(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113254665 A
(43)申请公布日 2021.08.13

(21)申请号 202110610082.1

(22)申请日 2021.06.01

(71)申请人 北京爱奇艺科技有限公司
地址 100080 北京市海淀区海淀北一街2号鸿城拓展大厦11层1101

(72)发明人 毋安安

(74)专利代理机构 北京柏杉松知识产权代理事务所(普通合伙) 11413
代理人 孟维娜 高莺然

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/28(2019.01)

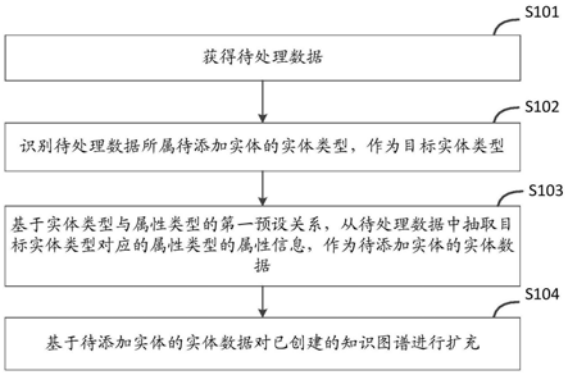
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种知识图谱扩充方法、装置、电子设备及存储介质

(57)摘要

本发明实施例提供了一种知识图谱扩充方法、装置、电子设备及存储介质，涉及数据处理技术领域，包括：获得待处理数据；识别所述待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据；基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。应用本发明实施例提供的方案，可以提高扩充知识图谱的效率。



CN 113254665 A

1.一种知识图谱扩充方法,其特征在于,所述方法包括:

获得待处理数据;

识别所述待处理数据所属待添加实体的实体类型,作为目标实体类型;

基于实体类型与属性类型的第一预设关系,从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息,作为所述待添加实体的实体数据;

基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。

2.根据权利要求1所述的方法,其特征在于,所述基于所述待添加实体的实体数据对已创建的知识图谱进行扩充,包括:

查找已创建的知识图谱中是否存在与所述待添加实体为同一实体的目标实体;

若存在,将所述待添加实体的实体数据与所述目标实体的实体数据进行合并,实现对所述知识图谱的扩充;

若不存在,在所述知识图谱中创建所述待添加实体,并在所述知识图谱中添加所述待添加实体的实体数据。

3.根据权利要求1所述的方法,其特征在于,所述查找已创建的知识图谱中是否存在与所述待添加实体为同一实体的目标实体,包括:

基于实体类型与查找方式的第二预设关系,确定在原始实体库中查找与所述待添加实体为同一实体的目标实体的查找方式,其中,所述原始实体库中包含所述知识图谱中实体的实体数据;

按照所确定的查找方式,查找所述原始实体库是否存在与所述待添加实体为同一实体的目标实体,若为是,确定所述知识图谱中存在所述目标实体。

4.根据权利要求3所述的方法,其特征在于,所述方法还包括:

将所述待添加实体的实体数据添加至所述原始实体库中。

5.根据权利要求3所述的方法,其特征在于,所述按照所确定的查找方式,查找所述原始实体库是否存在与所述待添加实体为同一实体的目标实体,包括:

利用预设的模糊查找方式,从所述原始实体库中查找与所述待添加实体相似的实体,作为候选实体;

按照所确定的查找方式,从所述候选实体中查找是否存在与所述待添加实体为同一实体的目标实体。

6.根据权利要求5所述的方法,其特征在于,所述利用预设的模糊查找方式,从所述原始实体库中查找与所述待添加实体相似的实体,作为候选实体,包括:

查找所述原始实体库中索引信息与目标属性信息相匹配的实体,作为候选实体,其中,每一实体的索引信息包含该实体的实体数据中预设属性信息的属性字段,所述目标属性信息为:所述待添加实体的属性信息中的名称信息。

7.根据权利要求3所述的方法,其特征在于,每一实体类型对应的查找方式为:

对该实体类型所对应属性类型的属性信息进行一一匹配的查找方式。

8.根据权利要求1-7中任一项所述的方法,其特征在于,所述基于所述待添加实体的实体数据对已创建的知识图谱进行扩充,包括:

对所述待添加实体的实体数据进行格式转化,得到所述待添加实体的实体数据的三元组信息,其中,所述三元组信息包含所述待添加实体的名称信息、属性类型信息、属性信息;

将所述三元组信息导入所述知识图谱的图数据库中，基于所述图数据库实现将所述待添加实体的实体数据添加至所述知识图谱中。

9.根据权利要求1-7中任一项所述的方法，其特征在于，所述基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据，包括：

确定所述待处理数据的格式类型，作为目标格式；

从预先设计的数据抽取脚本中，选择用于对所述目标格式的数据进行数据抽取的脚本，作为目标脚本；

基于实体类型与属性类型的第一预设关系，确定所述目标实体类型对应的属性类型，作为目标属性类型；

利用所述目标脚本，从所述待处理数据中抽取所述目标属性类型的属性信息，作为所述待添加实体的实体数据。

10.一种知识图谱扩充装置，其特征在于，所述装置包括：

数据获得模块，用于获得待处理数据；

类型识别模块，用于识别所述待处理数据所属待添加实体的实体类型，作为目标实体类型；

数据抽取模块，用于基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据；

图谱扩充模块，用于基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。

11.一种电子设备，其特征在于，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

存储器，用于存放计算机程序；

处理器，用于执行存储器上所存放的程序时，实现权利要求1-9任一所述的方法。

12.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质内存储有计算机程序，所述计算机程序被处理器执行时实现权利要求1-9任一所述的方法步骤。

一种知识图谱扩充方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及数据处理技术领域，特别是涉及一种知识图谱扩充方法、装置、电子设备及存储介质。

背景技术

[0002] 随着信息时代的到来，各种类型的信息呈爆炸式增长。这些信息所属的不同实体之间可能会存在直接关系或者间接关系，各种应用基于上述直接关系或者间接关系可以为用户提供各种服务。例如，一些应用可以基于上述直接关系或者间接关系为用户推送商品信息、影视信息等等。

[0003] 现有技术中，通常通过构建知识图谱，来存储上述具有直接关系或者间接关系的实体的信息。在构建知识图谱时，一般由工作人员基于已有信息手动构建知识图谱，并且随着信息的增加，可以基于新增的信息对已构建的知识图谱进行扩充，以丰富知识图谱中包含的信息。相关技术中，一般也是由工作人员手动对已构建的知识图谱进行扩充的，从而导致扩充知识图谱的效率较低。

发明内容

[0004] 本发明实施例的目的在于提供一种知识图谱扩充方法、装置、电子设备及存储介质，以提高扩充知识图谱的效率。具体技术方案如下：

[0005] 在本发明实施的第一方面，首先提供了一种知识图谱扩充方法，所述方法包括：

[0006] 获得待处理数据；

[0007] 识别所述待处理数据所属待添加实体的实体类型，作为目标实体类型；

[0008] 基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据；

[0009] 基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。

[0010] 本申请的一个实施例中，所述基于所述待添加实体的实体数据对已创建的知识图谱进行扩充，包括：

[0011] 查找已创建的知识图谱中是否存在与所述待添加实体为同一实体的目标实体；

[0012] 若存在，将所述待添加实体的实体数据与所述目标实体的实体数据进行合并，实现对所述知识图谱的扩充；

[0013] 若不存在，在所述知识图谱中创建所述待添加实体，并在所述知识图谱中添加所述待添加实体的实体数据。

[0014] 本申请的一个实施例中，所述查找已创建的知识图谱中是否存在与所述待添加实体为同一实体的目标实体，包括：

[0015] 基于实体类型与查找方式的第二预设关系，确定在原始实体库中查找与所述待添加实体为同一实体的目标实体的查找方式，其中，所述原始实体库中包含所述知识图谱中实体的实体数据；

[0016] 按照所确定的查找方式，查找所述原始实体库是否存在与所述待添加实体为同一实体的目标实体，若为是，确定所述知识图谱中存在所述目标实体。

[0017] 本申请的一个实施例中，所述方法还包括：

[0018] 将所述待添加实体的实体数据添加至所述原始实体库中。

[0019] 本申请的一个实施例中，所述按照所确定的查找方式，查找所述原始实体库是否存在与所述待添加实体为同一实体的目标实体，包括：

[0020] 利用预设的模糊查找方式，从所述原始实体库中查找与所述待添加实体相似的实体，作为候选实体；

[0021] 按照所确定的查找方式，从所述候选实体中查找是否存在与所述待添加实体为同一实体的目标实体。

[0022] 本申请的一个实施例中，所述利用预设的模糊查找方式，从所述原始实体库中查找与所述待添加实体相似的实体，作为候选实体，包括：

[0023] 查找所述原始实体库中索引信息与目标属性信息相匹配的实体，作为候选实体，其中，每一实体的索引信息包含该实体的实体数据中预设属性信息的属性字段，所述目标属性信息为：所述待添加实体的属性信息中的名称信息。

[0024] 本申请的一个实施例中，每一实体类型对应的查找方式为：

[0025] 对该实体类型所对应属性类型的属性信息进行一一匹配的查找方式。

[0026] 本申请的一个实施例中，所述基于所述待添加实体的实体数据对已创建的知识图谱进行扩充，包括：

[0027] 对所述待添加实体的实体数据进行格式转化，得到所述待添加实体的实体数据的三元组信息，其中，所述三元组信息包含所述待添加实体的名称信息、属性类型信息、属性信息；

[0028] 将所述三元组信息导入所述知识图谱的图数据库中，基于所述图数据库实现将所述待添加实体的实体数据添加至所述知识图谱中。

[0029] 本申请的一个实施例中，所述基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据，包括：

[0030] 确定所述待处理数据的格式类型，作为目标格式；

[0031] 从预先设计的数据抽取脚本中，选择用于对所述目标格式的数据进行数据抽取的脚本，作为目标脚本；

[0032] 基于实体类型与属性类型的第一预设关系，确定所述目标实体类型对应的属性类型，作为目标属性类型；

[0033] 利用所述目标脚本，从所述待处理数据中抽取所述目标属性类型的属性信息，作为所述待添加实体的实体数据。

[0034] 在本发明实施的第二方面，还提供了一种知识图谱扩充装置，所述装置包括：

[0035] 数据获得模块，用于获得待处理数据；

[0036] 类型识别模块，用于识别所述待处理数据所属待添加实体的实体类型，作为目标实体类型；

[0037] 数据抽取模块，用于基于实体类型与属性类型的第一预设关系，从所述待处理数

据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据；

[0038] 图谱扩充模块，用于基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。

[0039] 在本发明实施的第三方面，还提供了一种电子设备，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

[0040] 存储器，用于存放计算机程序；

[0041] 处理器，用于执行存储器上所存放的程序时，实现第一方面任一所述的方法。

[0042] 在本发明实施的又一方面，还提供了一种计算机可读存储介质，所述计算机可读存储介质内存储有计算机程序，所述计算机程序被处理器执行时实现上述任一所述的知识图谱扩充方法。

[0043] 在本发明实施的又一方面，还提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述任一所述的知识图谱扩充方法。

[0044] 本发明实施例提供的知识图谱扩充方案中，可以获得待处理数据；识别待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；基于待添加实体的实体数据对已创建的知识图谱进行扩充。这样在获得待处理数据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用本发明实施例提供的方案，可以提高扩充知识图谱的效率。

附图说明

[0045] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍。

[0046] 图1为本发明实施例提供的一种知识图谱扩充方法的流程示意图；

[0047] 图2为本发明实施例提供的一种获得实体数据的过程示意图；

[0048] 图3为本发明实施例提供的一种实体匹配的过程示意图；

[0049] 图4为本发明实施例提供的一种实体数据合并的过程示意图；

[0050] 图5为本发明实施例提供的一种知识图谱扩充的过程示意图；

[0051] 图6为本发明实施例提供的一种知识图谱扩充装置的结构示意图；

[0052] 图7为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0053] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行描述。

[0054] 为了提高扩充知识图谱的效率，本发明实施例提供了一种知识图谱扩充方法、装置、电子设备及存储介质。

[0055] 本发明的一个实施例中，提供了一种知识图谱扩充方法，该方法包括：

[0056] 获得待处理数据；

[0057] 识别待处理数据所属待添加实体的实体类型，作为目标实体类型；

[0058] 基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；

[0059] 基于待添加实体的实体数据对已创建的知识图谱进行扩充。

[0060] 这样在获得待处理数据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用本发明实施例提供的方案，可以提高扩充知识图谱的效率。

[0061] 下面通过具体实施例对本发明实施例提供的知识图谱扩充方法、装置、电子设备及存储介质进行详细介绍。

[0062] 参见图1，图1为本发明实施例提供的一种知识图谱扩充方法的流程示意图，该方法包括如下步骤S101至S104：

[0063] S101，获得待处理数据。

[0064] 其中，上述待处理数据可以为：用于描述实体的属性信息的数据，上述实体可以是明星、运动员、影片、游戏、商品、专辑等。例如，在实体为明星的情况下，上述属性信息可以包括：姓名信息、年龄信息、身高信息、作品信息等，在实体为专辑的情况下，上述属性信息可以包括：专辑名称信息、发行时间信息、出版方信息等。

[0065] 本发明的一个实施例中，上述待处理数据可以是文本类型的数据、图像类型的数据、语音类型的数据等。

[0066] 本发明的一个实施例中，在获得待处理数据时，可以从预设的数据库中获得数据，作为待处理数据。还可以从公开的数据平台中获取数据，将所获取的数据作为待处理数据。

[0067] 其中，上述数据平台可以是百科数据平台，如百度百科数据平台、搜狗百科数据平台、谷歌百科数据平台等。由于百科数据平台中的数据准确度高、信息丰富，而且信息排版清晰，便于后续从百科数据中抽取实体的实体数据，因此从百科数据平台中获取待处理数据，有利于提高后续所扩充的知识图谱的准确度。

[0068] 除此之外，上述数据平台还可以是新闻数据平台、视频数据平台、图像数据平台等。

[0069] S102，识别待处理数据所属待添加实体的实体类型，作为目标实体类型。

[0070] 其中，上述实体类型指：实体所属的类型，可以包括：明星、运动员、影片、游戏、商品、专辑等类型。例如，在待处理数据所属的待添加实体为“刘德华”的情况下，上述目标实体类型即为“明星”；在待处理数据所属的待添加实体为“十面埋伏”的情况下，上述目标实体类型即为“影片”。

[0071] 本发明的一个实施例中，在识别待添加实体所属的实体类型时，可以包括多种识别方式，下面分别进行介绍：

[0072] 一种方式中，可以从待处理数据中抽取部分数据字段，然后将所抽取的部分数据字段与预设的分类规则进行匹配，得到该待处理数据所属的待添加实体属于各个实体类型的得分，选择得分最大的实体类型，作为该待添加实体所属的目标实体类型。

[0073] 其中，上述分类规则为：预先设定的、用于根据字段对数据所属实体的类别进行区分的规则，上述部分数据字段可以是待处理数据的标题部分的字段、开头部分的字段、结束

部分的字段等。

[0074] 这样可以按照分类规则对待处理数据所属的实体进行类型划分，识别类型的方式简单，获得目标实体类型的效率高。

[0075] 另一种方式中，可以利用预先训练完成的二分类模型，对待处理数据所属待添加实体的实体类型进行分类。

[0076] 具体的，每一二分类模型用于针对一种实体类型的待处理数据进行分类，例如，假设一二分类模型用于针对“明星”类型的待处理数据进行分类，则将待处理数据输入该二分类模型后，二分类模型可以输出表征“是”或“非”的输出结果，表示所输入的待处理数据所述实体的实体类型是“明星”或不是“明星”。

[0077] 可以设计多个二分类模型，不同的二分类模型用于针对不同的实体类型的待处理数据进行分类，这样在获得待处理数据后，可以依次将待处理数据输入各个二分类模型，直至确定该待处理数据所属实体的实体类型。

[0078] 本申请的一个实施例中，针对每一二分类模型，在训练该二分类模型时，可以预先获得样本数据，并人工判断上述样本数据所描述的实体的实体类型是否为目标类型，得到标注信息，然后将上述样本数据输入该二分类模型，得到模型输出结果，在模型输出结果与上述标注信息不一致的情况下，对二分类模型进行参数调整，然后重新将样本数据输入参数调整后的二分类模型中，直至达到预设的训练结束条件，最终得到用于针对上述目标类型的数据进行分类的二分类模型。

[0079] 其中，上述目标类型可以理解为：期望训练后的二分类模型所能够分类的数据对应的实体类型。

[0080] 上述训练结束条件可以是训练次数达到预设的次数阈值，如50000次、100000次等，也可以是连续预设数量个样本数据对应的模型输出结果与标注信息一致，上述预设数量可以是20、30、50等。

[0081] 这样利用人工智能模型的方式识别待处理数据所属实体的实体类型，得到的识别结果的准确度更高。

[0082] 又一种方式中，也可以采用人工识别的方式，对所获得的各个待处理数据所属实体的实体类型进行识别。

[0083] 本发明的一个实施例中，可以结合上述三种方式识别待处理数据所属实体的实体类型。例如，可以首先利用预设的分类规则识别待处理数据对应的实体类型，在识别失败的情况下，再利用二分类模型进行识别，若再一次识别失败，则可以转由人工进行识别。

[0084] S103，基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据。

[0085] 具体的，不同的实体类型与不同的属性类型存在对应关系。例如，以实体类型为“影片”为例，对应的属性类型包括影片名称、发行时间、主演、导演、影片分类等；以实体类型为“游戏”为例，对应的属性类型包括游戏名称、发行方、游戏分类等；以实体类型为“明星”为例，对应的属性类型包括明星姓名、艺名、出生日期、作品、亲朋关系等。

[0086] 例如，假设识别到待添加实体所属的实体类型为明星，第一预设关系中“明星”对应的属性类型包括明星姓名、艺名、出生日期、作品、亲朋关系，则可以从待处理数据中抽取上述属性类型的属性信息，作为明星的实体数据。

[0087] 本发明的一个实施例中，在抽取待添加实体的实体数据时，可以将待处理数据和所属实体的实体类型输入预先训练完成的数据抽取模型，得到上述模型输出的、待添加实体的实体数据。

[0088] 除此之外，还可以将待处理数据转化至文本格式，然后对文本格式的待处理数据进行语义分析，从分析结果中选择用于描述上述目标实体类型对应的属性类型的信息的分析结果，作为待添加实体的实体数据。

[0089] 本发明的一个实施例中，可以确定待处理数据的格式类型，作为目标格式；从预先设计的数据抽取脚本中，选择用于对目标格式的数据进行数据抽取的脚本，作为目标脚本；基于实体类型与属性类型的第一预设关系，确定目标实体类型对应的属性类型，作为目标属性类型；利用目标脚本，从待处理数据中抽取目标属性类型的属性信息，作为待添加实体的实体数据。

[0090] 其中，上述格式类型包括结构化数据、半结构化数据、文本数据等。由于不同格式类型的数据对应的数据抽取方式存在差异，因此可以设计不同的数据抽取脚本，用于对不同格式类型的待处理数据进行数据抽取。

[0091] 具体的，可以首先获得待处理数据的目标格式，然后确定用于对该目标格式的数据进行数据抽取的目标脚本，最后可以利用该目标脚本抽取待处理数据中目标属性类型的信息，作为待添加实体的实体数据。

[0092] 本申请的一个实施例中，上述目标脚本可以对目标格式的待处理数据进行语义分析，得到待处理数据所描述的信息，然后从上述信息中查找用于描述目标属性类型的信息，作为抽取结果，从而得到待添加实体的实体数据。

[0093] 参见图2，图2为本发明实施例提供的一种获得实体数据的过程示意图。如图2所示，可以通过一些公开的数据接口OpenApi，从百度百科数据平台获取百度百科数据，作为待处理数据，然后利用预先训练完成的二分类器识别上述百度百科数据所属实体的实体类型，该二分类器即为上述的二分类模型，可以包括明星分类器、游戏分类器等，分别用于识别明星类型的实体、游戏类型的实体等，在识别得到百度百科数据所属实体的实体类型后，可以利用Extractor(抽取)脚本，从百度百科数据中抽取所属实体的实体数据，其中，每一实体类型对应一Extractor脚本。最终可以得到属于不同实体类型的实体的实体数据Raw Data。

[0094] S104，基于待添加实体的实体数据对已创建的知识图谱进行扩充。

[0095] 具体的，可以预先基于已有的实体的实体数据，创建一知识图谱，所创建的知识图谱中包含不同实体的实体类型、对应的属性类型、属性信息等。在得到待添加实体的实体数据后，可以将上述待添加实体的实体数据添加进知识图谱中，从而实现对知识图谱的扩充。

[0096] 本发明的一个实施例中，可以对待添加实体的实体数据进行格式转化，得到待添加实体的实体数据的三元组信息，将三元组信息导入知识图谱的图数据库中，基于图数据库实现将待添加实体的实体数据添加至知识图谱中。

[0097] 其中，三元组信息包含待添加实体的名称信息、属性类型信息、属性信息。上述三元组的格式可以为json-ld格式。

[0098] 假设待添加实体为“周杰伦”，所属的实体类型为明星，“明星”对应的属性类型包括明星姓名、别称、出生日期、作品、配偶，则该待添加实体的三元组信息可以如下表1所示：

[0099] 表1

[0100]

名称信息	属性类型信息	属性信息
周杰伦	别称	周董
	出生日期	1979.1.18
	作品	龙卷风
	配偶	昆凌

[0101] 具体的，可以对待添加实体的实体数据进行处理，得到上述实体数据json-ld格式的三元组信息，然后将上述三元组信息导入图数据库JanusGraph中，该图数据库会将上述三元组信息添加至知识图谱中，实现对知识图谱的扩充，并且能够提供实体的查询、游走等服务。

[0102] 上述实施例提供的知识图谱扩充方案中，可以获得待处理数据；识别待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；基于待添加实体的实体数据对已创建的知识图谱进行扩充。这样在获得待处理数据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用上述实施例提供的方案，可以提高扩充知识图谱的效率。

[0103] 本发明的一个实施例中，在对知识图谱进行扩充时，可以查找已创建的知识图谱中是否存在与待添加实体为同一实体的目标实体，若存在，将待添加实体的实体数据与目标实体的实体数据进行合并，实现对知识图谱的扩充；若不存在，在知识图谱中创建待添加实体，并在知识图谱中添加待添加实体的实体数据。

[0104] 具体的，若知识图谱中存在与上述待添加实体为同一实体的目标实体，则无需再次在知识图谱中创建该待添加实体，只需要将待添加实体的实体数据合并至知识图谱中目标实体的实体数据内，即可实现对知识图谱的扩充；

[0105] 若知识图谱中不存在与上述待添加实体为同一实体的目标实体，则需要在知识图谱中创建该待添加实体，并添加该实体的实体数据。

[0106] 本发明的一个实施例中，在查找知识图谱中是否存在目标实体时，可以基于实体类型与查找方式的第二预设关系，确定在原始实体库中查找与待添加实体为同一实体的目标实体的查找方式，按照所确定的查找方式，查找原始实体库是否存在与待添加实体为同一实体的目标实体，若为是，确定知识图谱中存在目标实体。

[0107] 其中，原始实体库中包含知识图谱中实体的实体数据。原始实体库可以是非关系型分布式数据库Hbase。

[0108] 具体的，由于不同实体类型的数据具有不同的属性类型，而针对不同属性类型的数据，可以按照不同的匹配方式进行查找。因此不同的实体类型对应于不同的查找方式。

[0109] 首先可以根据待添加实体的实体类型，从上述第二预设关系中确定该实体类型对

应的查找方式，然后按照所确定的查找方式从原始实体库中查找与待添加实体为同一实体的目标实体，若能够在原始实体库中查找到目标实体，则说明知识图谱中存在目标实体，若未能在原始实体库中查找到目标实体，则说明知识图谱中不存在目标实体。

[0110] 本发明的一个实施例中，每一实体类型对应的查找方式为：对该实体类型所对应属性类型的属性信息进行一一匹配的查找方式。

[0111] 这样在查找目标实体时，可以首先确定待添加实体的实体类型，进而确定该实体类型对应的属性类型，然后对每一属性类型的属性信息进行一一匹配，最终查找得到与待添加实体为同一实体的目标实体。

[0112] 本发明的一个实施例中，还可以将待添加实体的实体数据添加至原始实体库中。这样可以保证原始实体库中包含知识图谱中所有实体的实体数据，从而便于后续从原始实体库内查找知识图谱中实体的实体数据。

[0113] 本发明的一个实施例中，在查找目标实体时，可以利用预设的模糊查找方式，从原始实体库中查找与待添加实体相似的实体，作为候选实体，然后按照所确定的查找方式，从候选实体中查找是否存在与待添加实体为同一实体的目标实体。

[0114] 具体的，首先利用模糊查找方式，从原始实体库中查找是否存在与待添加实体相似的实体，若存在，则说明原始实体库中可能存在与待添加实体为同一实体的目标实体，因此可以将所查找到的实体作为候选实体，后续可以进一步地从候选实体中查找目标实体；

[0115] 若不存在候选实体，则说明原始实体库中不存在与待添加实体为同一实体的目标实体，因此无需进行进一步的查找。

[0116] 这样首先查找候选实体，再从候选实体中确定目标实体，由于查找候选实体的方式较为简单，在查找目标实体时可以降低所查找对象的数量，从而可以提高查找目标实体的效率。

[0117] 本发明的一个实施例中，在查找候选实体时，可以查找原始实体库中索引信息与目标属性信息相匹配的实体，作为候选实体。

[0118] 其中，每一实体的索引信息包含该实体的实体数据中预设属性信息的属性字段，目标属性信息为：待添加实体的属性信息中的名称信息。

[0119] 具体的，可以为原始实体库中每一实体建立索引，每一实体的索引信息中包含其预设属性信息的属性字段，上述预设属性信息可以包括：名称属性信息、关系属性信息等。其中，可以利用ElasticSearch程序建立各个实体的索引信息。

[0120] 例如，假设实体为“刘翔”，预设属性信息包括名称属性信息“刘翔”和关系属性信息“配偶：吴莎”，则该实体的索引信息可以为“刘翔；配偶：吴莎”。

[0121] 由于目标属性信息为待添加实体的名称信息，可以理解为，采用待添加实体的名称信息与原始实体库中各个实体的名称信息、关系信息进行匹配的模糊查找方式，从原始实体库中查找与待添加实体存在关联的候选实体。

[0122] 本发明的一个实施例中，在从候选实体中查找目标实体时，可以利用预先训练完成的实体匹配模型进行查找。

[0123] 具体的，参见图3，图3为本发明实施例提供的一种实体匹配的过程示意图。如图3所示，可以将待添加实体的实体数据与上述候选实体的实体数据输入上述实体匹配模型，该实体匹配模型首先可以利用结构转换模块对输入的实体数据进行结构转换，得到向量形

式的实体数据，然后利用训练的匹配分析模块，对结构转换后的实体数据进行匹配度分析，根据匹配度分析结果，判断候选实体与待添加实体是否为同一实体。

[0124] 参见图4，图4为本发明实施例提供的一种实体数据合并的过程示意图。如图4所示，在获得待添加实体的实体数据后，可以利用原始实体库中各个实体的索引信息，以模糊查找的方式从原始实体库中查找与待添加实体相似的候选实体，然后利用上述实体匹配模型，从候选实体中确定与待添加实体为同一实体的目标实体，然后按照属性类型，将待添加实体的实体数据与知识图谱内目标实体的实体数据中属于同一属性类型的数据进行合并，从而实现对知识图谱的扩充。

[0125] 参见图5，图5为本发明实施例提供的一种知识图谱扩充的过程示意图，如图5所示：

[0126] 可以首先基于爱奇艺数据，如角色、专辑、明星、游戏等实体的数据，以人工干预的方式识别爱奇艺数据所属实体的实体类型，并从爱奇艺数据中抽取上述实体的实体数据，从而构建得到知识图谱；

[0127] 然后通过一些公开的数据接口OpenApi，从百度百科数据平台获取百度百科数据，作为待处理数据，然后利用预先训练完成的二分类器识别上述百度百科数据所属实体的实体类型，该二分类器可以包括明星分类器、游戏分类器等，在识别得到百度百科数据所属实体的实体类型后，可以利用Extractor脚本，从百度百科数据中抽取所属实体的实体数据，最终可以得到属于不同实体类型的实体的实体数据Raw Data；

[0128] 之后针对不同实体类型的实体数据，可以利用去重器识别实体库内各个实体的实体数据中与待添加实体属于同一实体的实体数据，然后对属于同一实体的实体数据进行合并，将去重后的实体数据导入图数据库中，实现对知识图谱的扩充，例如，上述去重器可以包括明星去重器，用于对明星类型的实体数据进行去重，还可以包括游戏去重器，用于对游戏类型的实体数据进行去重。

[0129] 上述实施例提供的知识图谱扩充方案中，可以获得待处理数据；识别待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；基于待添加实体的实体数据对已创建的知识图谱进行扩充。这样在获得待处理数据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用上述实施例提供的方案，可以提高扩充知识图谱的效率。

[0130] 参见图6，图6为本发明实施例提供的一种知识图谱扩充装置的结构示意图，所述装置包括：

[0131] 数据获得模块601，用于获得待处理数据；

[0132] 类型识别模块602，用于识别所述待处理数据所属待添加实体的实体类型，作为目标实体类型；

[0133] 数据抽取模块603，用于基于实体类型与属性类型的第一预设关系，从所述待处理数据中抽取所述目标实体类型对应的属性类型的属性信息，作为所述待添加实体的实体数据；

[0134] 图谱扩充模块604，用于基于所述待添加实体的实体数据对已创建的知识图谱进行扩充。

[0135] 本申请的一个实施例中，所述图谱扩充模块604，包括：

[0136] 目标实体查找子模块，用于查找已创建的知识图谱中是否存在与所述待添加实体为同一实体的目标实体，若存在，触发第一扩充子模块，若不存在，触发第二扩充子模块；

[0137] 所述第一扩充子模块，用于将所述待添加实体的实体数据与所述目标实体的实体数据进行合并，实现对所述知识图谱的扩充；

[0138] 所述第二扩充子模块，用于在所述知识图谱中创建所述待添加实体，并在所述知识图谱中添加所述待添加实体的实体数据。

[0139] 本申请的一个实施例中，所述目标实体查找子模块，包括：

[0140] 查找方式确定单元，用于基于实体类型与查找方式的第二预设关系，确定在原始实体库中查找与所述待添加实体为同一实体的目标实体的查找方式，其中，所述原始实体库中包含所述知识图谱中实体的实体数据；

[0141] 目标实体查找单元，用于按照所确定的查找方式，查找所述原始实体库是否存在与所述待添加实体为同一实体的目标实体，若为是，确定所述知识图谱中存在所述目标实体。

[0142] 本申请的一个实施例中，所述装置还包括数据添加单元，用于：

[0143] 将所述待添加实体的实体数据添加至所述原始实体库中。

[0144] 本申请的一个实施例中，所述目标实体查找单元，包括：

[0145] 候选实体查找子单元，用于利用预设的模糊查找方式，从所述原始实体库中查找与所述待添加实体相似的实体，作为候选实体；

[0146] 目标实体查找子单元，用于按照所确定的查找方式，从所述候选实体中查找是否存在与所述待添加实体为同一实体的目标实体。

[0147] 本申请的一个实施例中，所述候选实体查找子单元，具体用于：

[0148] 查找所述原始实体库中索引信息与目标属性信息相匹配的实体，作为候选实体，其中，每一实体的索引信息包含该实体的实体数据中预设属性信息的属性字段，所述目标属性信息为：所述待添加实体的属性信息中的名称信息。

[0149] 本申请的一个实施例中，每一实体类型对应的查找方式为：

[0150] 对该实体类型所对应属性类型的属性信息进行一一匹配的查找方式。

[0151] 本申请的一个实施例中，所述图谱扩充模块604，具体用于：

[0152] 对所述待添加实体的实体数据进行格式转化，得到所述待添加实体的实体数据的三元组信息，其中，所述三元组信息包含所述待添加实体的名称信息、属性类型信息、属性信息；

[0153] 将所述三元组信息导入所述知识图谱的图数据库中，基于所述图数据库实现将所述待添加实体的实体数据添加至所述知识图谱中。

[0154] 本申请的一个实施例中，所述数据抽取模块603，具体用于：

[0155] 确定所述待处理数据的格式类型，作为目标格式；

[0156] 从预先设计的数据抽取脚本中，选择用于对所述目标格式的数据进行数据抽取的脚本，作为目标脚本；

[0157] 基于实体类型与属性类型的第一预设关系，确定所述目标实体类型对应的属性类型，作为目标属性类型；

[0158] 利用所述目标脚本，从所述待处理数据中抽取所述目标属性类型的属性信息，作为所述待添加实体的实体数据。

[0159] 上述实施例提供的知识图谱扩充方案中，可以获得待处理数据；识别待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；基于待添加实体的实体数据对已创建的知识图谱进行扩充。这样在获得待处理数据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用上述实施例提供的方案，可以提高扩充知识图谱的效率。

[0160] 本发明实施例还提供了一种电子设备，如图7所示，包括处理器701、通信接口702、存储器703和通信总线704，其中，处理器701，通信接口702，存储器703通过通信总线704完成相互间的通信，

[0161] 存储器703，用于存放计算机程序；

[0162] 处理器701，用于执行存储器703上所存放的程序时，实现知识图谱扩充方法。

[0163] 上述终端提到的通信总线可以是外设部件互连标准(Peripheral Component Interconnect，简称PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture，简称EISA)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0164] 通信接口用于上述终端与其他设备之间的通信。

[0165] 存储器可以包括随机存取存储器(Random Access Memory，简称RAM)，也可以包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。可选的，存储器还可以是至少一个位于远离前述处理器的存储装置。

[0166] 上述的处理器可以是通用处理器，包括中央处理器(Central Processing Unit，简称CPU)、网络处理器(Network Processor，简称NP)等；还可以是数字信号处理器(Digital Signal Processor，简称DSP)、专用集成电路(Application Specific Integrated Circuit，简称ASIC)、现场可编程门阵列(Field－Programmable Gate Array，简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0167] 在本发明提供的又一实施例中，还提供了一种计算机可读存储介质，所述计算机可读存储介质内存储有计算机程序，所述计算机程序被处理器执行时实现上述实施例中任一所述的知识图谱扩充方法。

[0168] 在本发明提供的又一实施例中，还提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述实施例中任一所述的知识图谱扩充方法。

[0169] 上述实施例提供的知识图谱扩充方案中，可以获得待处理数据；识别待处理数据所属待添加实体的实体类型，作为目标实体类型；基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据；基于待添加实体的实体数据对已创建的知识图谱进行扩充。这样在获得待处理数

据后，可以首先确定待处理数据所属实体的实体类型，然后从待处理数据中抽取该实体的实体数据，从而将该实体的实体数据添加至知识图谱中，实现对知识图谱的扩充，无需由人工对知识图谱进行扩充。由此可见，应用上述实施例提供的方案，可以提高扩充知识图谱的效率。

[0170] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘Solid State Disk(SSD))等。

[0171] 需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、物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、物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、物品或者设备中还存在另外的相同要素。

[0172] 本说明书中的各个实施例均采用相关的方式描述，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，对于装置实施例、电子设备实施例、计算机可读存储介质实施例、计算机程序产品实施例而言，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。

[0173] 以上所述仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内所作的任何修改、等同替换、改进等，均包含在本发明的保护范围内。

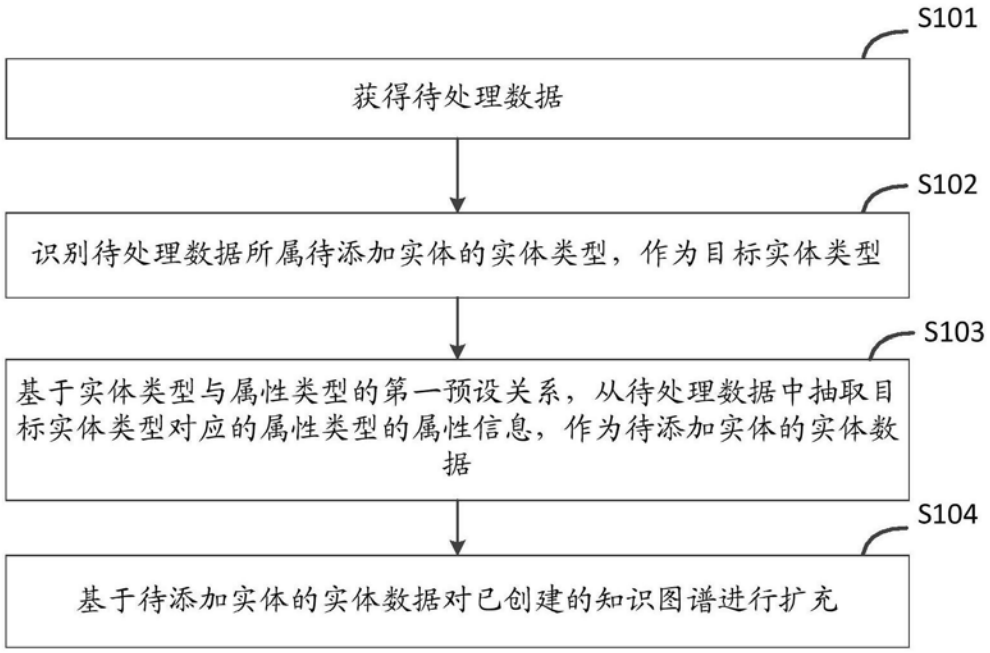

S101
获得待处理数据
S102
识别待处理数据所属待添加实体的实体类型，作为目标实体类型
S103
基于实体类型与属性类型的第一预设关系，从待处理数据中抽取目标实体类型对应的属性类型的属性信息，作为待添加实体的实体数据
S104
基于待添加实体的实体数据对已创建的知识图谱进行扩充

图1

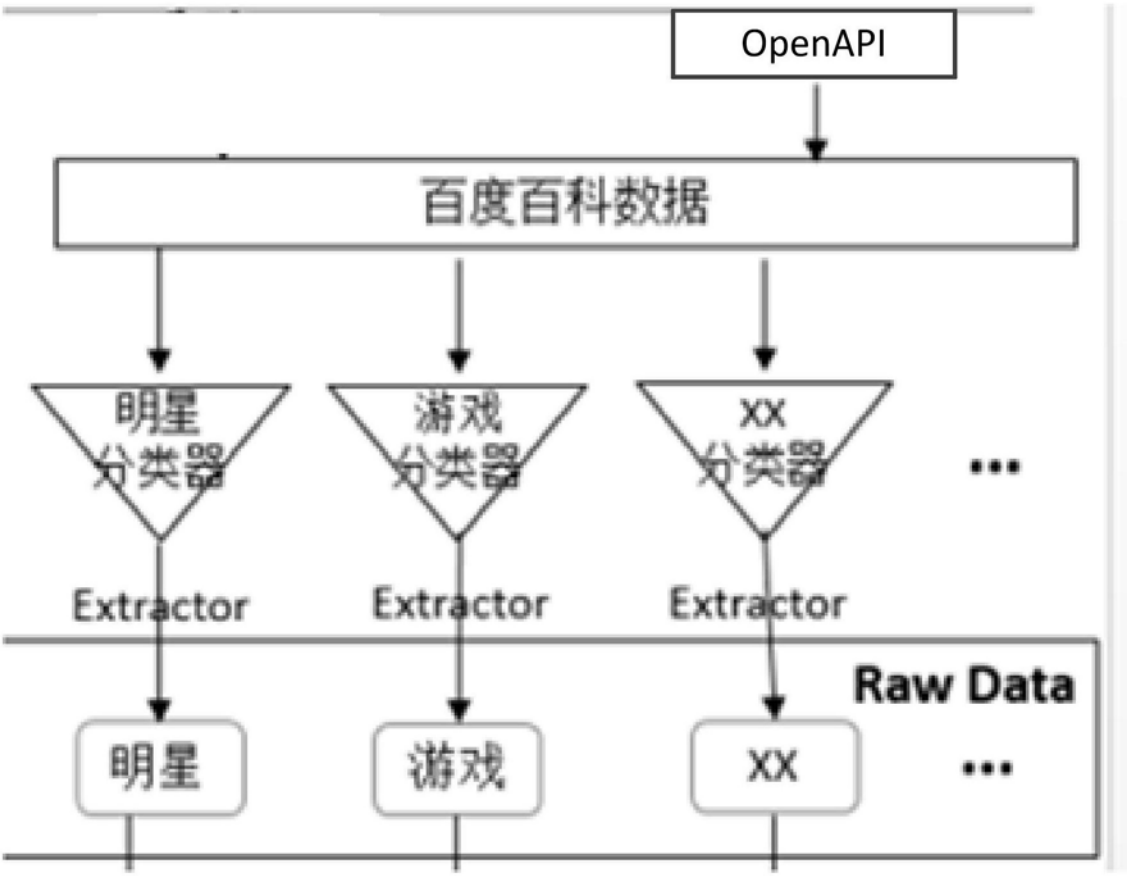

OpenAPI
百度百科数据
明星
分类器
游戏
分类器
XX
分类器
...
Extractor
Extractor
Extractor
Raw Data
明星
游戏
XX
...

图2

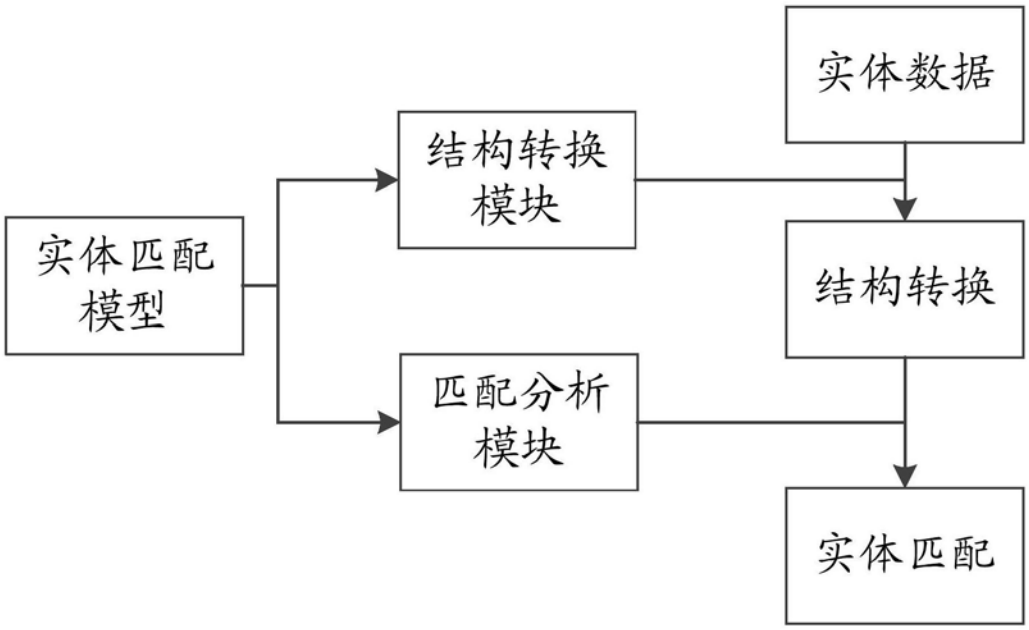

实体数据
结构转换模块
实体匹配模型
结构转换
匹配分析模块
实体匹配

图3

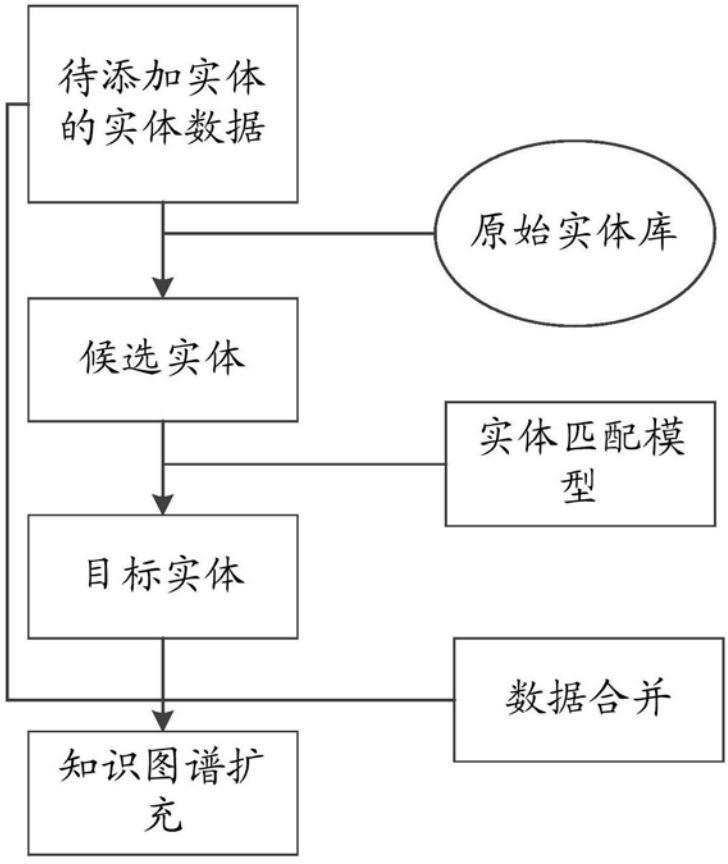

待添加实体的实体数据
原始实体库
候选实体
实体匹配模型
目标实体
数据合并
知识图谱扩充

图4

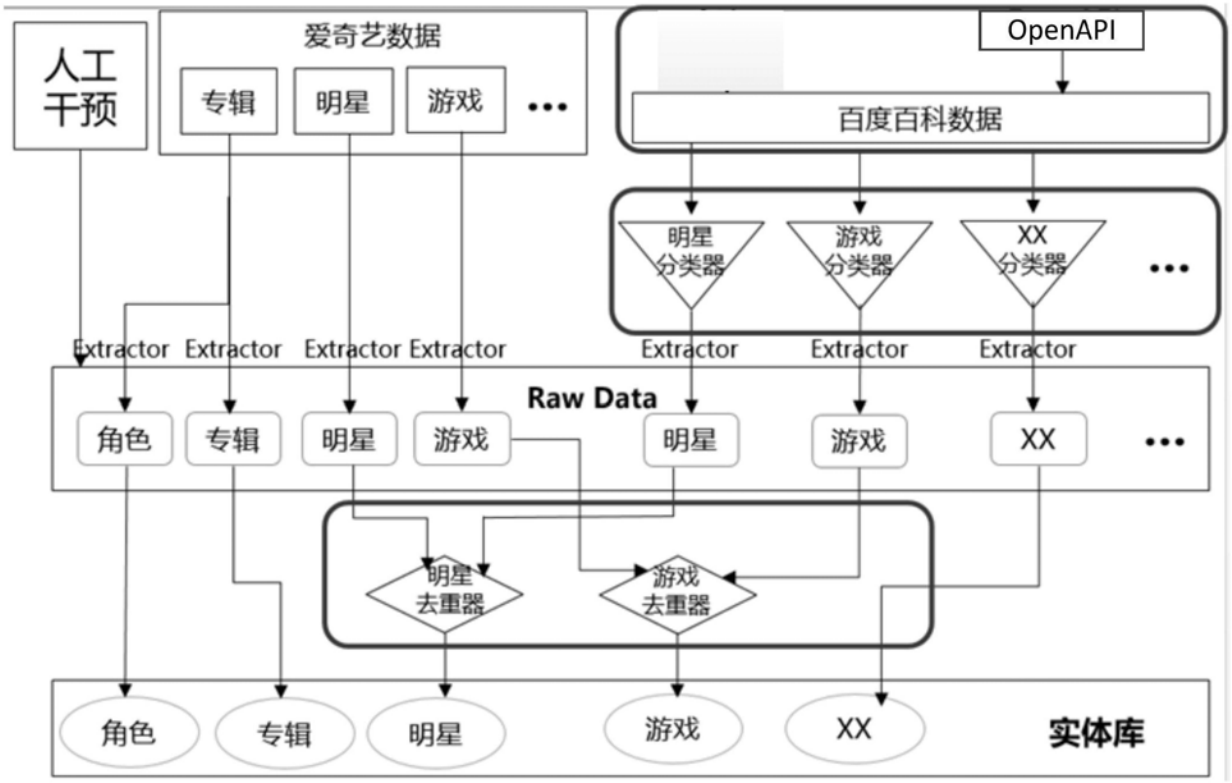

人工干预
爱奇艺数据
专辑
明星
游戏
...
OpenAPI
百度百科数据
明星分类器
游戏分类器
XX分类器
...
Extractor
Extractor
Extractor
Extractor
Extractor
Extractor
Extractor
Raw Data
角色
专辑
明星
游戏
明星
游戏
XX
...
明星去重器
游戏去重器
角色
专辑
明星
游戏
XX
实体库

图5

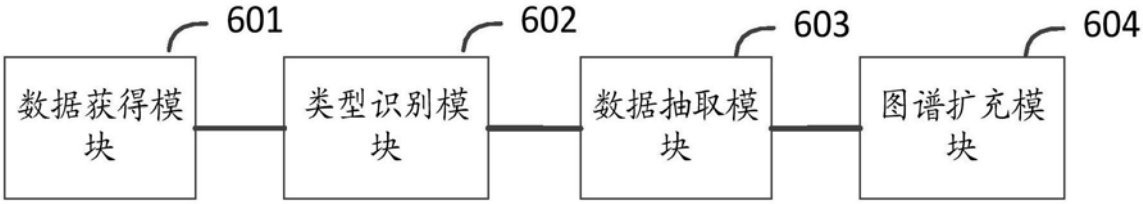

601
602
603
604
数据获得模块
类型识别模块
数据抽取模块
图谱扩充模块

图6

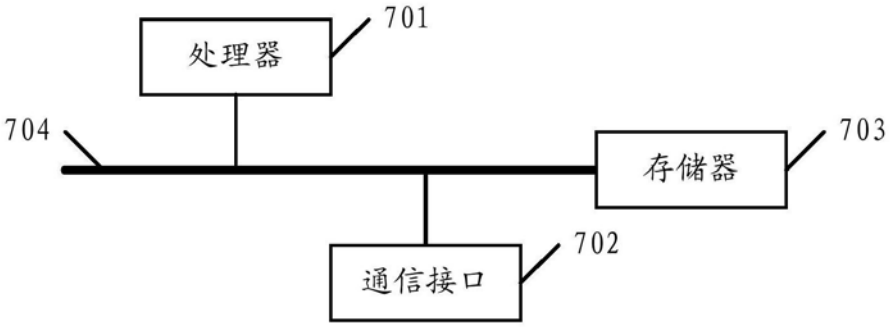

701
处理器
704
703
存储器
702
通信接口

图7