说 明 书

一种类目预测方法、装置、存储介质及计算机系统

技术领域

本说明书一个或多个实施例涉及人工智能领域，尤其涉及一种类目预测方法、装置、存储介质及计算机系统。

背景技术

在互联网服务中，搜索功能在获取信息、寻找资源、解决问题等方面发挥了重要的作用。例如，在电子商务网站中，搜索功能是用户寻找和购买商品或服务的重要入口。用户在搜索栏中输入想要搜索的内容，系统会返回自身数据库中与搜索内容相匹配的项目列表。

当一个系统中的数据库中的项目数量较少时，例如，在开拓新市场初期时，用户在网站中搜索某些内容时，现有的搜索方法可能无法在自身数据库中匹配到相应的结果，进而返回一个空列表，造成空搜。空搜会降低用户的使用意愿，进而影响网站的用户活跃度和用户留存率。因此，需要一种更好的方法，将用户的搜索内容匹配到数据库中相应的项目上。

发明内容

本说明书实施例提供了一种类目预测方法、装置、存储介质及计算机系统，通过对输入的搜索词进行类目预测，根据得到的类目预测结果匹配相应的项目作为搜索结果，以降低空搜率。

本说明书提供了一种类目预测方法，所述方法包括：获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

本说明书提供了一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如下步骤：获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；根据各个空搜词及其对

应的第一类目预测结果，构建类目数据库；根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

本说明书提供了一种类目预测装置，所述装置包括：获取模块，用于获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；预测模块，用于将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；构建模块，用于根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；训练模块，用于根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

本说明书提供了一种计算机可读非易失性存储介质，其特征在于，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述的类目预测方法。

本说明书实施例采用的上述至少一个技术方案能够达到以下有益效果中的至少一项：通过大语言模型预测类目，以扩充类目数据库，进而训练类目预测模型。使用类目数据库与类目预测模型相结合的离线加实时的方案，可以同时兼顾类目预测的效率和准确性。进而降低搜索过程中的空搜率，提高用户活跃度和用户留存率。同时，在训练过程中无需人工标注，降低了模型训练的成本。

附图说明

此处所说明的附图用来提供对本说明书的进一步理解，构成本说明书的一部分，本说明书的示意性实施例及其说明用于解释本说明书，并不构成对本说明书的不当限定。在附图中：

图 1 为本说明书中一种类目预测方法的实施场景示意图；

图 2 为本说明书中一种类目预测方法的流程示意图；

图 3 为本说明书中一种类目预测装置示意图；

图 4 为本说明书实施例提供的计算机系统的结构示意图。

具体实施方式

为使本说明书的目的、技术方案和优点更加清楚，下面将结合本说明书具体实施例及相应的附图对本说明书技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本说明书一部分实施例，而不是全部的实施例。基于本说明书中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本说明书保护的范围。

如前所述，用户搜索时返回空的结果列表的空搜问题会影响网站的用户活跃度和用户

留存率。例如，网站可以是一个外卖网站，当用户输入“搜索词 A”时，搜索词 A 例如是一家餐厅的名称，但是外卖网站的合作商中并没有该餐厅。如果此时直接返回空列表，则会造成空搜情况，用户在发现自己搜索不到想要的结果时，可能会降低继续使用本网站的意愿。此时，如果对该“搜索词 A”进行分析，可以发现该“搜索词 A”是一家餐厅，进一步地，例如是一家主要经营炸鸡和汉堡的餐厅，则外卖网站的搜索引擎可以在网站的合作商中搜索卖炸鸡和汉堡的餐厅，作为本次搜索的结果返回给用户。如此一来，通过根据搜索词分析用户的意图，并返回相关的内容，一方面可以降低空搜率，另一方面也可以丰富搜索结果列表中的内容。根据用户输入的搜索词预测用户可能想要查找的项目类别或类目的过程，被称为类目预测（Category Prediction）。

通过上述的分析可知，类目预测是搜索系统中的一个非常重要的功能。类目预测可以提高搜索结果的质量和相关性，同时帮助用户快速定位感兴趣的项目范围，以提升用户的搜索体验。

基于此，本说明书实施例提出一种类目预测方法。图 1 为本说明书中一种类目预测方法的实施场景示意图。如图 1 所示，类目预测总共分为如（a）、（b）、（c）示出的三个部分，图 1（a）示出了构建类目数据库与训练类目预测模型的过程；图 1（b）示出了使用类目数据库和类目预测模型进行类目预测的过程；图 1（c）示出了根据类目预测结果，匹配搜索词对应的搜索结果的过程。

如图 1（a）所示，首先，获取空搜词集合，其中包含多个空搜词，空搜词指的是，对应搜索结果为空的搜索词。然后，将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的对应的类目预测结果。接下来，根据各个空搜词及其对应的类目预测结果，构建类目数据库。类目数据库中保存的内容例如可以是，以空搜词-类目预测结果组成的对子（pair）。然后，以类目数据库中的空搜词为样本，以类目预测结果为标签，训练类目预测模型。

在训练得到类目预测模型后，接下来，如图 1（b）所示，首先，获取用户输入的搜索词，将搜索词在所述类目数据库中进行搜索，当搜索成功时，将搜索结果确定为该搜索词对应的目标类目；当搜索失败时，将该搜索词输入到类目预测模型中进行类目预测，将其预测结果确定为该搜索词对应的目标类目。

在得到搜索词对应的目标类目后，如图 1（c）所示，将目标类目输入到查询解析服务引擎中，使其在预设的项目数据库中查询与目标类目匹配度排名靠前的若干项目，作为该搜索词对应的搜索结果，并返回给用户。

以下结合具体的实施例，描述上述类目预测的具体实施步骤。

图 2 为本说明书中一种类目预测方法的流程示意图，所述方法的执行主体可以为任何具有计算、处理能力的平台或服务器或设备集群等。如图 2 所示，所述方法至少包括：步骤 202，获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；步骤 204，将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；步骤 206，根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；步骤 208，根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

下面描述上述各个步骤的具体执行过程。

首先，在步骤 202，获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词。

空搜词即为，在执行本说明书所述的类目预测方法的平台上调用搜索功能，对搜索词进行搜索时，返回空的搜索结果列表的搜索词。

然后，在步骤 204，将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果。

可以使用任意的公开的或者私人训练的大语言模型（Large Language Models，LLMs）。例如，可以使用 ChatGPT、Gemini、Claude 等等，这里不做限定。

通过组装相应的提示词（prompt），并将提示词输入到大语言模型中，可以使大语言模型进行类目预测。

在一个实施例中，步骤 204 具体包括：将各个空搜词中的目标空搜词输入到第一提示词模板中，得到目标提示词；所述目标提示词还包含多个预设类目，用于指示所述大语言模型基于所述多个预设类目对所述目标空搜词进行类目预测；将所述目标提示词输入到所述大语言模型中，得到所述目标空搜词对应的第一类目预测结果。

在一个更加具体的实施例中，所述目标提示词中还包含各个预设类目的描述信息、以及各个预设类目之间的层级关系。

第一提示词模板中可以包含，用于指示大语言模型进行类目预测的语句，以及第一占位符（placeholder）和第二占位符，其中，第一占位符中需填入多个预设类目和各个类目之间的层级关系，可选地，还可以填入各个类目的描述信息；第二占位符中需填入空搜词。可选地，第一提示词模板中还可以包含第三占位符，第三占位符中需填入执行本说明书所述的类目预测方法的平台所在的国家或地区。

示例性地，第一提示词模板可以如下所示：

“you are expert in the ecommerce, and experienced in the search, now I have several

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

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

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

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

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

其中，{placeholder1}可以是第一占位符，其中需要填入多个预设类目和各个类目之间的层级关系。例如，{placeholder1}中可以填入如下内容：

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

{placeholder2}可以是第二占位符，其中需要填入一个或多个空搜词。

将空搜词和预设类目的信息输入到第一提示词模板之后，可以得到目标提示词。将目标提示词输入到大语言模型中，可以得到各个空搜词的第一类目预测结果。

示例性地，针对某个空搜词的第一类目预测结果可以如下所示：

“{"token1": "1&2, 1&5"}”。

使用大语言模型进行类目预测，可以得到各个空搜词的类目预测结果。然而，由于大语言模型的参数量过大，导致其推理速度会相对较慢；同时，由于大语言模型被设计为通用目的，当直接使用大语言模型进行类目预测时，则需要与前述过程类似地，组装相对复杂的提示词才能实现类目预测的效果。

为了克服以上两点问题，本说明书实施例基于知识蒸馏的思想，利用大语言模型的回答训练一个参数量相对较小的、专用的类目预测模型，以提高推理速度，并降低使用难度。

接下来，回到图 2，在步骤 206，根据各个空搜词及其对应的第一类目预测结果，构建类目数据库。

类目数据库中可以包括，各个空搜词和对应的第一类目预测结果所组成的空搜词和-第一类目预测结果对子；或者，可以是以空搜词为主键，第一类目预测结果为字段的数据库条目。可以使用多种数据库软件来构造类目数据库，例如可以使用 HBASE 等等，这里不做限定。

然后，在步骤 208，根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

具体地，将所述类目数据库中的各个空搜词作为训练样本，将各个空搜词对应的第一类目预测结果作为标签，训练类目预测模型。

在一个实施例中，将任一空搜词输入到类目预测模型中，得到衡量该空搜词与各个预设类目的匹配程度的第一预测分数；根据该空搜词对应的第一类目预测结果中包含的一个或多个类目，确定衡量其标签与各个预设类目的匹配程度的第二预测分数；根据所述第一预测分数和第二预测分数之间的差异所确定的损失值，调整所述类目预测模型中的参数值。

类目预测模型输出的预测结果可以是，衡量该空搜词与各个预设类目的匹配程度的第一预测分数。空搜词与某个预设类目的匹配程度越高，则该预设类目对应的预测分数值越高。

由于标签中包含的是大语言模型针对空搜词输出的一个或多个类目，其并不直接包含预测分数。因此，需要使用特定的映射方法，将标签映射为预测分数。例如，可以将标签中包含的各个类目对应的预测分数都设置为 1，而将未包含在标签中的各个类目对应的预测分数都设置为 0；又例如，当标签中的各个类目之间有先后顺序时，可以将标签中的各个类目按照先后顺序递减地设置预测分数，将未包含在标签中的各个类目对应的预测分数都设置为 0。例如，将标签中第一个类目的预测分数设置为 1，第二个类目的预测分数设置为 0.8 等等。

可以使用多种现有的模型作为类目预测模型。例如，可以使用 BERT（Bidirectional Encoder Representations from Transformers）模型与多层神经网络 MLP（MultiLayer Perceptron）的组合，作为类目预测模型。

在训练类目预测模型的过程中，可以使用多种损失函数以计算类目预测模型针对于各个预设类目的预测分数与标签中各个预设类目的预测分数之间的损失值，在反向传播过通过梯度下降法，调整类目预测模型中的各个参数的值。例如，可以使用基于 sigmoid 函数的交叉熵损失函数，例如 sigmoid_cross_entropy。这里不做限定。

通过如步骤 202 至步骤 208 所述的方法，收集大语言模型针对于各个空搜词的类目预测结果，并保存到类目数据库中。同时，利用类目数据库中的数据进行知识蒸馏，训练参数量小且专用的类目预测模型，通过类目数据库与类目预测模型的结合，可以针对用户输入的搜索词进行高效且准确的类目预测。

在一些可能的实施方式中，在训练得到类目预测模型后，所述方法还包括步骤 210 至

步骤 214，用于对用户输入的搜索词进行类目预测。

首先，在步骤 210，获取用户输入的搜索词。

用户输入的搜索词可以是，用户在网站、APP 等界面的搜索框内输入的搜索词，也可以是用户在对话聊天界面调用搜索功能时输入的搜索词。

在接收到搜索词后，在一些实施例中，还可以先对搜索词进行预处理，为搜索词与前述的空搜词统一格式。例如，针对于搜索词中的中文部分，可以将繁体转换为简体，以及将句子进行分词；针对于搜索词中的英文部分，可以将大写字母转换为小写字母。

然后，在步骤 212，将所述搜索词在所述类目数据库中进行搜索，当搜索成功时，将搜索得到的第二类目预测结果确定为所述搜索词对应的目标类目。

将所述搜索词在所述类目数据库中进行搜索可以是，搜索包含所述搜索词的空搜词-第一类目预测结果的对子，将搜索得到的对子中包含的第一类目预测结果作为第二类目预测结果；或者，搜索以所述搜索词为主键的数据库条目，将搜索得到的条目中的第一类目预测结果字段，作为第二类目预测结果。

当步骤 212 搜索成功时，可以将搜索结果作为该搜索词的类目预测结果，或者称为目标类目。当步骤 212 搜索失败时，则接着在步骤 214，将所述搜索词输入到所述类目预测模型中，将其输出的第三类目预测结果确定为所述搜索词对应的目标类目。

类目预测模型输出的第三类目预测结果可以包含一个类目，也可以包含多个类目。例如，可以将类目预测模型针对于各个预设类目输出的预测分数大于预设阈值的各个类目，作为第三类目预测结果。

通过步骤 210 至步骤 214 所述的方法，对于类目数据库中已有的搜索词，直接输出保存的类目预测结果；对于无法在类目数据库中搜索到的搜索词，则调用类目预测模型进行类目预测。通过类目数据库与类目预测模型的结合，可以针对用户输入的搜索词进行高效且准确的类目预测。

在一些可能的实施方式中，在通过步骤 210 至步骤 214 确定所述搜索词对应的目标类目后，所述方法还包括：

在步骤 216，根据所述目标类目，确定所述搜索词对应的搜索结果。

通过搜索词对应的目标类目，可以从平台的项目数据库中查询与目标类目匹配度高的项目，作为搜索词对应的搜索结果。

具体地，将所述目标类目输入到查询解析（Query Parse）服务引擎中，使其在预设的项目数据库中查询与所述目标类目匹配度排名靠前的若干项目；然后，将所述若干项目确

定为所述搜索词对应的搜索结果。

通过步骤 202 至步骤 216，可以为用户的搜索词匹配到相关度更高的项目，同时降低空搜率。同时，由于无需人工标注，本说明书示出的方法可以自动化地完成，节约了人工成本。

由于在步骤 212 中搜索失败的搜索词，实际上是没有参与训练过类目预测模型的样本。因此，在一些可能的实施方式中，在通过步骤 210 至步骤 214 确定所述搜索词对应的目标类目后，所述方法还包括：

在步骤 218，获取在所述类目数据库中搜索失败的若干目标搜索词。

由于类目预测模型是使用类目数据库进行训练的，如果一个搜索词不包含在类目数据库中，它便也不会作为训练样本参与到类目预测模型的训练中。因此，可以在搜索功能上线后，持续收集这部分搜索词，然后定期更新类目数据库和类目预测模型。

在步骤 220，将各个目标搜索词分别输入到大语言模型中进行类目预测，得到各个目标搜索词的第四类目预测结果。

步骤 220 的具体实施过程可以参照步骤 204，这里不再赘述。

在步骤 224，将各个目标搜索词及其对应的第四类目预测结果添加到所述类目数据库中，得到更新的类目数据库。

将各个目标搜索词及其对应的第四类目预测结果添加到所述类目数据库中时，可以使用与类目数据库中已有数据相同的保存形式。

在步骤 226，根据更新的类目数据库，更新所述类目预测模型。

步骤 226 的具体实施过程可以参照步骤 208，这里不再赘述。

通过上述步骤，可以在事实本说明书实施例中示出的类目预测方法的搜索功能上线后，继续定期更新类目数据库和类目预测模型，使得类目预测的整体效果不断提升。

在步骤 202 至步骤 208 训练得到类目预测模型后，还可以对类目预测模型的准确性进行评估。在一些可能的实施方式中，在通过步骤 220 得到各个目标搜索词的第四类目预测结果后，所述方法还包括：

在步骤 222，根据各个目标搜索词对应的第三类目预测结果和第四类目预测结果之间的相似度，评估所述类目预测模型的预测准确度。

类目预测模型的预测准确度与第三类目预测结果和第四类目预测结果之间的相似度可以是正相关，即相似度越高，则类目预测模型的准确度越高。

可以使用多种方法衡量第三类目预测结果和第四类目预测结果之间的相似度，例如，

可以使用欧氏距离、余弦相似度等，这里不做限定。

以上为本说明书实施例提供的类目预测方法，基于同样的思路，本说明书还提供了相应的类目预测装置。如图 3 所示，该类目预测装置包括：获取模块 302，用于获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；预测模块 304，用于将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；构建模块 306，用于根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；训练模块 308，用于根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

为了描述的方便，描述以上装置时以功能分为各种模块分别描述。当然，在实施本说明书时可以把各模块的功能在同一个或多个软件和/或硬件中实现。

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

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

may encompass one or more computer systems, where appropriate.

This disclosure contemplates any suitable number of computer systems **1000**. This disclosure contemplates computer system **1000** taking any suitable physical form. As example and not by way of limitation, computer system **1000** may be an embedded computer system, a system-on-chip (SOC), a single-board computer system (SBC) (such as, for example, a computer-on-module (COM) or system-on-module (SOM)), a desktop computer system, a laptop or notebook computer system, an interactive kiosk, a mainframe, a mesh of computer systems, a mobile telephone, a personal digital assistant (PDA), a server, a tablet computer system, or a combination of two or more of these. Where appropriate, computer system **1000** may include one or more computer systems **1000**; be unitary or distributed; span multiple locations; span multiple machines; span multiple data centers; or reside in a cloud, which may include one or more cloud components in one or more networks. Where appropriate, one or more computer systems **1000** may perform without substantial spatial or temporal limitation one or more steps of one or more methods described or illustrated herein. As an example and not by way of limitation, one or more computer systems **1000** may perform in real time or in batch mode one or more steps of one or more methods described or illustrated herein. One or more computer systems **1000** may perform at different times or at different locations one or more steps of one or more methods described or illustrated herein, where appropriate.

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

In particular embodiments, processor **1002** includes hardware for executing instructions, such as those making up a computer program. As an example and not by way of limitation, to execute instructions, processor **1002** may retrieve (or fetch) the instructions from an internal register, an internal cache, memory **1004**, or storage **1006**; decode and execute them; and then write one or more results to an internal register, an internal cache, memory **1004**, or storage **1006**. In particular embodiments, processor **1002** may include one or more internal caches for data, instructions, or addresses. This disclosure contemplates processor **1002** including any suitable number of any suitable internal caches, where appropriate. As an example and not by way of limitation, processor **1002** may include one or more instruction caches, one or more data caches, and one or more translation lookaside buffers (TLB s). Instructions in the instruction caches may

be copies of instructions in memory **1004** or storage **1006**, and the instruction caches may speed up retrieval of those instructions by processor **1002**. Data in the data caches may be copies of data of **80** in memory **1004** or storage **1006** for instructions executing at processor **1002** to operate on; the results of previous instructions executed at processor **1002** for access by subsequent instructions executing at processor **1002** or for writing to memory **1004** or storage **1006**; or other suitable data. The data caches may speed up read or write operations by processor **1002**. The TLBs may speed up virtual-address translation for processor **1002**. In particular embodiments, processor **1002** may include one or more internal registers for data, instructions, or addresses. This disclosure contemplates processor **1002** including any suitable number of any suitable internal registers, where appropriate. Where appropriate, processor **1002** may include one or more arithmetic logic units (ALUs); be a multi-core processor; or include one or more processors **1002**. Although this disclosure describes and illustrates a particular processor, this disclosure contemplates any suitable processor.

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

memories **1004**, where appropriate. Although this disclosure describes and illustrates particular memory, this disclosure contemplates any suitable memory.

In particular embodiments, storage **1006** includes mass storage for data or instructions. As an example and not by way of limitation, storage **1006** may include a hard disk drive (HDD), a floppy disk drive, flash memory, an optical disc, a magneto-optical disc, magnetic tape, or a Universal Serial Bus (USB) drive or a combination of two or more of these. Storage **1006** may include removable or non-removable (or fixed) media, where appropriate. Storage **1006** may be internal or external to computer system **1000**, where appropriate. In particular embodiments, storage **1006** is non-volatile, solid-state memory. In particular embodiments, storage **1006** includes read-only memory (ROM). Where appropriate, this ROM may be mask-programmed ROM, programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), electrically alterable ROM (EAROM), or flash memory or a combination of two or more of these. This disclosure contemplates mass storage **1006** taking any suitable physical form. Storage **1006** may include one or more storage control units facilitating communication between processor **1002** and storage **1006**, where appropriate. Where appropriate, storage **1006** may include one or more storages **1006**. Although this disclosure describes and illustrates particular storage, this disclosure contemplates any suitable storage.

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

In particular embodiments, communication interface **1010** includes hardware, software, or both providing one or more interfaces for communication (such as, for example, packet-based communication) between computer system **1000** and one or more other computer systems **1000** or one or more networks. As an example and not by way of limitation, communication interface

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

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

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互

相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

以上所述仅为本说明书的实施例而已，并不用于限制本说明书。对于本领域技术人员来说，本说明书可以有各种更改和变化。凡在本说明书的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

权 利 要 求 书

1. 一种类目预测方法，所述方法包括：

获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；

将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；

根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；

根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

2. 如权利要求 1 所述的方法，所述方法还包括：

获取用户输入的搜索词；

将所述搜索词在所述类目数据库中进行搜索；

如果搜索成功，将搜索得到的第二类目预测结果确定为所述搜索词对应的目标类目；

如果搜索失败，将所述搜索词输入到所述类目预测模型中，将其输出的第三类目预测结果确定为所述搜索词对应的目标类目。

3. 如权利要求 2 所述的方法，所述方法还包括：

根据所述目标类目，确定所述搜索词对应的搜索结果。

4. 如权利要求 2 所述的方法，所述方法还包括：

获取在所述类目数据库中搜索失败的若干目标搜索词；

将各个目标搜索词分别输入到大语言模型中进行类目预测，得到各个目标搜索词的第四类目预测结果；

将各个目标搜索词及其对应的第四类目预测结果添加到所述类目数据库中，得到更新的类目数据库；

根据更新的类目数据库，更新所述类目预测模型。

5. 如权利要求 4 所述的方法，所述方法还包括：

根据各个目标搜索词对应的第三类目预测结果和第四类目预测结果之间的相似度，评估所述类目预测模型的预测准确度。

6. 如权利要求 1 所述的方法，将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果，包括：

将各个空搜词中的目标空搜词输入到第一提示词模板中，得到目标提示词；所述目标提示词还包含多个预设类目，用于指示所述大语言模型基于所述多个预设类目对所述目标

空搜词进行类目预测；

将所述目标提示词输入到所述大语言模型中，得到所述目标空搜词对应的第一类目预测结果。

7. 如权利要求 6 所述的方法，所述目标提示词中还包含各个预设类目的描述信息、以及各个预设类目之间的层级关系。

8. 如权利要求 1 所述的方法，根据所述类目数据库，训练类目预测模型，包括：

将所述类目数据库中的各个空搜词作为训练样本，将各个空搜词对应的第一类目预测结果作为标签，训练类目预测模型。

9. 如权利要求 8 所述的方法，训练类目预测模型，包括：

将任一空搜词输入到类目预测模型中，得到衡量该空搜词与各个预设类目的匹配程度的第一预测；

根据该空搜词对应的第一类目预测结果中包含的一个或多个类目，确定衡量其标签与各个预设类目的匹配程度的第二预测分数；

根据所述第一预测分数和第二预测分数之间的差异所确定的损失值，调整所述类目预测模型中的参数值。

10. 如权利要求 3 所述的方法，根据所述目标类目，确定所述搜索词对应的搜索结果，包括：

将所述目标类目输入到查询解析服务引擎中，使其在预设的项目数据库中查询与所述目标类目匹配度排名靠前的若干项目；

将所述若干项目确定为所述搜索词对应的搜索结果。

11. 一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如下步骤：

获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；

将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；

根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；

根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

12. 如权利要求 11 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

获取用户输入的搜索词；

将所述搜索词在所述类目数据库中进行搜索；

如果搜索成功，将搜索得到的第二类目预测结果确定为所述搜索词对应的目标类目；

当搜索失败时，将所述搜索词输入到所述类目预测模型中，将其输出的第三类目预测结果确定为所述搜索词对应的目标类目。

13. 如权利要求 12 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

根据所述目标类目，确定所述搜索词对应的搜索结果。

14. 如权利要求 12 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

获取在所述类目数据库中搜索失败的若干目标搜索词；

将各个目标搜索词分别输入到大语言模型中进行类目预测，得到各个目标搜索词的第四类目预测结果；

将各个目标搜索词及其对应的第四类目预测结果添加到所述类目数据库中，得到更新的类目数据库；

根据更新的类目数据库，更新所述类目预测模型。

15. 如权利要求 14 所述的计算机系统，所述处理器执行所述程序时还实现如下步骤：

根据各个目标搜索词对应的第三类目预测结果和第四类目预测结果之间的相似度，评估所述类目预测模型的预测准确度。

16. 如权利要求 11 所述的计算机系统，将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果，包括：

将各个空搜词中的目标空搜词输入到第一提示词模板中，得到目标提示词；所述目标提示词还包含多个预设类目，用于指示所述大语言模型基于所述多个预设类目对所述目标空搜词进行类目预测；

将所述目标提示词输入到所述大语言模型中，得到所述目标空搜词对应的第一类目预测结果。

17. 如权利要求 16 所述的计算机系统，所述目标提示词中还包含各个预设类目的描述信息、以及各个预设类目之间的层级关系。

18. 如权利要求 11 所述的计算机系统，根据所述类目数据库，训练类目预测模型，包括：

将所述类目数据库中的各个空搜词作为训练样本，将各个空搜词对应的第一类目预测结果作为标签，训练类目预测模型。

19. 如权利要求 18 所述的计算机系统，训练类目预测模型，包括：

将任一空搜词输入到类目预测模型中，得到衡量该空搜词与各个预设类目的匹配程度的第一预测分数；

根据该空搜词对应的第一类目预测结果中包含的一个或多个类目，确定衡量其标签与各个预设类目的匹配程度的第二预测分数；

根据所述第一预测分数和第二预测分数之间的差异所确定的损失值，调整所述类目预测模型中的参数值。

20. 如权利要求 13 所述的计算机系统，根据所述目标类目，确定所述搜索词对应的搜索结果，包括：

将所述目标类目输入到查询解析服务引擎中，使其在预设的项目数据库中查询与所述目标类目匹配度排名靠前的若干项目；

将所述若干项目确定为所述搜索词对应的搜索结果。

21. 一种类目预测装置，所述装置包括：

获取模块，用于获取空搜词集合；所述空搜词为对应搜索结果为空的搜索词；

预测模块，用于将各个空搜词分别输入到大语言模型中进行类目预测，得到各个空搜词的第一类目预测结果；

构建模块，用于根据各个空搜词及其对应的第一类目预测结果，构建类目数据库；

训练模块，用于根据所述类目数据库，训练类目预测模型；所述类目预测模型和类目数据库用于，针对用户输入的搜索词，预测其对应的类目。

22. 一种计算机可读非易失性存储介质，其特征在于，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述权利要求 1-10 任一项所述的方法。