(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 7 月 2 日 (02.07.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/135048 A1

(51) 国际专利分类号:
G06F 16/28 (2019.01)

(21) 国际申请号: PCT/CN2019/124552

(22) 国际申请日: 2019 年 12 月 11 日 (11.12.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811635696.X 2018年12月29日 (29.12.2018) CN

(71) 申请人: 颖投信息科技（上海）有限公司(YINGTOU INFORMATION & TECHNOLOGY (SHANGHAI) CO., LTD) [CN/CN]; 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。

(72) 发明人: 刘涛(LIU, Tao); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。 朱宏明 (ZHU, Hongming); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。 顾江 (GU, Jiang); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。 姜逸之 (JIANG, Yizhi); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。 王晓文(WANG, Xiaowen); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。 周游 (ZHOU, You); 中国上海市静安区南京西路1717号会德丰广场2706室, Shanghai 200040 (CN)。

(74) 代理人: 北京恒都律师事务所(BEIJING HENGDU LAW FIRM); 中国北京市朝阳区建国门外大街1号国贸三期B座50层, Beijing 100020 (CN)。

(54) **Title:** DATA MERGING METHOD AND APPARATUS FOR KNOWLEDGE GRAPH

(54) **发明名称:** 知识图谱的数据融合方法和装置

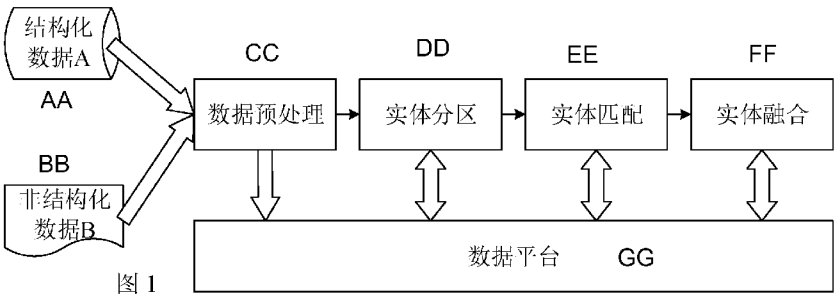

图 1

AA Structured data A
BB Unstructured data B
CC Data preprocessing
DD Division of subjects into blocks
EE Subject matching
FF Subject merging
GG Data platform

(57) **Abstract:** A data merging method and apparatus for a knowledge graph. A system for implementing the method comprises a data platform configured with a unified access interface. The method comprises: processing data from different data sources and then converting same to a subject-property-object format, storing same in the data platform by means of the unified access interface, and receiving graph data index information returned by the data platform; according to the graph data index information, dividing subjects stored in the data platform into one or more sub-blocks according to the attribute; performing similarity calculation on candidate subjects classified into the same sub-block, and screening matching subject pairs that meet a preset similarity condition; and supplementing

WO 2020/135048 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (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)。

本国际公布：
— 包括国际检索报告(条约第21条(3))。

and/or replacing subject attribute values of the matching subject pairs to generate unified subject representation. By the abovementioned method, the data merging problem that existing data merging techniques cannot flexibly adapt to different knowledge graphs can be effectively solved.

(57) 摘要：一种知识图谱的数据融合方法和装置，执行所述方法的系统包括配置有统一访问接口的数据平台，所述方法包括：将来自不同数据源的数据进行处理后转换为三元组格式，通过所述统一访问接口存储到数据平台，并接收所述数据平台返回的图数据索引信息；根据所述图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区；对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对；对所述匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。通过上述方法，可有效解决现有数据融合技术不能灵活适应不同知识库的数据融合问题。

知识图谱的数据融合方法和装置

本申请要求 2018 年 12 月 29 日提交的申请号为 201811635696.X、发明名称为“知识图谱的数据融合方法和装置”的中国发明专利申请的优先权，其全文引用在此供参考。

技术领域

本申请涉及知识图谱技术领域，特别地，涉及一种知识图谱的数据融合方法和装置。

背景技术

知识图谱是一种描述现实世界中存在的各种实体或概念及其关系而构成的一张巨大的语义网络图，其节点表示实体或概念，边则由属性或关系构成。现在的知识图谱已被用来泛指各种大规模的知识库。其中：实体是指具有可区别性且独立存在的某种事物，比如某个国家、某家公司、某个人等。属性是指一个实体的内在特性，比如国家具有“人口”、“面积”等不同属性（如图 4 所示），公司具有“名称”、“法定代表人”等属性。关系是一个实体与另一个实体的关联特征，比如某个公司注册在某个国家，某个人就职于某个公司等。

知识图谱的节点和边一般用三元组（S-P-O，Subject-Property-Object）的形式来定义，包括（实体 1-关系-实体 2）和（实体-属性-属性值）等形式， 知识图谱可以表示为三元组的集合，在数据模型上可以表现为图的形式（如图 4 所示），并采用图数据库来进行数据的存储和管理。

现实世界中知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、知识库层次结构缺失等问题；另外，不同的数据源对于同一实体可能有不同的知识表示，比如，在百度百科中某个公司实体具有名称属性‘阿里巴巴’，而从 google 搜索中抓取到的某个公司实体的名称属性是‘alibaba’，这两个实体在现实世界中有可能指向同一个实体，因此需要将他们的属性以及延伸的关系进行互相融合，从而在知识图谱中生成唯一的实体节点，消除歧义，生成高质量的知识库。

现有数据融合方案一般包括分区索引、相似度计算和实体融合等主要步骤，但在具体实现时会根据数据源以及知识库的特点选择对应的分区算法、相似度

匹配算法和实体对齐算法，并将上述方案集成为一个完整的系统，当数据源或知识库的范围发生变化时，为适应新的需求，需要重新构建数据融合系统。

发明内容

本申请提供一种知识图谱的数据融合方法和装置，用于解决现有数据融合技术不能灵活适应不同知识库的数据融合问题。

本申请公开的一种知识图谱的数据融合方法，执行所述方法的系统包括配置有统一访问接口的数据平台，所述方法包括：将来自不同数据源的数据进行处理后转换为三元组格式，通过所述统一访问接口存储到数据平台，并接收所述数据平台返回的图数据索引信息；根据所述图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区；对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对；对所述匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。

优选地，在步骤根据所述图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区之前，还包括：将来自多个数据源的数据转换为三元组格式之后存储在数据平台中的实体根据其属性的实际含义进行对齐。

优选地，所述子分区划分方式为根据实体属性产生的全局唯一分区键进行等值划分，或基于预设聚类模型进行划分。

优选地，对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对，具体为：为实体本身的属性以及与该实体相关的其他实体的属性分别设置不同的权重，加权求和计算候选实体对的总体相似度；若相同子分区中的候选实体对的总体相似度超过预设相似度阈值，则将该候选实体对作为匹配实体对。

优选地，对缺失的实体属性值进行补充的方法为通过爬虫从网络获取或进行人工填充。

优选地，所述图数据索引信息为三元组格式的图数据在所述数据平台的存储地址及其元数据。

本申请公开的一种知识图谱的数据融合装置，包括数据平台、数据预处理模块、实体分区模块、实体匹配模块和实体融合模块，其中：所述数据平台配置有统一访问接口；所述数据预处理模块配置为将来自不同数据源的数据进行

处理后转换为三元组格式，通过所述统一访问接口存储到数据平台，并接收所述数据平台返回的图数据索引信息；所述实体分区模块配置为根据所述数据预处理模块输出的图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区；所述实体匹配模块配置为将所述实体分区模块划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对；所述实体融合模块配置为对所述实体匹配模块筛选出的匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。

优选地，所述实体分区模块包括等值分区子模块和/或聚类分区子模块；所述等值分区子模块配置为根据实体属性产生的全局唯一分区键对存储在数据平台中的实体进行等值划分；所述聚类分区子模块配置为基于预设聚类模型对存储在数据平台中的实体进行划分。

优选地，所述实体匹配模块具体包括相似度计算子模块和比较子模块；所述相似度计算子模块配置为为实体本身的属性以及与该实体相关的其他实体的属性分别设置不同的权重，加权求和计算候选实体对的总体相似度；所述比较子模块配置为判断相同子分区中的候选实体对的总体相似度是否超过预设相似度阈值，若是，则将该候选实体对作为匹配实体对。

优选地，所述装置还包括数据处理模块和/或属性对齐模块；所述数据处理模块配置为通过所述统一访问接口对数据平台中的节点实体数据和边实体数据进行处理，并返回数据处理结果传递给下一个模块；所述属性对齐模块配置为将来自多个数据源的数据经所述数据预处理模块处理后存储在数据平台中的实体根据其属性的实际含义进行对齐。

本申请还公开了一种在其上记录有配置为执行上述方法的程序的存储介质。

与现有技术相比，本申请具有以下优点：

本申请优选实施例方案中的各阶段在流水线上具有上下游依赖关系，但不同阶段之间仅通过数据格式约束，通过数据平台提供的统一接口实现相互解耦，可独立开发完成。各阶段的算法本身可以灵活替换，通过实现自定义阶段，可以在不同阶段之间插入新的流程阶段，自由编制完成自定义需求。另外，本申请对数据平台的架构没有限制，例如可以采用 Hadoop 分布式文件系统或云计算架构，以方便在数据量增长的情况下扩展计算和存储资源。

附图说明

附图仅用于示出优选实施方式的目的，而并不认为是对本申请的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

图 1 为本申请知识图谱的数据融合方法第一实施例的流程示意图；

图 2 为本申请知识图谱的数据融合方法第二实施例的流程示意图；

图 3 为本申请知识图谱的数据融合装置一实施例的结构示意图；

图 4 为知识图谱的图数据模型示意图。

具体实施方式

为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

在本申请的描述中，需要理解的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。“多个”的含义是两个或两个以上，除非另有明确具体的限定。术语“包括”、“包含”及类似术语应该被理解为是开放性的术语，即“包括/包含但不限于”。术语“基于”是“至少部分地基于”。术语“一实施例”表示“至少一个实施例”；术语“另一实施例”表示“至少一个另外的实施例”。其他术语的相关定义将在下文描述中给出。

参照图 1，示出了本申请知识图谱的数据融合方法第一实施例的流程，执行本方法实施例的系统设置有为各阶段提供运行环境和计算资源的数据平台，各阶段均可以通过数据平台的统一访问接口实现交互。在具体实施时，数据平台可以构建在 Hadoop 分布式文件系统、云计算架构 （如亚马逊 AWS EMR）或其他架构上，对此，本申请不予限制。所述方法实施例包括以下几个阶段：

1. 数据预处理阶段（InputStage）：将多个同构或者异构数据源中（如结构化数据 A 和非结构化数据 B）的数据处理成相同的实体及其属性的格式（SPO 格式），作为后续阶段的输入。

通过配置不同的数据源信息以及数据模型，将数据从数据源抽取、清洗、变形后以统一的数据格式在数据平台上存储。例如对于关系型数据库数据源，通过配置连接信息、实体类型和实体表、关系类型和关系表，就可以抽取出需

要的 SPO 数据。对于图数据库，节点（实体-属性-属性值）和边（实体-关系-实体）是天然的 SPO 结构。

数据预处理阶段的部分配置参数如下表所示。

名称	备注	是否必选
sourceType	数据源类型	是
sourceConfig	数据源连接信息。对于数据库，一般是连接信息，对于文件数据，是可以访问的文件地址	是
dataModel	指定输入数据映射成SPO格式的方法： 1.对于关系数据库，需要指定： 实体：实体表（或sql），实体唯一id列 关系：关系表（或sql），两个实体id列 2.对于文件，需要指定： 实体：实体文件地址，实体唯一id的列名 关系：关系文件地址，两个实体id列名 3.图数据库 所有节点标签类型（vertex label） 所有边类型（edge label） 4.其他：	是
output	保到数据平台相关配置: output.target 保存的地址 output.entities 保存的实体类型名 output.relations 保存的关系类型名	否

具体实施时，可以采用自定义（CustomInputStage）方式实现对不同数据源预处理，接口形式如下：

```
class CustomInputStage(InputStage):
    def read_input(config):
        data = read_from_source_input(config)
```

```
output = write_data_to_platform(data)
return output
```

通过读取上表中定义的配置，实现读取远端数据后解析、存储数据。例如对非结构化数据源，可以调用机器学习接口、网络接口等完成知识提取，保存成三元组信息，返回保存数据的地址和元数据信息。

2. 实体分区阶段（BlockingStage）：将来自于多个数据源的实体根据其属性，划分到不同的子分区（Block），以降低候选匹配对的数据规模。

对于需要匹配的数据源 S 和 T，假设数据源 S 的实体数据规模是 m，数据源 T 的实体数据规模是 n，其需要检验匹配的数据规模是 m*n。在大数据场景下，这个数据规模基本上是无法实现的，必须降低需要匹配的数据对规模。

具体实施时，可以预先将两个数据源中不可能匹配的实体对划分到不同的数据分区中，使每个数据分区中的数据规模大大降低，多个数据分区可以并行计算完成。

例如，对于 S 和 T 中需要匹配的公司实体，一般注册在不同国家的实体在现实世界不太可能是同一家公司，那么可以根据公司的国家属性，划分为 220 多个（国家或地区）数据分区。对于每个分区，可以进一步根据相同或者相似属性，继续划分子分区。比如，同在‘美国’分区下面的公司，可以继续根据相同的‘州’属性分到新的分区。最后需要匹配的数据规模等于所有数据分区的和，在后续的计算中，所有的数据分区可以并行计算，从而可以较大程度地降低了整体匹配时间。

实体分区阶段的部分配置参数如下表所示。

名称	备注	是否必选
input	实体分区阶段（BlockingStage）的输入数据配置需要子分区（block）的实体类型和文件地址	否，默认使用上一阶段输出的 output
blockingProperties	进行分区计算时使用的属性列表，每个属性需要指定： 1.属性名称	是

	2.属性优先级 3.属性执行分区算法 Match：属性完全匹配时划分到同一个分区 Similarity：属性相似时划分到同一个分区	
minBlockSamples	每个block至少包含的匹配实体个数	是
maxBlockSamples	每个block至多包含的匹配实体个数	是
minBlockSize	最少分区个数	
maxBlockSize	最多分区个数	
output	blockingStage输出配置	否

另外，可以通过自定义的分区算法扩展实体分区阶段（BlockingStage）的分区方式，例如，通过如下接口形式：

```
class CustomBlockingStage(BlockingStage):
    def block(config, propKey, entity, currentBlockKey):
        key = generate_key(config, propValue)
        return key
```

可以根据当前实体所在分区和下一次分区所用的属性产生全局唯一的分区键（block key），从而将数据分入下一个分区。当该分区的可能匹配实体对数达到最低值或者总分区数达到最大值时，该分区不再继续划分。

对基于聚类的分区算法，可以利用已经训练好的聚类模型实现，接口形式如下：

```
class CustomBlockingStage(BlockingStage):
    def block(config, propKey, entity, currentBlockKey):
        key = clusteringModel.predict(entity)
        return key
```

聚类模型可以对当前实体直接进行预测，并对应到某个类当中，此时分区数量等于聚类模型的类数量。当然还可以在聚类的基础上继续对分区划分。

3. 实体匹配阶段（MatchStage）：对于同一分区内的候选实体对，可根据实体本身的属性以及与它有关联的实体的属性分别设置不同的权重，并通过加权求和计算该候选实体对的总体相似度；将超过一定相似度阈值的候选实体对筛选出，作为匹配实体对。

需要说明的是，本流程设计允许插入一些基于强关联的规则来直接完成匹配，如两个数据源中的公司数据，若其都是上市公司并且上市的股票代码完全相同，将可以被直接匹配，从而跳过相似度计算的流程，从而降低匹配阶段的计算复杂度。

当提供验证数据集时，可以通过匹配算法产生的结果与验证数据集进行对比，验证匹配算法的准确度。通过调整属性和权重参数，以及相似度阈值，多次对比计算结果，以不断提高准确度。例如两个公司实体通过名称和股票代码相似度加权和来比较，如果名称在不同数据源用不同语言表示，其相似度权重就较低，需要将其权重调低一些，而股票代码的相似度相对权重应该设的更高一些。

本申请的实体匹配算法可以通过调整参数多次迭代，以提高匹配结果的准确性。

实体匹配阶段（MatchStage）的部分配置参数如下表所示。

名称	备注	是否必选
input	MatchStage所用的input配置 生成可能匹配的实体id对	否，默认使用上一阶段输出的output
validationSet	验证数据集，用于计算匹配算法的准确度	否
matchConfigs	计算实体相似度时需要计算的属性相似度和关系实体相似度 1.属性 2.属性相似度算法	是

	3.属性相似度权重 4.关系实体类型 5.关系相似度算法 6.关系相似度权重	
similarityThreshold	相似度阈值。超过阈值范围的可以认作实际匹配的实体对	是
output	matchStage输出配置	否

通过自定义的实体匹配算法，可以比较两个实体是否指向同一个知识表示。接口形式如下：

```
class CustomMatchStage(MatchStage):
    def match(config, sourceEntity, targetEntity):
        is_equal = classifier.predict(sourceEntity,targetEntity)
        return is_euqal
```

上面例子中，使用预先训练的机器学习二分类模型，以两个实体的各属性相似度向量作为输入，推断是否能归为同一个实体的概率（是则为1）。

最后匹配的实体对将被输出到结果集合当中。

4. 实体融合阶段（MergeStage）：对实际指向同一实体的不同数据源中的数据，根据融合算法，对实体属性值进行补充、替换和规范化，最终生成统一的实体表示。

一般需要自定义的融合算法，接口形式如下:

```
class CustomMergeStage(MergeStage):
    def merge(config, propKey, sourcePropValue, targetPropValue):
        currentPropValue = sourcePropValue + targetPropValue
        return currentPropValue
```

数据融合时可结合不同的业务规则实现，比如名称可设置多个匿名，邮箱、地址等可以采用标准化格式。而对缺失的属性数据可以通过爬虫或者人工进行填充，构建高质量的数据，方便知识图谱的搜索、分析等应用。

在进一步的实施例中，除了以上定义的几个阶段，还可以编排入不同功能

的阶段（如数据处理阶段）。可采用以下形式的接口：

```
class CustomStage:
        def run(config) :
            input = config.input
            output = read_and_compute(input)
            return output
```

对需要处理的数据通过 input 配置参数传递，处理完成后写入 output，并传递给下一个阶段，实现系统功能的扩展。

本申请通过上述手段，实现了大数据场景下实体融合的通用流水线（Pipeline）。流水线由多个阶段（Stage）构成，每个阶段可以通过配置的方式灵活扩展，并且可以将自定义阶段（CustomStage）编排到流水线以适应不同的应用场景。除了数据预处理阶段（InputStage）只有 output 输出，其他各阶段都具有 input 输入配置。Input 配置可指定该阶段运行需要获取的来自不同数据源的实体列表、关系列表、数据地址以及相关数据元信息（schema 包括表名、列名等）。等到该阶段读取完输入数据，运行算法，写入到数据平台，并将所有数据地址和元数据通过 output 输出。因此各阶段可以通过 input 和 output 串联运行，也可以单独指定 input 参数运行。

参照图 2，示出了本申请知识图谱的数据融合方法第二实施例的流程，与上述第一方法实施例的区别在于，在数据预处理阶段和实体分区阶段之间增加一个属性对齐阶段（Attribute Matching）：用于将来自多个数据源的经预处理后存储在数据平台中的实体根据其属性的实际含义进行对齐，如将数据源 A 的『地址』字段与数据源 B 的『Address』字段进行对齐，在后续分区和匹配阶段中被对齐的字段将当作同一含义的字段来处理。

具体实施时，实体属性的实际含义可以人工设定，也可以通过在系统中设置一个属性含义对照表的形式实现，对此，本申请不予限制。

本申请还公开了一种在其上记录有用于执行上述方法的程序的存储介质。所述存储介质包括配置为以计算机（以计算机为例）可读的形式存储或传送信息的任何机制。例如，存储介质包括只读存储器（ROM）、随机存取存储器（RAM）、磁盘存储介质、光存储介质、闪速存储介质、电、光、声或其他形式的传播信号（例如，载波、红外信号、数字信号等）等。

参照图 3，示出了本申请知识图谱的数据融合装置一实施例的结构框图，包括数据平台 10、数据预处理模块 11、实体分区模块 12、实体匹配模块 13 和实体融合模块 14，其中：

数据平台 10 配置有统一访问接口，为其他模块提供计算和存储服务。本申请对数据平台的架构没有限制，为方便在数据量增长的情况下扩展计算和存储资源，可以采用 Hadoop 分布式文件系统或云计算架构。

数据预处理模块 11 用于将来自不同数据源的数据进行处理后转换为三元组（S-P-O）格式，通过所述统一访问接口存储到数据平台 10，并接收数据平台 10 返回的图数据索引信息。其中，图数据索引信息可以是三元组格式的图数据在数据平台 10 的存储地址及其元数据。

实体分区模块 12 用于根据所述图数据索引信息，通过所述统一访问接口将数据平台 10 中存储的实体按属性划分为一个或多个子分区。具体实施时，实体分区模块 12 可以包括根据实体属性产生的全局唯一分区键对存储在数据平台中的实体进行等值划分的等值分区子模块，基于预设聚类模型对存储在数据平台中的实体进行划分的聚类分区子模块，和/或其他分区方式的子模块。

实体匹配模块 13 用于对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对。

实体融合模块 14 用于对所述匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。

本申请装置实施例的各功能模块在流水线上具有上下游依赖关系，但不同模块之间仅通过数据格式约束，通过数据平台提供的统一接口实现相互解耦，可独立开发完成。各模块的算法本身可以灵活替换，通过实现自定义阶段，可以在不同模块之间插入新的模块，自由编制完成自定义需求。例如，为了提高对各种不同数据源的适应能力以及后续实体分区、匹配和融合的准确性，可以在数据预处理模块 11 和实体分区模块 12 之间插入属性对齐模块 15，用于将来自不同数据源的经数据预处理模块 11 处理后存储在数据平台 10 中的实体根据其属性的实际含义进行对齐。如将数据源 A 的『地址』字段与数据源 B 的『Address』字段进行对齐，在后续分区和匹配阶段中被对齐的字段将当作同一含义的字段来处理。

在进一步的优选装置实施例中，实体匹配模块 13 具体可以包括相似度计算

子模块和比较子模块；其中的相似度计算子模块用于为实体本身的属性以及与该实体相关的其他实体的属性分别设置不同的权重，加权求和计算候选实体对的总体相似度；比较子模块用于判断相同子分区中的候选实体对的总体相似度是否超过预设相似度阈值，若是，则将该候选实体对作为匹配实体对。

在另一优选装置实施例中，所述装置可以还包括数据处理模块，用于通过所述统一访问接口对数据平台中的节点实体数据和边实体数据进行处理，并返回数据处理结果传递给下一个模块。

上述数据处理模块可以采用以下形式实现：

```
class CustomStage:
        def run(config）:
            input = config.input
            output = read_and_compute(input)
            return output
```

其中，对需要处理的数据通过 input 配置参数传递，数据处理完成后将结果写入 output，并传递给下一个阶段的功能模块，实现装置功能的扩展。

本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。对于本申请的装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例部分的说明即可。以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，既可以位于一个地方或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求 书

1．一种知识图谱的数据融合方法，执行所述方法的知识图谱系统包括配置有统一访问接口的数据平台，所述方法包括：

将来自不同数据源的数据进行处理后转换为三元组格式，通过所述统一访问接口存储到数据平台，并接收所述数据平台返回的图数据索引信息；

根据所述图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区；

对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对；

对所述匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。

2．根据权利要求 1 所述的方法，其中，在步骤根据所述图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区之前，还包括：将来自多个数据源的数据转换为三元组格式之后存储在数据平台中的实体根据其属性的实际含义进行对齐。

3．根据权利要求 1 所述的方法，其中，所述子分区划分方式为根据实体属性产生的全局唯一分区键进行等值划分，或基于预设聚类模型进行划分。

4．根据权利要求 1 所述的方法，其中，对划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对，具体为：

为实体本身的属性以及与该实体相关的其他实体的属性分别设置不同的权重，加权求和计算候选实体对的总体相似度；

若相同子分区中的候选实体对的总体相似度超过预设相似度阈值，则将该候选实体对作为匹配实体对。

5．根据权利要求 1 所述的方法，其中，对缺失的实体属性值进行补充的方法为通过爬虫从网络获取或进行人工填充。

6．根据权利要求 1 所述的方法，其中，所述图数据索引信息为三元组格式的图数据在所述数据平台的存储地址及其元数据。

7．一种知识图谱的数据融合装置，包括数据平台、数据预处理模块、实体分区模块、实体匹配模块和实体融合模块，其中：

所述数据平台包括统一访问接口；

所述数据预处理模块配置为将来自不同数据源的数据进行处理后转换为三元组格式，通过所述统一访问接口存储到数据平台，并接收所述数据平台返回的图数据索引信息；

所述实体分区模块配置为根据所述数据预处理模块输出的图数据索引信息，将所述数据平台中存储的实体按属性划分为一个或多个子分区；

所述实体匹配模块配置为将所述实体分区模块划分到相同子分区中的候选实体对进行相似度计算，筛选出满足预设相似度条件的匹配实体对；

所述实体融合模块配置为对所述实体匹配模块筛选出的匹配实体对的实体属性值进行补充和/或替换，生成统一的实体表示。

8．根据权利要求 7 所述的装置，其中，

所述实体分区模块包括等值分区子模块和/或聚类分区子模块；

所述等值分区子模块配置为根据实体属性产生的全局唯一分区键对存储在数据平台中的实体进行等值划分；

所述聚类分区子模块配置为基于预设聚类模型对存储在数据平台中的实体进行划分；

所述实体匹配模块具体包括相似度计算子模块和比较子模块；

所述相似度计算子模块配置为为实体本身的属性以及与该实体相关的其他实体的属性分别设置不同的权重，加权求和计算候选实体对的总体相似度；

所述比较子模块配置为判断相同子分区中的候选实体对的总体相似度是否超过预设相似度阈值，若是，则将该候选实体对作为匹配实体对。

9．根据权利要求 7 所述的装置，其中，所述装置还包括数据处理模块和/或属性对齐模块；

所述数据处理模块配置为通过所述统一访问接口对数据平台中的节点实体数据和边实体数据进行处理，并返回数据处理结果传递给下一个模块；

所述属性对齐模块配置为将来自多个数据源的数据经所述数据预处理模块处理后存储在数据平台中的实体根据其属性的实际含义进行对齐。

10．一种存储介质，所述存储介质存储有配置为执行权利要求 1~6 任一所述的方法的程序。

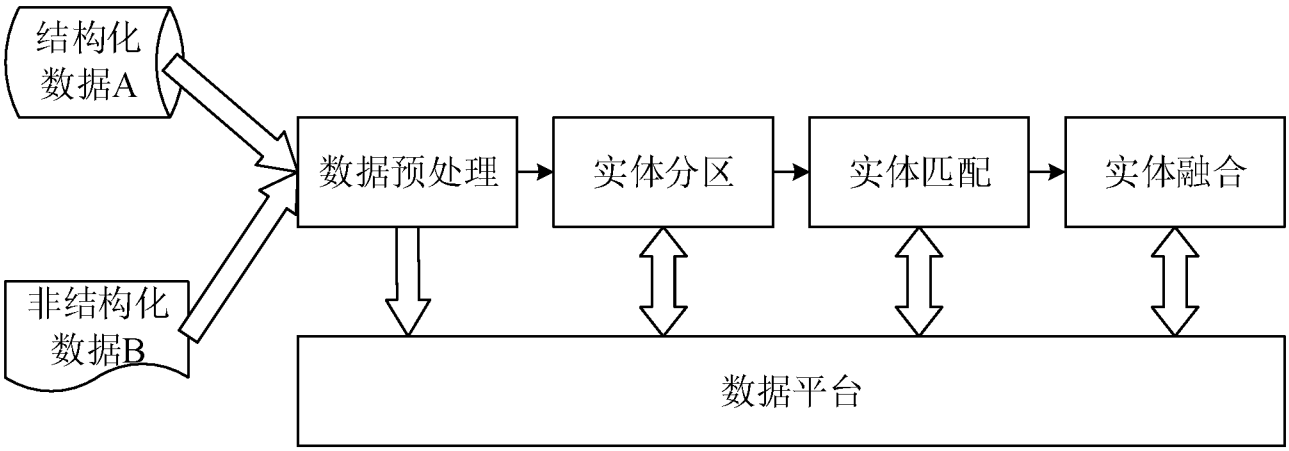
结构化
数据A
非结构化
数据B
数据预处理
实体分区
实体匹配
实体融合
数据平台

图 1

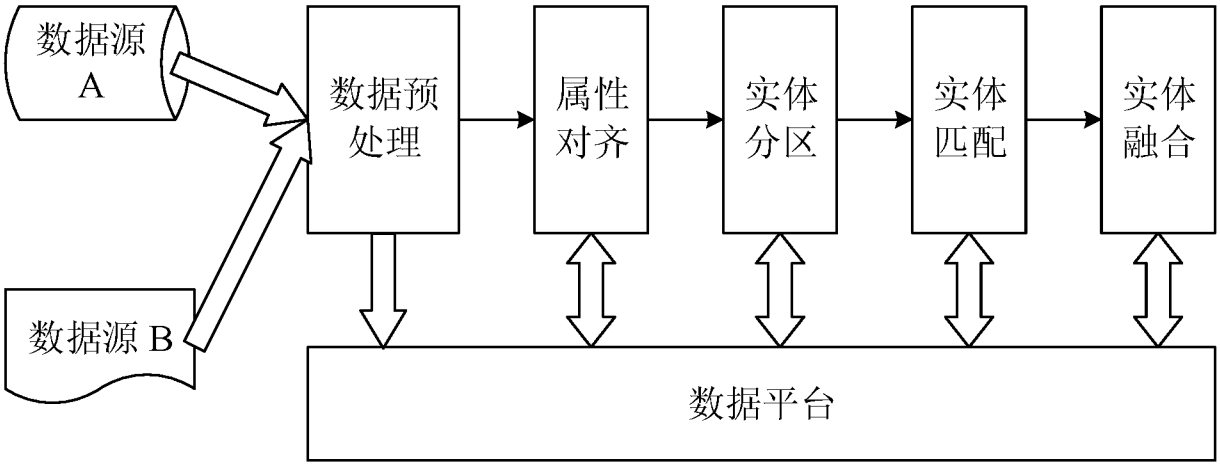
数据源
A
数据源 B
数据预
处理
属性
对齐
实体
分区
实体
匹配
实体
融合
数据平台

图 2

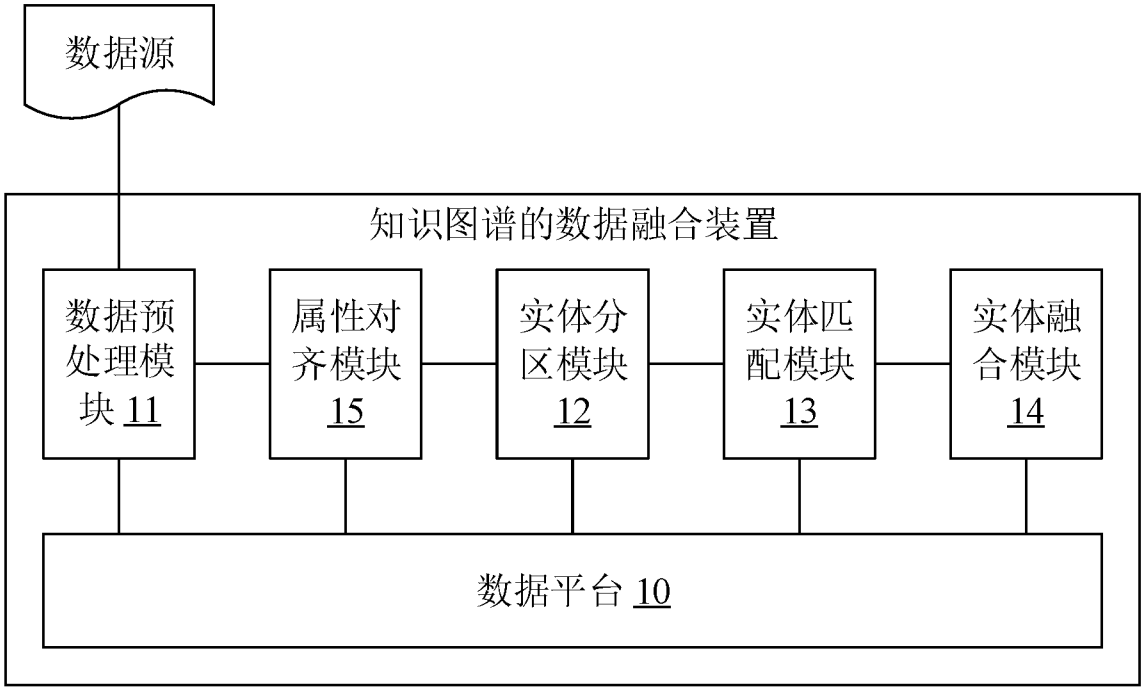
数据源
知识图谱的数据融合装置
数据预处理模块 11
属性对齐模块 15
实体分区模块 12
实体匹配模块 13
实体融合模块 14
数据平台 10

图 3

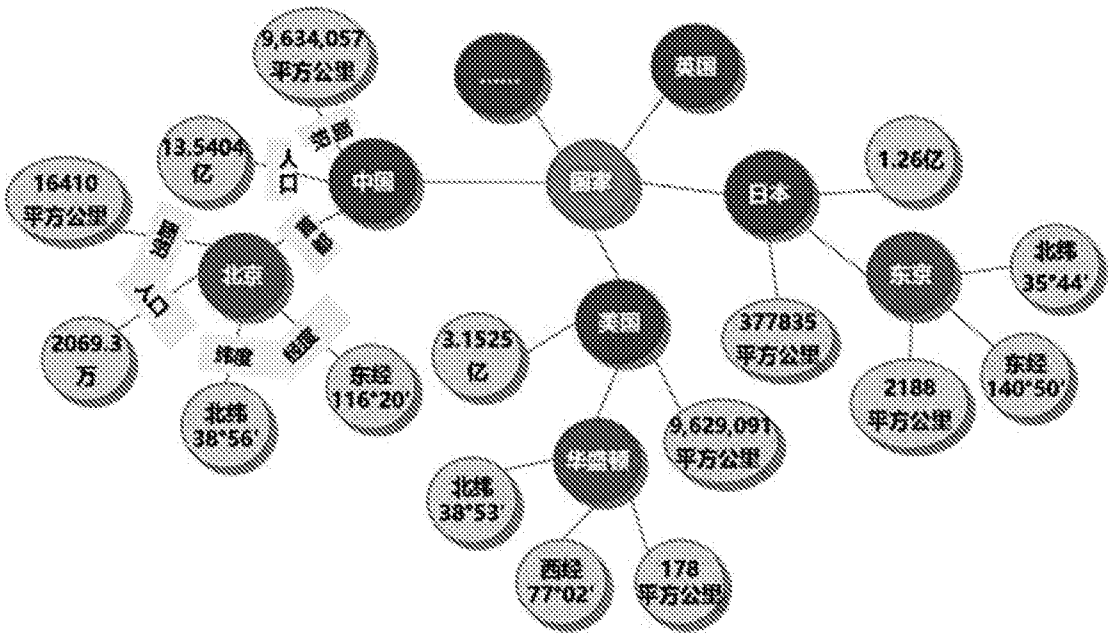
9,634,057平方公里
13.5404亿
人口
16410平方公里
2069.3万
北纬38°56'
东经116°20'
中国
北京
日本
1.26亿
377835平方公里
北纬35°44'
东经140°50'
2188平方公里
3.1525亿
9,629,091平方公里
北纬38°53'
西经77°02'
178平方公里

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/124552

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/28(2019.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

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

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

CNPAT, WPI, EPODOC, CNKI, IEEE: 知识图谱, 数据融合, 数据源, 三元组, 数据平台, 图, 索引, 实体, 属性, 划分, 相似度, 筛选, 补充, 替换, 对齐, knowledge w graph, data w fusion, data w source, triple, platform, graph, index, entity, attribute, partition, similarity, filter, supplement, replacement, alignment

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109739939 A (MIOTECH INFORMATION TECH SHANGHAI CO., LTD.) 10 May 2019 (2019-05-10) claims 1-10	1-10
X	CN 107958086 A (BEIJING RUILI TECH CO LTD) 24 April 2018 (2018-04-24) description, paragraphs [0003]-[0045], [0058], figure 3	1-10
A	CN 105956015 A (SICHUAN ZHONGRUAN TECHNOLOGY CO., LTD.) 21 September 2016 (2016-09-21) entire document	1-10
A	CN 107545046 A (BEIJING QIANXIN TECHNOLOGY CO., LTD.) 05 January 2018 (2018-01-05) entire document	1-10
A	CN 109033129 A (GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY) 18 December 2018 (2018-12-18) entire document	1-10
A	US 2015142829 A1 (FUJITSU LIMITED) 21 May 2015 (2015-05-21) entire document	1-10

☐ 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
25 February 2020	**16 March 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/124552

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109739939 A	10 May 2019	None	
CN 107958086 A	24 April 2018	None	
CN 105956015 A	21 September 2016	None	
CN 107545046 A	05 January 2018	None	
CN 109033129 A	18 December 2018	None	
US 2015142829 A1	21 May 2015	JP 2015099586 A	28 May 2015
		EP 2874073 A1	20 May 2015

国际检索报告

国际申请号
PCT/CN2019/124552

A. 主题的分类

G06F 16/28(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, WPI, EPODOC, CNKI, IEEE: 知识图谱，数据融合，数据源，三元组，数据平台，图，索引，实体，属性，划分，相似度，筛选，补充，替换，对齐，knowledge w graph, data w fusion, data w source, triple, platform, graph, index, entity, attribute, partition, similarity, filter, supplement, replacement, alignment

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 109739939 A (颖投信息科技上海有限公司) 2019年 5月 10日 (2019-05-10) 权利要求1-10	1-10
X	CN 107958086 A (北京睿力科技有限公司) 2018年 4月 24日 (2018-04-24) 说明书第[0003]-[0045], [0058]段, 图3	1-10
A	CN 105956015 A (四川中软科技有限公司) 2016年 9月 21日 (2016-09-21) 全文	1-10
A	CN 107545046 A (北京奇安信科技有限公司) 2018年 1月 5日 (2018-01-05) 全文	1-10
A	CN 109033129 A (桂林电子科技大学) 2018年 12月 18日 (2018-12-18) 全文	1-10
A	US 2015142829 A1 (FUJITSU LIMITED) 2015年 5月 21日 (2015-05-21) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 2月 25日	2020年 3月 16日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 李玉坤 电话号码 86-(10)-53961358

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/124552

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 109739939 A	2019年 5月 10日	无	
CN 107958086 A	2018年 4月 24日	无	
CN 105956015 A	2016年 9月 21日	无	
CN 107545046 A	2018年 1月 5日	无	
CN 109033129 A	2018年 12月 18日	无	
US 2015142829 A1	2015年 5月 21日	JP 2015099586 A	2015年 5月 28日
		EP 2874073 A1	2015年 5月 20日