improvements, and modifications are intended to be suggested by this disclosure, and are within the spirit and scope of the exemplary embodiments of this disclosure.

[0135] Moreover, certain terminology has been used to describe embodiments of the present disclosure. For example, the terms “one embodiment,” “an embodiment,” and/or “some embodiments” mean that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present disclosure. Therefore, it is emphasized and should be appreciated that two or more references to “an embodiment,” “one embodiment,” or “an alternative embodiment” in various portions of this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures or characteristics may be combined as suitable in one or more embodiments of the present disclosure.

[0136] Further, it will be appreciated by one skilled in the art, aspects of the present disclosure may be illustrated and described herein in any of a number of patentable classes or context including any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof. Accordingly, aspects of the present disclosure may be implemented entirely hardware, entirely software (including firmware, resident software, micro-code, etc.) or combining software and hardware implementation that may all generally be referred to herein as a “block,” “module,” “engine,” “unit,” “component,” or “system.” Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable media having computer readable program code embodied thereon.

[0137] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband

or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including electro-magnetic, optical, or the like, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that may communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device. Program code embodied on a computer readable signal medium may be transmitted using any appropriate medium, including wireless, wireline, optical fiber cable, RF, or the like, or any suitable combination of the foregoing.

[0138] Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Scala, Smalltalk, Eiffel, JADE, Emerald, C++, C#, VB. NET, Python or the like, conventional procedural programming languages, such as the "C" programming language, Visual Basic, Fortran 1703, Perl, COBOL 1702, PHP, ABAP, dynamic programming languages such as Python, Ruby and Groovy, or other programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider) or in a cloud computing environment or offered as a service such as a software as a service (SaaS).

[0139] Furthermore, the recited order of processing elements or sequences, or the use of numbers, letters, or other designations, therefore, is not intended to limit the claimed processes and methods to any order except as may be specified in the claims. Although the above disclosure discusses through various examples what is currently considered to be a variety of useful embodiments of the disclosure, it is to

be understood that such detail is solely for that purpose, and that the appended claims are not limited to the disclosed embodiments, but, on the contrary, are intended to cover modifications and equivalent arrangements that are within the spirit and scope of the disclosed embodiments. For example, although the implementation of various components described above may be embodied in a hardware device, it may also be implemented as a software-only solution—e.g., an installation on an existing server or mobile device.

[0140] Similarly, it should be appreciated that in the foregoing description of embodiments of the present disclosure, various features are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure aiding in the understanding of one or more of the various embodiments. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed subject matter requires more features than are expressly recited in each claim. Rather, claimed subject matter may lie in less than all features of a single foregoing disclosed embodiment.

[0141] In some embodiments, the numbers expressing quantities or properties used to describe and claim certain embodiments of the application are to be understood as being modified in some instances by the term “about,” “approximate,” or “substantially.” For example, “about,” “approximate,” or “substantially” may indicate $\pm 20\%$ variation of the value it describes, unless otherwise stated. Accordingly, in some embodiments, the numerical parameters set forth in the written description and attached claims are approximations that may vary depending upon the desired properties sought to be obtained by a particular embodiment. In some embodiments, the numerical parameters should be construed in light of the number of reported significant digits and by applying ordinary rounding techniques.

Notwithstanding that the numerical ranges and parameters setting forth the broad scope of some embodiments of the application are approximations, the numerical values set forth in the specific examples are reported as precisely as practicable.

[0142] Each of the patents, patent applications, publications of patent applications, and other material, such as articles, books, specifications, publications, documents,

things, and/or the like, referenced herein is hereby incorporated herein by this reference in its entirety for all purposes, excepting any prosecution file history associated with same, any of same that is inconsistent with or in conflict with the present document, or any of same that may have a limiting affect as to the broadest scope of the claims now or later associated with the present document. By way of example, should there be any inconsistency or conflict between the descriptions, definition, and/or the use of a term associated with any of the incorporated material and that associated with the present document, the description, definition, and/or the use of the term in the present document shall prevail.

[0143] In closing, it is to be understood that the embodiments of the application disclosed herein are illustrative of the principles of the embodiments of the application. Other modifications that may be employed may be within the scope of the application. Thus, by way of example, but not of limitation, alternative configurations of the embodiments of the application may be utilized in accordance with the teachings herein. Accordingly, embodiments of the present application are not limited to that precisely as shown and describe.

WE CLAIM:

1. An artificial intelligent system for semantic-based search, comprising:
 - at least one storage medium including a set of instructions;
 - a data exchange port communicatively connected to a network; and
 - at least one processor in communication with the storage medium, wherein when executing the set of instructions, the at least one processor is directed to:
 - obtain a query from a user terminal via the data exchange port;
 - transform the query to a query vector using a pre-generated model;
 - retrieve a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and
 - determine at least one POI based on the set of POI vectors to recommend to a user in response to the query.
2. The system of claim 1, wherein the pre-generated model is trained based on correlations between query semantics and POIs by performing:
 - obtaining a preliminary model;
 - obtaining a plurality of historical query-POI pairs, each of the plurality of historical query-POI pairs including a historical query and a corresponding historical POI;
 - for each of the plurality of historical query-POI pairs,
 - parsing at least one of the historical query or the corresponding historical POI into one or more segments, and
 - generating one or more text denotations for the historical query-POI pair based on the parsing; and
 - training the preliminary model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method to obtain the pre-generated model.
3. The system of claim 2, wherein the pre-generated model includes a deep POI semantic model (DPSM) model.

4. The system of any one of claims 1-3, wherein to retrieve the set of POI vectors from a search library based on the query vector using a Product Quantization (PQ) method, the at least one processor is further directed to:

obtain an index to a plurality of POIs in a POI database;

obtain at least one target location from the query; and

determine the set of POI vectors from the search library based on the index, the query vector and the at least one target location, the set of POI vectors being relevant to the query.

5. The system of claim 4, wherein to obtain an index to a plurality of POIs in a POI database, the at least one processor is further directed to:

obtain the plurality of POIs from the POI database;

transform the plurality of POIs to the plurality of POI vectors using the pre-generated model; and

generate the index to the plurality of POIs based on the plurality of POI vectors using graphics processing unit (GPU).

6. The system of any one of claims 1-5, wherein to determine the at least one POI based on the set of POI vectors to recommend to a user in response to the query, the at least one processor is further directed to:

generate an identifier (ID) list based on the set of POI vectors, the ID list including a set of IDs each directing to a candidate POI;

determine a set of similarity scores for a set of candidate POIs with respect to the set of POI vectors using a trained quality model, each similarity score indicating a relevance level of a corresponding POI to the query;

rank the set of similarity scores of the set of candidate POIs; and

determine the at least one POI from the set of candidate POIs based on the ranking of the set of similarity scores.

7. The system of claim 6, wherein the trained quality model is a Gradient Boosted Decision Tree (GBDT) model.

8. The system of claim 6, wherein to rank the set of similarity scores of the set of candidate POIs, the at least one processor is further directed to:

for each of the set of candidate POIs, determine whether the corresponding similarity score is greater than a score threshold;

obtain one or more relevant candidate POIs with similarity scores greater than the score threshold; and

rank the one or more relevant candidate POIs based on a trained ranking model.

9. The system of claim 8, wherein the trained ranking model is obtained by a learning to rank method.

10. An artificial intelligent method for semantic-based search implemented on a computing device including at least one storage medium including a set of instructions, a data exchange port communicatively connected to a network, and at least one processor in communication with the storage medium, the method comprising:

obtaining a query from a user terminal via the data exchange port;

transforming the query to a query vector using a pre-generated model;

retrieving a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and

determining at least one POI based on the set of POI vectors to recommend to a user in response to the query.

11. The method of claim 10, wherein the pre-generated model is trained based on correlations between query semantics and POIs by performing:

obtaining a preliminary model;

obtaining a plurality of historical query-POI pairs, each of the plurality of historical query-POI pairs including a historical query and a corresponding historical POI;

for each of the plurality of historical query-POI pairs,

parsing at least one of the historical query or the corresponding historical POI into one or more segments, and

generating one or more text denotations for the historical query-POI pair based on the parsing; and

training the preliminary model based on the text denotations for the plurality of historical query-POI pairs using a deep learning method to obtain the pre-generated model.

12. The method of claim 11, wherein the pre-generated model includes a deep POI semantic model (DPSM) model.

13. The method of any one of claims 10-12, wherein the retrieving of the set of POI vectors from a search library based on the query vector using a Product Quantization (PQ) further includes:

obtaining an index to a plurality of POIs in a POI database;

obtaining at least one target location from the query; and

determining the set of POI vectors from the search library based on the index, the query vector and the at least one target location, the set of POI vectors being relevant to the query.

14. The method of claim 13, wherein the obtaining of an index to a plurality of POIs in a POI database further includes:

obtaining the plurality of POIs from the POI database;

transforming the plurality of POIs to the plurality of POI vectors using the pre-generated model; and

generating the index to the plurality of POIs based on the plurality of POI vectors using graphics processing unit (GPU).

15. The method of any one of claims 10-14, wherein the determining of the at least one POI based on the set of POI vectors to recommend to a user in response to the query further includes:

generating an identifier (ID) list based on the set of POI vectors, the ID list including a set of IDs each directing to a candidate POI;

determining a set of similarity scores for a set of candidate POIs with respect to the set of POI vectors using a trained quality model, each similarity score indicating a relevance level of a corresponding POI to the query;

ranking the set of similarity scores of the set of candidate POIs; and

determining the at least one POI from the set of candidate POIs based on the ranking of the set of similarity scores.

16. The method of claim 15, wherein the trained quality model is a Gradient Boosted Decision Tree (GBDT) model.

17. The method of claim 15, wherein the ranking of the set of similarity scores of the set of candidate POIs further includes:

for each of the set of candidate POIs, determining whether the corresponding similarity score is greater than a score threshold;

obtaining one or more relevant candidate POIs with similarity scores greater than the score threshold; and

ranking the one or more relevant candidate POIs based on a trained ranking model.

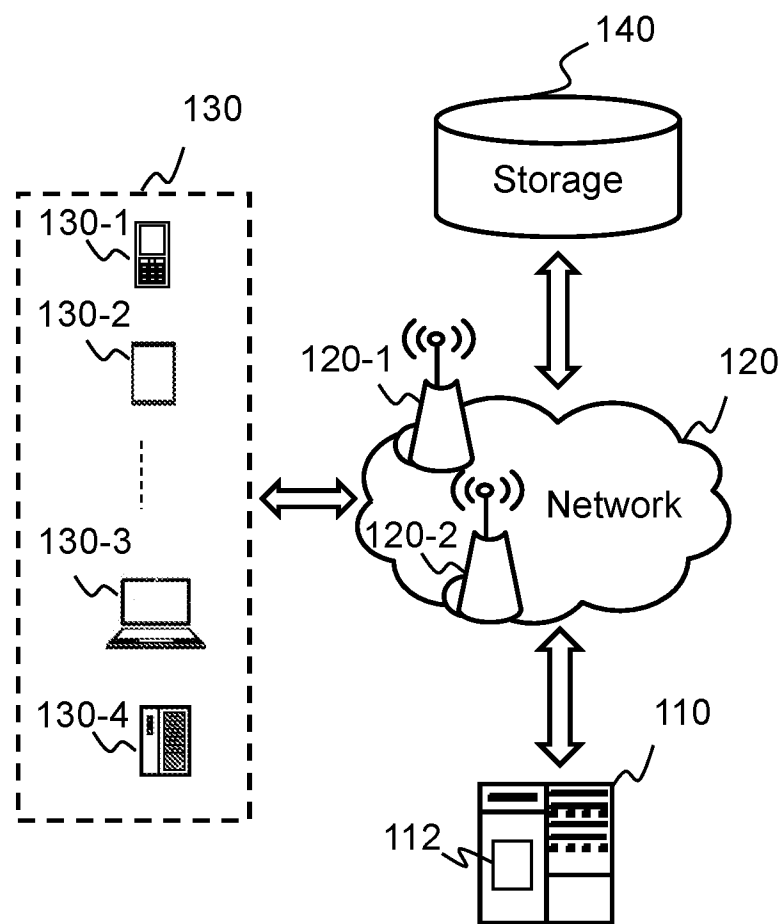
18. The method of claim 17, wherein the trained ranking model is obtained by a learning to rank method.

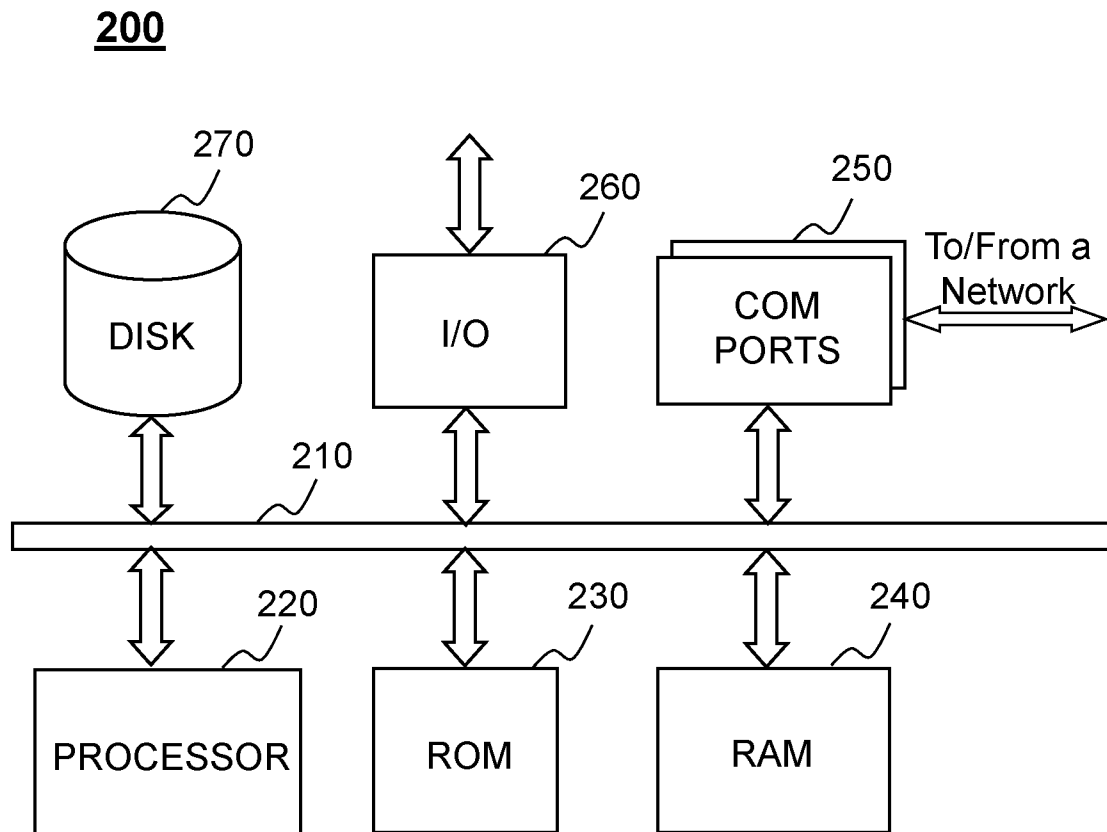
19. A non-transitory readable medium, comprising at least one set of instructions for semantic-based search, wherein when executed by at least one processor of an electrical device, the at least one set of instructions directs the at least one processor to perform a method, the method comprising:

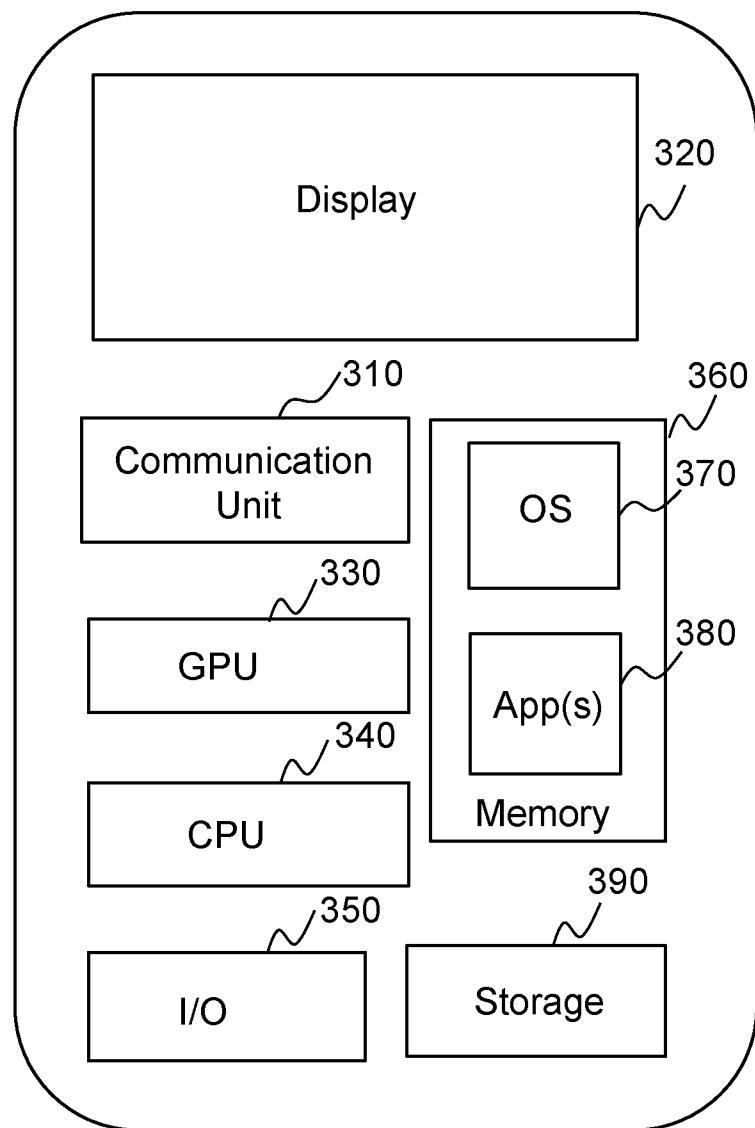
- obtaining a query from a user terminal;
- transforming the query to a query vector using a pre-generated model;
- retrieving a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and
- determining at least one POI based on the set of POI vectors to recommend to a user in response to the query.

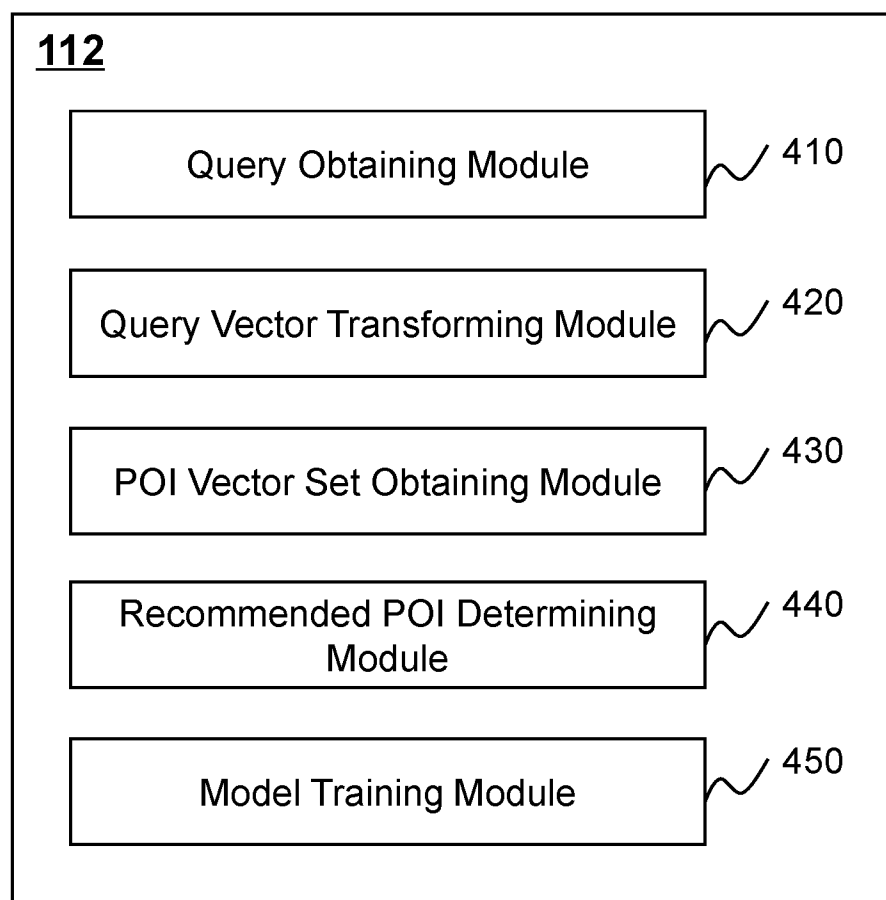
20. A system for semantic-based search, comprising:

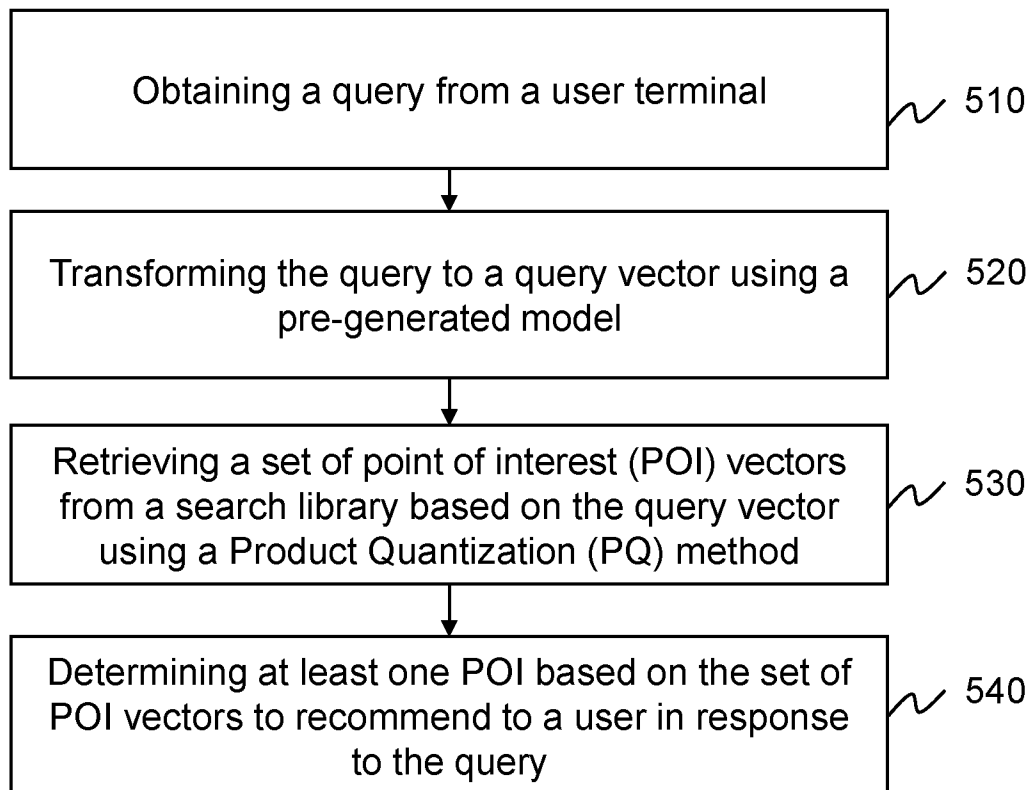
- a query obtaining module configured to obtain a query from a user terminal;
- a query vector transforming module configured to transform the query to a query vector using a pre-generated model;
- a POI vector set obtaining module configured to retrieve a set of point of interest (POI) vectors from a search library based on the query vector using a Product Quantization (PQ) method; and
- a recommended POI determining module configured to determine at least one POI based on the set of POI vectors to recommend to a user in response to the query.

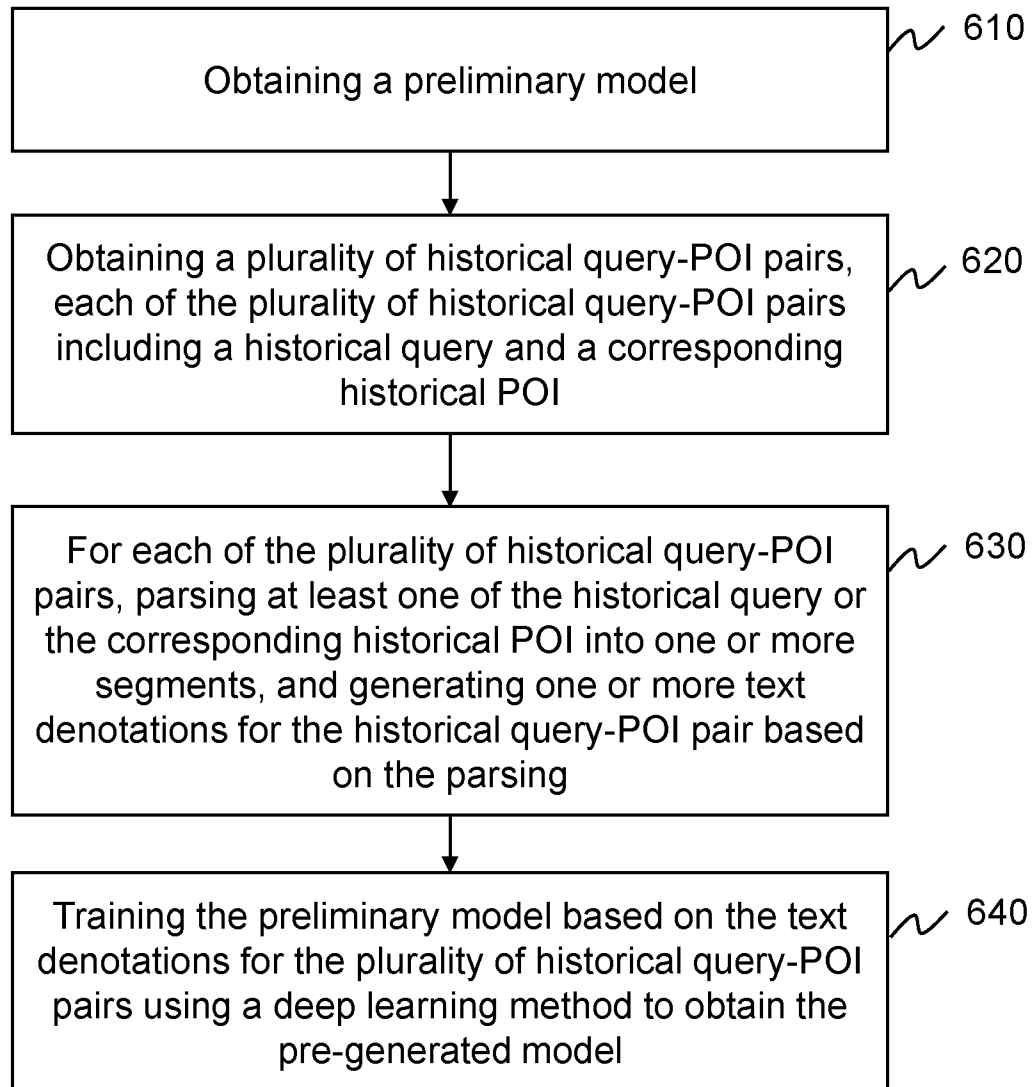
100**FIG. 1**

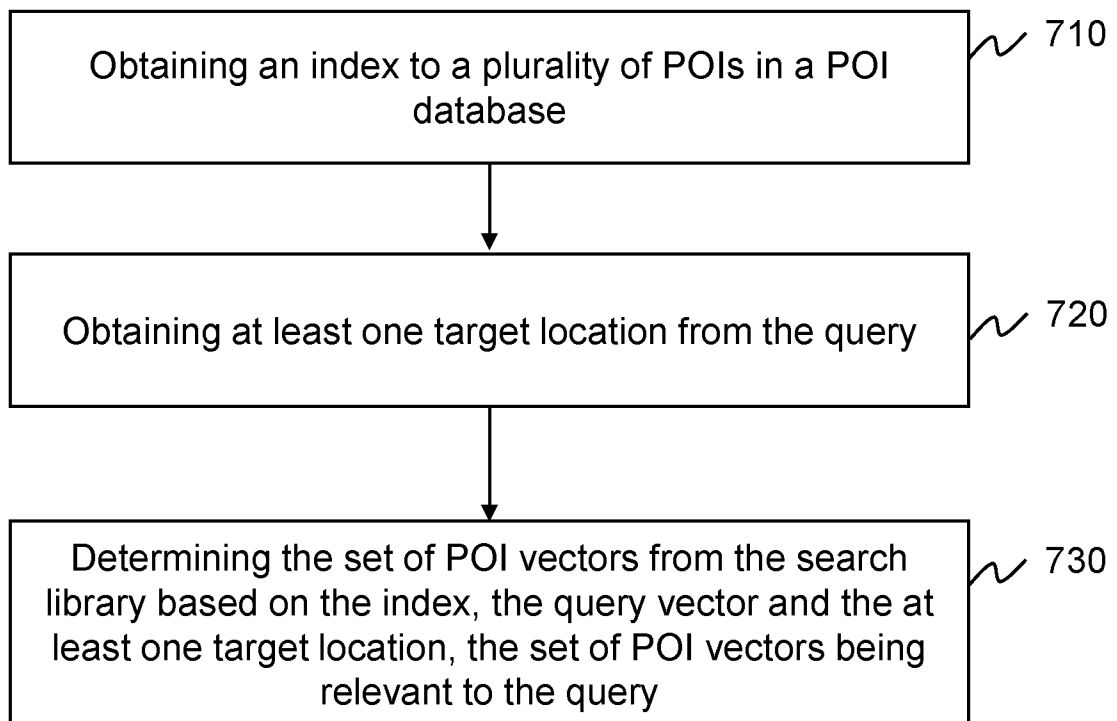
**FIG. 2**

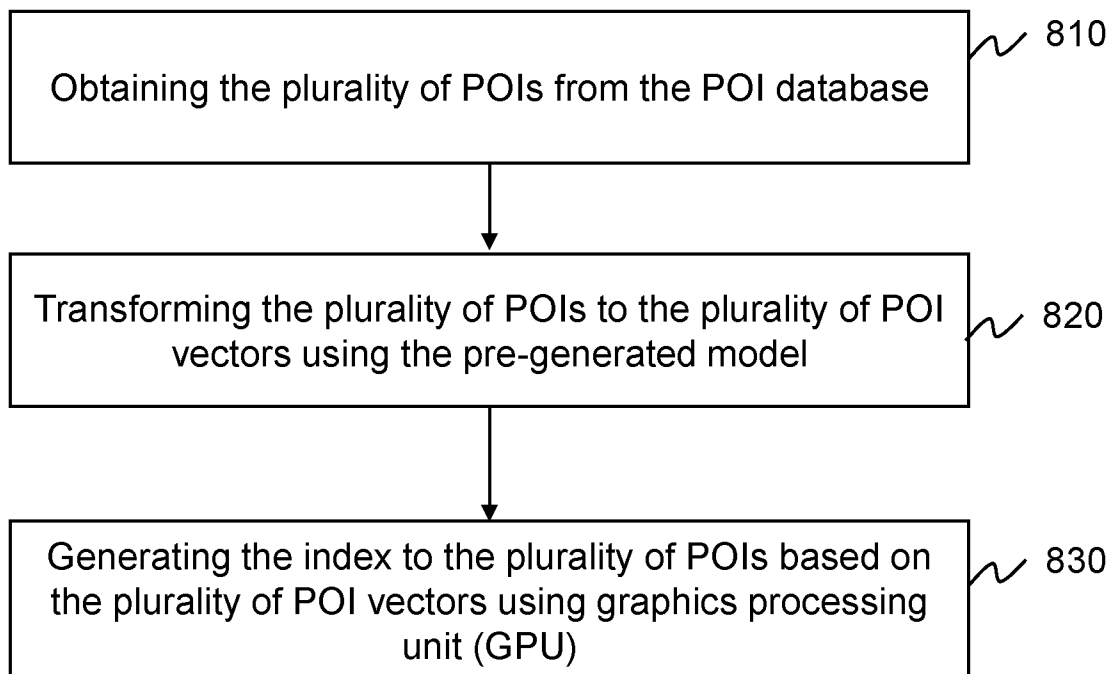
300**FIG. 3**

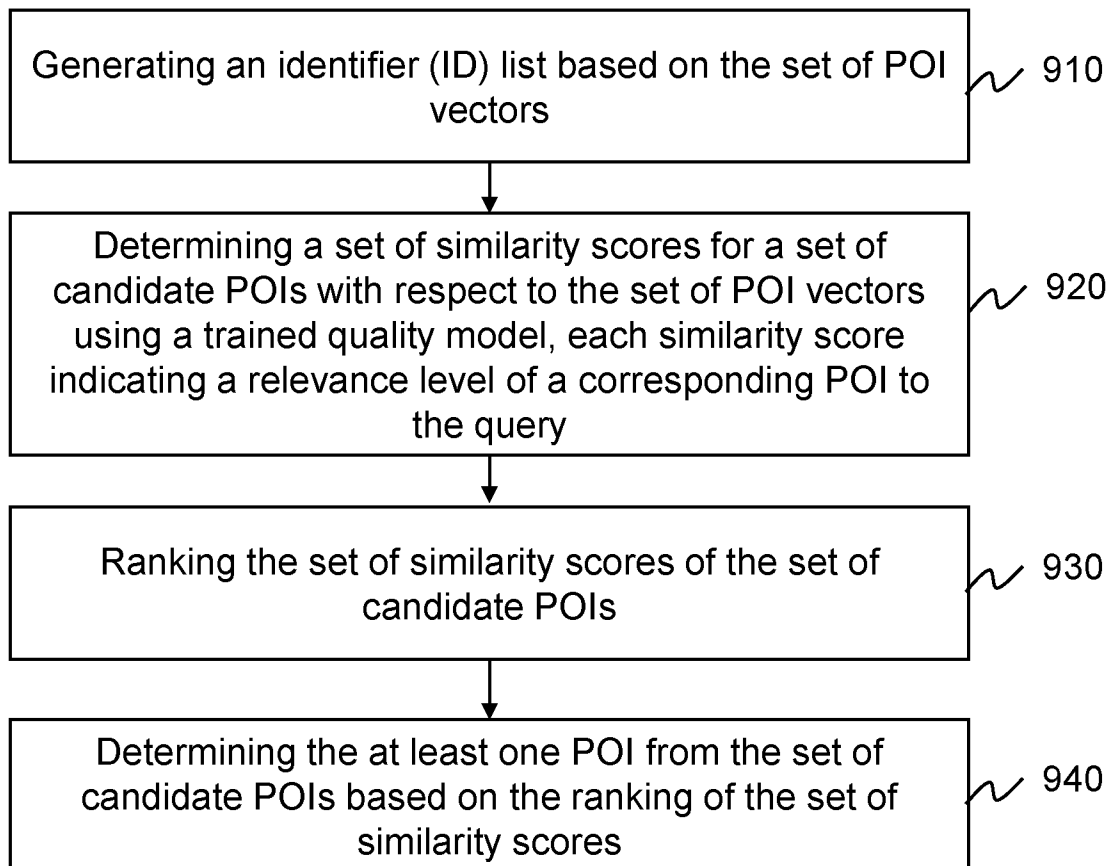
**FIG. 4**

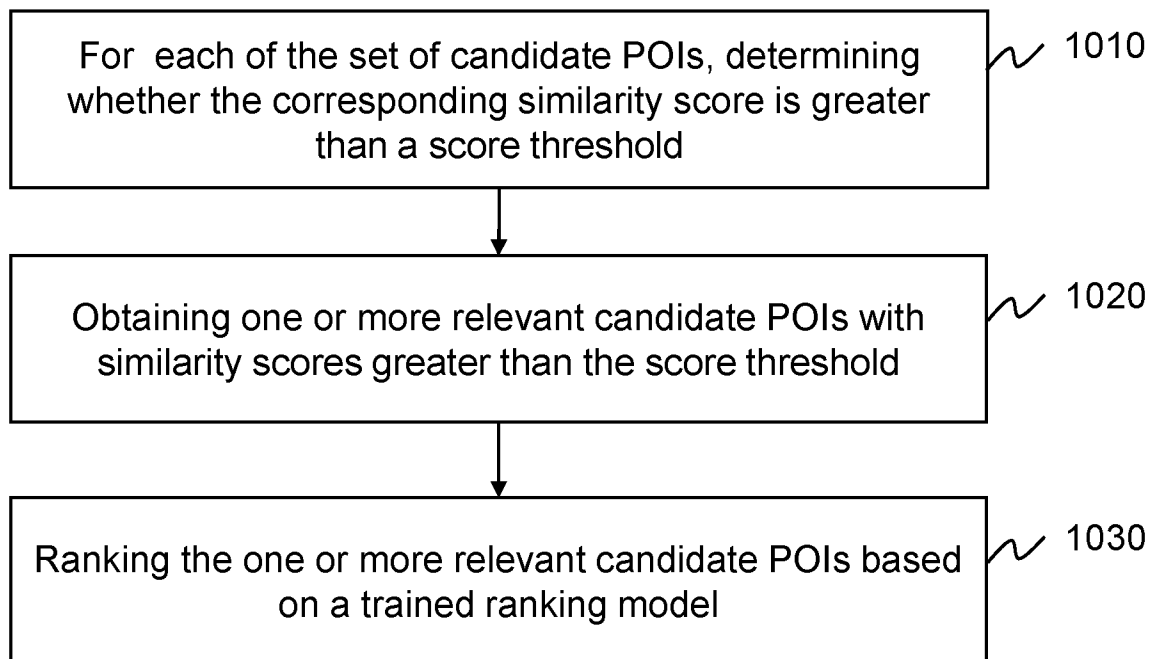
500**FIG. 5**

600**FIG. 6**

700**FIG. 7**

800**FIG. 8**

900**FIG. 9**

1000**FIG. 10**

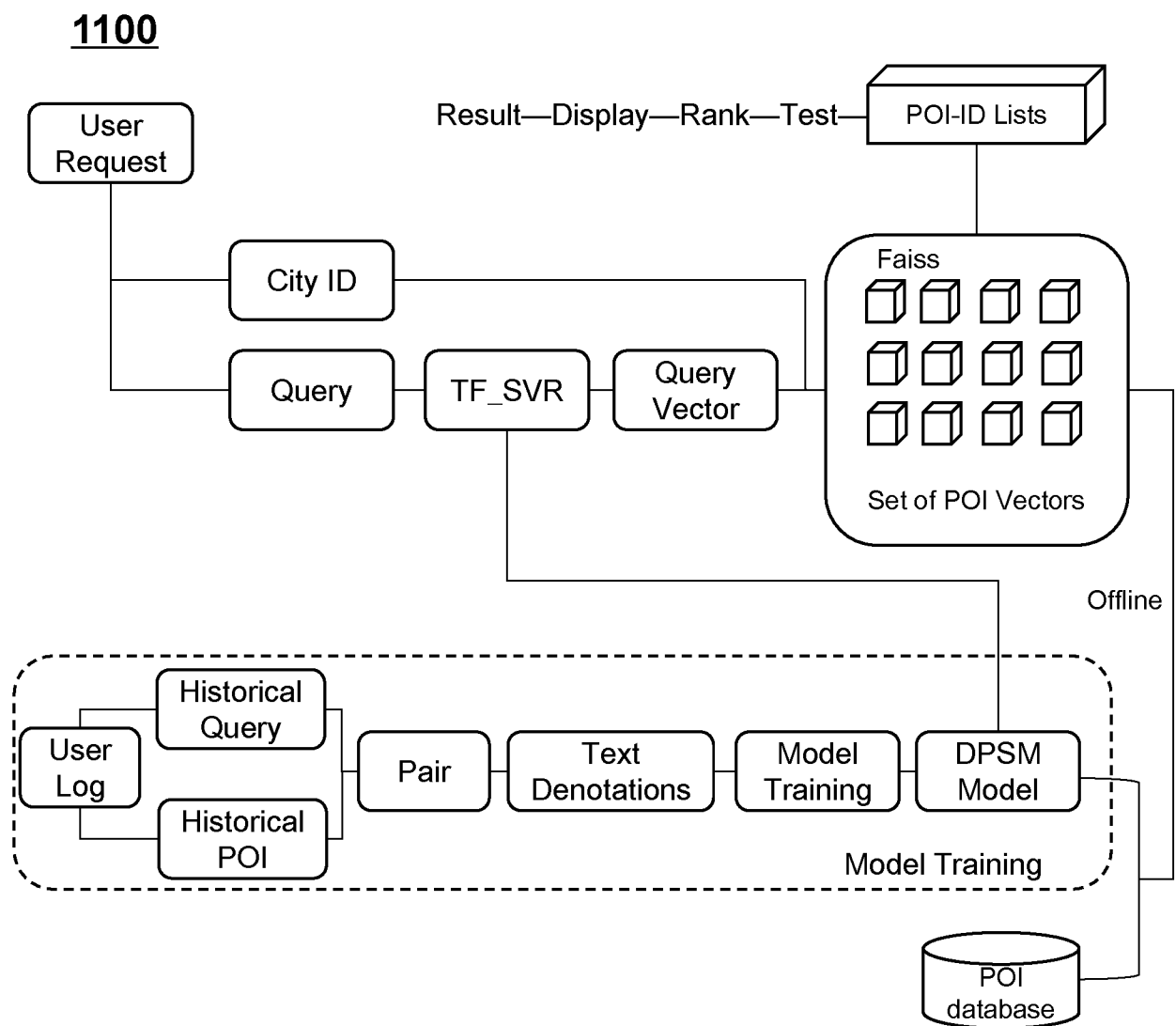


FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/125993

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT,WPI,EPODOC,IEEE,CNKI,GOOGLE: semantic+, search+, artificial, intelligent, AI, POI, vector, recommend+, deep w neural

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 105279288 A (SHENZHEN UNIVERSITY) 27 January 2016 (2016-01-27) description, paragraphs[0006]-[0012], [0041], [0050], [0058]	1-20
Y	CN 108733669 A (YOULU BEIJING INFORMATION TECHNOLOGY CO., LTD.) 02 November 2018 (2018-11-02) description, paragraphs[0023], [0061]-[0065], [0076], [0084]-[0088]	1-20
A	KR 20180001155 A (HUMANWORKS) 04 January 2018 (2018-01-04) the whole document	1-20
A	CN 106095983 A (BEIJING BAIDU NETCOM SCI. & TECHNOLOGY CO., LTD.) 09 November 2016 (2016-11-09) the whole document	1-20
A	CN 106528676 A (BEIJING BAIDU NETCOM SCI. & TECHNOLOGY CO., LTD.) 22 March 2017 (2017-03-22) the whole document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 September 2019

Date of mailing of the international search report

26 September 2019

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088
China

Authorized officer

GAO,Dandan

Facsimile No. (86-10)62019451

Telephone No. 86-(10)-53961301

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/125993

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105279288	A	27 January 2016	None	
CN	108733669	A	02 November 2018	None	
KR	20180001155	A	04 January 2018	None	
CN	106095983	A	09 November 2016	None	
CN	106528676	A	22 March 2017	None	