

Description

Title of Invention: METHOD, DEVICE AND SYSTEM FOR RETRIEVING RELEVANT POINT-OF-INTEREST DATA IN RESPONSE TO A SEARCH REQUEST TECHNICAL FIELD

[0001] Various aspects of this disclosure relate to methods, devices, and systems for retrieving relevant point-of-interest data in response to search requests.

BACKGROUND

[0002] The following discussion of the background is intended to facilitate an understanding of the present disclosure only. It should be appreciated that the discussion is not an acknowledgement or admission that any of the material referred to was published, known or is part of the common general knowledge of the person skilled in the art in any jurisdiction as of the priority date of the disclosure.

[0003] In response to a user query for e-commerce transactions, such as transport services, searching for POIs (Point-Of-Interest) in relation to the user query is usually carried out. However, POI searches in response to user queries have become increasingly challenging due to the exponential growth in POI corpus data sizes.

[0004] A conventional POI search system may comprise two functions: a candidate generation function and a ranker function. The candidate generation function aims at generating a set of POI candidates in response to a search query. The ranker function ranks or re-ranks the retrieved POI candidates to be displayed to the end-users. A known method for the POI candidate generation is the use of a search engine, such as the Elasticsearch (ES) system, which utilizes an inverted index structure to retrieve the POIs with high string similarity scores.

[0005] However, the exponential increase in the number of POIs may cause a relatively heavy computation load for the retrieval stage, resulting in a big increment of overall latency for the entire POI search system. Such increment of overall latency may adversely impact response to search queries.

[0006] Accordingly, there exists a need for an improved POI retrieval system, method, and/or device.

SUMMARY

[0007] The technical solution seeks to provide a method, device, and/or system for retrieving relevant point-of-interest (POI) data based on the selection of one or more POI candidates in response to a search request. Particularly, a category-based method for fast retrieval of relevant POI during POI searching process is disclosed.

The technical solution seeks to improve search latency, which is a crucial factor that could affect consumers' expectations. The inventors have discovered that about 100 milliseconds (ms) latency improvement in responsiveness could result in about 10.1% increment in travel conversions.

- [0008] A category-based method for retrieving one or more relevant point-of-interest data in response to a search request is proposed. The method may be implemented in e-commerce systems, user devices, and servers for facilitating the provision of services, such as transport on-demand services. In some embodiments, categories associated with POI data may be predicted by one or more category prediction modules, and the predicted categories may be used to reduce the search space based on a search engine or system, such as a distributed search and analytics engine. In some embodiment, the distributed search and analytics engine may be an Elasticsearch engine (ES). Advantageously, the use of predicted categories to reduce the search space can also reduce ES round-trip time (RTT) greatly while improving the POI search recall simultaneously.
- [0009] In one aspect of the disclosure, there is provided a method for retrieving relevant point-of-interest data in response to a search request, the method comprising the steps of: associating at least one category to the search request; retrieving, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and determining the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on comparing the search request with each of the plurality of point-of-interest candidate, and calculating a corresponding similarity score for each point-of-interest candidate.
- [0010] In some embodiments, the method further comprises ranking each of the plurality of point-of-interest candidates, based on the corresponding similarity score for each point-of-interest candidate.
- [0011] In some embodiments, the database comprises pre-categorized point-of-interest data, and wherein each entry of the database comprises a point-of-interest and at least one category.
- [0012] In some embodiments, the associating at least one category to the search request comprises parsing the search request; predicting, using a category prediction and detection module, one or more categories based on the parsed search request, and indexing the one or more categories using the category prediction and detection module.
- [0013] In some embodiments, parsing the search request comprises correlating the search request with one or more point-of-interest data within the database.
- [0014] In some embodiments, the correlating includes predicting a relationship between the search request and the one or more point-of-interest data.

- [0015] In some embodiments, the indexing comprises generating at least one index of a keyword type.
- [0016] In some embodiments, the method further comprises generating a category structure for the one or more categories, wherein the category structure comprises a three-level hierarchical category tree.
- [0017] In some embodiments, the predicting of one or more categories based on the parsed search request includes training the category prediction and detection module using a set of POI training data and a large language model.
- [0018] In some embodiments, the training further comprises adopting a natural language processing (NLP) model and a masked language modeling model.
- [0019] In some embodiments, the method further comprises applying a finetuning model, the finetuning model comprises a first finetuning step of generating weak-supervision data by mapping whitelisted categories to a first set of POI data, and a second finetuning step of collecting a predetermined number of second set of POI data and labelling the collected POI data in the second set of POI data based on the whitelisted categories.
- [0020] In some embodiments, the category prediction and detection module comprises a distributed search and analytics engine.
- [0021] According to another aspect of the disclosure there is provided a server apparatus, the server apparatus configured to retrieve one or more relevant point-of-interest data in response to a search request, the server apparatus comprises at least one processor, the at least one processor configured to: associate at least one category to the search request; retrieve, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and determine the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on a comparison of the search request with each of the plurality of point-of-interest candidate, and calculate a corresponding similarity score for each point-of-interest candidate.
- [0022] In some embodiments, the at least one processor is configured to rank each of the plurality of point-of-interest candidates, based on the corresponding similarity score for each point-of-interest candidate.
- [0023] In some embodiments, the database comprises pre-categorized point-of-interest data, and wherein each entry of the database comprises a point-of-interest and at least one category.
- [0024] In some embodiments, in associating at least one category to the search request, the at least one processor is configured to: parse the search request; predict, using a category prediction and detection module, one or more categories based on the parsed search request, and index the one or more categories using the category and detection prediction module.

- [0025] In some embodiments, in parsing the search request, the at least one processor is configured to correlate the search request with one or more point-of-interest data within the database.
- [0026] In some embodiments, the at least one processor is configured to predict a relationship between the search request and the one or more point-of-interest data.
- [0027] In some embodiments, in indexing the one or more categories, the at least one processor is configured to generate at least one index of a keyword type.
- [0028] In some embodiments, the at least one processor is configured to generate a category structure for the one or more categories, wherein the category structure comprises a three-level hierarchical category tree.
- [0029] In some embodiments, in the prediction of one or more categories based on the parsed search request, the at least one processor is configured to train the category prediction and detection module using a set of POI training data and a large language model.
- [0030] In some embodiments, the at least one processor is further configured to adopt a natural language processing (NLP) model and a masked language modeling model in training the category prediction and detection module.
- [0031] In some embodiments, the at least one processor is further configured to apply a finetuning model as follows: generate weak-supervision data by mapping whitelisted categories to a first set of POI data, and collect a predetermined number of second set of POI data and label the collected POI data in the second set of POI data based on the whitelisted categories.
- [0032] In some embodiments, the category prediction and detection module comprises a distributed search and analytics engine.
- [0033] According to another aspect of the disclosure there is provided a system for providing an on-demand transport service to a user, the system comprising a user device configured to receive a search request for point-of-interest data, the point-of-interest data associated with at least one of a pick-up point and a drop-off point; at least one processor configured to: associate at least one category to the search request; retrieve, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and determine the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on a comparison of the search request with each of the plurality of point-of-interest candidate, and calculate a corresponding similarity score for each point-of-interest candidate.
- [0034] According to another aspect of the disclosure there is provided a non-transitory computer-readable storage medium comprising instructions, which, when executed by one or more processors, cause the execution of the method for building a dataset for automatic speech recognition according to any one of described methods.

[0035] According to another aspect of the disclosure there is provided a data processing device configured to perform the method of any one of the described methods.

[0036] According to another aspect of the disclosure there is provided a computer executable code comprising instructions building a dataset for automatic speech recognition according to any one of the described methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0037] The disclosure will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

[0038] - FIG. 1 is a schematic flowchart of a method for retrieving one or more relevant point-of-interest data in response to a search request in accordance with various embodiments;

[0039] - FIG. 2A is a schematic block diagram of a system for retrieving one or more relevant point-of-interest data in response to a search request in accordance with various embodiments;

[0040] - FIG. 2B shows an exemplary database entry comprising a POI information and associated categories;

[0041] - FIG. 3 is a schematic diagram of an embodiment of a category-based POI candidate retrieval data pipeline;

[0042] - FIG. 4 is a schematic diagram of an embodiment of finetuning of a POI category prediction module;

[0043] - FIG. 5 shows an example of a search request and POIs used in the training for the prediction of one or more associated categories based on the method of the present disclosure;

[0044] - FIG. 6A shows an example of an implementation in JavaScript Object Notation (JSON) , presenting an example of a category-based ES search template construction;

[0045] - FIG. 6B shows an example illustrating the potentiality of Category based retrieval to boost POI search performance and relevancy;

[0046] - FIG. 7A is a table showing a comparison of the top-k accuracy performance measure obtained using a naïve finetuning strategy, a two-step finetuning strategy, and an active learning strategy on the POI category prediction module;

[0047] - FIG. 7B is a table showing the online A/B experiments that lasted for 2 weeks in various countries indicating search performance difference between control and treatment data groups;

[0048] - FIG. 7C is a table showing the reduction of online round trip time (RTT) for searching, using the category-based method of the present disclosure; and

[0049] - FIG. 8 shows a server computer/server apparatus according to some embodiments.

DETAILED DESCRIPTION

- [0050] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the disclosure may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure. Other embodiments may be utilized and structural, and logical changes may be made without departing from the scope of the disclosure. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.
- [0051] Embodiments described in the context of one of the enclosure systems, devices, or methods are analogously valid for the other systems, devices, or methods. Similarly, embodiments described in the context of a system are analogously valid for a device or a method, and vice-versa.
- [0052] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.
- [0053] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements.
- [0054] As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.
- [0055] As used herein, the term “data” may be understood to include information in any suitable analogue or digital form, for example, provided as a file, a portion of a file, a set of files, a signal or stream, a portion of a signal or stream, a set of signals or streams, waveforms, and the like. The term data, however, is not limited to the aforementioned examples and may take various forms and represent any information as understood in the art.
- [0056] As used herein, the term “first” and “second” are used to distinguish one element/feature from another, and, unless otherwise stated, does not denote order, priority or sequence.
- [0057] As used herein, the term “module” refers to, forms part of, or includes an application Specific Integrated Circuit (ASIC) ; an electronic circuit; a combinational logic circuit; a field programmable gate array (FPGA) ; a processor (shared, dedicated, or group)

that executes code; other suitable hardware components that provide the described functionality; or a combination of some or all of the above, such as in a system-on-chip. The term module may include memory (shared, dedicated, or group) that stores code executed by the processor. A single module or a combination of modules may be regarded as a device. A processor may include one or more modules. For example, multiple modules described in this disclosure may form a processor.

- [0058] As used herein, the term “associate”, “associated”, and “associating” indicate a defined relationship (or cross-reference) between two items. For example, a point-of-interest data (e.g. a hotel) may be associate with one or more categories (e.g. accommodation, food and beverages, shopping) .
- [0059] As used herein, “memory” may be understood as a non-transitory computer-readable medium in which data or information can be stored for retrieval. References to “memory” included herein may thus be understood as referring to volatile or non-volatile memory, including random access memory (“RAM”) , read-only memory (“ROM”) , flash memory, solid-state storage, magnetic tape, hard disk drive, optical drive, etc., or any combination thereof. Furthermore, it is appreciated that registers, shift registers, processor registers, data buffers, etc., are also embraced herein by the term memory. It is appreciated that a single component referred to as “memory” or “amemory” may be composed of more than one different type of memory, and thus may refer to a collective component including one or more types of memory. It is readily understood that any single memory component may be separated into multiple collectively equivalent memory components, and vice versa. Furthermore, while memory may be depicted as separate from one or more other components (such as in the drawings) , it is understood that memory may be integrated within another component, such as on a common integrated chip.
- [0060] As used herein, the term “point-of-interest” (POI) data refers to information about a particular location, for example, a location identifiable on a digital map. Such data may include the information needed to identify and define the particular location, such as a street address and GPS coordinates. The POI data may further include other information. As a non-limiting example, the POI data may include a probability of whether it is a likely drop-off point or a likely pickup point at different time periods. For instance, in the morning, a POI in a residential area is more likely to be a pickup point while with low probability to be a drop-off point. The pickup and drop-off time distribution of a POI may be pre-computed offline.
- [0061] As used herein, the term “category” refers to a class associated with a POI information or data. One or more categories in the present disclosure may be used to classify POI data into groups to facilitate searches and retrieval. Non-limiting

examples of such categories include “shopping mall”, “nature park”, “airport”, “hotel”, “food and beverage”, etc.

- [0062] According to an aspect of the disclosure and with reference to FIG. 1, there is provided a method 100 for retrieving relevant POI data in response to a search request. The method 100 comprises the steps of:
- [0063] Step 102: associating at least one category to the search request.
- [0064] Step 104: retrieving, from a database, a plurality of POI candidates, based on the associated at least one category.
- [0065] Step 106: determining the one or more relevant POI data from the plurality of POI candidates based on comparing the search request with each of the plurality of POI candidates, and calculating a corresponding similarity score for each POI candidate.
- [0066] Optionally, the plurality of POI candidates may be ranked to determine the relevance of each POI candidate. The method may further comprising ranking (step 108) each of the plurality of POI candidates, based on the corresponding similarity score for each POI candidate.
- [0067] The method 100 may be implemented in one or more server apparatus or device for an e-commerce service, such as a ride-hailing service, on-demand taxi service, food delivery service, and/or on demand logistic service. The method 100 may be implemented as executable codes or instructions in a non-transitory computer-readable storage medium or a data processing device, which, when executed by one or more processors, cause the execution of the method 100.
- [0068] The database may comprise pre-categorized point-of-interest data, such that each entry of the database comprises a point-of-interest and at least one category. The database may form part of a module for prediction of point-of-interest category, which will be subsequently elaborated.
- [0069] In some embodiments, the step of associating at least one category to the search request may comprise parsing the search request; predicting, using a category prediction and detection module (which may comprise an Elasticsearch engine), one or more categories based on the parsed search request, and indexing the one or more categories using the category prediction and detection module.
- [0070] In some embodiments, the parsing of the search request may comprise correlating the search request with one or more point-of-interest data within the database.
- [0071] In some embodiments, the correlating may include predicting a relationship between the search request and the one or more point-of-interest data.
- [0072] In some embodiments, the indexing may comprises generating at least one index of a keyword type.

- [0073] In some embodiments, the method further comprises generating a category structure for the one or more categories, wherein the category structure comprises a three-level hierarchical category tree.
- [0074] In some embodiments, the predicting of one or more categories based on the parsed search request includes training the category prediction and detection module using a set of POI training data and a large language model. The training may further comprises adopting a natural language processing (NLP) model and a masked language modeling model.
- [0075] In some embodiments, the method further comprises applying a finetuning model, the finetuning model comprises a first finetuning step of generating weak-supervision data by mapping whitelisted category keywords to a first set of POI data, and a second finetuning step of collecting a predetermined number of second set of POI data and labelling the collected POI data in the second set of POI data based on the whitelisted categories.
- [0076] As shown in FIG. 2A, the method 100 and embodiments mentioned above may be implemented on/in a transport service system 200, such as an on-demand transport service system or ride-hailing system. The system may comprise a user device 202, and a server 204. The user device 202 may be a smart phone device configured to communicate with the server 204 via a dedicated software application (colloquially referred to as an app) . The user device 202 may be configured to send a search request 206 to the server 204. In some embodiments, the search request 206 may be in the form of a text data or string data, and may contain a search query associated with one or more points-of-interest stored in a database 208. The server 204 may return relevant POI data 214 to the user device 202. In some embodiments, the relevant POI data 214 may be displayed as a list of POI for the user's selection.
- [0077] In some embodiments, the search request 206 may be provided by a user via a user interface. The user interface may form part of an on-demand transportation or a ride-hailing application for the user to input a search request for a pick-up point and/or a drop-off point. In some embodiments, the search request 206 may be formed from words from different languages. Non-limiting examples include English, Chinese, Indonesian, Malay, Thai, Vietnamese. In some embodiments, the search request 206 may be parsed by the user device 202 before being sent to the server 204.
- [0078] The server 204 may comprise one or more processors for handling the search request 206. The server 204 may be arranged in data communication with the database 208 for storing categorized POI data. As shown in FIG. 2B, in relation to a search request "Botanic Gardens Singapore", each entry 210 of the database 208 may comprise a POI information/data and at least one category. For example, and entry 210a may comprise a category "wedding photo venue" and a corresponding POI "Botanic

Gardens bandstand”, and an entry 210b may comprise a category “nature trail” and the corresponding POI “Botanic Gardens Rainforest trail”, etc. It is appreciable that one POI may be associated with multiple categories, and vice versa.

- [0079] In some embodiments, the server 204 may be a remote server, for example, a cloud server. In some embodiments, the server 204 may be configured in a distributed server architecture.
- [0080] In some embodiments, the server 204 may include a category prediction and detection module 212. The category prediction and detection module 212 may be configured to predict the categories of the POI data and modify the database 208 based on any historical POI data, or, new POI data entry. In some embodiments, the category prediction and detection module 212 may be configured to execute a prediction at every pre-determined time interval, or on-demand. In some embodiments, the category prediction and detection module 212 may be configured to predict the categories of the POI data offline, i.e., not within the operational time of the system 200 for receiving search request 206. In some embodiments, the category prediction and detection module 212 may be operational in an offline process, to distinguish from the online search query process by users. The offline process may take place at scheduled maintenance time period, or as and when required.
- [0081] In some embodiments, online automatic category prediction may be enabled for creating and updating POIs data.
- [0082] Upon receiving the search request 206, the server 204 may proceed to parse the search request; predict, using a category and detection search engine. The search engine may be a distributed search and analytics engine, and may include an Elasticsearch (ES) engine. The outcome of the prediction may be one or more categories based on the parsed search request, and index the one or more categories using the ES algorithm.
- [0083] In the parsing process, the search request 206 may be correlated with one or more POI data within the database 208, for example, the parsing of the search request 206 may be based on identifying words within the search request 206 that may correspond to one or more POI within the database 208. In some embodiments, the correlating may include predicting a relationship between the search request 206 and the POI data.
- [0084] The ES algorithm may be utilised to organise a relatively large amount of POI data in the database 208. In some embodiments, the database 208 may be a Redis data-store which may be used to shorten the response time of each request as cache. In some embodiments, the ES algorithm may be used to store POI data and execute POI queries.
- [0085] In some embodiments, each category may be indexed as a keyword type in the ES algorithm (i.e. ES index), to facilitate fast term filtering in the querying stage.

- [0086] In some embodiments, any new categories arising from new POI data may be updated/inserted into the ES index in real time.
- [0087] FIG. 3 shows an embodiment of a data pipeline 300 for processing the search request 206 and returning the relevant POI data from the server 204. The category prediction and detection module 212 may comprise an offline POI prediction module 314, and an online query-category detection module 322.
- [0088] The data pipeline 300 comprises an offline process 310 wherein a POI information store 312 provides uncategorized POI information from various sources, the uncategorized POI information is then fed into the POI prediction module 314, and the POI categories in the form of ES indexes are stored in the database 208.
- [0089] The data pipeline 300 also comprise an online process 320. The online process 320 may comprise the search request 206 being fed into a query-category detection module 322 as input, the output from the query-category detection module 322, which may be in the form of one or more associated categories, are then sent to a category-based search engine module, i.e. Elasticsearch (ES) module 324. The category-based ES module 324 may then be used to retrieve a plurality of POI candidates associated with the one or more categories.
- [0090] The plurality of POI candidates may then be fed into a ranker module 326 for ranking the POI candidates determine the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on comparing the search request with each of the plurality of point-of-interest candidate, and calculating a corresponding similarity score for each point-of-interest candidate.
- [0091] In some embodiments, a higher similarity score may be assigned to a point-of-interest candidate having a higher degree of similarity when compared to the search request.
- [0092] The online pipeline 320 may be regarded as a category-based POI candidate retrieval mechanism, wherein the query-category detection module 322 is used to detect one or more categories of each search request 206. The detected/predicted one or more categories may then be passed to the ES module 324. It is appreciable the ES module 324 is applied to a reduced search space to search and calculate the text similarity between the search request and the POIs belonging to the one or more categories. Therefore, the category-based POI candidate retrieval may reduce ES round-trip time compared to an uncategorized search space. In addition, the category-based POI candidate retrieval also has the potential to improve search recall as the pipeline may filter out irrelevant POIs from other categories.
- [0093] In some embodiments, a category structure comprising a three-level hierarchical category tree for the POIs-L1, L2, and L3, wherein L1 denotes the highest hierarchy level of category and L2, L3 denotes lower hierarchy level categories, may be

implemented. In some embodiments, there comprises 25 L1 categories, 685 L2 categories and 849 L3 categories.

- [0094] As an example, “food and beverage” may be one of the L1 categories, while “alcohol” is an L2 category belonging to the “food and beverage” L1 category. Within “alcohol”, there may be several L3 categories, for example, “breweries”, “distilleries”, and so on.
- [0095] In some embodiments, artificial intelligence, such as machine learning, may be used for the prediction of POI category in the offline process 310. Particularly, the POI category prediction module 314 may be trained using machine learning. In some embodiments, training dataset comprising existing or historical POI data from the POI information store 312 may be input to the artificial intelligence (AI) module, and the output from the AI module is the predicted one or more associated categories. Such machine learning may be trained via supervised learning, un-supervised learning, or hybrid learning model. Where supervised machine is used in the prediction, a suitable, and sufficient number of high-quality labeled data may be required to build an effective supervised machine learning system.
- [0096] In some embodiments, large pretrained language models (LLMs) may be used to pretrain the POI data 312. Such LLMs may be advantageous to manage data scarcity. With different self-supervised learning strategies, LLMs can obtain supervisory signals from the POI training data, which leverage the underlying structure in the data so as to yield higher performance in the prediction tasks even with limited labeled data.
- [0097] In some embodiments, the POI training data comprises POI information in multiple languages. A method for training the multi-lingual dataset may comprise pretraining a large language model on the POI training data (also referred to as POI corpus) and fine-tune the training on the limited labeled POI category data.
- [0098] In some embodiments, the training may adopt a natural language processing (NLP) model for machine learning, such as BERT NLP model, and/or A Lite BERT (ALBERT) NLP model. ALBERT is a modified version of the BERT NLP model. The ALBERT NLP model may be suited as a backbone model in training the POI category classification task, as it has relatively lower training computational cost and smaller inference latency. Compared to BERT, ALBERT utilizes Factorized Embedding Parameterization and Cross-Layer Parameter Sharing to achieve higher performance with only 10% parameters of BERT.
- [0099] In some embodiments, and in the context where the POI data comprise multiple languages including English, Malay, Indonesian, Thai, and Vietnamese, the ALBERT is trained from scratch based on the multi-lingual POI data, using various search

request data, before a finetuning step shown in FIG. 4. The POI data may be in various South East Asian languages, in addition to English.

[0100] It may be appreciable that the pretrained language models may adopt masked language modeling (MLM) to predict the “masked” tokens. The BERT NLP model may also use the next-sentence prediction (NSP) loss to learn the consecutive sentence relationship. The ALBERT NLP model uses the sentence-order prediction (SOP) loss to learn the finer-grained distinctions. According to some embodiments of the present disclosure, the ALBERT model may be modified, in that a POI relationship prediction is adopted together with MLM tasks to train what is known as a POI-ALBERT model. The POI-ALBERT model is based on domain knowledge obtained through observing historical search requests (queries) and the corresponding POI data. It was discovered that the search requests are typically highly connected with their selected POIs data, and the different attributes of the same POI data. In some embodiments, the search requests and the POI data may be pre-processed as part of a POI relationship predication task. The pre-processing relationship prediction task may include constructing segment pairs as follows:

[0101] ● Search query -Selected POI pairs:

[0102] ○ Search query -Selected POI name /native name

[0103] ○ Search query -Selected POI address /native address

[0104] ● POI attribute pairs:

[0105] ○ POI name -POI address

[0106] ○ POI native name -POI address

[0107] ○ POI name -POI native address

[0108] ○ POI native address -POI native address

[0109] In some embodiments of the POI relationship prediction task, positive examples (i.e. positive reinforcement) may be created by taking two segments from the same segment pair, while negative examples are from derived from different segment pairs. For example, a search request query ‘**airport’ and the POI name ‘Changi Airport’ may be regarded as a positive paired segment, whereas ‘airport’ and the POI ‘Orchard Road’ are regarded as a negative paired segment. Together with the MLM task, the POI relationship prediction task may help the POI relationship prediction model to understand the correlation between search requests and POI attributes.

[0110] In some embodiments, the POI-ALBERT model may be trained with 12 hidden layers. The size of each hidden layer may be 768 and the embedding size may be 128. As ALBERT adopts factorized embedding and cross-layer parameter sharing, there may be 12 million parameters which is much smaller than if the BERT model is adopted. In some embodiments, the training may be performed using a WordPiece tokenizer on our search query -POI corpus with a 100,000-token vocabulary. An

AdamW optimizer, which is a stochastic optimization method that modifies the typical implementation of weight decay in the adaptive optimizer Adam, by decoupling weight decay from the gradient update, may be selected with a learning rate of $5e^{-5}$.

- [0111] FIG. 4 shows an exemplary model 400, wherein the different POI information may be concatenated as an input, e.g. POI name, native name and fed into the POI-ALBERT model 402. The output of the POI-ALBERT model 402, in the form of one or more categories, may be input into a pooling layer 404. The pooling layer 404 provides an approach to down sampling feature data by summarizing the presence of features in patches of the feature map.
- [0112] The pooling layer 404 may implement a pooling strategy, such as using a mean vector of all token embeddings of last two layers from the POI-ALBERT layer. A fixed-size embedding output from the pooling layer 404 may then be fed into the final softmax layer 406 to get the category predictions. A softmax loss function may be adopted in the softmax layer 406 to fine-tune the entire model. The softmax function converts a vector of K real numbers into a probability distribution of K possible outcomes. In some embodiments, the input of the SoftMax layer may be a mean vector of all token embeddings of last two layers from the POI-ALBERT layer, and the output of the SoftMax layer/function is the predicted probability of each category.
- [0113] In some embodiments, a finetuning model may be to apply the pre-trained MLM language model directly to the downstream data (i.e. a *naïve* finetuning strategy) .
- [0114] In another embodiment, a two-step finetuning strategy may be modelled. The finetuning model comprises a first finetuning step of generating weak-supervision data by mapping whitelisted categories to a first set of POI data, and a second finetuning step of collecting a predetermined number of second set of POI data and labelling the collected POI data in the second set of POI data based on the whitelisted categories.
- [0115] In a first finetuning step, weak-supervision data are generated. Weak-supervision data are based on whitelisted category keywords, based on the assumption that whenever users search for these whitelisted keywords, the POIs selected actually belong to the corresponding category. The weak-supervision data may be generated by mapping whitelisted category keywords to the POIs. By finetuning exclusively on these weak-supervision data, the POI-ALBERT model can become more suitable for the category prediction task. In particular, this initial pretraining round may be regarded as a warm-up step for the final dense-softmax layer, which is only randomly initialized at the beginning of finetuning.
- [0116] In a second finetuning step, 20,000 (or other predefined number) of POIs across 8 countries in south east Asia (SEA) may be collected and labelled into the predefined whitelisted categories. The POI-ALBERT model may then be fine-tuned on the manually labeled data.

- [0117] In both the first finetuning step and the second finetuning step, an early stopping criteria may be adopted to fine-tune the model. The two-step finetuning strategy may be shown to improve the model performance compared to the naïve finetuning strategy.
- [0118] In some embodiments, the output from the softmax layer 406 may be configured as a first version of the POI category prediction model. An entropy-based active learning algorithm may then be used to sample the most valuable unlabeled POI for a 2nd round of labeling to further improve the model's performance. The entropy-based active learning algorithm may be based on the following steps as follows: (a) Use the finetuned POI category prediction model to predict the category of all POIs; (b) Calculate the entropy of category prediction probabilities, where higher entropy possibly indicates that the model is less confident about the category prediction; (c) Sample additional POI data, for example 20,000 POI data for manual labeling based on the sampling weights proportional to the entropies. Assign different sampling weights to the POIs based on the corresponding entropies (POIs with larger entropies will be assigned with relatively larger sampling weights) .
- [0119] With the POI category prediction model generated using the POI-ALBERT and finetuned, the offline process 310 may be ready.
- [0120] In the online process 320, the query-category detection module 322 may be trained so as to properly predict or classify each search request or query into predefined categories. These predicted categories will be passed to the ES module 324 together with the search requests. The ES module 324 may be configured to retrieve the relevant POIs within these categories, which could save the computational cost of ES so as to reduce its latency.
- [0121] In some embodiments, the training data for training the query-category detection module 322 may be prepared via labelling the training data. In one embodiment, such label data may be obtained through crowdsourcing. In another embodiment, since the POI categories have been obtained using the query-category detection module 322, each POI category may be associated or map to its corresponding search queries by assuming that the search queries and their associated POIs share the same categories. Advantageously, this embodiment is relatively less expensive and less time-consuming compared to the crowdsourcing method.
- [0122] FIG. 5 shows an example of where one or more users input a search request 206 in the form of a text "Botanic garden gallop" . The corresponding POI 504 may be selected as "Visitor Centre, Gallop Extension Botanic Gardens" . Based on the corresponding POI 504, the category 506 may be predicted as art: : garden: : botanicalgardern and attractions: : parks: : . The categories of this query are assumed to be the above.

- [0123] In some embodiments, the query-category detection module 322 may also adopt the POI-ALBERT algorithm. As another embodiment, for the sake of inference time and resource consumption, a lightweight convolutional neural network (textCNN) model may be used. As the training data (corpus) comprises multiple languages, a character tokenizer may be used. The embedding size for each token is set as 200. In some embodiments, 5 kernels with 1024 filters and 5 convolutional layers may be chosen. The P99 latency of the query category model in the production environment achieves less than 10 milliseconds (ms) , with almost the same online performance of the POI-ALBERT model.
- [0124] As may be appreciated, the POI category prediction module 314 of the offline process 310, and the query-category detection module 322 of the online process 320 of the data pipeline 300, may be trained separately.
- [0125] FIG. 6A and FIG. 6B show the details of the entire category retrieval pipeline, which includes the offline POI category prediction and online category-based POI retrieval. The POI category prediction may be utilized to obtain top 5 (or any other number as desired) predicted categories of each POI in the database 208 in an offline manner. The categories may be indexed as described and the keyword type in ES to facilitate fast term filtering.
- [0126] In operation, the categories may function as a filter when searching for POI candidates in ES. Referring to FIG. 6A, an example in JavaScript Object Notation (JSON) json syntax presents an example of a category-based ES search template construction. A search request 206 "Marina Bay Sands" is predicted to be associated with the categories "food and beverage" and "hotel" . The aforementioned two categories may serve as filters sent to the ES module 324, which may help to limit the search space during the search process in the ES module 324, such that only the POI candidates belonging to these two categories will be retrieved, to reduce the latency of ES retrieval.
- [0127] FIG. 6B shows an example of presenting the potentiality of Category based retrieval to boost POI search performance. A user (who may be a customer of the e-commerce platform) who sends a search request 206 may wish to select a related POI 'Main Entrance, City Gate' when inputting the search request 'Coty gate en' . However, based on the original ES without category-based search, the POI 'Main Entrance, City Gate' may be ranked as 22, which may not be ranked in a top 5 related POI and consequently may not be displayed as a relevant POI for the user's selection. In contrast, under the category-based retrieval of the present disclosure, the categories associated with the search request may be of "commercial" , "general merchandise" , "industrial" , etc. These categories may help ES to filter the irrelevant POIs of other categories, which can boost the intended POI's ranking from 22 to 4 in terms of ES

rank. Therefore, the category-based retrieval of the present disclosure can also be used to potentially improve the POI search recall.

- [0128] The efficacy of the method 100 and the POI category prediction model may be evaluated in terms of performance measures such as top-k accuracy-see FIG. 7A, which indicates how many times the correct category appears in the top-k predicted classes; FIG. 7B, which shows the online A/B experiments that lasted for 2 weeks in Vietnam, Philippines, Malaysia, and Thailand; and FIG. 7C, which shows the reduction in prediction round trip time (RTT) when compared with an ES module without category prediction.
- [0129] Referring to FIG. 7A, the top-1, top-3 and top-5 accuracy metrics are displayed. The results are based on the two-step finetuning strategy to fine-tune our POI-ALBERT pretrained model. The table shown in FIG. 7A compares the results of the *naïve* finetuning strategy with the two-step finetuning strategy, and the active learning strategy. It may be appreciable that the two-step finetuning strategy can improve the category prediction performance compared to the *naïve* finetuning strategy by around 9% or more. Additionally, the active learning strategy also improves top-5 accuracy by 2% compared to the two-step finetuning strategy. Hence, using two-step finetuning strategy and active learning strategy, to predict the categories of the existing POIs may yield the most desired results. For each POI, the top-5 POI predictions as its corresponding categories may be displayed to the user.
- [0130] Referring to FIG. 7B, the effectiveness of the proposed data pipeline and method 100 may be validated. Online A/B experiments which lasted for 2 weeks in Vietnam, Philippines, Malaysia, and Thailand were performed. Online A/B experiments' results based on recall@k and NDCG@k metrics were evaluated. For recall@k, it computes a percentage of POIs among top-k results have been selected by end-users, whereas NDCG@k metric takes into account the position of relevant items in the ranked list. As can be seen, category retrieval led to a considerable increment of POI search metrics in all countries. In summary, the online A/B results provide evidence that implementing category retrieval can enhance POI search performance, while simultaneously decreasing search latency.
- [0131] The results of the ES RTT comparisons are shown in FIG. 7C, which shows the use of category-based method of the present disclosure results in a significant reduction of ES RTT in all countries.
- [0132] In some embodiments, various indexes may have been proposed for the management of large volume of POI data, such as Rtree, k-d tree. Another index method is Geohash, which encodes a geographic location into a short string of letters and digits and ensure that nearby places share similar prefix. In various embodiments, when a search request of user location (e.g., which may be expressed as coordinates such as

latitude and longitude) is received, the Elasticsearch engine/system of the category prediction module may make use of Geohash to return relevant POIs close to the given location for further association with categories or next step processing.

- [0133] In some embodiments, the backend architecture of the system and data pipeline 300 is implemented in Golang. As its parallel processing is based on goroutine and channel, each http request may be allocated to an independent goroutine to process.
- [0134] FIG. 8 shows a server computer 800 according to an embodiment.
- [0135] The server computer 800 includes a communication interface 801 (e.g. configured to receive interaction data, i.e. information about interactions) . The server computer 800 further includes a processing unit 802 and a memory 803. The memory 803 may be used by the processing unit 802 to store, for example, data to be processed, such as search request data files, POI data, and category-related data files. The server computer may be configured to perform the method of FIG. 1.
- [0136] The methods described herein may be performed and the various processing or computation units and the devices and computing entities described herein may be implemented by one or more circuits. In an embodiment, a "circuit" may be understood as any kind of a logic implementing entity, which may be hardware, software, firmware, or any combination thereof. Thus, in an embodiment, a "circuit" may be a hard-wired logic circuit or a programmable logic circuit such as a programmable processor, e.g. a microprocessor. A "circuit" may also be software being implemented or executed by a processor, e.g. any kind of computer program, e.g. a computer program using a virtual machine code. Any other kind of implementation of the respective functions which are described herein may also be understood as a "circuit" in accordance with an alternative embodiment.
- [0137] While the disclosure has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the disclosure as defined by the appended claims. The scope of the disclosure is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

Claims

- [Claim 1] A method for retrieving relevant point-of-interest data in response to a search request, the method comprising the steps of:
associating at least one category to the search request;
retrieving, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and
determining the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on comparing the search request with each of the plurality of point-of-interest candidate, and calculating a corresponding similarity score for each point-of-interest candidate.
- [Claim 2] The method of claim 1, further comprising ranking each of the plurality of point-of-interest candidates, based on the corresponding similarity score for each point-of-interest candidate.
- [Claim 3] The method of claim 1 or 2, wherein the database comprises pre-categorized point-of-interest data, and wherein each entry of the database comprises a point-of-interest and at least one category.
- [Claim 4] The method of any one of the preceding claims, wherein the associating at least one category to the search request comprises
parsing the search request;
predicting, using a category prediction and detection module, one or more categories based on the parsed search request, and
indexing the one or more categories using the category prediction and detection module.
- [Claim 5] The method of claim 4, wherein parsing the search request comprises correlating the search request with one or more point-of-interest data within the database.
- [Claim 6] The method of claim 5, wherein the correlating includes predicting a relationship between the search request and the one or more point-of-interest data.
- [Claim 7] The method of any one of claims 4 to 6, wherein the indexing comprises generating at least one index of a keyword type.
- [Claim 8] The method of any one of claims 4 to 7, further comprising generating a category structure for the one or more categories, wherein the category structure comprises a three-level hierarchical category tree.
- [Claim 9] The method of any one of claims 4 to 8, wherein the predicting of one or more categories based on the parsed search request includes

- training the category prediction and detection module using a set of POI training data and a large language model.
- [Claim 10] The method of claim 9, wherein the training further comprises adopting a natural language processing (NLP) model and a masked language modeling model.
- [Claim 11] The method of claim 10, further comprising applying a finetuning model, the finetuning model comprises a first finetuning step of generating weak-supervision data by mapping whitelisted categories to a first set of POI data, and a second finetuning step of collecting a predetermined number of second set of POI data and labelling the collected POI data in the second set of POI data based on the whitelisted categories.
- [Claim 12] The method of any one of claims 4 to 11, wherein the category prediction and detection module comprises a distributed search and analytics engine.
- [Claim 13] A server apparatus configured to retrieve one or more relevant point-of-interest data in response to a search request, the server apparatus comprises at least one processor, the at least one processor configured to:
- associate at least one category to the search request;
 - retrieve, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and
 - determine the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on a comparison of the search request with each of the plurality of point-of-interest candidate, and calculate a corresponding similarity score for each point-of-interest candidate.
- [Claim 14] The server apparatus of claim 13, wherein the at least one processor is configured to rank each of the plurality of point-of-interest candidates, based on the corresponding similarity score for each point-of-interest candidate.
- [Claim 15] The server apparatus of claim 13 or 14, wherein the database comprises pre-categorized point-of-interest data, and wherein each entry of the database comprises a point-of-interest and at least one category.
- [Claim 16] The server apparatus of any one of claims 13 to 15, wherein, in associating at least one category to the search request, the at least one processor is configured to:
- parse the search request;

predict, using a category prediction and detection module, one or more categories based on the parsed search request, and
index the one or more categories using the category and detection prediction module.

- [Claim 17] The server apparatus of claim 16, wherein, in parsing the search request, the at least one processor is configured to correlate the search request with one or more point-of-interest data within the database.
- [Claim 18] The server apparatus of claim 17, wherein the at least one processor is configured to predict a relationship between the search request and the one or more point-of-interest data.
- [Claim 19] The server apparatus of any one of claims 16 to 18, wherein, in indexing the one or more categories, the at least one processor is configured to generate at least one index of a keyword type.
- [Claim 20] The server apparatus of any one of claims 16 to 19, wherein the at least one processor is configured to generate a category structure for the one or more categories, wherein the category structure comprises a three-level hierarchical category tree.
- [Claim 21] The server apparatus of any one of claims 16 to 20, wherein in the prediction of one or more categories based on the parsed search request, the at least one processor is configured to train the category prediction and detection module using a set of POI training data and a large language model.
- [Claim 22] The server apparatus of claim 21, wherein the at least one processor is further configured to adopt a natural language processing (NLP) model and a masked language modeling model in training the category prediction and detection module.
- [Claim 23] The server apparatus of claim 22, wherein the at least one processor is further configured to apply a finetuning model as follows:
generate weak-supervision data by mapping whitelisted categories to a first set of POI data, and
collect a predetermined number of second set of POI data and label the collected POI data in the second set of POI data based on the whitelisted categories.
- [Claim 24] The server apparatus of any one of claims 16 to 23, wherein the category prediction and detection module comprises a distributed search and analytics engine.
- [Claim 25] A system for providing a on-demand transport service to a user, the system comprising

a user device configured to receive a search request for point-of-interest data, the point-of-interest data associated with at least one of a pick-up point and a drop-off point;
at least one processor configured to:
associate at least one category to the search request;
retrieve, from a database, a plurality of point-of-interest candidates, based on the associated at least one category; and
determine the one or more relevant point-of-interest data from the plurality of point-of-interest candidates based on a comparison of the search request with each of the plurality of point-of-interest candidate, and calculate a corresponding similarity score for each point-of-interest candidate.

- [Claim 26] A non-transitory computer-readable storage medium comprising instructions, which, when executed by one or more processors, cause the execution of the method for building a dataset for automatic speech recognition according to any one of claims 1 to 8.
- [Claim 27] A data processing device configured to perform the method of any one of claims 1 to 12.
- [Claim 28] A computer executable code comprising instructions building a dataset for automatic speech recognition according to any one of claims 1 to 12.