

(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



(10) 国际公布号

WO 2020/108641 A1

(43) 国际公布日
2020 年 6 月 4 日 (04.06.2020)

(51) 国际专利分类号:
G06F 16/36 (2019.01)

省 杭州市 滨江区 阡陌路 555 号, Zhejiang
310051 (CN)。

(21) 国际申请号: PCT/CN2019/122159

(22) 国际申请日: 2019 年 11 月 29 日 (29.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811459195.0 2018 年 11 月 30 日 (30.11.2018) CN

(71) 申请人: 杭州海康威视数字技术股份有限公司 (HANGZHOU HIKVISION DIGITAL TECHNOLOGY CO., LTD.) [CN/CN]; 中国浙江

(72) 发明人: 吴小飞 (WU, Xiaofei); 中国浙江省杭州市滨江区阡陌路 555 号, Zhejiang 310051 (CN)。浦世亮 (PU, Shiliang); 中国浙江省杭州市滨江区阡陌路 555 号, Zhejiang 310051 (CN)。姜伟浩 (JIANG, Weihao); 中国浙江省杭州市滨江区阡陌路 555 号, Zhejiang 310051 (CN)。闫春 (YAN, Chun); 中国浙江省杭州市滨江区阡陌路 555 号, Zhejiang 310051 (CN)。

(74) 代理人: 北京三高永信知识产权代理有限公司 (BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市

(54) Title: ONTOLOGY ESTABLISHMENT METHOD AND APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 本体建立方法、装置及存储介质

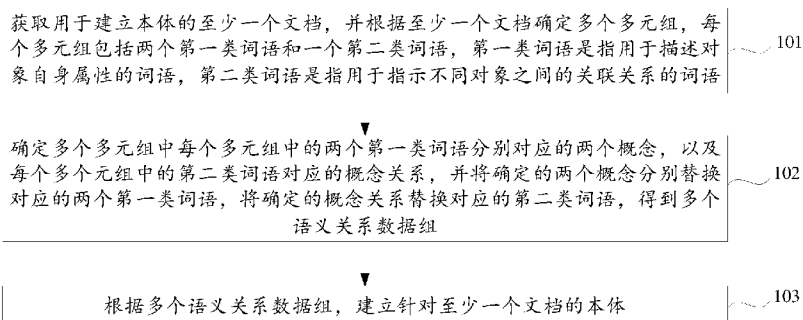


图 1

- 101 Acquire at least one document for establishing an ontology, and determine a plurality of tuples according to the at least one document, each tuple comprising two first-type words and one second-type word, the first-type word being a word for describing the attribute of an object itself, and the second-type word being a word for indicating an association relationship between different objects
- 102 Determine two concepts respectively corresponding to the two first-type words in each tuple among the plurality of tuples, and a concept relationship corresponding to the second-type word in each tuple among the plurality of tuples, and replace the corresponding two first-type words with the determined two concepts and replace the corresponding second-type word with the determined concept relationship, so as to obtain a plurality of semantic relationship data sets
- 103 Establish, according to the plurality of semantic relationship data sets, an ontology for the at least one document

(57) Abstract: The present application discloses an ontology establishment method and apparatus, and a storage medium, belonging to the technical field of big data processing. Said method comprises: for at least one document used for establishing an ontology, determining a plurality of tuples according to the at least one document, determining a plurality of semantic relationship data sets according to the plurality of tuples, and then establishing, according to the plurality of semantic relationship data sets, an ontology for the at least one document. The whole ontology establishment process involves no manual operation, improving the ontology establishing efficiency.

(57) 摘要: 本申请公开了一种本体建立方法、装置及存储介质，属于大数据处理技术领域。所述方法包括：对于用于建立本体的至少一个文档，根据至少一个文档确定多个多元组，根据多个多元组确定多个语义关系数据组，然后根据多个语义关系数据组，建立针对至少一个文档的本体。整个建立本体的过程中，无需人工参与，提高了建立本体的效率。

海淀区学院路蓟门里和景园A座1单元
102室, Beijing 100088 (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, 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))。

本体建立方法、装置及存储介质

本申请要求于 2018 年 11 月 30 日提交的申请号为 201811459195.0、发明名称为“本体建立方法、装置及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及大数据处理技术领域，特别涉及一种本体建立方法、装置及存储介质。

背景技术

本体是一种数据结构，该数据结构用于对某些文档采用标准的、规范的方式进行说明，以便于根据该本体对这些文档进行共享。其中，本体包括多个概念和多个概念中存在关联的两个概念之间的概念关系，且概念关系是有方向的，概念关系的方向用于指示这两个概念之间的归属关系。比如，针对描述雨伞的多个文档建立一个本体，该本体中包括多个概念，比如雨伞、用具、伞把等。其中，概念“雨伞”和概念“用具”之间存在一个概念关系，且该概念关系由“雨伞”指向“用具”，用于指示概念“雨伞”归属于概念“用具”。概念“雨伞”和概念“伞把”之间也存在一个概念关系，且该概念关系由“伞把”指向“雨伞”，用于指示概念“伞把”归属于概念“雨伞”。

相关技术中，当需要建立本体时，获取用于建立本体的至少一个文档，通过人工方式对至少一个文档中每个文档中的词语进行归类，然后根据归类之后的多个词语建立本体。该建立本体的过程人工参与度较高，严重影响了建立本体的效率。

发明内容

本申请实施例提供了一种本体建立方法、装置及存储介质，可以提高建立本体的效率。所述技术方案如下：

第一方面，提供了一种本体建立方法，所述方法包括：

获取用于建立本体的至少一个文档，根据所述至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，所述第一类词语是指用于指示对象自身属性的词语，所述第二类词语是指用于指示不同对象之间的关联关系的词语；

确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，采用确定的两个概念分别替换对应的两个第一类词语，采用确定的概念关系替换对应的第二类词语，得到多个语义关系数据组；

根据所述多个语义关系数据组，建立针对所述至少一个文档的本体。

可选地，所述确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，包括：

对于所述多个多元组中的多元组 A，从参考数据库中查找与所述多元组 A 中的两个第一类词语分别对应的两个概念，所述参考数据库包括多个词语中每个词语所代表的概念以及不同概念之间的概念关系，所述多元组 A 为所述多个多元组中任一个；

根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系。

可选地，所述从参考数据库中查找所述多元组 A 中的两个第一类词语分别对应的两个概念，包括：

对于所述多元组 A 中两个第一类词语中的第一类词语 B，确定所述至少一个文档中与所述第一类词语 B 的词义相同的词语，所述第一类词语 B 为所述两个第一类词语中任一个；

如果所述至少一个文档中不存在词义与所述第一类词语 B 的词义相同的词语，则从所述参考数据库中查找所述第一类词语 B 对应的概念。

可选地，所述确定所述至少一个文档中词义与所述第一类词语 B 的词义相同的词语之后，还包括：

如果所述至少一个文档中存在词义与所述第一类词语 B 的词义相同的词语，则确定所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语分别在所述至少一个文档中的出现次数；

从所述第一类词语 B，以及与所述第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；

从所述参考数据库中查找所述出现次数最大的词语对应的概念，将查找到的概念作为所述第一类词语 B 对应的概念。

可选地，所述方法还包括：

如果没有从所述参考数据库中查找到所述第一类词语 B 对应的概念，则通过 LDA（Latent Dirichlet Allocation，潜在狄利克雷分布）算法为所述第一类词语 B 创建一个概念。

可选地，所述根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系，包括：

从所述参考数据库中确定所述查找到的两个概念之间的路径，得到多条路径；

根据所述多条路径的路径长度，从所述多条路径中选择一条目标路径，所述目标路径中包括至少一个概念关系；

从所述至少一个概念关系中选择与所述多元组 A 中的第二类词语之间相似度最大的概念关系，将选择的概念关系确定为与所述多元组 A 中的第二类词语对应的概念关系。

可选地，所述根据所述多个语义关系数据组，建立针对所述至少一个文档的本体，包括：

根据所述多个语义关系数据组，建立语义关系图，所述语义关系图中的一个结点对应所述多个语义关系组中的一个概念，所述语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且所述两个结点之间的方向为对应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向相应结点的其他结点的数量，每个结点的出度是指相应结点指向的其他结点的数量；

对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪；

当对所述语义关系图中所有入度等于 0 且出度大于 0 的结点均裁剪完时，将所述语义关系图中所有入度等于 0 且出度大于 0 的结点指向的结点的入度设置为 0；

对除了已经裁剪的结点之外的其他节点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的步骤，直至遍历完所述语义关系图中的所有结点，将最终得到的语义关系图作为针对所述至少一个文档建立的本体。

可选地, 所述对所述第一结点进行裁剪, 包括:

确定所述语义关系图中的至少一个结点, 所述至少一个结点对应的概念所指示的词语在所述至少一个文档中的文档位置与所述第一结点对应的概念所指示的词语在所述至少一个文档中的文档位置相邻;

确定所述至少一个结点中与所述第一结点连接度最大的结点, 删除所述至少一个结点中除确定出的结点之外的其他结点, 以及所述其他结点与所述第一结点之间的关系, 任意两个结点之间的连接度用于指示所述两个结点之间的关联程度。

可选地, 所述确定所述至少一个结点中与所述第一结点连接度最大的结点, 包括:

基于第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度;

其中, 所述第一公式为:
$$Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

所述 W_i 和所述 W_j 分别为用于确定连接度的两个结点, 所述 $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度, 所述 $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度, 所述 α 和所述 β 分别为针对相似度配置的加权系数和针对相关度配置的加权系数, 所述 α 和所述 β 的和为 1;

根据按照所述第一公式确定的连接度, 确定所述至少一个结点中与所述第一结点连接度最大的结点。

可选地, 所述根据按照所述第一公式确定的连接度, 确定所述至少一个结点中与所述第一结点连接度最大的结点, 包括:

如果按照所述第一公式确定的连接度中的最大连接度小于连接度阈值, 则调整所述 α 和所述 β 的数值, 得到更新之后的第一公式;

基于更新之后的第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度;

如果按照所述更新之后的第一公式确定的连接度中的最大连接度小于所述连接度阈值, 则返回执行调整所述 α 和所述 β 的数值的步骤, 直至确定出的最大连接度大于或等于所述连接度阈值, 将最后一次确定出的最大连接度对应的结点确定为所述至少一个结点中与所述第一结点连接度最大的结点。

可选地, 所述对于入度等于 0 且出度大于 0 的任一第一结点, 对所述第一结点进行裁剪之前, 还包括:

如果所述语义关系图中存在入度等于 0 且出度也等于 0 的孤立结点，将所述孤立结点配置为指向所述语义关系图中的入度等于 0 且出度大于 0 的一个结点上，得到更新之后的语义关系图；

基于更新之后的语义关系图，执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的操作。

可选地，所述根据所述至少一个文档确定多个多元组，包括：

对所述至少一个文档中每个文档进行分词处理，得到多个词语；

确定所述多个词语中每个词语的词性；

根据所述多个词语中每个词语的词性和每个词语在所述至少一个文档中的文档位置，确定所述多个多元组。

第二方面，提供了一种本体建立装置，所述装置包括：

获取模块，用于获取用于建立本体的至少一个文档，根据所述至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，所述第一类词语是指用于指示对象自身属性的词语，所述第二类词语是指用于指示不同对象之间的关联关系的词语；

确定模块，用于确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，采用确定的两个概念分别替换对应的两个第一类词语，采用确定的概念关系替换对应的第二类词语，得到多个语义关系数据组；

建立模块，用于根据所述多个语义关系数据组，建立针对所述至少一个文档的本体。

可选地，所述确定模块包括：

第一查找单元，用于对于所述多个多元组中的多元组 A，从参考数据库中查找与所述多元组 A 中的两个第一类词语分别对应的两个概念，所述参考数据库包括多个词语中每个词语所代表的概念以及不同概念之间的概念关系，所述多元组 A 为所述多个多元组中任一个；

第二查找单元，用于根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系。

可选地，所述第一查找单元，用于：

对于所述多元组 A 中两个第一类词语中的第一类词语 B，确定所述至少一

个文档中词义与所述第一类词语 B 的词义相同的词语，所述第一类词语 B 为所述两个第一类词语中任一个；

如果所述至少一个文档中不存在与所述第一类词语 B 的词义相同的词语，则从所述参考数据库中查找所述第一类词语 B 对应的概念。

可选地，所述第一查找单元，还用于：

如果所述至少一个文档中存在词义与所述第一类词语 B 的词义相同的词语，则确定所述第一类词语 B，以及与所述第一类词语 B 的词义相同的词语分别在所述至少一个文档中的出现次数；

从所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；

从所述参考数据库中查找所述出现次数最大的词语对应的概念，并将查找到的概念作为所述第一类词语 B 对应的概念。

可选地，所述装置还包括：

创建单元，用于如果没有从所述参考数据库中查找到所述第一类词语 B 对应的概念，则通过 LDA 算法为所述第一类词语 B 创建一个概念。

可选地，所述第二查找单元，用于：

从所述参考数据库中确定所述查找到的两个概念之间的路径，得到多条路径；

根据所述多条路径的路径长度，从所述多条路径中选择一条目标路径，所述目标路径中包括至少一个概念关系；

从所述至少一个概念关系中选择与所述多元组 A 中的第二类词语之间相似度最大的概念关系，将选择的概念关系确定为与所述多元组 A 中的第二类词语对应的概念关系。

可选地，所述建立模块包括：

建立单元，用于根据所述多个语义关系数据组，建立语义关系图，所述语义关系图中的一个结点对应所述多个语义关系组中的一个概念，所述语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且所述两个结点之间的方向为对应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向相应结点的其他结点的数量，每个结点的出度是指相应结点指向的其他结点的数量；

裁剪单元，用于对于入度等于 0 且出度大于 0 的任一第一结点，对所述第

一结点进行裁剪；

设置单元，用于当对所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点均裁剪完时，将所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点指向的其他结点的入度设置为 0；

所述裁剪单元，还用于对除已经裁剪的第一结点之外的其他结点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的步骤，直至遍历完所述语义关系图中的所有结点，将最终得到的语义关系图作为针对所述至少一个文档建立的本体。

可选地，所述裁剪单元，用于：

确定所述语义关系图中的至少一个结点，所述至少一个结点对应的概念所指示的词语在所述至少一个文档中的文档位置与所述第一结点对应的概念所指示的词语在所述至少一个文档中的文档位置相邻；

确定所述至少一个结点中与所述第一结点连接度最大的结点，删除所述至少一个结点中除确定出的结点之外的其他结点，以及所述其他结点与所述第一结点之间的关系，任意两个结点之间的连接度用于指示所述两个结点之间的关联程度。

可选地，所述裁剪单元，还用于：

基于第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度；

其中，所述第一公式为：
$$Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

所述 W_i 和所述 W_j 分别为用于确定连接度的两个结点，所述 $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度，所述 $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度，所述 α 和所述 β 分别为针对相似度配置的加权系数和针对相关度配置的加权系数，所述 α 和所述 β 的和为 1；

根据按照所述第一公式确定的连接度，确定所述至少一个结点中与所述第一结点连接度最大的结点。

可选地，所述裁剪单元，还用于：

如果按照所述第一公式确定的连接度中的最大连接度小于连接度阈值，则调整所述 α 和所述 β 的数值，得到更新之后的第一公式；

基于更新之后的第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度；

如果按照所述更新之后的第一公式确定的连接度中的最大连接度小于所述连接度阈值，则返回执行调整所述 α 和所述 β 的数值的步骤，直至确定出的最大连接度大于或等于所述连接度阈值，将最后一次确定出的最大连接度对应的结点确定为所述至少一个结点中与所述第一结点连接度最大的结点。

可选地，所述建立模块还包括：

配置单元，用于如果所述语义关系图中存在入度等于 0 且出度也等于 0 的孤立结点，将所述孤立结点配置为指向所述语义关系图中的入度等于 0 且出度大于 0 的一个结点上，得到更新之后的语义关系图；

所述裁剪单元，还用于基于更新之后的语义关系图，执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的操作。

可选地，所述获取模块包括：

分词处理单元，用于对所述至少一个文档中每个文档进行分词处理，得到多个词语；

第一确定单元，用于确定所述多个词语中每个词语的词性；

第二确定单元，用于根据所述多个词语中每个词语的词性和每个词语在所述至少一个文档中的文档位置，确定所述多个多元组。

第三方面，提供了一种本体建立装置，所述装置包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行上述第一方面所述的任一项方法的步骤。

第四方面，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上存储有指令，所述指令被处理器执行时实现上述第一方面所述的任一项方法的步骤。

第五方面，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第一方面所述的任一方法的步骤。

本申请实施例提供的技术方案带来的有益效果是：

在本申请实施例中，对于用于建立本体的至少一个文档，根据至少一个文

档确定多个多元组，根据多个多元组确定多个语义关系数据组，然后根据多个语义关系数据组，建立针对至少一个文档的本体。整个建立本体的过程中，无需人工参与，提高了建立本体的效率。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 是本申请实施例提供的一种本体建立方法流程图；

图 2 是本申请实施例提供的一种查询概念和概念关系的方法流程图；

图 3 是本申请实施例提供的一种概念图的示意图；

图 4 是本申请实施例提供的一种本体建立装置示意图；

图 5 是本申请实施例提供的一种终端的结构示意图；

图 6 是本申请实施例提供的一种服务器的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

图 1 是本申请实施例提供的一种本体建立方法流程图。如图 1 所示，该方法包括如下步骤：

步骤 101：获取用于建立本体的至少一个文档，并根据至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，第一类词语是指用于指示对象自身属性的词语，第二类词语是指用于指示不同对象之间的关联关系的词语。

在本申请实施例中，为了确保建立的本体能够最大可能描述至少一个文档中的内容，在建立本体时，先根据至少一个文档确定多个多元组。其中，每个多元组包括两个第一类词语和一个第二类词语，第一类词语是指用于指示对象自身属性的词语，比如名词。第二类词语是指用于指示不同对象之间的关联关系的词语，比如动词。且每个多元组中包括的词语均是至少一个文档中的词语，如此可以确保后续建立的本体是根据至少一个文档中的信息得到的，提高了建

立的本体的准确性。

其中，本申请实施例提供的本体建立方法可以由终端来执行，此时，用于建立本体的至少一个文档可以是管理员预先输入至终端的。本申请实施例提供的本体建立方法也可以由服务器来执行，此时，用于建立本体的至少一个文档可以是管理员预先通过终端上传至服务器的。本申请实施例在此不做限定。

在一种可能的实现方式中，根据至少一个文档确定多个多元组可以为：对至少一个文档中每个文档进行分词处理，得到多个词语；确定多个词语中每个词语的词性；根据多个词语中每个词语的词性和每个词语在至少一个文档中的文档位置，确定多个多元组。

其中，对至少一个文档进行分词处理可以采用基于字符串匹配中的正向最大匹配分词方法，当然也可以采用其他的分词方法，本申请实施例在此不做限定。

另外，由于每个词语可能有不同的词性，因此可以将多个词语按照各个词语在文档中的位置进行划分，得到多个词语序列，比如，可以将一段文本中的所有词语组合为一个词语序列。此时，确定多个词语中每个词语的词性的实现方式可以为：对于任一词语序列，按照该词语序列包括的每个词语对应的不同的词性，将不同的词性进行组合，得到针对该词语序列的多个词性序列，每个词性序列包括与该词语序列中的词语一一对应的词性。通过评估函数判断该多个词性序列中每个词性序列的概率，从多个词性序列中选择概率最大的词性序列，并将概率最大的词性序列中的词性确定为对应的词语的词性。

另外，由于按照分词处理之后的多个词语是按照在文档中的位置顺序排列的，因此，根据多个词语中每个词语的词性和每个词语在至少一个文档中的文档位置，确定多个多元组的实现方式可以为：对于任一第一类词语，在多个词语中查找排序在该第一类之后且与该第一类词语相邻的第二类词语，再继续查找排序在该第二类词语之后且与该第二类词语相邻的一个第一类词语，将这三个词语组合，便可得到一个多元组。

可选地，由于至少一个文档除了第一类词语和第二类词语之外，可能还存在其他词性的词语，比如副词或介词等，而这些其他词性的词语对建立本体的贡献度不大，因此在上述对至少一个文档中每个文档进行分词处理之前，还可以先对至少一个文档进行无用词过滤操作，以过滤掉至少一个文档中无用的词语。

另外，在本申请实施例中，由于用于建立本体的文档通常数量较多，为了提高建立本体的效率，可以通过基于分布式计算框架的计算模型来执行步骤 101。示例性地，将至少一个文档划分到不同的分区，各个分区之间可以并行地执行上述步骤 101，进而提高了建立本体的速度。其中，将至少一个文档划分到不同的分区可以是指将每个文档划分到一个分区，也可以是指相邻的几个段落划分到一个分区，或者是一个段落划分到一个分区，本申请实施例在此不做限定。

其中，当至少一个文档被划分到不同的分区时，下述步骤 102 是以分区为单位执行的，也即是，各个分区并行地根据划入自身的文档执行步骤 102。

步骤 102: 确定多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，并将确定的两个概念分别替换对应的两个第一类词语，将确定的概念关系替换对应的第二类词语，得到多个语义关系数据组。

由于目前已经存在有大量用于表征概念的数据库，因此，在本申请实施例中，可以基于参考数据库执行步骤 102。也即是，基于参考数据库确定每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系。由于确定每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系的实现方式基本相同，因此，下面以多个多元组中多元组 A 为例说明如何确定一个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系。其中，多元组 A 为多个多元组中任一多元组。本申请实施例提供的一种查询概念和概念关系的方法流程图，如图 2 所示，该方法包括如下两个步骤。

步骤 1021: 从参考数据库中查找与多元组 A 中的两个第一类词语分别对应的两个概念，参考数据库包括多个词语中每个词语所代表的概念以及不同概念之间的概念关系。

其中，参考数据库可以为 HowNet（知网）数据库。为了后续便于说明，在此对 HowNet 数据库进行简单介绍：HowNet 数据库是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系的数据库。HowNet 数据库采用 KDML (Knowledge database mark-up language, 知识系统描述语言)，利用嵌套式的结构，对概念以及概念的属性进行描述。也即是，对于一个复杂的概念用较简单的概念进行解释，较简单的概

念再用更简单的概念解释，直到能够用义原表示为止，义原是指能够表达词义的最基本的单元。这种结构其实质是一种隐含的图结构，称之为概念图。

图 3 是本申请实施例提供的一种概念图的示意图。如图 3 所示，概念“场馆”在 HowNet 数据库中的描述可以为：NO.（编号）=129348，W_C（概念名称）=奥运场馆，DEF（定义文件扩展名，Definitions File Name Extension）={facilities|设施，domain（领域）={sports|体育}，location（地方）={compete|比赛、exercise|锻炼}。在“场馆”的概念描述中，KDML 表示了这样的含义：场馆是一个设施，这个设施所属的领域(Domain)是体育领域，这个设施是比赛的地方 (Location)，这个设施也是锻炼的地方 (Location)。也就是说，场馆是一个用来进行体育比赛和健身锻炼的场所。

基于上述参考数据库的介绍可知，参考数据库中的一个概念可能对应至少一个文档中的一个词语，也可以对应至少一个文档中的多个词语。比如，词语“饭碗”、“职业”、“差事”就是多个词义相近的词，它们代表的是同一个概念。因此，从参考数据库中查找多元组 A 中的两个第一类词语分别对应的两个概念的实现方式可以为：对于多元组 A 中两个第一类词语中的第一类词语 B，确定至少一个文档中词义与第一类词语 B 的词义相同的词语；如果至少一个文档中不存在词义与第一类词语 B 的词义相同的词语，则从参考数据库中查找第一类词语 B 对应的概念。其中，第一类词语 B 为多元组 A 中两个第一类词语中任一个。

如果至少一个文档中存在词义与第一类词语 B 的词义相同的词语，则确定第一类词语 B，以及词义与第一类词语 B 的词义相同的词语分别在至少一个文档中的出现次数；从第一类词语 B，以及词义与第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；从参考数据库中查找出现次数最大的词语对应的概念，并将查找到的概念作为第一类词语 B 对应的概念。

其中，当至少一个文档被划分到不同的分区时，由于各个分区并行地根据划入自身的文档执行步骤 102，因此，上述实现方式中，与第一类词语 B 的词义相同的词语是指：在第一类词语 B 所在的分区内词义与第一类词语 B 的词义相同的词语。

另外，在按照上述方式从参考数据库中查找与第一类词语 B 对应的概念时，可能存在没有查找到的情况，此时则通过 LDA（Latent Dirichlet Allocation，潜在狄利克雷分布）算法为第一类词语 B 创建一个概念。其中，LDA 算法是文本

语义分析的一个重要模型。

其中，当至少一个文档被划分到不同的分区时，由于各个分区并行地根据划入自身的文档执行步骤 102，此时，为每个分区配置一个 LDA 列表，对于任一分区，该 LDA 列表用于存放该分区内没有从参考数据库中查找到对应的概念的词。因此，通过 LDA 算法为第一类词语 B 创建一个概念的实现方式可以为：对第一类词语 B 所在的分区的 LDA 列表中的所有词语进行概念建模，得到针对该 LDA 列表的一个概念树，将该概念树添加到参考数据库中，然后再从参考数据库中查找第一类词语 B 对应的概念，相当于为第一类词语 B 创建了一个概念。

步骤 1022：根据查找到的两个概念，从参考数据库中继续查找与多元组 A 中的第二类词语对应的概念关系。

由图 3 可知，参考数据库中的不同的两个概念之间存在的的多条路径，每条路径上存在不止一个概念关系，因此，步骤 1022 的实现方式可以为：从参考数据库中确定查找到的两个概念之间的路径，得到多条路径；根据多条路径的路径长度，从多条路径中选择一条目标路径，目标路径中包括至少一个概念关系；从至少一个概念关系中选择与多元组 A 中的第二类词语之间相似度最大的概念关系，并将选择的概念关系确定为与多元组 A 中的第二类词语对应的概念关系。

其中，根据多条路径的路径长度，从多条路径中选择一条目标路径的实现方式可以为：从多条路径中路径长度最大的一条路径作为目标路径。当然，本申请实施例中还可以根据其他原则确定目标路径，在此不做限定。

另外，由于在文档中通常存在一些通用的动词，比如“是”“为”等动词，这些动词对于建立本体的贡献度并不大，因此，在从参考数据库中确定查找到的两个概念之间的路径之前，还可以确定多元组 A 包括的第二类词语的 TF-IDF (term frequency-inverse document frequency, 信息检索数据挖掘的常用加权技术) 值，根据 TF-IDF 值确定该第二类词语是否为通用的动词，如果是，则过滤掉该通用的动词，此时则不再执行“从参考数据库中确定查找到的两个概念之间的路径”的步骤，而是将多元组 A 中的第二类词语对应的概念关系设置为空值。如果根据 TF-IDF 值确定该第二类词语不是通用的动词，则执行上述“从参考数据库中确定查找到的两个概念之间的路径”，以从参考数据库中确定为与多元组 A 中的第二类词语对应的概念关系。

通过上述步骤 101 和步骤 102 可以得到多个语义关系数据组，由于每个语

义关系数据组中包括两个概念和这两个概念之间的概念关系，因此可以通过下述步骤 103 建立针对至少一个文档的本体。

步骤 103: 根据多个语义关系数据组，建立针对至少一个文档的本体。

在本申请实施例汇中，由于每个语义关系数据组中包括两个概念和这两个概念之间的概念关系，因此可以直接将多个语义关系数据组所指示的网状结构数据作为针对至少一个文档的本体。

可选地，当通过分区模式执行上述步骤 101 和步骤 102 时，不同的分区之间可能存在相同的概念，导致多个语义关系数据组所指示的网状结构数据比较臃肿，不利于其他用户从多个语义关系数据组所指示的网状结构数据中查询信息。因此，在本申请实施例中，还可以对多个语义关系数据组所指示的网状结构数据进行裁剪，以提高建立的本体的通用性。

其中，对多个语义关系数据组所指示的网状结构数据进行裁剪的实现方式可以为：根据多个语义关系数据组，建立语义关系图，语义关系图中的一个结点对应多个语义关系组中的一个概念，语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且两个结点之间的方向为对应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向该结点的其他结点的数量，每个结点的出度是指该结点指向的其他结点的数量。对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪；当对语义关系图中所有入度等于 0 且出度大于 0 的第一结点均裁剪完时，将语义关系图中所有入度等于 0 且出度大于 0 的结点指向的结点的入度设置为 0；对除了已经裁剪的第一结点之外的其他结点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪的步骤，直至遍历完语义关系图中的所有结点，将最终得到的语义关系图作为针对至少一个文档建立的本体。

其中，根据多个语义关系数据组，建立语义关系图是指将多个语义关系数据组中的各个概念按照概念关系进行关联，得到的数据体系称之为语义关系图。

另外，在对语义关系图进行裁剪时，采用的是广度优先的方式遍历语义关系图中所有的结点。其中，广度优先是指先处理入度等于 0 且出度大于 0 的结点，再处理入度等于 0 且出度大于 0 的结点直接指向的结点，该过程如上述实现方式。当然，也可以其他的遍历方式来遍历语义关系图中的结点，比如深度优先的遍历方式，本申请实施例在此不再一一说明。

在一种可能的实现方式中，对第一结点进行裁剪可以为：确定语义关系图

中的至少一个结点，这至少一个结点对应的概念所指示的词语在至少一个文档中的文档位置与第一结点对应的概念所指示的词语在至少一个文档中的文档位置相邻；确定至少一个结点中与第一结点连接度最大的结点，删除至少一个结点中除确定出的结点之外的其他结点，并删除其他结点与第一结点之间的关系。

其中，当至少一个文档被划分到不同的分区时，至少一个结点则是指对应的概念所指示的词语与第一结点对应的概念所指示的词语在同一分区内。

另外，任意两个结点之间的连接度用于指示这两个结点之间的关联程度，在本申请实施例中，确定至少一个结点中与第一结点连接度最大的结点的实现方式可以为：基于第一公式确定至少一个结点中每个结点与第一结点之间的连接度，根据按照第一公式确定的连接度，确定至少一个结点中与第一结点连接度最大的结点。

$$\text{其中，第一公式为： } Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

W_i 和 W_j 分别为用于确定连接度的两个结点， $Joint(W_i, W_j)$ 表示结点 W_i 和结点 W_j 之间的连接度， $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度， $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度， α 和 β 分别为针对相似度配置的加权系数和针对相关度配置的加权系数， α 和 β 的和为 1。

其中，结点 W_i 和结点 W_j 之间的相似度可以通过结点 W_i 和结点 W_j 在语义关系图中的路径的长度来确定，本申请实施例在此不再详细阐述。 W_i 和结点 W_j 之间的相关度可以通过结点 W_i 对应的词语和结点 W_j 对应的词语在至少一个文档中的出现次数来确定。并且，当至少一个文档被划分到不同的分区时，结点 W_i 对应的词语和结点 W_j 对应的词语在至少一个文档中的出现次数是指结点 W_i 对应的词语和结点 W_j 对应的词语在同一分区中出现次数。因此，通过本申请实施例提供的连接度的确定方式，可以使得连接度结合了词句上下文语境和参考数据库中包含的相似度关系，从而使得确定出的概念关系更加符合两个词语之间的关系。

在上述实现方式中，在确定至少一个结点中每个结点与第一结点之间的连接度之后，可以直接确定连接度最大的结点，这种实现方式操作简单，易于实现。但是实际应用中，可能存在确定至少一个结点中每个结点与第一结点之间的连接度之后，最大的连接度不满足指定要求，这样按照上述实现方式直接确定出连接度最大的结点可能不能代表与第一节点关联度最大的结点。

因此，可选地，如果按照第一公式确定的连接度中的最大连接度小于连接

度阈值，则调整 α 和 β 的数值，得到更新之后的第一公式；基于更新之后的第一公式确定至少一个结点中每个结点与第一结点之间的连接度；如果按照更新之后的第一公式确定的连接度中的最大连接度小于连接度阈值，则返回执行调整 α 和 β 的数值的步骤，直至确定出的最大连接度大于或等于连接度阈值，将最后一次确定出的最大连接度对应的结点确定为至少一个结点中与第一结点连接度最大的结点。

另外，由于在步骤 102 中，多元组 A 中的第二类词语对应的概念关系可能为空值，因此，语义关系图中可能存在入度等于 0 且出度也等于 0 的孤立结点。对于这些孤立结点，可以先将这些孤立结点并入到其他结点中，然后在语义关系图进行裁剪。也即是，在对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪之前，还可以执行如下操作：如果语义关系图中存在入度等于 0 且出度也等于 0 的孤立结点，将孤立结点配置为指向语义关系图中的入度等于 0 且出度大于 0 的一个结点上，得到更新之后的语义关系图；基于更新之后的语义关系图，执行对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪的操作。通过这种合并的方式，可以确保将步骤 102 建立的本体中的每个概念都与其他概念之间存在关系。

在本申请实施例中，对于用于建立本体的至少一个文档，根据至少一个文档确定多个多元组，根据多个多元组确定多个语义关系数据组，然后根据多个语义关系数据组，建立针对至少一个文档的本体。整个建立本体的过程中，无需人工参与，提高了建立本体的效率。

图 4 是本申请实施例提供的一种本体建立装置。如图 4 所示，该装置 400 包括：

获取模块 401，用于获取用于建立本体的至少一个文档，并根据至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，第一类词语是指用于描述对象自身属性的词语，第二类词语是指用于指示不同对象之间的关联关系的词语；

确定模块 402，用于确定多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，并将确定的两个概念分别替换对应的两个第一类词语，将确定的概念关系替换对应的第二类词语，得到多个语义关系数据组；

建立模块 403，用于根据多个语义关系数据组，建立针对至少一个文档的本体。

可选地，确定模块 402 包括：

第一查找单元，用于对于多个多元组中任一多元组 A，从参考数据库中查找与多元组 A 中的两个第一类词语分别对应的两个概念，参考数据库是指用于描述多个词语中每个词语所代表的概念以及不同概念之间的概念关系的数据库；

第二查找单元，用于根据查找到的两个概念，从参考数据库中继续查找与多元组 A 中的第二类词语对应的概念关系。

可选地，第一查找单元，用于：

对于多元组 A 中两个第一类词语中的任一第一类词语 B，确定至少一个文档中与第一类词语 B 的词义相同的词语；

如果至少一个文档中不存在与第一类词语 B 的词义相同的词语，则从参考数据库中查找第一类词语 B 对应的概念。

可选地，第一查找单元，还用于：

如果至少一个文档中存在与第一类词语 B 的词义相同的词语，则确定第一类词语 B，以及与第一类词语 B 的词义相同的词语分别在至少一个文档中的出现次数；

从第一类词语 B，以及与第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；

从参考数据库中查找出现次数最大的词语对应的概念，并将查找到的概念作为第一类词语 B 对应的概念。

可选地，装置 400 还包括：

创建单元，用于如果没有从参考数据库中查找到第一类词语 B 对应的概念，则通过 LDA 算法为第一类词语 B 创建一个概念。

可选地，第二查找单元，用于：

从参考数据库中确定查找到的两个概念之间的路径，得到多条路径；

根据多条路径的路径长度，从多条路径中选择一条目标路径，目标路径中包括至少一个概念关系；

从至少一个概念关系中选择与多元组 A 中的第二类词语之间相似度最大的概念关系，并将选择的概念关系确定为与多元组 A 中的第二类词语对应的概念

关系。

可选地，建立模块 403 包括：

建立单元，用于根据多个语义关系数据组，建立语义关系图，语义关系图中的一个结点对应多个语义关系组中的一个概念，语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且两个结点之间的方向为对应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向每个结点的结点的数量，每个结点的出度是指每个结点指向的结点的数量；

裁剪单元，用于对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪；

设置单元，用于当对语义关系图中所有入度等于 0 且出度大于 0 的结点均裁剪完时，将语义关系图中所有入度等于 0 且出度大于 0 的结点指向的结点的入度设置为 0；

裁剪单元，还用于对除了已经裁剪的结点之外的其他节点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对第一结点进行裁剪的步骤，直至遍历完语音关系图中的所有结点，将最终得到的语义关系图作为针对至少一个文档建立的本体。

可选地，裁剪单元，用于：

确定语义关系图中的至少一个结点，至少一个结点对应的概念所指示的词语在至少一个文档中的文档位置与第一结点对应的概念所指示的词语在至少一个文档中的文档位置相邻；

确定至少一个结点中与第一结点连接度最大的结点，删除至少一个结点中除确定出的结点之外的其他结点，并删除其他结点与第一结点之间的关系。

可选地，裁剪单元，还用于：

基于第一公式确定至少一个结点中每个结点与第一结点之间的连接度；

其中，第一公式为：
$$Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

W_i 和 W_j 分别为用于确定连接度的两个结点， $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度， $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度， α 和 β 分别为针对相似度配置的加权系数和针对相关度配置的加权系数， α 和 β 的和为 1；

根据按照第一公式确定的连接度，确定至少一个结点中与第一结点连接度

最大的结点。

可选地，裁剪单元，还用于：

如果按照第一公式确定的连接度中的最大连接度小于连接度阈值，则调整 α 和 β 的数值，得到更新之后的第一公式；

基于更新之后的第一公式确定至少一个结点中每个结点与第一结点之间的连接度；

如果按照更新之后的第一公式确定的连接度中的最大连接度小于连接度阈值，则返回执行调整 α 和 β 的数值的步骤，直至确定出的最大连接度大于或等于连接度阈值，将最后一次确定出的最大连接度对应的结点确定为至少一个结点中与第一结点连接度最大的结点。

可选地，建立模块还包括：

配置单元，用于如果语义关系图中存在入度等于0且出度也等于0的孤立结点，将孤立结点配置为指向语义关系图中的入度等于0且出度大于0的一个结点上，得到更新之后的语义关系图；

裁剪单元，还用于基于更新之后的语义关系图，执行对于入度等于0且出度大于0的任一第一结点，对第一结点进行裁剪的操作。

可选地，获取模块401包括：

分词处理单元，用于对至少一个文档中每个文档进行分词处理，得到多个词语；

第一确定单元，用于确定多个词语中每个词语的词性；

第二确定单元，用于根据多个词语中每个词语的词性和每个词语在至少一个文档中的文档位置，确定多个多元组。

在本申请实施例中，对于用于建立本体的至少一个文档，根据至少一个文档确定多个多元组，根据多个多元组确定多个语义关系数据组，然后根据多个语义关系数据组，建立针对至少一个文档的本体。整个建立本体的过程中，无需人工参与，提高了建立本体的效率。

需要说明的是：上述实施例提供的本体建立装置在建立本体时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的本体建立装置与本体建立方法实施例属于同一构思，其实现过程详见方法实施例，这里不再赘述。

图 5 是本申请实施例提供的一种终端 500 的结构框图。该终端 500 可以是：智能手机、平板电脑、MP3 播放器（Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面 3）、MP4（Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面 4）播放器、笔记本电脑或台式电脑。终端 500 还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

通常，终端 500 包括有：处理器 501 和存储器 502。

处理器 501 可以包括一个或多个处理核心，比如 4 核心处理器、8 核心处理器等。处理器 501 可以采用 DSP（Digital Signal Processing，数字信号处理）、FPGA（Field — Programmable Gate Array，现场可编程门阵列）、PLA（Programmable Logic Array，可编程逻辑阵列）中的至少一种硬件形式来实现。处理器 501 也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称 CPU（Central Processing Unit，中央处理器）；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器 501 可以在集成有 GPU（Graphics Processing Unit，图像处理器），GPU 用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器 501 还可以包括 AI（Artificial Intelligence，人工智能）处理器，该 AI 处理器用于处理有关机器学习的计算操作。

存储器 502 可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器 502 还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，存储器 502 中的非暂态的计算机可读存储介质用于存储至少一个指令，该至少一个指令用于被处理器 501 所执行以实现本申请中方法实施例提供的本体建立方法。

在一些实施例中，终端 500 还可选包括有：外围设备接口 503 和至少一个外围设备。处理器 501、存储器 502 和外围设备接口 503 之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口 503 相连。示例性地，外围设备包括：射频电路 504、触摸显示屏 505、摄像头 506、音频电路 507、定位组件 508 和电源 509 中的至少一种。

外围设备接口 503 可被用于将 I/O（Input /Output，输入/输出）相关的至少

一个外围设备连接到处理器 501 和存储器 502。在一些实施例中，处理器 501、存储器 502 和外围设备接口 503 被集成在同一芯片或电路板上；在一些其他实施例中，处理器 501、存储器 502 和外围设备接口 503 中的任意一个或两个可以在单独的芯片或电路板上实现，本实施例对此不加以限定。

射频电路 504 用于接收和发射 RF (Radio Frequency, 射频) 信号，也称电磁信号。射频电路 504 通过电磁信号与通信网络以及其他通信设备进行通信。射频电路 504 将电信号转换为电磁信号进行发送，或者，将接收到的电磁信号转换为电信号。可选地，射频电路 504 包括：天线系统、RF 收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路 504 可以通过至少一种无线通信协议来与其它终端进行通信。该无线通信协议包括但不限于：城域网、各代移动通信网络 (2G、3G、4G 及 5G)、无线局域网和/或 WiFi(Wireless Fidelity, 无线保真) 网络。在一些实施例中，射频电路 504 还可以包括 NFC (Near Field Communication, 近距离无线通信) 有关的电路，本申请对此不加以限定。

显示屏 505 用于显示 UI (User Interface, 用户界面)。该 UI 可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏 505 是触摸显示屏时，显示屏 505 还具有采集在显示屏 505 的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器 501 进行处理。此时，显示屏 505 还可以用于提供虚拟按钮和/或虚拟键盘，也称软按钮和/或软键盘。在一些实施例中，显示屏 505 可以为一个，设置终端 500 的前面板；在另一些实施例中，显示屏 505 可以为至少两个，分别设置在终端 500 的不同表面或呈折叠设计；在再一些实施例中，显示屏 505 可以是柔性显示屏，设置在终端 500 的弯曲表面上或折叠面上。甚至，显示屏 505 还可以设置成非矩形的不规则图形，也即异形屏。显示屏 505 可以采用 LCD(Liquid Crystal Display, 液晶显示屏)、OLED(Organic Light-Emitting Diode, 有机发光二极管)等材质制备。

摄像头组件 506 用于采集图像或视频。可选地，摄像头组件 506 包括前置摄像头和后置摄像头。通常，前置摄像头设置在终端的前面板，后置摄像头设置在终端的背面。在一些实施例中，后置摄像头为至少两个，分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种，以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及 VR (Virtual Reality, 虚拟现实) 拍摄功能或者其它融合拍摄功能。在一些实施

例中，摄像头组件 506 还可以包括闪光灯。闪光灯可以是单色温闪光灯，也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合，可以用于不同色温下的光线补偿。

音频电路 507 可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波，并将声波转换为电信号输入至处理器 501 进行处理，或者输入至射频电路 504 以实现语音通信。出于立体声采集或降噪的目的，麦克风可以为多个，分别设置在终端 500 的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器 501 或射频电路 504 的电信号转换为声波。扬声器可以是传统的薄膜扬声器，也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时，不仅可以将电信号转换为人类可听见的声波，也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中，音频电路 507 还可以包括耳机插孔。

定位组件 508 用于定位终端 500 的当前地理位置，以实现导航或 LBS（Location Based Service，基于位置的服务）。定位组件 508 可以是基于美国的 GPS（Global Positioning System，全球定位系统）、中国的北斗系统、俄罗斯的格雷纳斯系统或欧盟的伽利略系统的定位组件。

电源 509 用于为终端 500 中的各个组件进行供电。电源 509 可以是交流电、直流电、一次性电池或可充电电池。当电源 509 包括可充电电池时，该可充电电池可以支持有线充电或无线充电。该可充电电池还可以用于支持快充技术。

在一些实施例中，终端 500 还包括有一个或多个传感器 510。该一个或多个传感器 510 包括但不限于：加速度传感器 511、陀螺仪传感器 512、压力传感器 513、指纹传感器 514、光学传感器 515 以及接近传感器 516。

加速度传感器 511 可以检测以终端 500 建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器 511 可以用于检测重力加速度在三个坐标轴上的分量。处理器 501 可以根据加速度传感器 511 采集的重力加速度信号，控制触摸显示屏 505 以横向视图或纵向视图进行用户界面的显示。加速度传感器 511 还可以用于游戏或者用户的运动数据的采集。

陀螺仪传感器 512 可以检测终端 500 的机体方向及转动角度，陀螺仪传感器 512 可以与加速度传感器 511 协同采集用户对终端 500 的 3D 动作。处理器 501 根据陀螺仪传感器 512 采集的数据，可以实现如下功能：动作感应（比如根据用户的倾斜操作来改变 UI）、拍摄时的图像稳定、游戏控制以及惯性导航。

压力传感器 513 可以设置在终端 500 的侧边框和/或触摸显示屏 505 的下层。当压力传感器 513 设置在终端 500 的侧边框时，可以检测用户对终端 500 的握持信号，由处理器 501 根据压力传感器 513 采集的握持信号进行左右手识别或快捷操作。当压力传感器 513 设置在触摸显示屏 505 的下层时，由处理器 501 根据用户对触摸显示屏 505 的压力操作，实现对 UI 界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

指纹传感器 514 用于采集用户的指纹，由处理器 501 根据指纹传感器 514 采集到的指纹识别用户的身份，或者，由指纹传感器 514 根据采集到的指纹识别用户的身份。在识别出用户的身份为可信身份时，由处理器 501 授权该用户执行相关的敏感操作，该敏感操作包括解锁屏幕、查看加密信息、下载软件、支付及更改设置等。指纹传感器 514 可以被设置终端 500 的正面、背面或侧面。当终端 500 上设置有物理按键或厂商 Logo 时，指纹传感器 514 可以与物理按键或厂商 Logo 集成在一起。

光学传感器 515 用于采集环境光强度。在一个实施例中，处理器 501 可以根据光学传感器 515 采集的环境光强度，控制触摸显示屏 505 的显示亮度。示例性地，当环境光强度较高时，调高触摸显示屏 505 的显示亮度；当环境光强度较低时，调低触摸显示屏 505 的显示亮度。在另一个实施例中，处理器 501 还可以根据光学传感器 515 采集的环境光强度，动态调整摄像头组件 506 的拍摄参数。

接近传感器 516，也称距离传感器，通常设置在终端 500 的前面板。接近传感器 516 用于采集用户与终端 500 的正面之间的距离。在一个实施例中，当接近传感器 516 检测到用户与终端 500 的正面之间的距离逐渐变小时，由处理器 501 控制触摸显示屏 505 从亮屏状态切换为息屏状态；当接近传感器 516 检测到用户与终端 500 的正面之间的距离逐渐变大时，由处理器 501 控制触摸显示屏 505 从息屏状态切换为亮屏状态。

本领域技术人员可以理解，图 5 中示出的结构并不构成对终端 500 的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

本申请实施例还提供了一种非临时性计算机可读存储介质，当所述存储介质中的指令由终端的处理器执行时，使得终端能够执行上述实施例提供的本体

建立方法。

本申请实施例还提供了一种包含指令的计算机程序产品，当其在终端上运行时，使得终端执行上述实施例提供的本体建立方法。

图 6 是本申请实施例提供的一种服务器结构示意图。该服务器可以是后台服务器集群中的服务器。

服务器 600 包括中央处理单元（CPU）601、包括随机存取存储器（RAM）602 和只读存储器（ROM）603 的系统存储器 604，以及连接系统存储器 604 和中央处理单元 601 的系统总线 605。服务器 600 还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统（I/O 系统）606，和用于存储操作系统 613、应用程序 614 和其他程序模块 615 的大容量存储设备 607。

基本输入/输出系统 606 包括有用于显示信息的显示器 608 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 609。其中显示器 608 和输入设备 609 都通过连接到系统总线 605 的输入输出控制器 610 连接到中央处理单元 601。基本输入/输出系统 606 还可以包括输入输出控制器 610 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器 610 还提供输出到显示屏、打印机或其他类型的输出设备。

大容量存储设备 607 通过连接到系统总线 605 的大容量存储控制器（未示出）连接到中央处理单元 601。大容量存储设备 607 及其相关联的计算机可读介质为服务器 600 提供非易失性存储。也就是说，大容量存储设备 607 可以包括诸如硬盘或者 CD-ROM 驱动器之类的计算机可读介质（未示出）。

不失一般性，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括 RAM、ROM、EPROM、EEPROM、闪存或其他固态存储其技术，CD-ROM、DVD 或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然，本领域技术人员可知计算机存储介质不局限于上述几种。上述的系统存储器 604 和大容量存储设备 607 可以统称为存储器。

根据本申请的各种实施例，服务器 600 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即服务器 600 可以通过连接在系统总线 605 上的网络接口单元 611 连接到网络 612，或者说，也可以使用网络接口单元 611 来

连接到其他类型的网络或远程计算机系统（未示出）。

上述存储器还包括一个或者一个以上的程序，一个或者一个以上程序存储于存储器中，被配置由 CPU 执行。所述一个或者一个以上程序包含用于进行上述实施例提供的本体建立方法的指令。

本申请实施例还提供了一种非临时性计算机可读存储介质，当所述存储介质中的指令由服务器的处理器执行时，使得服务器能够执行上述实施例提供的本体建立方法。

本申请实施例还提供了一种包含指令的计算机程序产品，当其在服务器上运行时，使得服务器执行上述实施例提供的本体建立方法。

本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

以上所述仅为本申请的较佳实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权 利 要 求 书

1、一种本体建立方法，其特征在于，所述方法包括：

获取用于建立本体的至少一个文档，根据所述至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，所述第一类词语是指用于指示对象自身属性的词语，所述第二类词语是指用于指示不同对象之间的关联关系的词语；

确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，采用确定的两个概念分别替换对应的两个第一类词语，采用确定的概念关系替换对应的第二类词语，得到多个语义关系数据组；

根据所述多个语义关系数据组，建立针对所述至少一个文档的本体。

2、如权利要求 1 所述的方法，其特征在于，所述确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，包括：

对于所述多个多元组中的多元组 A，从参考数据库中查找与所述多元组 A 中的两个第一类词语分别对应的两个概念，所述参考数据库包括多个词语中每个词语对应的概念以及不同概念之间的概念关系，所述多元组 A 为所述多个多元组中任一个；

根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系。

3、如权利要求 2 所述的方法，其特征在于，所述从参考数据库中查找所述多元组 A 中的两个第一类词语分别对应的两个概念，包括：

对于所述多元组 A 中两个第一类词语中的第一类词语 B，确定所述至少一个文档中词义与所述第一类词语 B 的词义相同的词语，所述第一类词语 B 为所述两个第一类词语中任一个；

如果所述至少一个文档中不存在词义与所述第一类词语 B 的词义相同的词语，则从所述参考数据库中查找所述第一类词语 B 对应的概念。

4、如权利要求 3 所述的方法，其特征在于，所述确定所述至少一个文档中词义与所述第一类词语 B 的词义相同的词语之后，还包括：

如果所述至少一个文档中存在词义与所述第一类词语 B 的词义相同的词语，则确定所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语分别在所述至少一个文档中的出现次数；

从所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；

从所述参考数据库中查找所述出现次数最大的词语对应的概念，将查找到的概念作为所述第一类词语 B 对应的概念。

5、如权利要求 3 或 4 所述的方法，其特征在于，所述方法还包括：

如果没有从所述参考数据库中查找到所述第一类词语 B 对应的概念，则通过潜在狄利克雷分布 LDA 算法为所述第一类词语 B 创建一个概念。

6、如权利要求 2 所述的方法，其特征在于，所述根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系，包括：

从所述参考数据库中确定所述查找到的两个概念之间的路径，得到多条路径；

根据所述多条路径的路径长度，从所述多条路径中选择一条目标路径，所述目标路径中包括至少一个概念关系；

从所述至少一个概念关系中选择与所述多元组 A 中的第二类词语之间相似度最大的概念关系，将选择的概念关系确定为与所述多元组 A 中的第二类词语对应的概念关系。

7、如权利要求 1 所述的方法，其特征在于，所述根据所述多个语义关系数据组，建立针对所述至少一个文档的本体，包括：

根据所述多个语义关系数据组，建立语义关系图，所述语义关系图中的一个结点对应所述多个语义关系组中的一个概念，所述语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且所述两个结点之间的方向为对

应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向相应结点的其他结点的数量，每个结点的出度是指相应结点指向的其他结点的数量；

对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪；

当对所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点均裁剪完时，将所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点指向的其他结点的入度设置为 0；

对除已经裁剪的第一结点之外的其他结点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的步骤，直至遍历完所述语音关系图中的所有结点，将最终得到的语义关系图作为针对所述至少一个文档建立的本体。

8、如权利要求 7 所述的方法，其特征在于，所述对所述第一结点进行裁剪，包括：

确定所述语义关系图中的至少一个结点，所述至少一个结点对应的概念所指示的词语在所述至少一个文档中的文档位置与所述第一结点对应的概念所指示的词语在所述至少一个文档中的文档位置相邻；

确定所述至少一个结点中与所述第一结点连接度最大的结点，删除所述至少一个结点中除确定出的结点之外的其他结点，以及所述其他结点与所述第一结点之间的关系，任意两个结点之间的连接度用于指示所述两个结点之间的关联程度。

9、如权利要求 8 所述的方法，其特征在于，所述确定所述至少一个结点中与所述第一结点连接度最大的结点，包括：

基于第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度；

其中，所述第一公式为：
$$Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

所述 W_i 和所述 W_j 分别为用于确定连接度的两个结点，所述 $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度，所述 $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度，所述 α 和所述 β 分别为针对相似度配置的加权系数和针对相关

度配置的加权系数, 所述 α 和所述 β 的和为 1;

根据按照所述第一公式确定的连接度, 确定所述至少一个结点中与所述第一结点连接度最大的结点。

10、如权利要求 9 所述的方法, 其特征在于, 所述根据按照所述第一公式确定的连接度, 确定所述至少一个结点中与所述第一结点连接度最大的结点, 包括:

如果按照所述第一公式确定的连接度中的最大连接度小于连接度阈值, 则调整所述 α 和所述 β 的数值, 得到更新之后的第一公式;

基于更新之后的第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度;

如果按照所述更新之后的第一公式确定的连接度中的最大连接度小于所述连接度阈值, 则返回执行调整所述 α 和所述 β 的数值的步骤, 直至确定出的最大连接度大于或等于所述连接度阈值, 将最后一次确定出的最大连接度对应的结点确定为所述至少一个结点中与所述第一结点连接度最大的结点。

11、如权利要求 7 所述的方法, 其特征在于, 所述对于入度等于 0 且出度大于 0 的任一第一结点, 对所述第一结点进行裁剪之前, 还包括:

如果所述语义关系图中存在入度等于 0 且出度也等于 0 的孤立结点, 将所述孤立结点配置为指向所述语义关系图中的入度等于 0 且出度大于 0 的一个结点上, 得到更新之后的语义关系图;

基于更新之后的语义关系图, 执行对于入度等于 0 且出度大于 0 的任一第一结点, 对所述第一结点进行裁剪的操作。

12、如权利要求 1 所述的方法, 其特征在于, 所述根据所述至少一个文档确定多个多元组, 包括:

对所述至少一个文档中每个文档进行分词处理, 得到多个词语;

确定所述多个词语中每个词语的词性;

根据所述多个词语中每个词语的词性和每个词语在所述至少一个文档中的文档位置, 确定所述多个多元组。

13、一种本体建立装置，其特征在于，所述装置包括：

获取模块，用于获取用于建立本体的至少一个文档，根据所述至少一个文档确定多个多元组，每个多元组包括两个第一类词语和一个第二类词语，所述第一类词语是指用于指示对象自身属性的词语，所述第二类词语是指用于指示不同对象之间的关联关系的词语；

确定模块，用于确定所述多个多元组中每个多元组中的两个第一类词语分别对应的两个概念，以及每个多元组中的第二类词语对应的概念关系，采用确定的两个概念分别替换对应的两个第一类词语，采用确定的概念关系替换对应的第二类词语，得到多个语义关系数据组；

建立模块，用于根据所述多个语义关系数据组，建立针对所述至少一个文档的本体。

14、如权利要求 13 所述的装置，其特征在于，所述确定模块包括：

第一查找单元，用于对于所述多个多元组中的多元组 A，从参考数据库中查找与所述多元组 A 中的两个第一类词语分别对应的两个概念，所述参考数据库包括多个词语中每个词语所代表的概念以及不同概念之间的概念关系，所述多元组 A 为所述多个多元组中任一个；

第二查找单元，用于根据查找到的两个概念，从所述参考数据库中继续查找与所述多元组 A 中的第二类词语对应的概念关系。

15、如权利要求 14 所述的装置，其特征在于，所述第一查找单元，用于：

对于所述多元组 A 中两个第一类词语中的第一类词语 B，确定所述至少一个文档中词义与所述第一类词语 B 的词义相同的词语，所述第一类词语 B 为所述两个第一类词语中任一个；

如果所述至少一个文档中不存在词义与所述第一类词语 B 的词义相同的词语，则从所述参考数据库中查找所述第一类词语 B 对应的概念。

16、如权利要求 15 所述的装置，其特征在于，所述第一查找单元，还用于：

如果所述至少一个文档中存在词义与所述第一类词语 B 的词义相同的词语，

则确定所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语分别在所述至少一个文档中的出现次数；

从所述第一类词语 B，以及词义与所述第一类词语 B 的词义相同的词语中，确定出现次数最大的词语；

从所述参考数据库中查找所述出现次数最大的词语对应的概念，并将查找到的概念作为所述第一类词语 B 对应的概念。

17、如权利要求 15 或 16 所述的装置，其特征在于，所述装置还包括：

创建单元，用于如果没有从所述参考数据库中查找到所述第一类词语 B 对应的概念，则通过潜在狄利克雷分布 LDA 算法为所述第一类词语 B 创建一个概念。

18、如权利要求 14 所述的装置，其特征在于，所述第二查找单元，用于：

从所述参考数据库中确定所述查找到的两个概念之间的路径，得到多条路径；

根据所述多条路径的路径长度，从所述多条路径中选择一条目标路径，所述目标路径中包括至少一个概念关系；

从所述至少一个概念关系中选择与所述多元组 A 中的第二类词语之间相似度最大的概念关系，将选择的概念关系确定为与所述多元组 A 中的第二类词语对应的概念关系。

19、如权利要求 13 所述的装置，其特征在于，所述建立模块包括：

建立单元，用于根据所述多个语义关系数据组，建立语义关系图，所述语义关系图中的一个结点对应所述多个语义关系组中的一个概念，所述语义关系图中两个结点之间的关系为对应的两个概念之间的概念关系，且所述两个结点之间的方向为对应的两个概念之间的概念关系所指示的方向，每个结点配置有入度和出度，且每个结点的入度是指指向相应结点的其他结点的数量，每个结点的出度是指相应结点指向的其他结点的数量；

裁剪单元，用于对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪；

设置单元，用于当对所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点均裁剪完时，将所述语义关系图中所有入度等于 0 且出度大于 0 的第一结点指向的其他结点的入度设置为 0；

所述裁剪单元，还用于对除已经裁剪的第一结点之外的其他结点，返回执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的步骤，直至遍历完所述语义关系图中的所有结点，将最终得到的语义关系图作为针对所述至少一个文档建立的本体。

20、如权利要求 19 所述的装置，其特征在于，所述裁剪单元，用于：

确定所述语义关系图中的至少一个结点，所述至少一个结点对应的概念所指示的词语在所述至少一个文档中的文档位置与所述第一结点对应的概念所指示的词语在所述至少一个文档中的文档位置相邻；

确定所述至少一个结点中与所述第一结点连接度最大的结点，删除所述至少一个结点中除确定出的结点之外的其他结点，以及所述其他结点与所述第一结点之间的关系，任意两个结点之间的连接度用于指示所述两个结点之间的关联程度。

21、如权利要求 20 所述的装置，其特征在于，所述裁剪单元，还用于：

基于第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度；

其中，所述第一公式为：
$$Joint(W_i, W_j) = \frac{\alpha}{Sim(W_i, W_j) + \alpha} + \frac{\beta}{Rel(W_i, W_j) + \beta};$$

所述 W_i 和所述 W_j 分别为用于确定连接度的两个结点，所述 $Sim(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相似度，所述 $Rel(W_i, W_j)$ 为结点 W_i 和结点 W_j 之间的相关度，所述 α 和所述 β 分别为针对相似度配置的加权系数和针对相关度配置的加权系数，所述 α 和所述 β 的和为 1；

根据按照所述第一公式确定的连接度，确定所述至少一个结点中与所述第一结点连接度最大的结点。

22、如权利要求 21 所述的装置，其特征在于，所述裁剪单元，还用于：

如果按照所述第一公式确定的连接度中的最大连接度小于连接度阈值，则

调整所述 α 和所述 β 的数值，得到更新之后的第一公式；

基于更新之后的第一公式确定所述至少一个结点中每个结点与所述第一结点之间的连接度；

如果按照所述更新之后的第一公式确定的连接度中的最大连接度小于所述连接度阈值，则返回执行调整所述 α 和所述 β 的数值的步骤，直至确定出的最大连接度大于或等于所述连接度阈值，将最后一次确定出的最大连接度对应的结点确定为所述至少一个结点中与所述第一结点连接度最大的结点。

23、如权利要求 19 所述的装置，其特征在于，所述建立模块还包括：

配置单元，用于如果所述语义关系图中存在入度等于 0 且出度也等于 0 的孤立结点，将所述孤立结点配置为指向所述语义关系图中的入度等于 0 且出度大于 0 的一个结点上，得到更新之后的语义关系图；

所述裁剪单元，还用于基于更新之后的语义关系图，执行对于入度等于 0 且出度大于 0 的任一第一结点，对所述第一结点进行裁剪的操作。

24、如权利要求 13 所述的装置，其特征在于，所述获取模块包括：

分词处理单元，用于对所述至少一个文档中每个文档进行分词处理，得到多个词语；

第一确定单元，用于确定所述多个词语中每个词语的词性；

第二确定单元，用于根据所述多个词语中每个词语的词性和每个词语在所述至少一个文档中的文档位置，确定所述多个多元组。

25、一种本体建立装置，其特征在于，所述装置包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行上述权利要求 1 至权利要求 12 中的任一项权利要求所述的方法的步骤。

26、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上存储有指令，所述指令被处理器执行时实现上述权利要求 1 至权利要求 12 中的

任一项权利要求所述的方法的步骤。

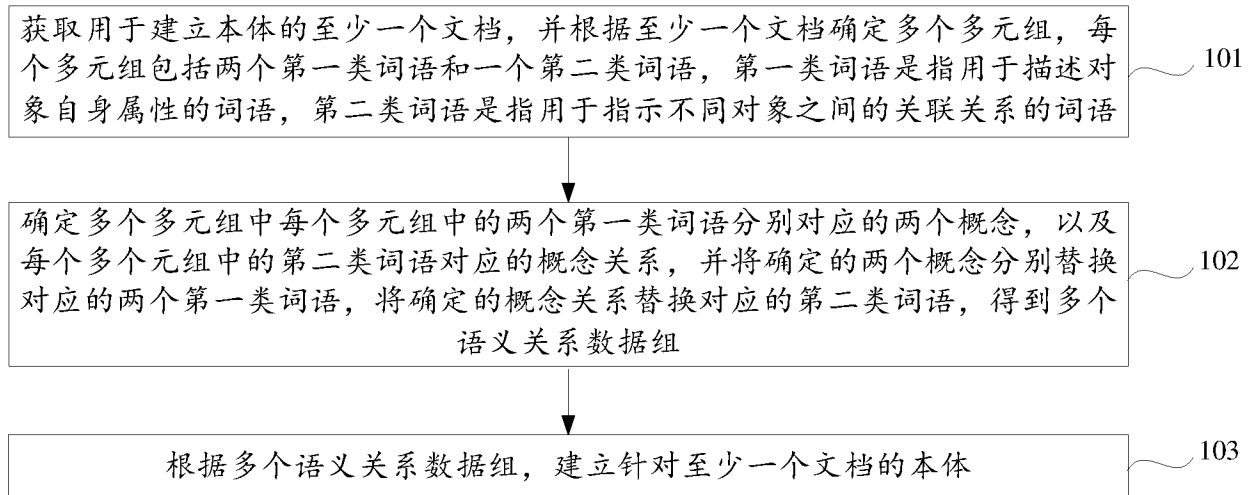


图 1

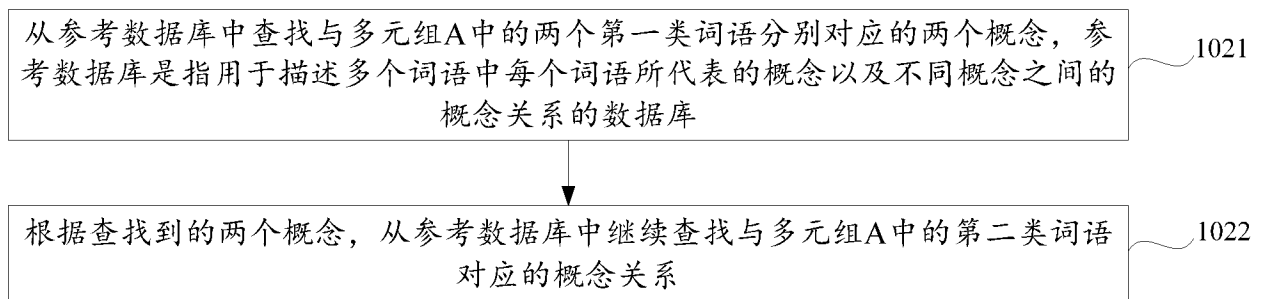


图 2

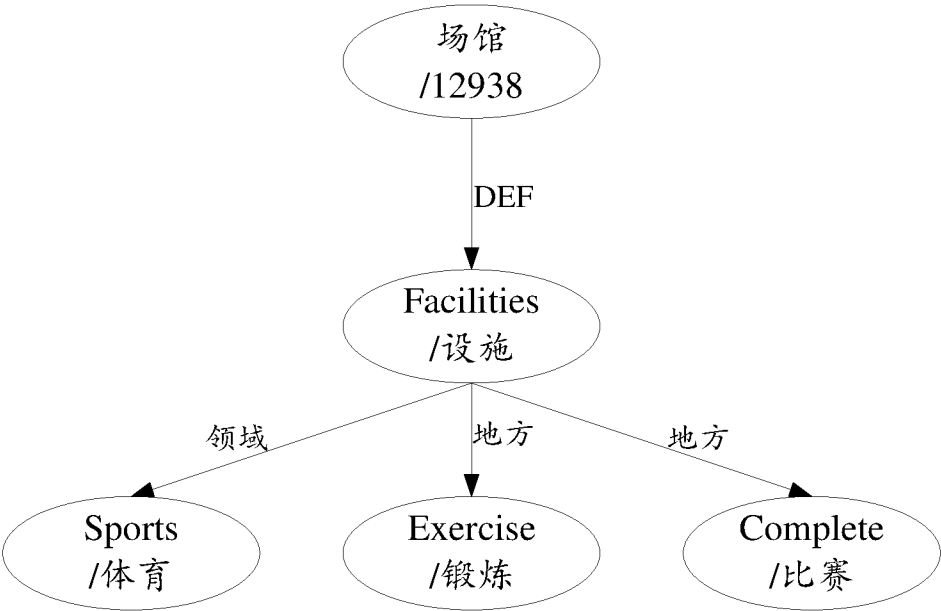


图 3

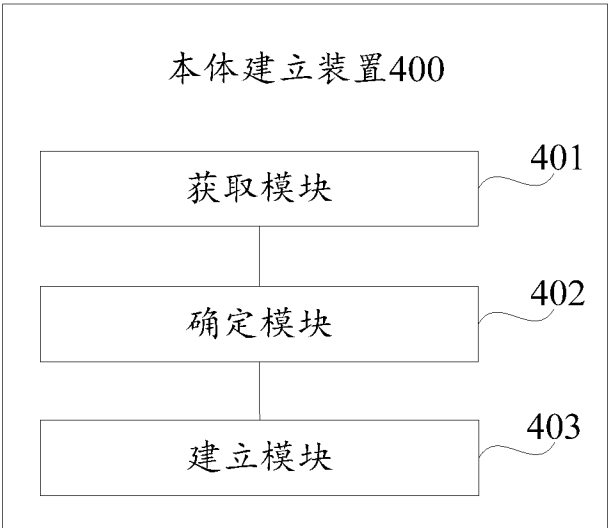


图 4

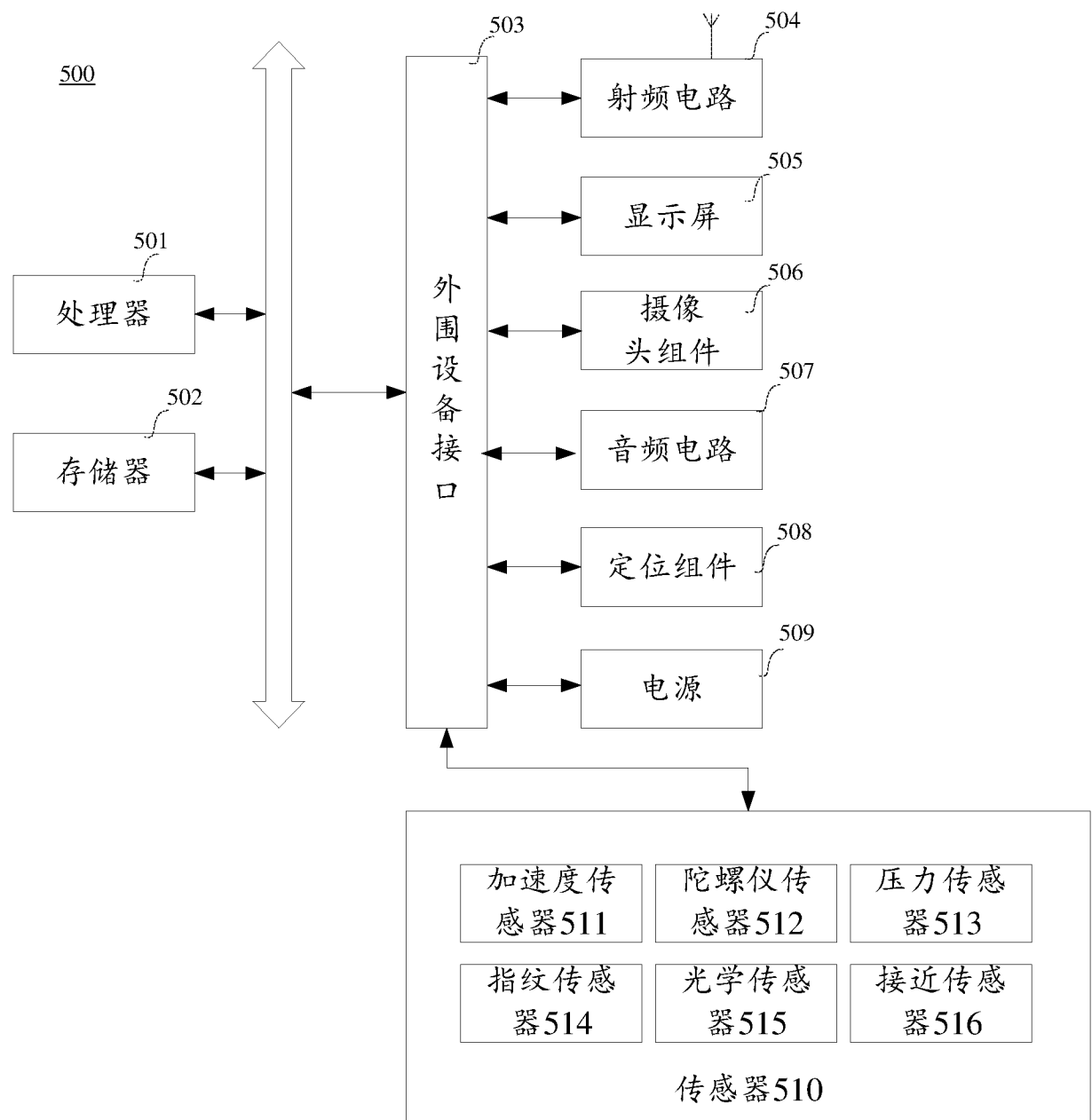


图 5

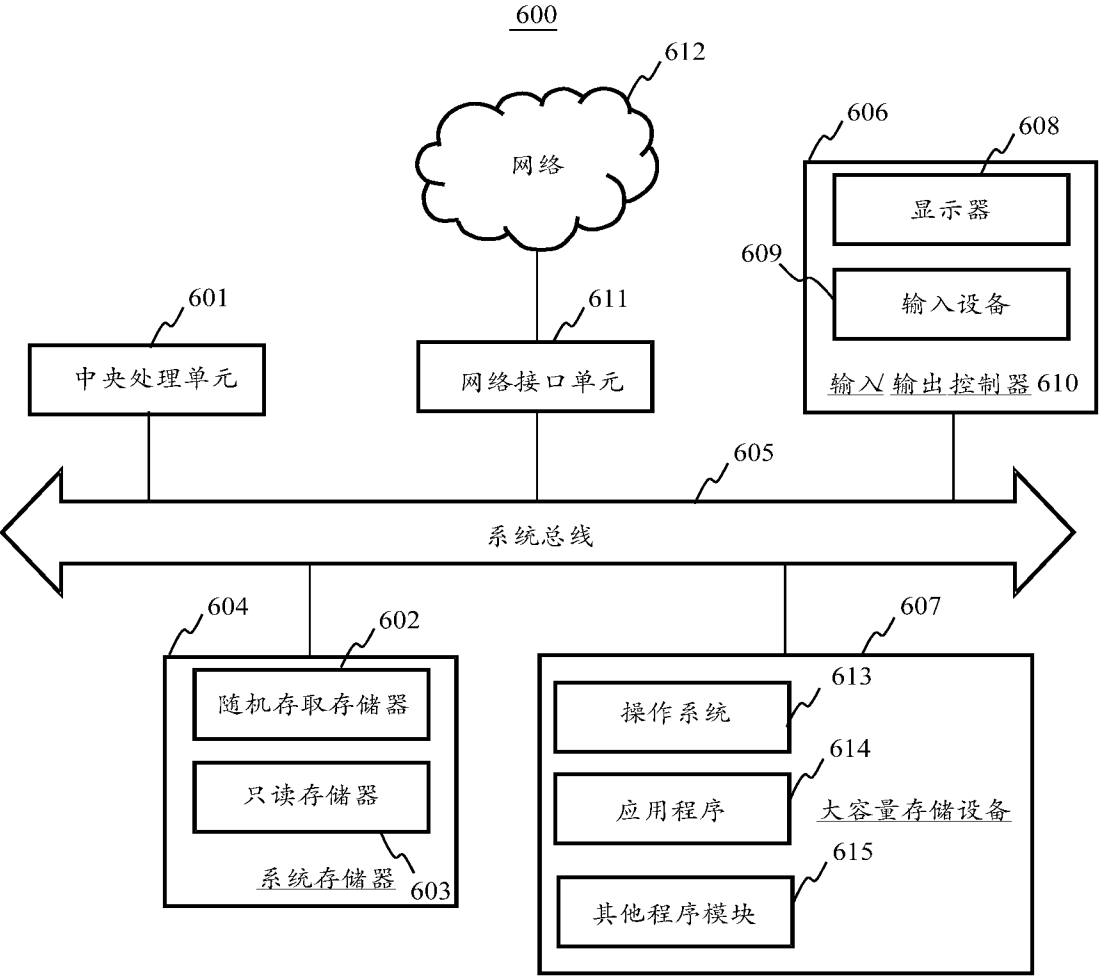


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/122159

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/36(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, CNKI: 海康威视, 文档, 文本, 分词, 词语, 短语, 关键词, 词汇, 关键字, 构建, 建立, 生成, 创建, 构造, 本体, 概念, 属性, 关系, 数据库, 语料库, 语义, 结点. EPODOC, WPI: ontology, property, properties, concept?, word?, vocabulary, text, document, creat, relation?, relationship?.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106445990 A (SHANGHAI UNIVERSITY) 22 February 2017 (2017-02-22) description, paragraphs [0045]-[0051], and figure 2	1-26
A	CN 101710343 A (BEIJING ZHONGJIKEHAI TECHNOLOGY & DEVELOPMENT CO., LTD.) 19 May 2010 (2010-05-19) entire document	1-26
A	CN 103207856 A (TONGJI UNIVERSITY) 17 July 2013 (2013-07-17) entire document	1-26
A	CN 108491385 A (GUANGXI NORMAL UNIVERSITY) 04 September 2018 (2018-09-04) entire document	1-26
A	CN 101681353 A (KNEWCO, INC.) 24 March 2010 (2010-03-24) entire document	1-26
A	US 2004054672 A1 (IBM CORPORATION) 18 March 2004 (2004-03-18) entire document	1-26



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

24 February 2020

Date of mailing of the international search report

02 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration
 No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
 100088
 China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/122159

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106445990	A	22 February 2017	None			
CN	101710343	A	19 May 2010	None			
CN	103207856	A	17 July 2013	None			
CN	108491385	A	04 September 2018	None			
CN	101681353	A	24 March 2010	EP	2143012	A2	13 January 2010
				CA	2682582	A1	09 October 2008
				JP	2010529518	A	26 August 2010
				CN	101681351	A	24 March 2010
				BR	PI0811415	A2	02 May 2017
				IL	201230	D0	31 May 2010
				CA	2682602	A1	09 October 2008
				AU	2008233078	A1	09 October 2008
				WO	2008121382	A1	09 October 2008
				IL	201232	D0	31 May 2010
				US	2010174739	A1	08 July 2010
				WO	2008121377	A2	09 October 2008
				AU	2008233083	A1	09 October 2008
				US	2010174675	A1	08 July 2010
				EP	2143011	A1	13 January 2010
				JP	2010532506	A	07 October 2010
				US	2008301174	A1	04 December 2008
				US	2008306918	A1	11 December 2008
				US	2009217179	A1	27 August 2009
US	2004054672	A1	18 March 2004	JP	2004094806	A	25 March 2004

国际检索报告

国际申请号

PCT/CN2019/122159

A. 主题的分类 G06F 16/36 (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, CNKI: 海康威视, 文档, 文本, 分词, 词语, 短语, 关键词, 词汇, 关键字, 构建, 建立, 生成, 创建, 构造, 本体, 概念, 属性, 关系, 数据库, 语料库, 语义, 结点. EP0DOC, WPI: ontology, property, properties, concept?, word?, vocabulary, text, document, creat, relation?, relationship?.																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>CN 106445990 A (上海大学) 2017年 2月 22日 (2017 - 02 - 22) 说明书第[0045]-[0051]段、图2</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 101710343 A (北京中机科海科技发展有限公司) 2010年 5月 19日 (2010 - 05 - 19) 全文</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 103207856 A (同济大学) 2013年 7月 17日 (2013 - 07 - 17) 全文</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 108491385 A (广西师范大学) 2018年 9月 4日 (2018 - 09 - 04) 全文</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>CN 101681353 A (纽科股份有限公司) 2010年 3月 24日 (2010 - 03 - 24) 全文</td> <td>1-26</td> </tr> <tr> <td>A</td> <td>US 2004054672 A1 (IBM CORPORATION) 2004年 3月 18日 (2004 - 03 - 18) 全文</td> <td>1-26</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	A	CN 106445990 A (上海大学) 2017年 2月 22日 (2017 - 02 - 22) 说明书第[0045]-[0051]段、图2	1-26	A	CN 101710343 A (北京中机科海科技发展有限公司) 2010年 5月 19日 (2010 - 05 - 19) 全文	1-26	A	CN 103207856 A (同济大学) 2013年 7月 17日 (2013 - 07 - 17) 全文	1-26	A	CN 108491385 A (广西师范大学) 2018年 9月 4日 (2018 - 09 - 04) 全文	1-26	A	CN 101681353 A (纽科股份有限公司) 2010年 3月 24日 (2010 - 03 - 24) 全文	1-26	A	US 2004054672 A1 (IBM CORPORATION) 2004年 3月 18日 (2004 - 03 - 18) 全文	1-26
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
A	CN 106445990 A (上海大学) 2017年 2月 22日 (2017 - 02 - 22) 说明书第[0045]-[0051]段、图2	1-26																					
A	CN 101710343 A (北京中机科海科技发展有限公司) 2010年 5月 19日 (2010 - 05 - 19) 全文	1-26																					
A	CN 103207856 A (同济大学) 2013年 7月 17日 (2013 - 07 - 17) 全文	1-26																					
A	CN 108491385 A (广西师范大学) 2018年 9月 4日 (2018 - 09 - 04) 全文	1-26																					
A	CN 101681353 A (纽科股份有限公司) 2010年 3月 24日 (2010 - 03 - 24) 全文	1-26																					
A	US 2004054672 A1 (IBM CORPORATION) 2004年 3月 18日 (2004 - 03 - 18) 全文	1-26																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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																			
* 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 2020年 2月 24日		国际检索报告邮寄日期 2020年 3月 2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 李慧 电话号码 86-10-53961720																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/122159

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	106445990	A	2017年 2月 22日	无			
CN	101710343	A	2010年 5月 19日	无			
CN	103207856	A	2013年 7月 17日	无			
CN	108491385	A	2018年 9月 4日	无			
CN	101681353	A	2010年 3月 24日	EP	2143012	A2	2010年 1月 13日
				CA	2682582	A1	2008年 10月 9日
				JP	2010529518	A	2010年 8月 26日
				CN	101681351	A	2010年 3月 24日
				BR	PI0811415	A2	2017年 5月 2日
				IL	201230	D0	2010年 5月 31日
				CA	2682602	A1	2008年 10月 9日
				AU	2008233078	A1	2008年 10月 9日
				WO	2008121382	A1	2008年 10月 9日
				IL	201232	D0	2010年 5月 31日
				US	2010174739	A1	2010年 7月 8日
				WO	2008121377	A2	2008年 10月 9日
				AU	2008233083	A1	2008年 10月 9日
				US	2010174675	A1	2010年 7月 8日
				EP	2143011	A1	2010年 1月 13日
				JP	2010532506	A	2010年 10月 7日
				US	2008301174	A1	2008年 12月 4日
				US	2008306918	A1	2008年 12月 11日
				US	2009217179	A1	2009年 8月 27日
US	2004054672	A1	2004年 3月 18日	JP	2004094806	A	2004年 3月 25日