

(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
2017年8月17日 (17.08.2017) WIPO | PCT



(10) 国际公布号

WO 2017/137000 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号:
PCT/CN2017/072995
- (22) 国际申请日:
2017年2月6日 (06.02.2017)
- (25) 申请语言:
中文
- (26) 公布语言:
中文
- (30) 优先权:
201610084741.1 2016年2月14日 (14.02.2016) CN
- (71) 申请人: 广州神马移动信息科技有限公司
(GUANGZHOU SHENMA MOBILE INFORMATION TECHNOLOGY CO., LTD.) [CN/CN]; 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元, Guangdong 510627 (CN)。
- (72) 发明人: 杨扬 (YANG, Yang); 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元, Guangdong 510627 (CN)。 穆冠宇 (MU, Guanyu); 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元, Guangdong 510627 (CN)。 华能威 (HUA, Nengwei); 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元,

Guangdong 510627 (CN)。 张伟 (ZHANG, Wei); 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元, Guangdong 510627 (CN)。 吴嘉 (WU, Jia); 中国广东省广州市天河区黄埔大道西平云路163号广电平云广场B塔12层自编01单元, Guangdong 510627 (CN)。

- (74) 代理人: 北京博雅睿泉专利代理事务所(特殊普通合伙) (BEYOND TALENT PATENT AGENT FIRM); 中国北京市朝阳区朝阳门外大街10号昆泰大厦1202单元, Beijing 100020 (CN)。
- (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, 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,

[见续页]

(54) Title: METHOD, DEVICE AND APPARATUS FOR COMBINING DIFFERENT INSTANCES DESCRIBING SAME ENTITY

(54) 发明名称: 对描述同一实体的不同实例进行合并的方法、装置及设备

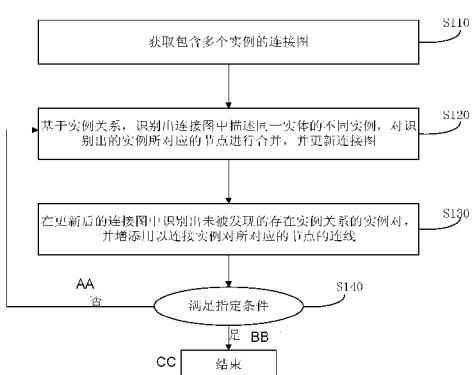


图 5

S110 Acquire a connection diagram comprising a plurality of instances
S120 Identify, from the connection diagram and on the basis of an instance relationship, different instances describing the same entity, merge the nodes corresponding to the identified instances and update the connection diagram
S130 Identify from the updated connection diagram a pair of instances having an undiscovered instance relationship existing therebetween, and add a connecting line so as to connect the nodes corresponding to the pair of instances
S140 Is specified condition satisfied?

AA No
BB Yes
CC Finish

(57) Abstract: A method, device and apparatus for combining different instances describing the same entity. The method comprises: acquiring a connection diagram comprising a plurality of instances (S110), wherein different nodes in the connection diagram indicate different instances, and connecting lines between the nodes indicates an instance relationship between the instances corresponding to the nodes; identifying, from the connection diagram and on the basis of the instance relationship, different instances describing the same entity, merging the nodes corresponding to the identified instances and updating the connection diagram (S120); identifying from the updated connection diagram a pair of instances having an undiscovered instance relationship existing therebetween, and adding a connecting line so as to connect the nodes corresponding to the pair of instances (S130); and iteratively executing the steps of updating the connection diagram on the basis of the instance relationships and adding connecting lines in the updated connection diagrams, until a specified condition is satisfied (S140). The method can more fully identify different instances describing the same entity.

(57) 摘要:

[见续页]



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))。

一种对描述同一实体的不同实例进行合并的方法、装置及设备。所述方法包括：获取包含多个实例的连接图(S110)，其中，连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；基于实例关系，识别出连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新连接图(S120)；在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接实例对所对应的节点的连线(S130)；迭代执行基于实例关系更新连接图的步骤和在更新后的连接图中增添连线的步骤，直到满足指定条件(S140)。能够较为充分地识别出描述同一实体的不同实例。

说 明 书

对描述同一实体的不同实例进行合并的方法、装置及设备

5

技术领域

本发明涉及计算机技术领域，更具体地，涉及一种对描述同一实体的不同实例进行合并的方法、装置及设备。

背景技术

知识图谱旨在描述真实世界中存在的各种实体或概念。知识图谱中的每个实体或概念用一个全局唯一确定的 ID 来标识，称为它们的标识符 (identifier)。每个属性-值对 (attribute-value pair, 又称 AVP) 用来刻画实体的内在特性，而关系 (relation) 用来连接两个实体，刻画它们之间的关联。

在知识图谱的构建过程中，需要用到不同来源的数据来构建图谱中的实体及关系，例如，为了使得构建的知识图谱可以更加全面，可以用来自百度百科、维基百科、搜狗百科等多种百科类站点来源的数据来构建知识图谱中的实体及关系。而实体在不同来源数据中往往会有存在差异化、表述不同的实例。直接使用未融合的实例数据将给知识图谱带来冗余和错误信息，因此对描述相同实体的不同实例进行融合是知识图谱构建中一个重要的任务和步骤。

目前常见的融合方法主要是通过计算不同实例间的属性相似度，将属性相似度超过阈值的实例对进行融合。这种融合方法虽然在一定程度上也能识别出描述同一实体的不同实例，但是由于这种融合方法仅以属性相似度作为融合实例的标准，使得其对融合过程中所使用的属性模糊匹配规则必须尽可能设置完善，才能有效识别同一实体的不同实例以进行融合，但这在实际应用中是很难实现的，因此很容易将表述同一实体的实例对识别为不同的实例，对知识图谱的构建带来冗余的数据。

由此，需要一种可以较为充分地识别出描述同一实体的不同实例的方案。

发明内容

本发明主要解决的技术问题是提供一种对描述同一实体的不同实例进行合并的方法、装置及设备，其能够较为充分地识别出描述同一实体的实例对。
5 实例对。

根据本发明的一个方面，提供了一种计算设备，包括：存储器，用于存储包含多个实例的连接图，其中，连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；处理器，与存储器相连接，处理器能够从存储器获取连接图，该处理器配置为：基于实例关系，识别出连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新连接图；在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接实例对所对应的节点的连线；迭代执行基于实例关系更新连接图的步骤和在更新后的连接图中增添连线的操作，直到满足指定条件。
10

由此，本发明的设备采用连接图的方式对多个待判定实例中的等价实例进行合并。而在合并的过程中又利用了实例关系，并基于合并后的连接图，扩充实例关系，然后迭代执行上述合并、扩充的步骤，使得可以较为充分地挖掘出连接图中存在的等价实例。
15

根据本发明的另一个方面，提供了一种对描述同一实体的不同实例进行合并的装置，包括：获取模块，用于获取包含多个实例的连接图，其中，连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；合并模块，用于基于实例关系，识别出连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新连接图；扩充模块，用于在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接实例对所对应的节点的连线；迭代模块，用于使得合并模块和扩充模块迭代执行更新连接图的操作和增添连线的操作，直到满足指定条件。
20
25

可选地，上述装置中述及的扩充模块可以包括：关联度计算模块，用于对于更新后的连接图中的任一节点，计算该节点所对应的实例和与该节

点通过 N 个节点进行连接的节点所对应的实例之间的关联度，其中 N 大于等于 1；第一识别模块，用于将关联度达到预定关联度阈值的两个节点所对应的实例对识别为存在实例关系的实例对，并增添连接这两个节点之间的连线。由于等价实例的合并，合并后的连接图中的实例关系也会发生一定的变化。此时，可以通过计算节点间的关联度，来发现存在实例关系的实例对。

可选地，上述装置中述及的指定条件可以设定为，扩充模块在更新后的连接图中识别出的未发现的存在实例关系的实例对的数目为零。

可选地，上述装置中述及的合并模块可以包括：分组模块，用于对多个实例进行分组；相似度计算模块，用于针对每个分组，基于实例关系计算组内任意两个实例之间的相似度；第二识别模块，用于将相似度达到预定相似度阈值的实例对识别为描述同一实体的实例对。

可选地，对于来自不同来源的两个实例，相似度计算模块可以根据以下公式计算这两个实例之间的相似度 Sim：

$$Sim = Jac_{ij} / Uniq$$

$$Jac_{ij} = \frac{C_i \cap C_j}{C_i \cup C_j}$$

$$Uniq = Log \left(Max \left(Cnt_{sourceA,i}, Cnt_{sourceB,j} \right) + 1 \right)$$

其中， C_i 为与实例 i 具有实例关系的实例集合， C_j 为与实例 j 具有实例关系的实例集合， Jac_{ij} 为实例 i 、 j 之间的实例关系相似度， $Uniq$ 为实例的唯一性度量， $Cnt_{sourceA,i}$ 为实例 i 在来源 A 中的同名实例的个数、 $Cnt_{sourceB,j}$ 为实例 j 在来源 B 中的同名实例的个数。

可选地，上述装置中述及的获取模块还可以包括：属性相似度计算模块可以计算连接图中任意两个节点所对应的实例之间的属性相似度；和第二合并模块，可以将属性相似度超过预定属性相似度阈值的两个实例所对应的节点合并为一个节点。

根据本发明的另一个方面，提供了一种对描述同一实体的不同实例进行合并的方法，该方法包括：获取包含多个实例的连接图，其中，连接图

中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；基于实例关系，识别出连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新连接图；在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接实例对所对应的节点的连线；迭代执行基于实例关系更新连接图的步骤和在更新后的连接图中增添连线的步骤，直到满足指定条件。
5

可选地，上述方法中述及的在更新后的连接图中识别出未发现的存在实例关系的实例对的步骤可以包括：对于更新后的连接图中的任一节点，计算该节点所对应的实例和与该节点通过 N 个节点进行连接的节点所对应的实例之间的关联度，其中 N 大于等于 1；将关联度达到预定关联度阈值
10 的两个节点所对应的实例对识别为存在实例关系的实例对，增添连接这两个节点之间的连线。

可选地，上述方法中述及的指定条件可以设定为，在更新后的连接图中识别出的未发现的存在实例关系的实例对的数目为零。

15 可选地，上述方法中述及的基于实例关系，识别出连接图中描述同一实体的不同实例的步骤可以包括：对多个实例进行分组；针对每个分组，基于实例关系计算组内任意两个实例之间的相似度；将相似度达到预定相似度阈值的实例对识别为描述同一实体的实例对。

可选地，对于来自不同来源的两个实例，可以根据以下公式计算这两个实例之间的相似度 Sim：
20

$$Sim = Jac_{ij} / Uniq$$

$$Jac_{ij} = \frac{C_i \cap C_j}{C_i \cup C_j}$$

$$Uniq = Log \left(Max(Cnt_{sourceA,i}, Cnt_{sourceB,j}) + 1 \right)$$

其中， C_i 为与实例 i 具有实例关系的实例集合， C_j 为与实例 j 具有实例关系的实例集合， Jac_{ij} 为实例 i 、 j 之间的实例关系相似度， $Uniq$ 为实例的唯一性度量， $Cnt_{sourceA,i}$ 为实例 i 在来源 A 中的同名实例的个数、
25

$Cnt_{sourceB,j}$ 为实例 j 在来源 B 中的同名实例的个数。

可选地，上述方法中述及的获取包含多个实例的连接图的步骤还可以包括：计算连接图中任意两个节点所对应的实例之间的属性相似度；和将属性相似度超过预定属性相似度阈值的两个实例所对应的节点合并为一个
5 节点。

本发明的对描述同一实体的不同实例进行合并的方法、装置及设备采用连接图的方式对多个实例中的等价实例进行合并，其中，在合并的过程中利用了连接图中存在的实例关系，并基于合并后的连接图扩充实例关系，然后再基于扩充的实例关系进一步发现连接图中存在的等价实例，以此类
10 推，迭代执行上述合并、扩充的步骤，使得基于本发明的方案可以较为充分地识别出描述同一实体的实例对。

通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

15 附图说明

被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连同其说明一起用于解释本发明的原理。

图 1 示出了本发明述及的连接图的示意图。

图 2 示出了根据本发明一实施例的计算设备的结构示意图。

20 图 3 示出了根据本发明一实施例的对描述同一实体的不同实例进行合并的装置的功能模块示意图。

图 4 示出了根据本发明另一实施例的对描述同一实体的不同实例进行合并的装置的功能模块示意图。

25 图 5 示出了根据本发明一实施例的对描述同一实体的不同实例进行合并的方法的示意性流程图。

图 6 示出了图 5 中的步骤 S110 可以包括的子步骤的示意性流程图。

图 7 示出了图 5 中的步骤 S120 可以包括的子步骤的示意性流程图。

图 8 示出了图 5 中的步骤 S130 可以包括的子步骤的示意性流程图。

具体实施方式

现在将参照附图来详细描述本发明的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

5 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。

在这里示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

10 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

在介绍本发明前，首先对本发明涉及的几个概念做以简要说明。

实体：知识图谱中的知识单元，具有唯一确定的 ID 标识。

15 实例：在构建知识图谱中的实体的过程中用到的各种来源的数据。

实例关系：实例间存在的关系，对于不同的数据来源，这里的关系可以是属性关系、引用关系、链接关系等多种关系。

同名实例：名称相同，但描述的实体（内容）不同的实例。

等价实例：描述同一实体（内容）的实例。

20 举例来说，百度百科中的各种词条就是实例。百度百科中的词条“李宁”是一个多义词，有指代著名体操运动员的李宁，也有指代魔术师的李宁。这里，指代魔术师的李宁和指代体操运动员的李宁就是一个同名实例。在指代著名体操运动员的李宁的词条下，还存在着“奥运冠军”、“金牌”等词条，这里，我们就可以认为“李宁”和“奥运冠军”、“金牌”存在实例关系。而百度百科中指代体操运动员的“李宁”和搜狗百科中的“体操王子”就属于等价实例。

25 本发明主要提出了一种在众多实例中识别等价实例的方案。该方案主要基于连接图的方式识别出等价实例，并不断对连接图进行更新，以识别出更多的等价实例。

具体地说，可以首先构建包含多个实例的连接图，如图 1 所示，连接图中的节点表示实例，节点间的连线表示实例关系。对于连接图中的多个实例，可以根据连接图中存在的实例关系，识别出存在的等价实例，对识别出的等价实例所对应的节点进行合并。其中，在根据实例关系对连接图 5 中的节点进行合并后，可以基于一定的识别规则，找出连接图中未发现的实例关系，根据所找到的实例关系更新连接图。然后重复执行上述基于实例关系找到等价实例的步骤和基于合并后的连接图，寻找未发现的实例关系的步骤，直到满足指定条件。

这里的指定条件可以是找不到新的实例关系或找不到新的等价实例 10 或重复步骤达到一定次数，当然还可以是其它指定条件。另外，对本方案中的对等价实例进行合并以及更新连接图的步骤来说，可以是在对连接图中的所发现的等价实例全部合并后再更新连接图。

本发明的方案可以实现为一种如图 2 所示的计算设备。该计算设备可以配置为包括存储器 1 和处理器 2。存储器 1 可以存储包含多个实例的连接图。处理器 2 与存储器 1 连接，可以从存储器 1 获取连接图，并可以执行实现上述方案中的相关步骤的操作。 15

具体地，处理器 2 例如可以是中央处理器 CPU、微处理器 MCU 等，存储器 1 例如包括 ROM（只读存储器）、RAM（随机存取存储器）、诸如硬盘的非易失性存储器等。并且，应当理解的是，尽管在图 2 中仅示出了计算设备包括存储器 1 和处理器 2，但是计算设备还可以包括接口装置、通信装置、显示装置、输入装置、扬声器、麦克风等，但是，这些部件与本发明无关，故在此省略。在本发明中，并不限制计算设备的实体实施形式， 20 计算设备可以是服务器，例如刀片服务器，也可以是计算机或者类似计算机的电子设备，例如笔记本电脑、平板电脑等。

本发明的方案还可以实现为一种包含多个功能模块的装置。其中，图 25 2 示出的处理器 2 的功能可以由该装置中相应的功能模块实现。

参见图 3，本发明的对描述同一实体的不同实例进行合并的装置可以包括获取模块 21、合并模块 22、扩充模块 23 以及迭代模块 24。其中，获取模块 21、合并模块 22、扩充模块 23 以及迭代模块 24 可以执行实现上述

方案中的相应步骤的操作。简单地说，获取模块 21 可以获取连接图。合并模块 22 可以基于连接图中存在的实例关系识别出连接图中存在的等价实例，并对识别到的等价实例所对应的节点进行合并。扩充模块 23 可以识别出连接图中未发现的实例关系。迭代模块 24 可以使得合并模块 22 和扩充模块 23 迭代执行相应的操作，直到满足指定条件。
5 模块 23 迭代执行相应的操作，直到满足指定条件。

参见图 4，获取模块 21 可以包括属性相似度计算模块 211 和第二合并模块 212。合并模块 22 可以包括分组模块 221、相似度计算模块 222 以及第二识别模块 223。扩充模块 23 可以包括关联度计算模块 231 和第一识别模块 232。对于图 4 所示的结构来说，获取模块 21、合并模块 22 及扩充模块 10 模块 23 的功能可以由其包括的相应子模块实现，此处暂不做具体描述。

图 5 至图 8 详细示出了执行本发明的方案的流程图。其中，图 5 至图 8 所示的各个步骤都可以由上文提及的处理器或装置中的相应功能模块实现，下面结合图 5 至图 8 对本发明的方案的工作流程进行详细说明。

参见图 5，在步骤 S110，由处理器 2 或者获取模块 21，获取包含多个 15 实例的连接图。

这里述及的获取连接图的步骤可以是获取事先构建好的连接图。例如，可以事先根据多个实例构建连接图，然后存储在存储器中，需要处理时，由处理器 2 或获取模块 21 从存储器获取。

也可以是根据需要判定的实例数据构建连接图。例如，可以根据待判定的多个实例数据及实例数据中存在的实例关系，构建连接图。对于构建好的连接图，可以将其存储在存储器中，需要处理时，再由处理器 2 或获取模块 21 从存储器获取连接图，当然也可以将构建好的连接图直接发送给处理器 2 或获取模块 21。
20

在执行步骤 S110 的过程中，还可以基于一定的识别规则识别出连接图中存在的等价实例，并合并等价实例所对应的节点。这里述及的识别规则可以是如图 6 所示的属性相似度的识别方式。
25

如图 6 所示，在步骤 S1110，由处理器 2 或者由获取模块 21 中的属性相似度计算模块 211 计算连接图中任意两个节点所对应的实例间的属性相似度。

在步骤 S1120，由处理器 2 或者由获取模块 21 中的第二合并模块 212 将属性相似度超过预定属性相似度阈值的实例所对应的节点合并为一个节点。

应该知道，在步骤 S110 中对连接图中的节点进行合并的步骤（步骤 5 S110、步骤 S1120）是本发明的一个可选方案，这样使得可以基于现有的计算方式初步发现连接图中存在的等价实例，并对其进行合并，以降低后续步骤的复杂度。

返回步骤 S110，在执行完步骤 S110 后，就可以执行步骤 S120，由处理器 10 2 或者由合并模块 21，基于实例关系，识别出连接图中描述同一实体的不同实例（即等价实例），对识别出的等价实例所对应的节点进行合并，并更新连接图。

其中，可以有多种基于实例关系识别连接图中的等价实例的方式。例如，可以在计算实例间的相似度的过程中，将与当前实例存在实例关系的实例参与到相似度的计算的过程中，然后将相似度超过阈值的实例识别 15 为等价实例。

图 7 示出了一种基于实例关系识别出等价实例的具体实施方式。

如图 7 所示，在步骤 S1210，由处理器 2 或者由合并模块 22 中的分组模块 221，对连接图中的多个实例进行分组。

其中，可以有多种分组方式，如可以根据名称进行分组，还可以根据 20 属性值进行。当然根据具体情况，还有其它分组方式，此处不再赘述。

在步骤 S1220，针对每个分组，可以由处理器 2 或者由分组模块 22 中的相似度计算模块 222，基于实例关系计算组内任意两个实例之间的相似度。

其中，对于来自不同数据来源的两个实例来说，可以根据下述公式计算这两个实例间的相似度 Sim：

$$Sim = Jac_{ij} / Uniq$$

$$Jac_{ij} = \frac{C_i \cap C_j}{C_i \cup C_j}$$

$$Uniq = Log \left(Max \left(Cnt_{sourceA,i}, Cnt_{sourceB,j} \right) + 1 \right)$$

其中, C_i 为与实例 i 具有实例关系的实例集合, C_j 为与实例 j 具有实例关系的实例集合, Jac_{ij} 为实例 i 、 j 之间的实例关系相似度, $Uniq$ 为实例的唯一性度量, $Cnt_{sourceA,i}$ 为实例 i 在来源 A 中的同名实例的个数、 $Cnt_{sourceB,j}$ 为实例 j 在来源 B 中的同名实体的个数。

5 其中, 对于不同来源的实例数据来说, 上述公式可以有不同形式的变形。以实例数据来源为百科词条来说, 可以基于下列公式计算来自不同百科的两个实例间的相似度 Sim :

$$Sim = (\alpha \times Jac_{out} + (1 - \alpha) \times Jac_{in}) / Uniq$$

$$Jac_{out} = \frac{C_{iout} \cap C_{jout}}{C_{iout} \cup C_{jout}}, Jac_{in} = \frac{C_{iin} \cap C_{jin}}{C_{iin} \cup C_{jin}}$$

$$Uniq = Log(\max(Cnt_{sourceA,i}, Cnt_{sourceB,j}) + 1)$$

其中, α 为权重系数, C_{iout} 为待判定实例 i 链出的实例的个数, C_{jout} 为待判定实例 j 链出的实例的个数, C_{iin} 为待判定实例 i 被链入的实例的个数, C_{jin} 为待判定实例 j 被链入的实例的个数, Jac_{out} 为待判定实例 i 、 j 链出的实例的相似度, Jac_{in} 为待判定实例 i 、 j 被链入的实体的相似度, $Uniq$ 为实例的唯一性度量, $Cnt_{sourceA,i}$ 为待判定实例 i 在来源 A 中的同名实例的个数、 $Cnt_{sourceB,j}$ 为待判定实例 j 在来源 B 中的同名实体的个数。

15 以百度百科和搜狗百科为例对上述变形公式加以说明。以百度百科中的词条“李宁”和搜狗百科中的词条“李宁”来说。在百度百科中, 词条“李宁”具有 60 个同名实例, 在搜狗百科中, 词条“李宁”具有 52 个同名实例。而对于表示体操运动员的“李宁”, 该词条在百度百科中存在着“奥运冠军”、“金牌”、“自由体操”等内链词条, 这些词条与“李宁”
20 就存在实例关系, 这些词条就可以看成是词条“李宁”的链出的词条(实例)。而词条“体操王子”下存在词条“李宁”, 此时, “体操王子”就是“李宁”被链入的词条(实例), 词条“体操王子”与词条“李宁”也存在实例关系。此时, 基于上述变形公式就可以计算出百度百科中的词条“李宁”和搜狗百科中的词条“李宁”之间的相似度。

其中，上述计算公式可以在分布式计算平台如 SPARK 上并行实现，达到大规模并行化图计算的目的。另外，应该知道，对于其它来源的实例数据来说，还可以有其它基于实例关系计算相似度的方式，此处不再赘述。

在步骤 S1230，由处理器 2 或者由分组模块 22 中的第二识别模块 223 5 将相似度达到预定相似度阈值的实例对识别为等价实例。由此，就可以基于实例关系识别出连接图中存在的等价实例。

下面返回步骤 S120，在执行完步骤 S120 后，就可以执行步骤 S130，由处理器 2 或者由扩充模块 23，在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接实例对所对应的节点的连线。

10 对于执行步骤 S120 后的连接图，由于等价实例的合并，合并后的连接图中的实例关系也会发生一定的变化。此时，可以使用一定的识别规则识别出连接图中新增的存在实例关系的实例对。

图 8 示出了一种识别出连接图中未发现的实例关系的具体实施方式。

15 如图 8 所示，在步骤 S1310，对于更新后的连接图中的任一节点，由处理器 2 或者由扩充模块 23 中的关联度计算模块 231 计算该节点所对应的实例和与该节点通过 N 个节点进行连接的节点所对应的实例之间的关联度，N 大于等于 1。

在步骤 S1320，由处理器 2 或者由扩充模块 23 中的第一识别模块 232 20 将关联度达到预定关联度阈值的两个节点所对应的实例对识别为存在实例关系的实例对，增添连接这两个节点之间的连线。

其中，可以有多种计算关联度的方式。例如，对于图 1 中的节点 D 和节点 L 来说，节点 D 和节点 L 通过节点 A、节点 E 两个节点进行连接，这样就可以通过分析节点 A 和节点 E 之间的相似度的大小，来判断节点 D 和 E 之间是否存在关联度。

25 下面返回步骤 S130，对于经过步骤 S130 扩充后实例关系的连接图，可以执行步骤 S140，由处理器 2 或者可以由迭代模块 24 判断是否满足指定条件，在不满足指定条件的情况下，返回步骤 S120，重复执行 S120、S130、S140 的步骤。直至满足指定条件，输出合并后的连接图。

其中，步骤 S140 中的指定条件可以是重复执行步骤 S120、S130、S140

的次数达到一定值。也可以是在重复执行步骤 S120、S130、S140 的过程中，在步骤 S120，在扩充后实例关系的连接图中找不到新的等价实例（作为优选，可以是步骤 S120 连续多次识别不到新的等价实例）。还可以是在 S130 的执行过程中，找不到新的实例关系。作为优选，可以将在 S130 的执行过程中，找不到新的实例关系作为指定条件。
5

至此，参考附图详细描述了根据本发明的对描述同一实体的不同实例进行合并的方法、装置及设备。通过上述描述可知，本发明的对描述同一实体的不同实例进行合并的方法、装置及设备采用连接图的方式对多个实例中的等价实例进行合并。其中，在合并的过程中利用了连接图中存在的实例关系，并基于合并后的连接图扩充实例关系，然后再基于扩充的实例关系进一步发现连接图中存在的等价实例，以此类推，迭代执行上述合并、扩充的步骤，使得连接图可以并行化传播，并使得基于本发明的方案可以更充分地挖掘出等价实例。
10
10

此外，根据本发明的方法还可以实现为一种计算机程序，该计算机程序包括用于执行本发明的上述方法中限定的上述各步骤的计算机程序代码指令。或者，根据本发明的方法还可以实现为一种计算机程序产品，该计算机程序产品包括计算机可读介质，在该计算机可读介质上存储有用于执行本发明的上述方法中限定的上述功能的计算机程序。本领域技术人员还将明白的是，结合这里的公开所描述的各种示例性逻辑块、模块、电路和
20 算法步骤可以被实现为电子硬件、计算机软件或两者的组合。

附图中的流程图和框图显示了根据本发明的多个实施例的系统和方法的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也
25 应当注意，在有些作为替换的实现中，方框中所标记的功能也可以以不同于附图中所标记的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实

现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

权利要求书

1. 一种计算设备，包括：

5 存储器，用于存储包含多个实例的连接图，其中，所述连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；以及

处理器，与所述存储器相连接，所述处理器能够从所述存储器获取所述连接图，该处理器配置为：

10 基于所述实例关系，识别出所述连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新所述连接图；

在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接所述实例对所对应的节点的连线；

15 迭代执行所述基于实例关系更新连接图的步骤和所述在更新后的连接图中增添连线的操作，直到满足指定条件。

2. 一种对描述同一实体的不同实例进行合并的装置，包括：

20 获取模块，用于获取包含多个实例的连接图，其中，所述连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；

合并模块，用于基于所述实例关系，识别出所述连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新所述连接图；

25 扩充模块，用于在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接所述实例对所对应的节点的连线；

迭代模块，用于使得所述合并模块和所述扩充模块迭代执行更新所述连接图的操作和增添连线的操作，直到满足指定条件。

3. 根据权利要求 2 所述的装置，其中，所述扩充模块包括：

关联度计算模块，用于对于更新后的连接图中的任一节点，计算该节点所对应的实例和与该节点通过 N 个节点进行连接的节点所对应的实例之间的关联度，其中 N 大于等于 1；

第一识别模块，用于将所述关联度达到预定关联度阈值的两个节点所对应的实例对识别为存在实例关系的实例对，并增添连接这两个节点之间的连线。

4. 根据权利要求 2 或 3 所述的装置，其中，所述指定条件被设定为，所述扩充模块在更新后的连接图中识别出的未发现的存在实例关系的实例对的数目为零。

5. 根据权利要求 2-4 中任意一项所述的装置，其中，所述合并模块包括：

分组模块，用于对所述多个实例进行分组；

相似度计算模块，用于针对每个分组，基于实例关系计算组内任意两个实例之间的相似度；

第二识别模块，用于将相似度达到预定相似度阈值的实例对识别为描述同一实体的实例对。

20 6. 根据权利要求 2-5 中任意一项所述的装置，其中，对于来自不同来源的两个实例，所述相似度计算模块根据以下公式计算这两个实例之间的相似度 Sim ：

$$Sim = Jac_{ij} / Uniq$$

$$Jac_{ij} = \frac{C_i \cap C_j}{C_i \cup C_j}$$

$$Uniq = Log \left(Max \left(Cnt_{sourceA,i}, Cnt_{sourceB,j} \right) + 1 \right)$$

其中， C_i 为与实例 i 具有实例关系的实例集合， C_j 为与实例 j 具有实例关系的实例集合， Jac_{ij} 为实例 i 、 j 之间的实例关系相似度， $Uniq$ 为实

例的唯一性度量， $Cnt_{sourceA,i}$ 为实例 i 在来源 A 中的同名实例的个数、 $Cnt_{sourceB,j}$ 为实例 j 在来源 B 中的同名实例的个数。

7. 根据权利要求 2-6 中任意一项所述的装置，其中，所述获取模块

5 还包括：

属性相似度计算模块，用于计算连接图中任意两个节点所对应的实例之间的属性相似度；和

第二合并模块，用于将所述属性相似度超过预定属性相似度阈值的两个实例所对应的节点合并为一个节点。

10

8. 一种对描述同一实体的不同实例进行合并的方法，包括：

获取包含多个实例的连接图，其中，所述连接图中的不同节点表示不同实例，节点间的连线表示节点所对应的实例之间的实例关系；

15 基于所述实例关系，识别出所述连接图中描述同一实体的不同实例，对识别出的实例所对应的节点进行合并，并更新所述连接图；

在更新后的连接图中识别出未发现的存在实例关系的实例对，并增添用以连接所述实例对所对应的节点的连线；

迭代执行所述基于所述实例关系更新所述连接图的步骤和所述在更新后的连接图中增添连线的步骤，直到满足指定条件。

20

9. 根据权利要求 8 所述的方法，其中，所述在更新后的连接图中识别出未发现的存在实例关系的实例对的步骤包括：

对于更新后的连接图中的任一节点，计算该节点所对应的实例和与该节点通过 N 个节点进行连接的节点所对应的实例之间的关联度，其中 N 大于等于 1；

将所述关联度达到预定关联度阈值的两个节点所对应的实例对识别为存在实例关系的实例对，增添连接这两个节点之间的连线。

10. 根据权利要求 8 或 9 所述的方法，其中，所述指定条件被设定为，

在更新后的连接图中识别出的未发现的存在实例关系的实例对的数目为零。

11. 根据权利要求 8-10 中任意一项所述的方法，其中，所述基于实例关系，识别出连接图中描述同一实体的不同实例的步骤包括：

5 对所述多个实例进行分组；

针对每个分组，基于所述实例关系计算组内任意两个实例之间的相似度；

10 将相似度达到预定相似度阈值的实例对识别为描述同一实体的实例对。

12. 根据权利要求 8-11 中任意一项所述的方法，其中，对于来自不同来源的两个实例，根据以下公式计算这两个实例之间的相似度 Sim ：

$$Sim = Jac_{ij} / Uniq$$

$$Jac_{ij} = \frac{C_i \cap C_j}{C_i \cup C_j}$$

$$Uniq = Log \left(Max \left(Cnt_{sourceA,i}, Cnt_{sourceB,j} \right) + 1 \right)$$

15 其中， C_i 为与实例 i 具有实例关系的实例集合， C_j 为与实例 j 具有实例关系的实例集合， Jac_{ij} 为实例 i 、 j 之间的实例关系相似度， $Uniq$ 为实例的唯一性度量， $Cnt_{sourceA,i}$ 为实例 i 在来源 A 中的同名实例的个数、 $Cnt_{sourceB,j}$ 为实例 j 在来源 B 中的同名实例的个数。

20

13. 根据权利要求 8-12 中任意一项所述的方法，其中，所述获取包含多个实例的连接图的步骤还包括：

计算连接图中任意两个节点所对应的实例之间的属性相似度；和

25 将所述属性相似度超过预定属性相似度阈值的两个实例所对应的节点合并为一个节点。

说 明 书 附 图

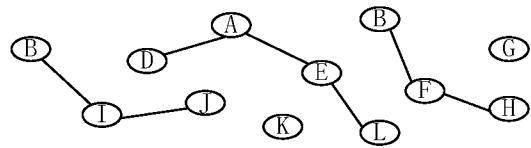


图 1

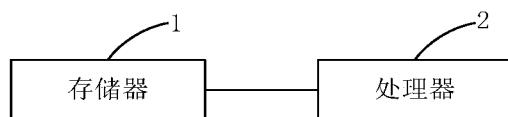


图 2

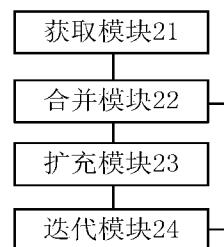


图 3

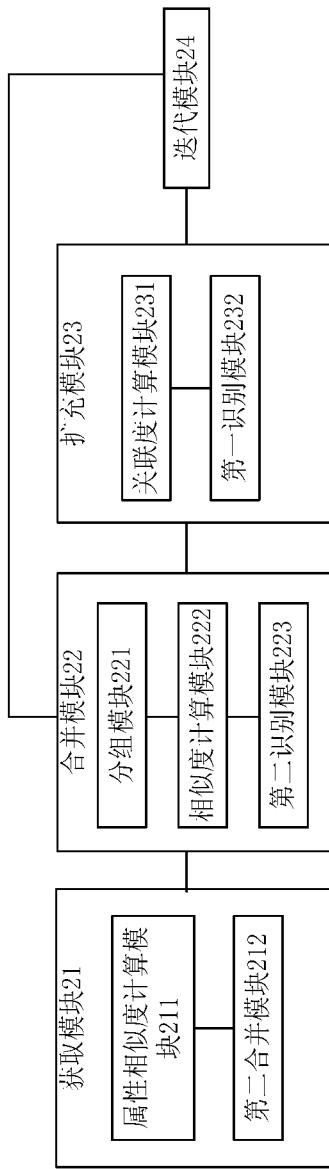


图4

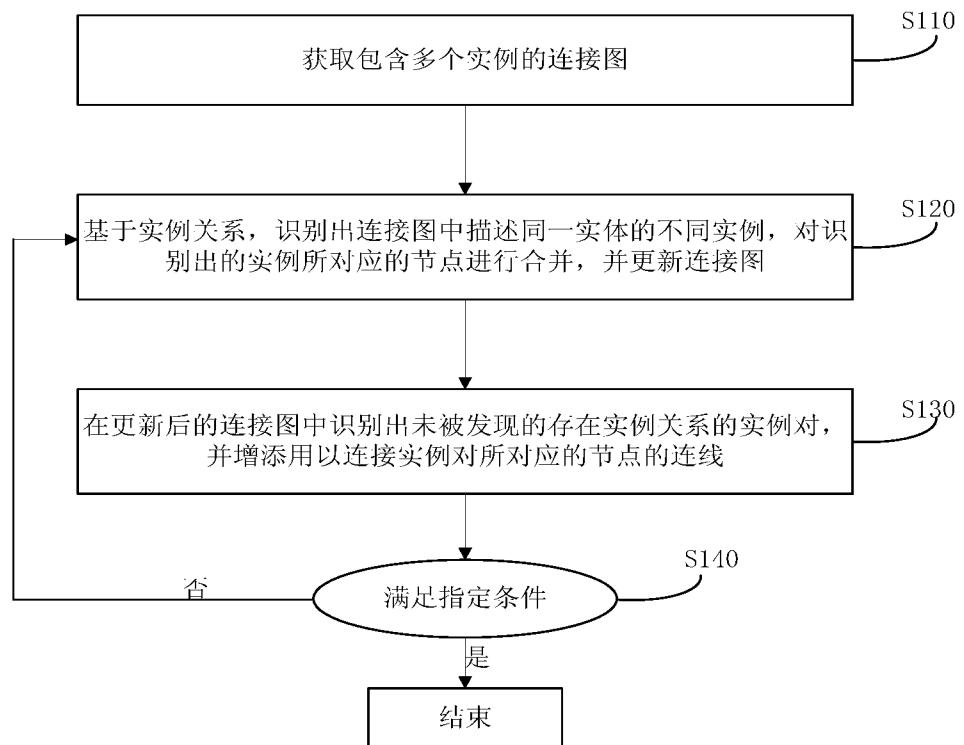


图 5

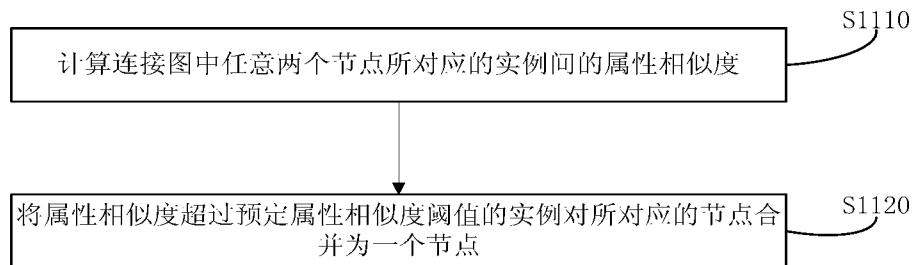


图 6

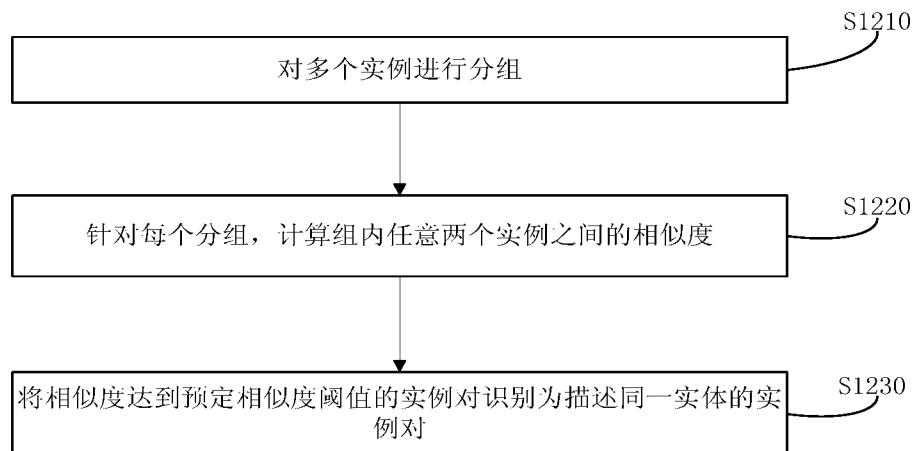


图 7

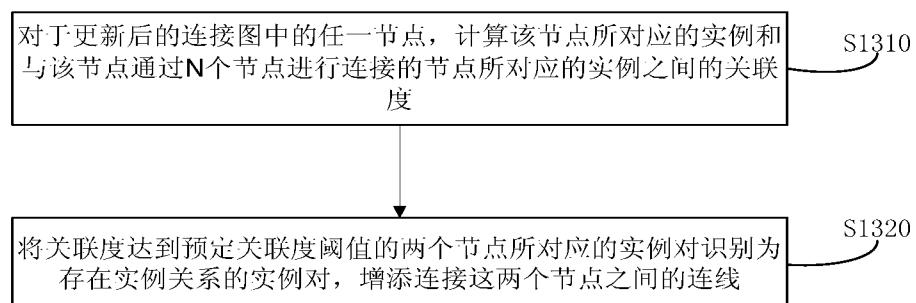


图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/072995

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

WPI, EPODOC, CNPAT, CNKI: entity, example, node?, tree, graph, merg+, combin+, similarit+, threshold, predetermined, preset, reference

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105786980 A (GUANGZHOU SHENMA MOBILE INFORMATION TECHNOLOGY CO., LTD.) 20 July 2016 (20.07.2016) claims 1-13	1-13
X	CN 101714142 A (ESOBI HOLDING CO., LTD.) 26 May 2010 (26.05.2010) description, paragraphs [0040]-[0052], and figure 2	1-13
A	CN 105045863 A (ZHANGJIAGANG INSTITUTE OF INDUSTRIAL TECHNOLOGIES SOOCHOW UNIVERSITY) 11 November 2015 (11.11.2015) the whole document	1-13
A	CN 101667201 A (ZHE JIANG UNIVERSITY) 10 March 2010 (10.03.2010) the whole document	1-13

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
06 April 2017

Date of mailing of the international search report
28 April 2017

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
FANG, Lei
Telephone No. (86-10) 61648120

INTERNATIONAL SEARCH REPORTInternational application No.
PCT/CN2017/072995

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015081656 A1 (SAP AG) 19 March 2015 (19.03.2015) the whole document	1-13
A	US 2009307213 A1 (DENG, XIAOTIE et al.) 10 December 2009 (10.12.2009) the whole document	1-13
A	US 2005192926 A1 (IBM) 01 September 2005 (01.09.2005) the whole document	1-13

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/072995

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105786980 A	20 July 2016	None	
CN 101714142 A	26 May 2010	CN 101714142 B	17 October 2012
CN 105045863 A	11 November 2015	None	
CN 101667201 A	10 March 2010	None	
US 2015081656 A1	19 March 2015	US 9430584 B2 CN 104462084 A	30 August 2016 25 March 2015
US 2009307213 A1	10 December 2009	US 2014304267 A1 US 8676815 B2	09 October 2014 18 March 2014
US 2005192926 A1	01 September 2005	CN 1658234 B CN 1658234 A	26 May 2010 24 August 2005

国际检索报告

国际申请号

PCT/CN2017/072995

A. 主题的分类

G06F 17/30 (2006. 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))

WPI, EPODOC, CNPAT, CNKI: 实体, 实例, 相似度, 近似度, 节点, 合并, 树, 图, 阈值, 门槛, 门坎, node?, tree, graph, merg+, combin+, similarit+, threshold, predetermined, preset, reference

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 105786980 A (广州神马移动信息科技有限公司) 2016年 7月 20日 (2016 - 07 - 20) 权利要求1-13	1-13
X	CN 101714142 A (易搜比控股公司) 2010年 5月 26日 (2010 - 05 - 26) 说明书第[0040]-[0052]段、附图2	1-13
A	CN 105045863 A (苏州大学张家港工业技术研究院) 2015年 11月 11日 (2015 - 11 - 11) 全文	1-13
A	CN 101667201 A (浙江大学) 2010年 3月 10日 (2010 - 03 - 10) 全文	1-13
A	US 2015081656 A1 (SAP AG) 2015年 3月 19日 (2015 - 03 - 19) 全文	1-13
A	US 2009307213 A1 (DENG, XIAOTIE等) 2009年 12月 10日 (2009 - 12 - 10) 全文	1-13
A	US 2005192926 A1 (IBM) 2005年 9月 1日 (2005 - 09 - 01) 全文	1-13

 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

2017年 4月 6日

国际检索报告邮寄日期

2017年 4月 28日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

受权官员

方蕾

传真号 (86-10)62019451

电话号码 (86-10)61648120

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/072995

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	105786980	A	2016年 7月 20日	无			
CN	101714142	A	2010年 5月 26日	CN	101714142	B	2012年 10月 17日
CN	105045863	A	2015年 11月 11日	无			
CN	101667201	A	2010年 3月 10日	无			
US	2015081656	A1	2015年 3月 19日	US	9430584	B2	2016年 8月 30日
				CN	104462084	A	2015年 3月 25日
US	2009307213	A1	2009年 12月 10日	US	2014304267	A1	2014年 10月 9日
				US	8676815	B2	2014年 3月 18日
US	2005192926	A1	2005年 9月 1日	CN	1658234	B	2010年 5月 26日
				CN	1658234	A	2005年 8月 24日

表 PCT/ISA/210 (同族专利附件) (2009年7月)