(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 June 2023 (15.06.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/104301 A1

(51) International Patent Classification:
G06F 16/53 (2019.01)

(21) International Application Number:
PCT/EP2021/084775

(22) International Filing Date:
08 December 2021 (08.12.2021)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HUAWEI TECHNOLOGIES CO., LTD.** [CN/CN]; Huawei Administration Building Bantian, Longgang District, Shenzhen, Guangdong 518129 (CN).

(72) Inventor; and

(71) Applicant *(for MN only)*: **XIA, Baiqiang** [CN/SE]; Huawei Technologies Sweden AB, Skalholtsgatan 9, 16440 Kista (SE).

(74) Agent: **KREUZ, Georg M.**; Huawei Technologies Duesseldorf GmbH, Riesstr. 25, 80992 Munich (DE).

(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, IT, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, 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))*

WO 2023/104301 A1

(54) **Title:** METHOD FOR RESPONDING TO A SEARCH QUERY

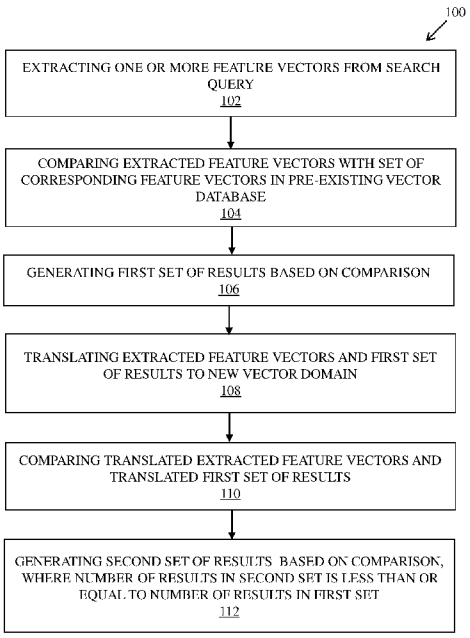

FIG. 1

(57) **Abstract:** A method for responding to a search query generated by a user. The method includes extracting one or more feature vectors from the search query and comparing the extracted feature vectors with a set of corresponding feature vectors in a pre-existing vector database. The method further includes generating a first set of results based on the comparison and translating the extracted feature vectors and the first set of results to a new vector domain. The method further includes comparing the translated extracted feature vectors and the translated first set of results and generating a second set of results based on the comparison, where the number of results in the second set is less than or equal to the number of results in the first set.

METHOD FOR RESPONDING TO A SEARCH QUERY

TECHNICAL FIELD

The present disclosure relates generally to the field of search engines; and more specifically to a method for responding to a search query.

BACKGROUND

Generally, various search engines, such as Google, Petal Search, Yandex, Google Lens, and the like, play a significant role in modern internet services. Such search engines retrieve the most relevant information within the available web content, in response to various queries made by users. The available web content is a large-scale dataset, which may include a billion or trillion raw data points. For example, as of the year 2021, there are 1.2 billion websites and 56.5 billion web pages, 130 million different books in the world, and 800 million videos hosted on YouTube® alone. The creation of web (or the internet) content is still ongoing at an unprecedented rate. Despite being large scale, the web content is usually created without an explicit intention to serve a search purpose (or a search query). To make a search successful, the search engines are required to understand the relevance between query inputs and raw data points, and make an efficient search from a large number of raw data points. In order to deliver an efficient and successful search experience, conventionally, raw data points of a large-scale dataset are firstly represented as feature vectors by employing a feature vector extraction model, to capture the similarity or relevance between the raw data points and various queries made by users. Thereafter, the feature vectors are organised into a structured database, for example, a backend vector database, to facilitate fast information retrieval.

The search engines are usually in long-term and continuous development and are frequently updated to perform effective searches and to adapt to new and emerging application scenarios. Typically, for an updated model, the corresponding vector database is regenerated. Therefore, new feature vectors have to be extracted from each data point of the large-scale dataset and organised into a new vector database. Thereafter, queries of the user are processed against the

new vector database and the old vector database is discarded. However, the regeneration of the new vector database is both costly and time-intensive, which may result in delay or even cancellation of a new service publication. For example, extracting feature vectors from one trillion images would take 2.8 million GPU hours, given a feature extraction speed estimation of 10 milliseconds per image, which is prohibitively slow and expensive.

Thus, there exists a technical problem of updating a feature extraction model in order to deliver a cost-effective and successful search experience.

SUMMARY

The present disclosure provides a method for responding to a search query. An aim of the present disclosure is to provide a solution that overcomes the problems encountered in the prior art and provides an improved method for responding to a search query.

According to an aspect of the present disclosure, there is provided a method for responding to a search query generated by a user. The method comprises extracting one or more feature vectors from the search query and comparing the extracted feature vectors with a set of corresponding feature vectors in a pre-existing vector database. The method further comprises generating a first set of results based on the comparison and translating the extracted feature vectors and the first set of results to a new vector domain. The method further comprises comparing the translated extracted feature vectors and the translated first set of results and generating a second set of results based on the comparison, where the number of results in the second set is less than or equal to the number of results in the first set.

The disclosed method enables enhanced search experience and takes less time to upgrade the search service using a new feature extraction model with less computational resources as compared to existing systems. The extraction of feature vectors from the search query and the translating of the extracted feature vectors and the first set of results to the new vector domain avoids generating a new vector database from scratch (i.e., from raw data points), which in turn makes the method computationally efficient.

In an implementation form, the method further comprises at least one of the following: providing the second set of results to the user, or generating a third set of results based on the second set of results, and providing the third set of results to the user.

It is advantageous to provide either the second set of results or the third set of results generated based on the second set of results to the user based on the relevance to the query made by the user.

In a further implementation form, the method further comprises ranking the results in the first set after the comparison with the translated extracted feature vectors.

The ranking (more specifically, re-ranking) the results in the first set after comparison with the translated extracted feature vectors results in better search accuracy.

In a further implementation form, the ranking is based on feature vectors extracted using an updated feature extraction model.

By virtue of performing the ranking based on feature vectors extracted using the updated feature extraction model, an efficient and successful search experience can be achieved.

In a further implementation form, the method further comprises aggregating the ranked results and the first set of results to generate the second set of results.

The aggregation of the ranked results in the first set and the first set of results generates the second set of results, which is likely to be more relevant to the query made by the user.

In a further implementation form, the method further comprises omitting the first set of results or the ranked results during aggregation.

The omission of the first set of results or the ranked results during aggregation may provide a more relevant set of search results to the user.

In a further implementation form, comparing comprises determining the similarity or relevance between the extracted feature vectors from the search query and the corresponding feature vectors in the pre-existing vector database.

The use of the pre-existing vector database for comparison with the extracted feature vectors from the search query leads to faster and more cost-efficient updates in the search service.

In a further implementation form, the method further comprises obtaining raw data points of the first set of results.

The obtaining of the raw data points of the first set of results leads to improving accuracy.

In a further implementation form, the method further comprises translating the pre-existing vector database in its entirety to construct a new vector database.

The translation of the pre-existing vector database to the new vector database reduces complexity and runtime overhead.

In a further implementation form, translating the feature vectors comprises training a machine learning model with a batch of training data.

The training of the machine learning model with the batch of training data imparts information obtained from the batch of training data to the translated feature vectors.

In another aspect, the present disclosure provides a computer program comprising instructions which, when executed by a computer, cause the computer to carry out the steps of the method.

In an implementation form, the computer program is stored on a non-transitory computer-readable medium.

It has to be noted that all devices, elements, circuitry, units, and means described in the present application could be implemented in the software or hardware elements or any kind of combination thereof. All steps which are performed by the various entities described in the present application as well as the functionalities described to be performed by the various entities are intended to mean that the respective entity is adapted to or configured to perform the respective steps and functionalities. Even if, in the following description of specific embodiments, a specific functionality or step to be performed by external entities is not reflected in the description of a specific detailed element of that entity which performs that specific step or functionality, it should be clear for a skilled person that these methods and functionalities can be implemented in respective software or hardware elements, or any kind of combination thereof. It will be appreciated that features of the present disclosure are susceptible to being combined in various combinations without departing from the scope of the present disclosure as defined by the appended claims.

Additional aspects, advantages, features, and objects of the present disclosure would be made apparent from the drawings and the detailed description of the illustrative implementations construed in conjunction with the appended claims that follow.

BRIEF DESCRIPTION OF THE DRAWINGS

The summary above, as well as the following detailed description of illustrative embodiments, is better understood when read in conjunction with the appended drawings. For the purpose of illustrating the present disclosure, exemplary constructions of the disclosure are shown in the drawings. However, the present disclosure is not limited to specific methods and instrumentalities disclosed herein. Moreover, those in the art will understand that the drawings are not to scale. Wherever possible, like elements have been indicated by identical numbers.

Embodiments of the present disclosure will now be described, by way of example only, with reference to the following diagrams wherein:

FIG. 1 is a flowchart of a method for responding to a search query generated by a user, in accordance with an embodiment of the present disclosure;

FIG. 2 illustrates a technical workflow for responding to a search query through feature translation, in accordance with an embodiment of the present disclosure;

FIG. 3 illustrates a technical workflow for responding to a search query through feature translation, in accordance with another embodiment of the present disclosure;

FIG. 4 illustrates a training process of a feature translation model, in accordance with an embodiment of the present disclosure; and

FIG. 5 illustrates a technical workflow of processing of an old vector database to a new vector database through feature translation, in accordance with an embodiment of the present disclosure.

In the accompanying drawings, an underlined number is employed to represent an item over which the underlined number is positioned or an item to which the underlined number is adjacent. When a number is non-underlined and accompanied by an associated arrow, the non-underlined number is used to identify a general item at which the arrow is pointing.

DETAILED DESCRIPTION OF EMBODIMENTS

The following detailed description illustrates embodiments of the present disclosure and ways in which they can be implemented. Although some modes of carrying out the present disclosure have been disclosed, those skilled in the art would recognize that other embodiments for carrying out or practicing the present disclosure are also possible.

FIG. 1 is a flowchart of a method for responding to a search query generated by a user, in accordance with an embodiment of the present disclosure. With reference to FIG. 1, there is shown a method **100** for responding to a search query generated by a user. The method **100** includes steps **102** to **112**.

The method **100** introduces an efficient way of generating search results according to one or more new artificial intelligence (AI) models through feature translation. The method **100** is described in detail, in following steps.

At step **102**, the method **100** comprises extracting one or more feature vectors from a search query. The search query corresponds to a query from a user for which the user desires to obtain relevant information by availing a search service provided by a search engine. In an implementation, the search query may be a textual input. In another implementation, the search query may be an audio input. After receiving the search query from the user, one or more feature vectors are extracted from the search query. The process of extracting the feature vectors from the search query may also be termed as query features extraction. Generally, the feature vectors can be defined as one or more vectors that stores the features for one or more particular observations in a specific order. The extracted feature vectors from the search query may also be termed as query feature vectors. It is to be understood by a person of ordinary skill in the art that the search query may be a query generated automatically from an Artificial intelligence (AI) system in an example, without limiting the scope of the disclosure.

At step **104**, the method **100** further comprises comparing the extracted feature vectors with a set of corresponding feature vectors in a pre-existing vector database. The pre-existing vector database may also be termed as an old vector database. The pre-existing vector database holds an existing feature vector database, which is generated using an old version of a feature extraction model. In this way, the method **100** uses the pre-existing vector database in contrast to conventional methods, which discard the pre-existing vector database.

In an implementation, the step of comparing comprises determining the similarity or relevance between the extracted feature vectors from the search query and the corresponding feature vectors in the pre-existing vector database. The comparison of the extracted feature vectors (or query feature vectors) with the set of corresponding feature vectors in the pre-existing vector database includes calculating a similarity or relevance index between the extracted feature vectors from the search query and the corresponding feature vectors in the pre-existing vector database. In an example, the similarity or relevance index may be measured by use of mathematical metrics, such as feature correlation, Euclidean distance, Hamming distance, and the like. In another example, the similarity or relevance index can be measured automatically by use of a metric learned by an AI model from certain data, for example, through metric learning.

At step **106**, the method **100** further comprises generating a first set of results based on the comparison. After comparing the extracted feature vectors (or query feature vectors) from the search query with the set of corresponding feature vectors in the pre-existing vector database, the first set of results is generated. The first set of results corresponds to the closest matches to the search query made by the user and the corresponding feature vectors in the pre-existing vector database. The first set of results may also be referred to as top-N closest matches, where N is an integer with a value more than the size of the final output of search results. The first set of results may also be referred to as initial ranking results.

In an implementation, the method **100** further comprises obtaining raw data points of the first set of results. In an implementation, the raw data points of the first set of results (e.g., the top-N closest matches) can be obtained, depending on the design of a feature translation model, described in detail, for example, in FIG. 4.

At step **108**, the method **100** further comprises translating the extracted feature vectors and the first set of results to a new vector domain. After generation of the first set of results, the extracted feature vectors from the search query and the first set of results are translated to the new vector domain by use of the feature translation model, described in detail, for example, in FIG. 4.

In an implementation, translating the feature vectors comprises training a machine learning model with a batch of training data. The translation of the feature vectors (or the extracted feature vectors) is performed by use of the feature translation model. The feature translation

model includes training the machine learning model using the batch of training data, described in detail, for example, in FIG. 4.

In an implementation, the method **100** further comprises translating the pre-existing vector database in its entirety to construct a new vector database. The translation of the feature vectors (or the extracted feature vectors) may result into a translation of the pre-existing vector database entirely to the new vector database.

At step **110**, the method **100** further comprises comparing the translated extracted feature vectors and the translated first set of results. After translation of the extracted feature vectors and the first set of results to the new vector domain, the translated extracted feature vectors and the translated first set of results are compared with each other. The comparison of the translated extracted feature vectors with the translated first set of results includes calculation of similarity or relevance between the translated extracted feature vectors and the translated first set of results.

In an implementation, the method **100** further comprises ranking the results in the first set after the comparison with the translated extracted feature vectors. Based on comparison of the translated extracted feature vectors and the translated first set of results, the results in the first set are ranked (i.e., re-ranked) accordingly. Alternatively stated, the top-N closest matches are re-ranked based on comparison of the translated extracted feature vectors and the translated first set of results.

In an implementation, the ranking is based on feature vectors extracted using an updated feature extraction model. In an implementation, the ranking (i.e., re-ranking) of the results in the first set can be done based on feature vectors which are extracted using the updated feature extraction model (or the new feature extraction model). The updated feature extraction model may be an artificial intelligence (AI) feature extraction model, trained by various machine learning algorithms.

At step **112**, the method **100** further comprises generating a second set of results based on the comparison, where the number of results in the second set is less than or equal to the number of results in the first set. After comparison of the translated extracted feature vectors and the translated first set of results, the second set of results is generated based on the comparison. However, the number of results in the second set is less than or equal to the number of results in the first set. The second set of results may also be referred to as top-K closest matches, where

K is an integer with a value less than or equal to the number of results in the first set (i.e., the top-N closest matches).

In an implementation, the method **100** further comprises aggregating the ranked results and the first set of results to generate the second set of results. The aggregation of the ranked results in the first set and the first set of results generates the second set of results, which may be more relevant to the query made by the user.

In an implementation, the method **100** further comprises omitting the first set of results or the ranked results during aggregation. The aggregation of the first set of results (i.e., top-N closest matches) and the ranked (i.e., re-ranked) results in the first set, can omit either the first set of results (i.e., top-N closest matches) or the ranked (i.e., re-ranked) results in the first set. The omission of the first set of results or the ranked results during aggregation may remove duplicate or non-relevant results and provide a more accurate set of results to the user.

In an implementation, the method **100** further comprises at least one of the following: providing the second set of results to the user, or generating a third set of results based on the second set of results, and providing the third set of results to the user. In an implementation, the second set of results may be provided to the user as the final output of the search query. In another implementation, the third set of results may be generated based on the second set of results and provided to the user as the final output of the search query.

Thus, the method **100** is based on updating the feature extraction model in the search engine at a fast rate and in a cost-effective manner. Typically, with an updated AI model, a new vector database is regenerated from the raw data points, which is very slow and cost-intensive. However, the method **100** updates the feature extraction model using the feature translation model which translates the pre-existing vector database into the new vector database instead of regenerating the new vector database from the raw data points.

In contrast to a conventional method of updating the feature extraction model, where the new vector database is considered and the pre-existing vector database (i.e., the old vector database) is discarded, the method **100** is based on using the pre-existing vector database (i.e., the old vector database) for generating the first set of results. This makes the method **100** computationally efficient and reliable as compared to existing methods and systems. Moreover, the method **100** is adaptive to migrations between the pre-existing vector database (i.e., the old vector database) and the updated feature extraction model (i.e., new feature extraction model).

Furthermore, the method **100** can be used to verify the effectiveness of the updated feature extraction model (i.e., new feature extraction model) on a search service in real time by either switching on or switching off the feature translation model.

The steps **102** to **112** are only illustrative, and other alternatives can also be provided where one or more steps are added, one or more steps are removed, or one or more steps are provided in a different sequence without departing from the scope of the claims herein.

FIG. 2 illustrates a technical workflow for responding to a search query through feature translation, in accordance with an embodiment of the present disclosure. FIG. 2 is described in conjunction with elements from FIG. 1. With reference to FIG. 2, there is shown a block diagram **200** of a technical workflow for responding to a search query through feature translation. The technical workflow includes a sequence of operations, such as operations **202**, **204**, **206**, **210**, **212**, **214**, **216**, **218**, **220**, and **222**.

At operation **202**, a search query of user is received as an input. The received input may also be referred to as a system input.

At operation **204**, features from the search query are extracted. Alternatively stated, feature vectors from the search query are extracted using a pre-existing version of a feature extraction model.

At operation **206**, the extracted feature vectors from the search query and a set of corresponding feature vectors from a pre-existing vector database **208** are compared with each other. The pre-existing vector database **208** refers to an old vector database.

At operation **210**, a first set of results (i.e., top-N closest matches) is generated based on the comparison between the extracted feature vectors from the search query and the set of corresponding feature vectors in the pre-existing vector database **208**.

At operation **212**, the extracted feature vectors are translated to new vector domain, using a feature translation model, described in detail, for example, in FIG. 4. After translation, the translated extracted feature vectors are obtained.

At operation **214**, the first set of results (i.e., the top-N closest matches) is translated to a new vector domain using the feature translation model. After translation, the translated first set of results is obtained.

At operation **216**, the translated extracted feature vectors and the translated first set of results are compared with each other.

At operation **218**, based on comparison between the translated extracted feature vectors and the translated first set of results, the results in the first set are re-ranked.

At operation **220**, the first set of results and the re-ranked results in the first set are aggregated using various aggregation algorithms. During aggregation, either the first set of results or the re-ranked results in the first set may be omitted.

At operation **222**, a second set of results (i.e., top-K closest matches) is generated based on an aggregation of the first set of results and the re-ranked results in the first set. The second set of results is provided to the user as final outputs of the search query.

The method **100** (of FIG. 1) may be executed on or by a device that may include but not limited to, a computer, a server, a portable electronic device, a smart phone, a tablet, and the like. The device may include a processor that may be configured to execute operations of the method **100**. In an implementation, the device may have either a single hardware server and/or a plurality of hardware servers operating in a parallel or distributed architecture.

FIG. 3 is a diagram that illustrates a technical workflow for responding to a search query through feature translation, in accordance with another embodiment of the present disclosure. FIG. 3 is described in conjunction with elements from FIGs. 1, and 2. With reference to FIG. 3, there is shown a block diagram **300** that illustrates a technical workflow for responding to a search query through feature translation.

The technical workflow of the block diagram **300** for responding to a search query is same as that of the technical workflow of the block diagram **200** (of FIG. 2) except that the operation **212 (**of the FIG. 2) is replaced by an operation **302** in FIG. 3. Similar to the block diagram **200** (of FIG. 2), the block diagram **300** illustrates the update of the search service with the new AI model in an online mode.

At operation **302**, feature vectors from the search query are extracted using an updated feature extraction model. The updated feature extraction model is obtained using a feature translation model, described in detail, for example, in FIG. 4.

Thereafter, the feature vectors extracted using the updated feature extraction model are compared with the translated first set of results at the operation **216**. Since, the query feature vectors are extracted using the updated feature extraction model, therefore, there is no need for translating the extracted feature vectors for carrying out the comparison with the translated first set of results. The translation of the extracted feature vectors may incorporate inaccuracies. Therefore, in the FIG. 3, the feature vectors extracted using the updated feature extraction model are used for the computation of the second set of results. Since, the second set of results (i.e., the top-K closest matches) are obtained using the updated feature extraction model hence, the second set of results (i.e., the top-K closest matches) are more accurate and reliable.

FIG. 4 illustrates a training process of a feature translation model, in accordance with an embodiment of the present disclosure. FIG. 4 is described in conjunction with elements from FIGs. 1, 2, and 3. With reference to FIG. 4, there is shown a block diagram **400** that illustrates a training process of a feature translation model. The block diagram **400** includes a training dataset **402**, a first feature extraction model **404**, a second feature extraction model **406**, a first feature vector **408**, a second feature vector **410**, a feature translation model **412**, a third feature vector **414** and a loss function **416**.

The feature translation model **412** can be trained as follows. The old feature vector extraction model can be represented as the first feature extraction model **404** (also represented as feature vector extraction model #1) and an updated feature vector extraction model can be represented as the second feature extraction model **406** (also represented as feature vector extraction model #2). The training dataset **402**, the first feature extraction model **404** (i.e., feature vector extraction model #1) and the second feature extraction model **406** (i.e., feature vector extraction model #2) are used to train the feature translation model **412**.

In order to train the feature translation model **412**, the feature vectors corresponding to the first feature extraction model **404** (i.e., feature vector extraction model #1) and the second feature extraction model **406** (i.e., feature vector extraction model #2) are extracted and are represented as the first feature vector **408** (also represented as feature vectors #1) and the second feature vector **410** (also represented as feature vectors #2), respectively. In this way, each training data point receives a pair of feature vectors that is the first feature vector **408** and the second feature vector **410**. The first feature vector **408** is considered as an input to the feature translation model **412** and an output of the feature translation model **412** can be represented as the third feature vector **414** (also represented as feature vectors #3). However, the third feature vector **414** (i.e.,

the feature vectors #3) has the same feature dimensionality with the second feature vector **410** (i.e., the feature vectors #2).

Optionally, the training dataset **402** can be fused with the first feature vector **408** (i.e, the feature vectors #1) to act as the input to the feature translation model **412**. The feature translation model **412** can be implemented using a machine learning model. Examples of implementation of the machine learning model may include but are not limited to, auto-encoders, generative adversarial networks (GAN), and the like. In order to converge the third feature vector **414** (i.e., the feature vectors #3) to the second feature vector **410** (i.e., the feature vectors #2), either the loss function **416** between the third feature vector **414** (i.e., the feature vectors #3) and the second feature vector **410** (i.e., the feature vectors #2) is enforced or relaxed criteria is set between the third feature vector **414** and the second feature vector **410** so that the third feature vector **414** generates same topological properties as the second feature vector **410** generates. For example, the third feature vector **414** may have the same closest neighbors for each data point within the training dataset **402**, after converging to the second feature vector **410**.

The feature translation model **412** has lower complexity and runtime overhead in comparison to the new feature extraction model. The feature translation model **412** can get full information from the training dataset **402** which is used to develop the new feature extraction model through the design of input. The feature translation model **412** considers the existing feature vectors (i.e., pre-existing vector database or the old vector database) as the input that is already in use for the search service, which means that the feature translation model **412** is not learning from scratch (i.e., from the raw data points) as the new feature extraction model is generated from scratch. Moreover, the training process of the feature translation model **412** can be seen as a model distillation procedure that learns directly from the new feature extraction model than from the training labels. The loss function **416** can be designed with relaxation to enforce the second feature vector **410** (i.e., the feature vectors #2) and the third feature vector **414** (i.e., the feature vectors #3) to produce the same search results, for example, ranking the results in the database inspite of strictly forcing the results to be the same representations through learning. The feature translation model **412** can be switched on or switched off to allow the A/B test without updating the vector database, depending on an application scenario.

FIG. 5 illustrates processing of an old vector database to a new vector database through feature translation, in accordance with an embodiment of the present disclosure. FIG. 5 is described in conjunction with elements from FIGs. 1, 2, 3, and 4. With reference to FIG. 5, there is shown a

processing pipeline **500** that illustrates processing of a pre-existing vector database **502** (i.e., an old vector database) to a new vector database **508** through feature translation. The processing pipeline **500** includes operations **504** and **506**.

Initially, the pre-existing vector database **502** (i.e., the old vector database) is configured. Thereafter, at operation **504**, the pre-existing vector database **502** is translated to a new vector domain using the feature translation model **412** (of FIG. 4). At operation **506**, a database is generated corresponding to the new vector domain of the pre-existing vector database **502**, which further results in the generation of the new vector database **508**.

The generated new vector database **508** replaces the existing one that is the pre-existing vector database **502** (i.e., the old vector database) and serves as the new backend vector database for search service. Thus, instead of processing the first set of results (i.e., the top-N closest matches) using the feature translation model **412** in the online mode as described in FIGs. 2 and 3, the old vector database **502** is translated and reorganized into the new vector database **508** in an offline mode. Since, the feature translation model **412** manifests low complexity and runtime overhead, the time and resource costs should be much lower than extracting new vector database **508** directly from the new feature extraction model. Thus, the processing of the old vector database **502** to the new vector database **508** using the feature translation model **412** in the offline mode provides a fast and cost-effective feature extraction model update in a search engine.

Modifications to embodiments of the present disclosure described in the foregoing are possible without departing from the scope of the present disclosure as defined by the accompanying claims. Expressions such as "including", "comprising", "incorporating", "have", "is" used to describe and claim the present disclosure are intended to be construed in a non-exclusive manner, namely allowing for items, components or elements not explicitly described also to be present. Reference to the singular is also to be construed to relate to the plural. The word "exemplary" is used herein to mean "serving as an example, instance or illustration". Any embodiment described as “exemplary” is not necessarily to be construed as preferred or advantageous over other embodiments and/or to exclude the incorporation of features from other embodiments. The word "optionally" is used herein to mean "is provided in some embodiments and not provided in other embodiments". It is appreciated that certain features of the present disclosure, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of

the present disclosure, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable combination or as suitable in any other described embodiment of the disclosure.

CLAIMS

1. A method (100) for responding to a search query generated by a user, the method (100) comprising:

extracting one or more feature vectors from the search query;

comparing the extracted feature vectors with a set of corresponding feature vectors in a pre-existing vector database (208, 502);

generating a first set of results based on the comparison;

translating the extracted feature vectors and the first set of results to a new vector domain;

comparing the translated extracted feature vectors and the translated first set of results; and

generating a second set of results based on the comparison, wherein the number of results in the second set is less than or equal to the number of results in the first set.

2. The method (100) of claim 1, further comprising at least one of the following:

providing the second set of results to the user; or

generating a third set of results based on the second set of results, and providing the third set of results to the user.

3. The method (100) of claim 1, further comprising ranking the results in the first set after the comparison with the translated extracted feature vectors.

4. The method (100) of claim 3, wherein the ranking is based on feature vectors extracted using an updated feature extraction model.

5. The method (100) of claim 3 or 4, further comprising aggregating the ranked results and the first set of results to generate the second set of results.

6. The method (100) of claim 5, further comprising omitting the first set of results or the ranked results during aggregation.

7. The method (100) of any preceding claim, wherein comparing comprises determining the similarity or relevance between the extracted feature vectors from the search query and the corresponding feature vectors in the pre-existing vector database (208, 502).

8. The method (100) of any preceding claim, further comprising obtaining raw data points of the first set of results.

9. The method (100) of any preceding claim, further comprising translating the pre-existing vector database (208, 502) in its entirety to construct a new vector database (508).

10. The method (100) of any preceding claim, wherein translating the feature vectors comprises training a machine learning model with a batch of training data.

11. A computer program comprising instructions which, when executed by a computer, cause the computer to carry out the steps of the method (100) of any one of claims 1 to 10.

12. The computer program of claim 11, stored on a non-transitory computer-readable medium.

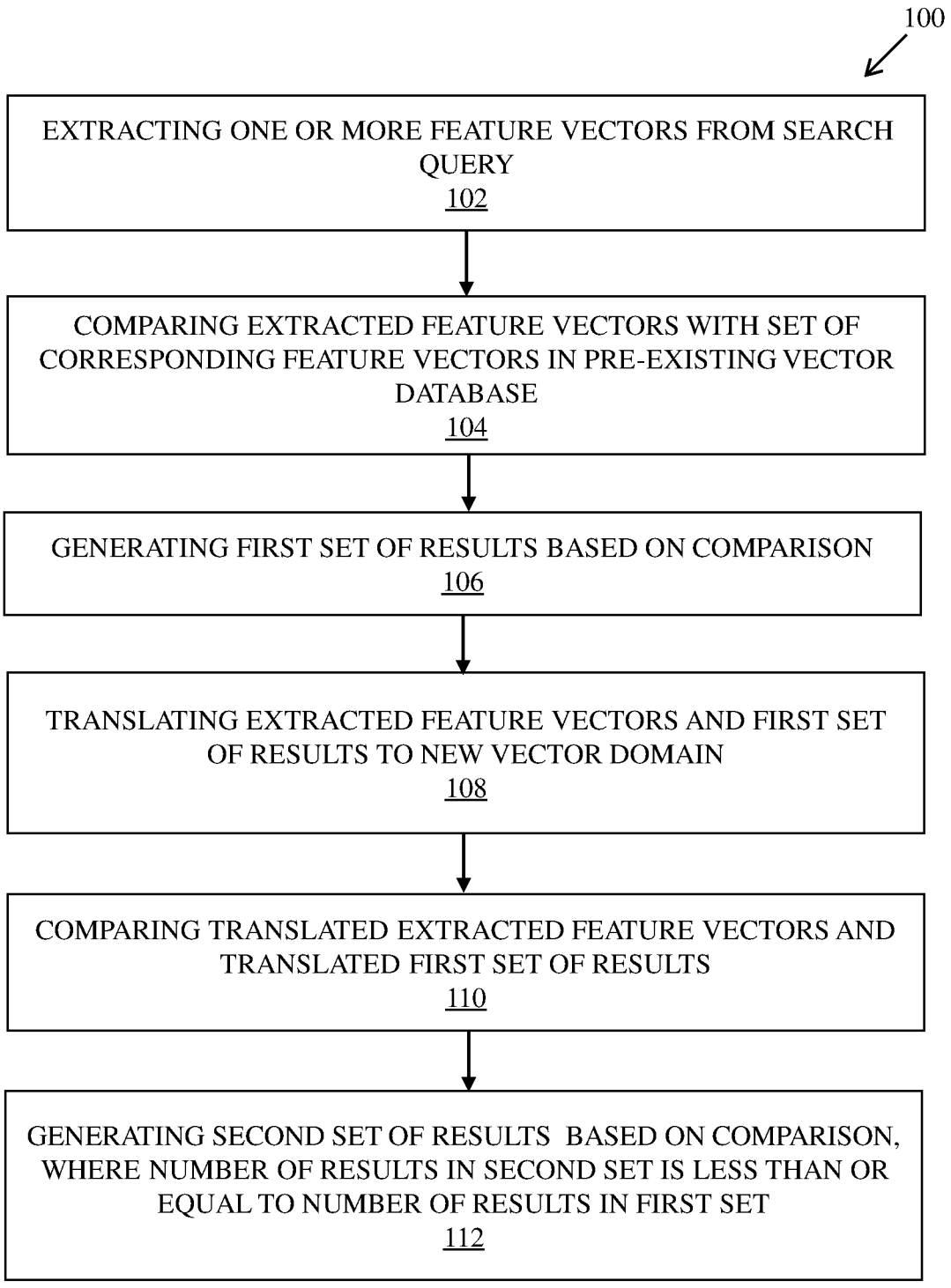
100
EXTRACTING ONE OR MORE FEATURE VECTORS FROM SEARCH QUERY
102
COMPARING EXTRACTED FEATURE VECTORS WITH SET OF CORRESPONDING FEATURE VECTORS IN PRE-EXISTING VECTOR DATABASE
104
GENERATING FIRST SET OF RESULTS BASED ON COMPARISON
106
TRANSLATING EXTRACTED FEATURE VECTORS AND FIRST SET OF RESULTS TO NEW VECTOR DOMAIN
108
COMPARING TRANSLATED EXTRACTED FEATURE VECTORS AND TRANSLATED FIRST SET OF RESULTS
110
GENERATING SECOND SET OF RESULTS BASED ON COMPARISON, WHERE NUMBER OF RESULTS IN SECOND SET IS LESS THAN OR EQUAL TO NUMBER OF RESULTS IN FIRST SET
112

FIG. 1

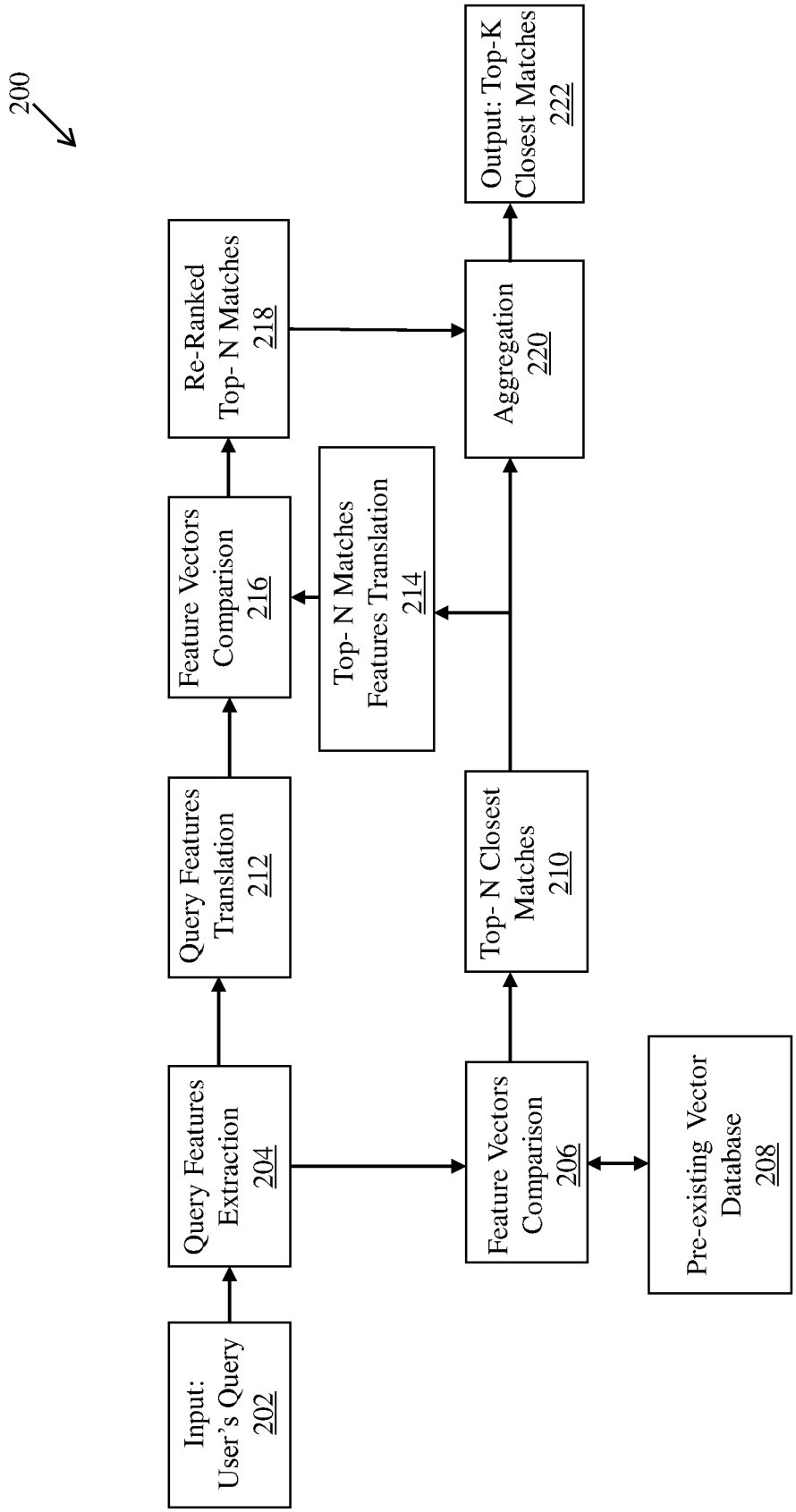
200
Input: User's Query 202
Query Features Extraction 204
Feature Vectors Comparison 206
Pre-existing Vector Database 208
Top- N Closest Matches 210
Query Features Translation 212
Top- N Matches Features Translation 214
Feature Vectors Comparison 216
Re-Ranked Top- N Matches 218
Aggregation 220
Output: Top-K Closest Matches 222

FIG. 2

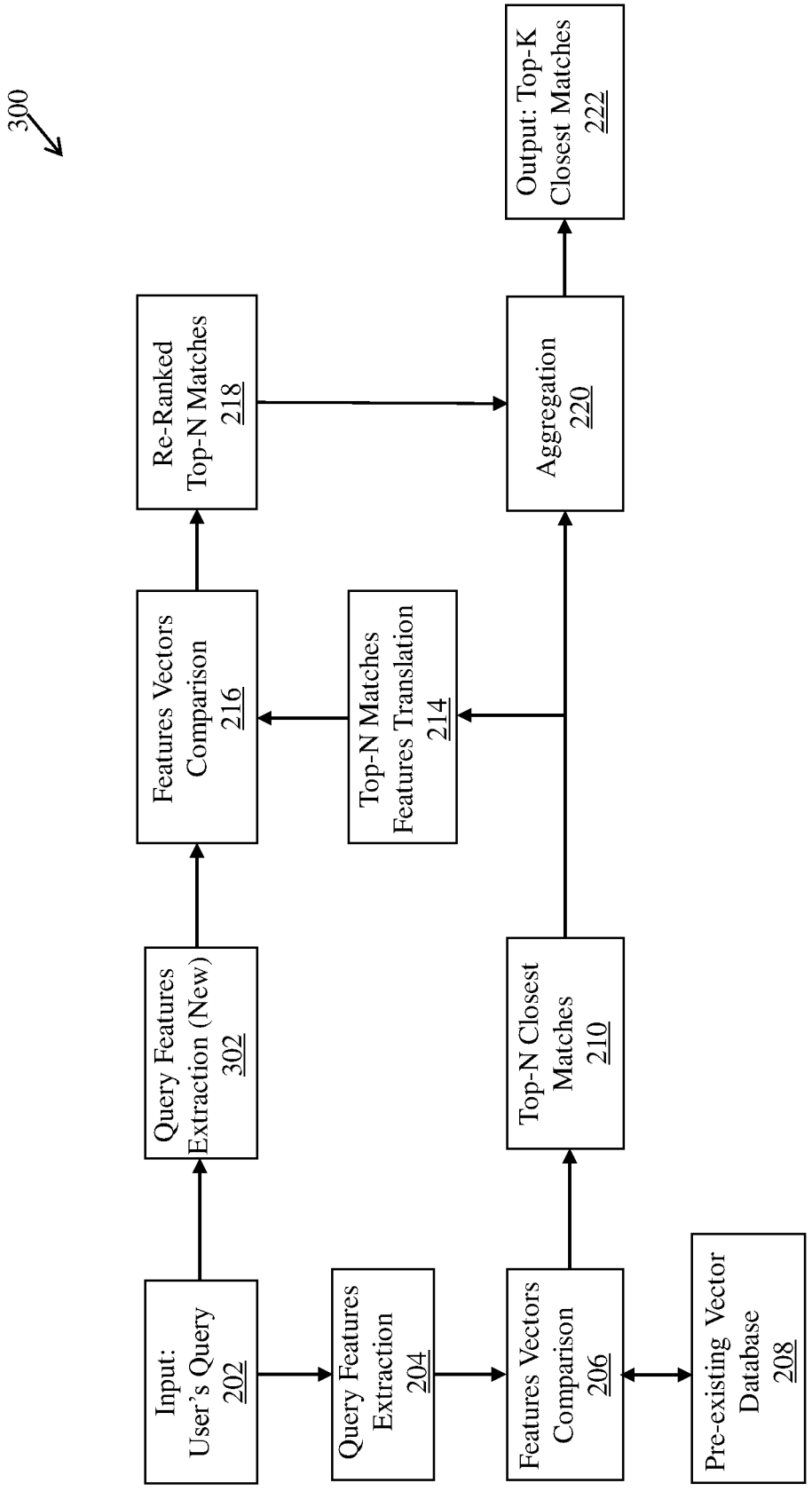
300
Input: User's Query 202
Query Features Extraction 204
Features Vectors Comparison 206
Pre-existing Vector Database 208
Query Features Extraction (New) 302
Top-N Closest Matches 210
Features Vectors Comparison 216
Top-N Matches Features Translation 214
Re-Ranked Top-N Matches 218
Aggregation 220
Output: Top-K Closest Matches 222

FIG. 3

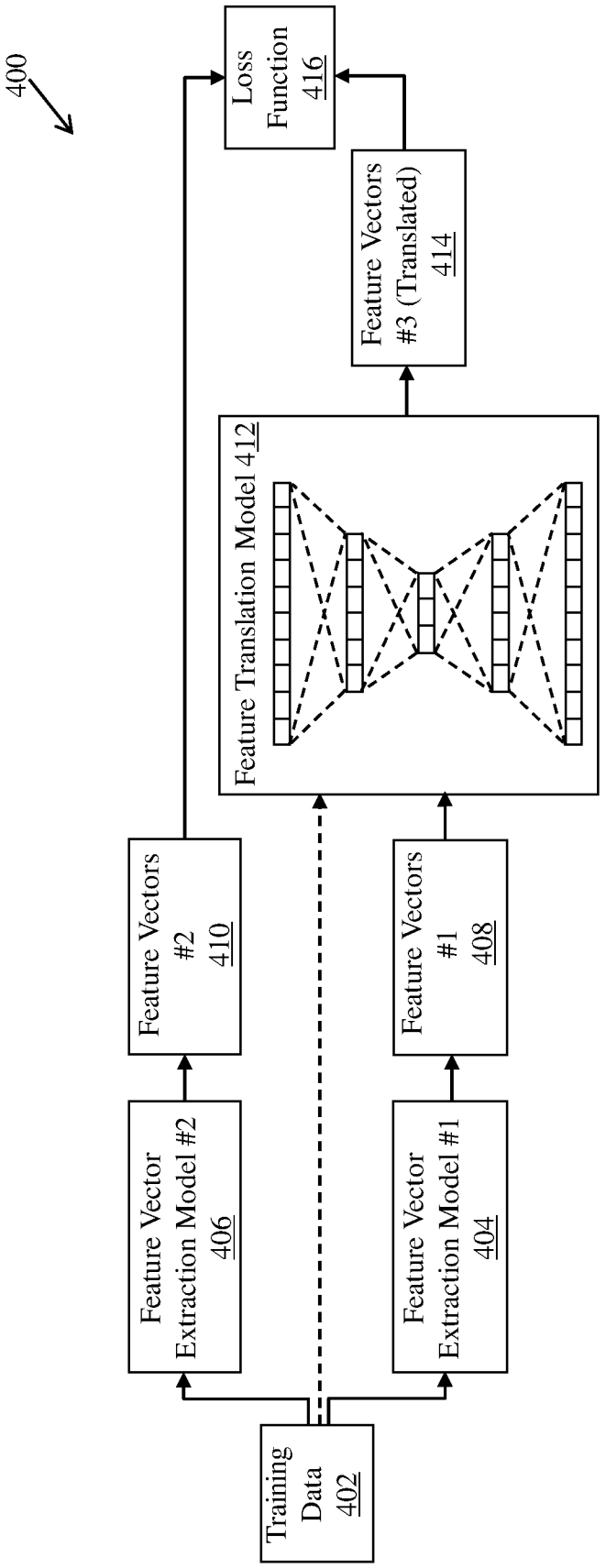
400
Loss Function 416
Feature Vectors #3 (Translated) 414
Feature Translation Model 412
Feature Vectors #2 410
Feature Vectors #1 408
Feature Vector Extraction Model #2 406
Feature Vector Extraction Model #1 404
Training Data 402

FIG. 4

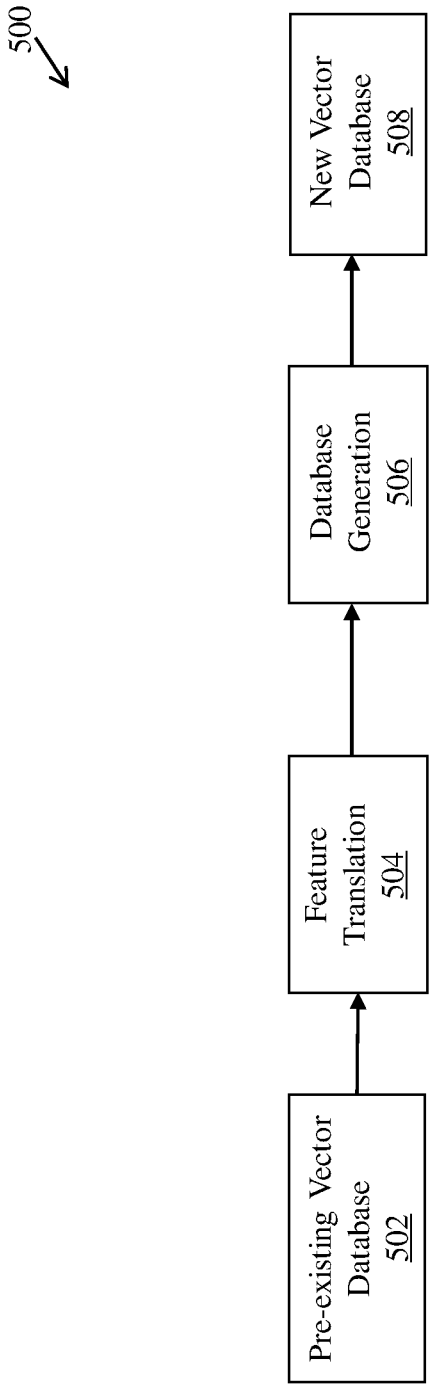
500
Pre-existing Vector Database 502
Feature Translation 504
Database Generation 506
New Vector Database 508

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2021/084775

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F16/53
ADD.

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)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017/068737 A1 (HO TIN K [US] ET AL) 9 March 2017 (2017-03-09) paragraph [0043] - paragraph [0049] paragraph [0078] - paragraph [0081] -----	1-12
A	DAY OSCAR ET AL: "A survey on heterogeneous transfer learning", JOURNAL OF BIG DATA, vol. 4, no. 1, 26 September 2017 (2017-09-26), pages 1-42, XP055822584, DOI: 10.1186/s40537-017-0089-0 Retrieved from the Internet: URL:http://link.springer.com/article/10.1186/s40537-017-0089-0/fulltext.html> abstract -----	1-12

☐ 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	Date of mailing of the international search report
22 July 2022	02/08/2022

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Correia Martins, F

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2021/084775

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2017068737 A1	09-03-2017	US 2017068734 A1	09-03-2017
		US 2017068737 A1	09-03-2017