(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 122489779 A
(43)申请公布日 2026.07.31

(21)申请号 202610387195.2

(22)申请日 2026.03.27

(71)申请人 中煤科工开采研究院有限公司
地址 101399 北京市顺义区中关村科技园区顺义园临空二路1号

(72)发明人 吕依濛

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 孟洋

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/10(2020.01)
G06F 40/205(2020.01)
G06F 40/30(2020.01)
G06F 18/22(2023.01)
G06F 18/26(2023.01)
G06N 5/025(2023.01)

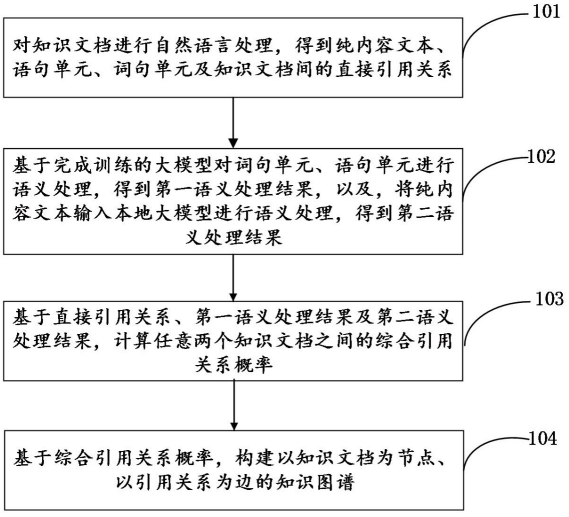
权利要求书3页 说明书13页 附图2页

(54)发明名称

针对知识文档的知识图谱构建方法、装置及系统

(57)摘要

本公开是关于一种针对知识文档的知识图谱构建方法、装置及系统。其中,方法包括:对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;基于完成训练的大模型对词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;基于直接引用关系、第一语义处理结果及第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率;基于综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。本方案可以实现提升了从海量非结构化文档中挖掘和构建知识关联的效率与准确性。



CN 122489779 A

1.一种针对知识文档的知识图谱构建方法,其特征在于,包括:

对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;

基于完成训练的大模型对所述词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将所述纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;所述第一语义处理结果包括词句在其所属语句中的重要程度评分;所述第二语义处理结果包括知识文档对之间的语义关联度评分;

基于所述直接引用关系、所述第一语义处理结果及所述第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率;

基于所述综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

2.根据权利要求1所述的针对知识文档的知识图谱构建方法,其特征在于,所述基于完成训练的大模型对所述词句单元、语句单元进行语义处理,得到第一语义处理结果,包括:

将所述词句单元和语句单元分别输入至多个已训练完成的通用大模型;

接收各通用大模型返回的语义处理结果,并对语义处理结果进行归一化处理;语义处理结果至少包括对应词句的重要性评价值、大模型检索的总语料数以及命中关联语料数;

采用以下公式对通用大模型返回的重要性评价进行加权融合,得到所述第一语义处理结果:

$$P\_api\_sentence(word, sentence)=\sum_i \gamma_i * importance(i) * \frac{hit(i)}{num(i)}$$

其中,$P_api_sentence(word, sentence)$为第一语义处理结果,$\gamma_i$为第i个大模型的权重,$importance(i)$为第i个大模型输出的重要性评价值,hit(i)为第i个大模型检索的总语料数,num(i)为第i个大模型的命中关联语料数。

3.根据权利要求1所述的针对知识文档的知识图谱构建方法,其特征在于,将所述纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果,包括:

将所述纯内容文本输入至已完成训练且针对特定领域优化的所述本地大模型进行语义处理;

获取所述本地大模型返回的第二语义处理结果;所述第二语义处理结果还包括专业词库中的词句在所述纯内容文本中的语义重要度评分。

4.根据权利要求1所述的针对知识文档的知识图谱构建方法,其特征在于,在所述对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系之后,还包括:

将所述词句单元与预设的煤炭开采领域的专业词库中的词句进行匹配,在匹配成功的情况下,计算词句单元在所属语句中的第一重要性概率,以及计算词句在所属文档中的第二重要性概率;

基于词句在其所属语句中的重要程度评分、所述第一重要性概率和第二重要性概率,采用以下公式计算词句的关键词概率 $P\_keyword(word,text)$:

$$P\_keyword(word,text)=P'\_sentence*\beta+(1-\beta)*P\_text(word,sentence)$$

$$P'\_sentence = \frac{\sum_{n} \delta * P\_api\_sentence(word, sentence) + (1-\delta)\, P\_sentence(word, sentence)}{n}$$

其中，P_api_*sentence*（word，sentence）为第一语义处理结果，P_sentence(word,sentence)为第一重要性概率，P_text(word,text)为第二重要性概率，为调节系数；

所述方法还包括：

基于所述关键词概率构建所述知识图谱中的属性信息。

5.根据权利要求4所述的针对知识文档的知识图谱构建方法，其特征在于，所述基于所述关键词概率构建所述知识图谱中的属性信息，包括：

在所述关键词概率大于或者等于1的情况下，将所述关键词概率更新为1，将所述词句单元配置以及更新后的关键词概率配置为所属知识文档对应节点的属性信息；

在所述关键词概率大于或者等于预设关键词概率并且小于1的情况下，将所述词句单元配置以及所述关键词概率配置为所属知识文档对应节点的属性信息。

6.根据权利要求1所述的针对知识文档的知识图谱构建方法，其特征在于，所述基于所述直接引用关系、所述第一语义处理结果及所述第二语义处理结果，计算任意两个知识文档之间的综合引用关系概率，包括：

采用以下公式计算任意两个知识文档之间的综合引用关系概率P(a,b)：

$$P(a,b) = (\sigma * P\_api(a,b) + (1-\sigma) * P\_pre(a,b)) * (1 - P\_quote(a,b)) + P\_quote(a,b)$$

其中，a和b为任意两个知识文档，$P_api(a,b)$为知识文档a与知识文档b之间的语义关联度评分，P_pre(a,b)为知识文档a与知识文档b之间的预处理关联度评分，P_quote(a,b)为知识文档a与知识文档b之间的直接引用关系。

7.根据权利要求6所述的针对知识文档的知识图谱构建方法，其特征在于，所述基于所述综合引用关系概率，构建以知识文档为节点、以引用关系为边的知识图谱，包括：

以知识文档为节点、以引用关系为边权构建概率关系图；

删除所述概率关系图中的节点自环；

计算所述概率关系图中任意两节点x与y间所有连通路径的综合概率均值，得到目标引用概率，其中，单条路径的概率为该路径上所有边的综合引用关系概率的乘积；

基于所述目标引用概率构建所述知识图谱。

8.根据权利要求7所述的针对知识文档的知识图谱构建方法，其特征在于，所述计算所述概率关系图中任意两节点x与y间所有连通路径的综合概率均值，得到目标引用概率，包括：

采用以下公式计算所述概率关系图中任意两节点x与y间所有连通路径的目标引用概率：

$$P\_build(x,y) = \sum \frac{P(x,i\ldots j,y)}{root(a,i,\ldots,j,b)}$$

$$P(x,i\ldots j,y) = P(x,i) * P(i,n) * \ldots * P(m,j) * P(j,y)$$

其中，P_build(x,y)为节点x与y之间的目标引用概率，$P(x,i\ldots j,y)$为单条路径的概率，a,i...j,b为遍历路径，root(a,i,...,j,b)为经过节点数。

9.根据权利要求7所述的针对知识文档的知识图谱构建方法，其特征在于，所述基于所

述目标引用概率构建所述知识图谱,包括:

针对任意两个知识文档之间的目标引用概率,若所述目标引用概率大于或等于第一阈值,则在所述知识图谱中建立从其中一个知识文档对应的第一节点指向另一个知识文档对应的第二节点的引用关系边;

若所述目标引用概率小于所述第一阈值并且大于第二阈值,则遍历所述概率关系图中从所述第一节点指向所述第二节点的所有可能路径,若遍历到存在至少一条子路径的目标引用概率大于或等于所述第一阈值的至少一个候选路径,则从至少一个候选路径中选取目标引用概率最高的目标路径,并沿所述目标路径在所述知识图谱中建立所述第一节点指向所述第二节点的引用关系链。

10.一种针对知识文档的知识图谱构建装置,其特征在于,包括:

自然语言处理单元,用于对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;

语义处理单元,用于基于完成训练的大模型对所述词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将所述纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;所述第一语义处理结果包括词句在其所属语句中的重要程度评分;所述第二语义处理结果包括知识文档对之间的语义关联度评分;

计算单元,用于基于所述直接引用关系、所述第一语义处理结果及所述第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率;

构建单元,用于基于所述综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

针对知识文档的知识图谱构建方法、装置及系统

技术领域

[0001] 本公开涉及煤矿开采技术领域,尤其涉及一种针对知识文档的知识图谱构建方法、装置及系统。

背景技术

[0002] 相关技术中,随着矿业领域科研与工程实践的快速发展,知识服务平台内积累的技术报告、专利、论文等非结构化文档数量急剧增长。这些文档之间蕴含的复杂技术关联,如引用脉络、主题交叉与语义相似性,难以通过传统的关键词检索或人工分类方式被有效挖掘与利用。当前,依赖人工整理和简单规则的方法不仅效率低下、成本高昂,且难以适应知识的动态更新与深层关系的发现,导致大量高价值知识处于分散和孤立的“数据孤岛”状态,无法形成系统化的知识体系以支撑高效的智能分析与决策。

发明内容

[0003] 为克服相关技术中存在的问题,本公开提供一种针对知识文档的知识图谱构建方法、装置及系统。

[0004] 根据本公开实施例的第一方面,提供一种针对知识文档的知识图谱构建方法,包括:

[0005] 对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;

基于完成训练的大模型对所述词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将所述纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;所述第一语义处理结果包括词句在其所属语句中的重要程度评分;所述第二语义处理结果包括知识文档对之间的语义关联度评分;

基于所述直接引用关系、所述第一语义处理结果及所述第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率;

基于所述综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

[0006] 根据本公开实施例的第二方面,提供一种针对知识文档的知识图谱构建装置,包括:

自然语言处理单元,用于对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;

语义处理单元,用于基于完成训练的大模型对所述词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将所述纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;所述第一语义处理结果包括词句在其所属语句中的重要程度评分;所述第二语义处理结果包括知识文档对之间的语义关联度评分;

计算单元,用于基于所述直接引用关系、所述第一语义处理结果及所述第二语义

处理结果,计算任意两个知识文档之间的综合引用关系概率;

构建单元,用于基于所述综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

[0007] 根据本公开实施例的第三方面,一种电子设备,包括:存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时,实现如第一方面中任一项所述的方法。

[0008] 根据本公开实施例的第四方面,提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如第一方面中任一项所述的方法。

[0009] 根据本公开实施例的第五方面,提供一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现如第一方面中任一项所述的方法。

[0010] 本公开的实施例提供的技术方案可以包括以下有益效果:通过自动化自然语言处理提取文档的纯内容文本、语句单元、词句单元及直接引用关系,并协同完成训练的通用大模型与本地领域大模型进行多层级语义处理,有效融合了直接引用关系与深层语义关联度,从而自动化、精准地计算出知识文档间的综合引用关系概率。在此基础上,依据该概率自动构建以知识文档为节点、引用关系为边的知识图谱,显著提升了从海量非结构化文档中挖掘和构建知识关联的效率与准确性,实现了知识体系的结构化、智能化组织。

[0011] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0012] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本发明的实施例,并与说明书一起用于解释本发明的原理。

[0013] 图1是根据一示例性实施例示出的一种针对知识文档的知识图谱构建方法的流程图。

[0014] 图2是根据一示例性实施例示出的一种针对知识文档的知识图谱构建装置的框图。

[0015] 图3是根据一示例性实施例示出的一种用于针对知识文档的知识图谱构建方法的装置的框图。

具体实施方式

[0016] 这里将详细地对示例性实施例进行说明,其示例表示在附图中。下面的描述涉及附图时,除非另有表示,不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0017] 在本公开实施例使用的术语是仅仅出于描述特定实施例的目的,而非旨在限制本公开实施例。在本公开实施例和所附权利要求书中所使用的单数形式的“一种”和“该”也旨在包括多数形式,除非上下文清楚地表示其他含义。

[0018] 应当理解,尽管在本公开实施例可能采用术语第一、第二、第三等来描述各种信息,但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如,

在不脱离本公开实施例范围的情况下,第一信息也可以被称为第二信息,类似地,第二信息也可以被称为第一信息。取决于语境,如在此所使用的词语“如果”及“若”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0019] 此外,可以使用本公开实施例所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本公开公开的技术方案所期望的结果,本文在此不进行限制。

[0020] 需要说明的是,本公开的技术方案中,所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0021] 图1是根据一示例性实施例示出的一种针对知识文档的知识图谱构建方法的流程图,如图1所示,需要说明的是,本公开实施例的针对知识文档的知识图谱构建方法应用于针对知识文档的知识图谱构建装置中。如图1所示,该方法可以包括以下步骤:

步骤101,对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系。

[0022] 在一个实施例中,可以以知识文档为单位进行处理,在此处利用小语言模型对于文本进行主要内容的提取处理,去除标题、摘要、引用文献、项目信息等知识主要内容外的内容,形成纯内容文本text,存入数据中台文本语料库,供大模型处理模块使用,同时此提取文本中的知识文档引用关系,共有6中引用关系:论文-论文引用、专利-专利引用、技术报告-技术报告引用、论文-专利引用、技术报告-论文引用、技术报告-专利引用。

[0023] 作为一种示例,对于两个知识文档a,b,如果a,b皆为知识服务平台中的知识文档,则两个文档存在引用关系,记为P_quote(a, b) = 1,该种引用关系为强关系,概率为1;当a或b其中一个不存在于知识服务平台中时,则这种引用和被引用关系可以作为该文档的属性值存入引用属性或者被引用属性,记为attribute[a，'quote'] = b,attribute[b，'cited'] = a。

[0024] 此外,可以将知识文档分解为能完成表达基本语义的符合语法的句子作为语义识别单元,记为语句单元sentence,存入数据中台语句语料库,并将这些句子作为索引链接到知识文档a的语句属性上,记为attribute[a,'sentence'] = sentence。

[0025] 另外,可以将词句进行NLP分析,分解为词句作为语义识别单元,记为词句单元word,并将这些词句作为索引链接到相应语句单元sentence上,如果文档明确存在关键字内容,如论文的关键字等,将这些关键字识别出来存入开采专业词库,存在该词句为文档的直接引用,记为:P_quote_text(word, text) = 1。

[0026] 在本实施例中,以单篇知识文档为处理单元,借助轻量级语言模型剥离标题、摘要、参考文献等外围信息,提取出核心的纯内容文本(text),并同步识别其中存在的多种文档间引用关系(如论文-专利、报告-论文等)。对于明确存在于知识服务平台内的文档对,直接建立强引用关联(概率为1);对于单方存在的引用,则将其记录为文档属性。接着,将纯文本切分为完整的句子单元(sentence)并建立索引;进一步地,对句子进行分词处理得到词句单元(word),若词句属于明确列出的关键词,则将其标记为直接引用并存入领域专业词库,从而完成从文档、句子到词句的多粒度结构化解析与语义资源建设。

[0027] 在本申请一些实施例中,步骤101之后,方法还可以包括以下步骤:

步骤a1,将词句单元与预设的煤炭开采领域的专业词库中的词句进行匹配,在匹配成功的情况下,计算词句单元在所属语句中的第一重要性概率,以及计算词句在所属文

档中的第二重要性概率;

步骤a2,基于词句在其所属语句中的重要程度评分、第一重要性概率和第二重要性概率,采用以下公式计算词句的关键词概率 $P\_keyword(word,text)$:

$$P\_keyword(word,text)=P'\_sentence*\beta+(1-\beta)*P\_text(word,sentence)$$

[0028] $$P'\_sentence=\frac{\sum_{n}^{\delta}*P\_api\_sentence(word,sentence)+(1-\delta)P\_sentence(word,sentence)}{n}$$

[0029] 其中,$P\_api\_sentence$(word, sentence)为第一语义处理结果,P_sentence(word, sentence)为第一重要性概率,$P\_text(word,sentence)$ 为第二重要性概率,β为调节系数;

方法还包括:

基于关键词概率构建知识图谱中的属性信息。

[0030] 在本申请一些实施例中,在步骤a2之后,方法还可以包括以下步骤:

将词句单元与知识文档中明确标明的关键字进行比对,若词句单元为关键字,则确定其直接引用概率为1;

基于直接引用概率、第一重要性概率、第二重要性概率及重要程度评分,采用以下公式计算关键词概率:

P_keywords(word,text)=P_keywords(word,text)*(1-P_quote_text(word,text))+P_quote_text(word,text)

其中,P_keywords(word,text)为关键词概率,P_quote_text(word,text)为直接引用概率。

[0031] 在本申请实施例中,通过引入直接引用概率并设定其明确规则(当词句单元为文档中明确标明的关键字时概率为1),确保了已被文档显式标记的关键信息在后续计算中得到充分且准确的体现;进而,通过所设计的概率融合公式,将直接引用概率与基于内容分析的第一重要性概率、第二重要性概率及重要程度评分进行有机结合,从而在关键词概率计算中实现了对显式标记与隐式重要性的综合平衡与修正,显著提升了关键词识别的准确性与鲁棒性,从而既能充分尊重文档自身的结构化标注,又能智能捕捉上下文中的潜在重点,最终生成更可靠、更贴合文档意图的关键词概率评估结果。

[0032] 在本申请实施例中,步骤a1具体可以包括:将解析出的词句单元(word)与该专业词库进行匹配。

[0033] 对于匹配成功的词句,进一步计算其在当前语句上下文及全文语境中的重要性。

[0034] 具体地,计算该词句在其所属语句单元(sentence)中的第一TF-IDF值TF-IDF(word, sentence),以及在其所属纯内容文本(text)中的第二TF-IDF值TF-IDF(word, sentence)。基于文本语料库的文档总数m和语句语料库的句子总数n,分别计算该词句的语句级重要性概率P_sentence(word, sentence) = TF-IDF(word, sentence)/lg(n),以及文档级重要性概率P_text(word, text) = TF-IDF(word, text)/lg(m)。

[0035] 对于未成功匹配专业词库的词句,通过一个自适应决策模型判断其是否应作为新术语纳入词库:

设定调节系数β(0<β≤1)和阈值θ,当满足P_sentence(word,text)β+$P\_text(word, sentence)$(1-β)≥θ时,将该词句存入专业词库并后续进行与比中词句相同的处

理;否则暂不处理。所有处理后的词句及概率信息存入数据中台,并与相应的语句单元及知识文档建立索引关联。

[0036] 在一个实施例中,将解析出的词句单元与专业词库进行匹配,可以采用基于字符串的匹配规则来实现。例如,可以通过将词句单元与专业词库中的词条进行精确字符串比对来完成;在实际应用中,亦可采用在精确匹配基础上允许一定字符容错的模糊匹配方式,或采用基于词边界的最长匹配策略,以处理实际文本中的表述变体或未登录词情况。本领域技术人员可根据实际需求选择合适的匹配算法,上述具体方式的变更均落入本发明的保护范围之内。

[0037] 在本申请实施例中,通过与权威领域专业词库的匹配,实现了对专业术语的精准初始筛选,确保了核心实体提取的准确性,为知识结构化奠定了可靠基础。其次,结合词句在微观语句语境与宏观文档语境中的TF-IDF统计特征,并通过对数归一化处理,生成了稳定、可比的语句级与文档级双重重要性概率,实现了对术语语义贡献度的细粒度、多维度量化评估,此外基于概率加权与自适应阈值的决策机制,能够智能判断并吸纳文本中涌现的新术语进入专业词库,使系统具备了持续学习与自我演进的能力,从根本上克服了静态词库难以适应技术词汇动态发展的瓶颈。

[0038] 在本申请一些实施例中,上述基于关键词概率构建知识图谱中的属性信息,具体可以包括:

在关键词概率大于或者等于1的情况下,将关键词概率更新为1,将词句单元配置以及更新后的关键词概率配置为所属知识文档对应节点的属性信息;

在关键词概率大于或者等于预设关键词概率并且小于1的情况下,将词句单元配置以及关键词概率配置为所属知识文档对应节点的属性信息。

[0039] 在本申请实施例中,对于关键词概率大于或等于1的情况(例如文档中作者明确标注的关键词),将其关键词概率重置为1后与词句一同作为节点属性,将此类“权威标注”转化为图谱中确定性最高、权重最强的属性信息,确保了核心知识标签的权威性。

[0040] 在本申请另一个实施例中,对于关键词概率介于预设阈值d与1之间的情况(例如算法综合推断出的重要术语),保留其原概率值作为属性,从而在知识图谱中构建出带有连续置信度谱系的知识关联,不仅精准反映了术语与文档之间真实、细微的关联强度差异,更使得后续基于图谱的智能应用能够利用这些概率值进行精细化计算。

[0041] 在本申请实施例中,通过“确定性处理”与“概率性保留”相结合的智能策略,在提升图谱构建自动化程度的同时,有效保障了其输出结果的准确性、丰富性与可计算性。

[0042] 步骤102,基于完成训练的大模型对词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果。

[0043] 其中,第一语义处理结果包括词句在其所属语句中的重要程度评分;第二语义处理结果包括知识文档对之间的语义关联度评分。

[0044] 在本申请实施例中,完成训练的大模型可以是指已在海量通用或领域数据上完成预训练过程、具备强大语义理解与生成能力的大型深度学习模型。

[0045] 在本申请实施例中,本地大模型可以是指部署于用户私有服务器或专属云计算环境中的、已完成预训练并经过特定垂直领域知识(如煤炭开采)进一步微调优化的大型深度

学习模型。其运行在封闭的本地网络环境中,确保所有敏感或专有的知识文档数据无需离境,从而在根本上保障了数据主权与知识产权安全。该模型可以具备通用语义理解能力,还可以深度融合领域术语、技术逻辑与行业规范,使其能够对完整的专业文档进行深层次的内容理解、主题提炼与语义关联分析,最终输出高质量的文档间语义关联度评分。

[0046] 在本申请实施例中,可以将细粒度的词句单元和语句单元提交至一个或多个已完成训练的通用大模型API进行并发处理,利用其强大的通用语义理解能力,快速、高效地评估每个词在特定语句上下文中的局部重要性,输出归一化的重要程度评分,形成第一语义处理结果。与此同时,将代表文档整体内容的纯内容文本,输入至本地部署且经过领域数据微调的专用大模型进行处理;该本地模型在确保敏感数据不外泄的前提下,能够深入理解文档的完整技术逻辑与主题脉络,并综合评估任意两篇文档在语义层面的整体关联强度,输出文档对之间的语义关联度评分,作为第二语义处理结果。

[0047] 在本申请实施例中,通过利用大模型API处理细粒度单元,本地领域模型处理完整文档的协同设计,既保障了处理效率与前沿语义能力的利用,又兼顾了数据安全、领域专业知识深度挖掘以及整体语义关联的精准判断,为后续的概率融合与图谱构建提供了兼具广度与深度的语义分析基础。

[0048] 在本申请一些实施例中,步骤102中的基于完成训练的大模型对词句单元、语句单元进行语义处理,得到第一语义处理结果,具体可以包括以下步骤:

将词句单元和语句单元分别输入至多个已训练完成的通用大模型;

接收各通用大模型返回的语义处理结果,并对语义处理结果进行归一化处理;语义处理结果至少包括对应词句的重要性评价值、大模型检索的总语料数以及命中关联语料数;

采用以下公式对通用大模型返回的重要性评价进行加权融合,得到第一语义处理结果:

$$P\_api\_sentence(word, sentence) = \sum_i \gamma_i * importance(i) * \frac{hit(i)}{num(i)}$$

[0049] 其中,$P_api_sentence(word, sentence)$为第一语义处理结果,$\gamma_i$为第i个大模型的权重,全部大模型的权重之和为1,$importance(i)$为第i个大模型输出的重要性评价值,hit(i)为第i个大模型检索的总语料数,num(i)为第i个大模型的命中关联语料数。

[0050] 在本申请实施例中,可以并行地向各模型提交待分析的语义单元,并接收其返回的包含词句重要性评价值、总检索语料数及命中关联语料数在内的处理结果。通过引入基于权重的融合公式,不仅综合了各模型的重要性评分,还以命中率作为客观校正因子,从而将多模型的输出结果进行归一化与加权融合,最终生成一个稳定、可靠且兼具主客观评估依据的第一语义处理结果。从而有效提升了语义理解的鲁棒性与准确性。

[0051] 在本申请一些实施例中,对未命中专业词库的词句单元,可以设定一个调节值δ $(0<\delta<1)$,其值可根据系统实际运行反馈动态调整。可以综合考量来自大模型的语义评价与基于统计的重要性概率,通过以下判断条件决定是否将该词句作为新术语纳入专业词库:

$$\delta*P\_api\_sentence(word,sentence)+(1-\delta)*[P\_sentence(word,sentence)*\beta+P\_text(word,text)*(1-\beta)]\geq\theta$$

[0052] 其中,P_api_sentence(word, sentence)为第一语义处理结果中该词句的重要性评分,P_sentence(word, sentence)和P_text(word, text)分别为计算得到的语句级与文档级重要性概率,β为另一调节系数,θ为预设的入库阈值。

[0053] 若满足上述条件,则判定该词句具有足够的领域相关性及重要性,将其存入专业词库,并随后按命中词库的流程进行处理与索引;否则,暂不将该词句作为专业术语处理。

[0054] 在本申请一些实施例中,步骤102中的将纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果,具体可以包括以下步骤:

将纯内容文本输入至已完成训练且针对特定领域优化的本地大模型进行语义处理;

获取本地大模型返回的第二语义处理结果;第二语义处理结果还包括专业词库中的词句在纯内容文本中的语义重要度评分。

[0055] 在本申请实施例中,可以将经过清洗和提炼的纯内容文本(即去除冗余格式与外围信息的核心内容)输入至本地部署且已针对特定领域(如煤炭开采)完成优化训练的大模型中。该本地大模型利用其深度融合的领域知识,执行两项核心分析:一是从整体上理解文档的语义主旨与技术逻辑,从而计算出任意两篇文档在领域语境下的深层语义关联度评分;二是基于其内部的知识表示,对专业词库中既有的术语进行情境化评估,判断并输出每个专业词在当前特定文档内容中的具体语义重要度评分。这一过程不仅实现了对文档间关联的精准量化,也完成了专业术语在具体技术上下文中的重要性校准,为后续构建高质量、高精度的知识关联提供了关键的语义理解基础。

[0056] 在本申请的一些实施例中,本地大模型返回的第二语义处理结果包含专业词库中的词句在纯内容文本中的语义重要度评分 P_api_word(word, text)。可以预设有一判断阈值 q。当语义重要度评分 P_api_word(word, text) 大于或等于阈值 q 时,判定该专业词对当前知识文档的内容具有显著的语义影响。此时,可以保存该词句与文档的关联信息,关联信息至少包括词句、知识文档的标识以及语义重要度评分,以供后续关键词概率计算或知识图谱属性构建等模块调用与集成。

[0057] 步骤103,基于直接引用关系、第一语义处理结果及第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率。

[0058] 在本申请一些实施例中,步骤103具体可以包括以下步骤:

采用以下公式计算任意两个知识文档之间的综合引用关系概率P(a,b):

$$P(a,b)=(\sigma*P\_api(a,b)+(1-\sigma)*P\_pre(a,b))*(1-P\_quote(a,b))+P\_quote(a,b)$$

[0059] 其中,a和b为任意两个知识文档,$P_api(a,b)$为知识文档a与知识文档b之间的语义关联度评分,σ为调节系数,$P_pre(a,b)$为知识文档a与知识文档b之间的预处理关联度评分,$P_quote(a,b)$为知识文档a与知识文档b之间的直接引用关系。

[0060] 在一个实施例中,预处理关联度评分$P_pre(a,b)$可以通过计算知识文档a与知识文档b之间的余弦相似度得到。

[0061] 在本申请实施例中,将直接引用关系(P_quote)作为确定性基础,确保显式关联不

被遗漏,同时利用调节系数σ动态整合基于整体文档深度语义的评分(P_api)与基于局部单元预处理的分析结果(P_pre),既兼顾了上下文整体相关性又捕捉了细粒度证据,有效克服了单一信息源的局限。这种融合机制大幅降低了误判风险,使系统能够自动化、精准地量化文档间潜在的复杂引用强度,为构建高质量、高可信度的知识图谱提供了可靠且可解释的概率化关联依据。

[0062] 步骤104,基于综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

[0063] 在一个实施例中,可以依据预设的阈值规则,对任意两篇文档间的综合引用关系概率进行判断:当综合引用关系概率值达到或超过“强关联”阈值时,则直接在对应的两个文档节点间建立一条有向的引用边;当概率值处于“中等关联”区间时,则启动路径分析,寻找通过中间文档形成的可靠引用链条,并据此构建间接的引用关系网络;当概率值低于“中等关联”区间时,不建立引用边。通过这一过程,海量离散的非结构化知识文档被自动、智能地组织成一个以文档为实体、以概率化引用关系为纽带的结构化知识网络,为智能检索、知识发现与价值评估提供了直接的数据基础。

[0064] 在本申请一些实施例中,步骤104具体可以包括以下步骤:

步骤b1,以知识文档为节点、以引用关系为边权构建概率关系图。

[0065] 在本申请实施例中,可以将前期计算得出的、量化的综合引用关系概率值设定为节点(知识文档)间有向边的权重,从而构建一个初始的概率关系图,将抽象的关联强度转化为具体的、可计算的网络拓扑。

[0066] 步骤b2,删除概率关系图中的节点自环。

[0067] 在本申请实施例中,可以通过删除所有节点指向自身的边(也即自环),主动修正数据中可能存在的逻辑悖论与计算噪声,确保了图的数学严谨性与业务合理性,为后续分析提供了纯净的结构基础。

[0068] 步骤b3,计算概率关系图中任意两节点x与y间所有连通路径的综合概率均值,得到目标引用概率。

[0069] 其中,单条路径的概率为该路径上所有边的综合引用关系概率的乘积。

[0070] 在本申请实施例中,可以通过遍历图中任意两个节点间所有可能的连通路径,并将每条路径上各边权重(即概率)的乘积作为该路径的整体关联强度,最后对所有路径强度取平均值,从而得到最终的目标引用概率,综合考虑了文档间所有直接与间接的关联证据,将局部的、二元的关系概率,融合为一个反映全局网络连通性强度的稳定指标,极大地增强了最终图谱构建决策的鲁棒性与准确性。

[0071] 在本申请一些实施例中,步骤b3具体可以包括以下步骤:

采用以下公式计算概率关系图中任意两节点x与y间所有连通路径的目标引用概率:

$$P\_build(x,y) = \sum \frac{P(x, i...j, y)}{root\ (a,i,...,j,b)}$$

[0072] $$P(x,i...j,y) = P(x,i) * P(i,n) * ... * P(m,j) * P(j,y)$$

[0073] 其中,$P_build(x, y)$为节点x与y之间的目标引用概率,$P(x, i...j, y)$为单条路径的概率,a,i...j,b为遍历路径,root(a,i,...,j,b)为经过节点数。

[0074] 在本申请实施例中,对于概率关系图中任意两个文档节点x和y,系统会遍历其间所有可能的连通路径;对于每条路径(例如x,i,...,j,y),其整体关联强度通过将该路径上每一段相邻节点间的基础引用关系概率连续相乘得到。可以将所有连通路径的强度计算结果进行求和,并除以所涉路径的节点总数进行归一化处理(也即对路径节点数root(a,i,...,j,b)的考量),最终得出一个综合反映直接与间接所有关联证据的、更为鲁棒的目标引用概率P_build(x,y),从而有效克服了单一直接关联可能存在的偶然性或噪声干扰,通过图网络的全局结构信息提升了最终知识关联判断的准确性与稳定性。

[0075] 步骤b4,基于目标引用概率构建知识图谱。

[0076] 在本申请一些实施例中,步骤a4具体可以包括以下步骤:

针对任意两个知识文档之间的目标引用概率,若目标引用概率大于或等于第一阈值,则在知识图谱中建立从其中一个知识文档对应的第一节点指向另一个知识文档对应的第二节点的引用关系边;

若目标引用概率小于第一阈值并且大于第二阈值,则遍历概率关系图中从第一节点指向第二节点的所有可能路径,若遍历到存在至少一条子路径的目标引用概率大于或等于第一阈值的至少一个候选路径,则从至少一个候选路径中选取目标引用概率最高的目标路径,并沿目标路径在知识图谱中建立第一节点指向第二节点的引用关系链。

[0077] 在本申请一个实施例中,可以预先设定两个关键阈值:第一阈值(高置信度阈值)与第二阈值(中置信度阈值)。

[0078] 对于任意两篇文档,若其综合计算得到的目标引用概率大于或等于第一阈值,表明二者之间存在强且直接的关联证据,系统随即在知识图谱中直接建立从前一文档节点指向后一文档节点的引用关系边。

[0079] 当目标引用概率介于第一阈值与第二阈值之间时,表明直接关联强度不足但存在潜在间接关联。此时,不直接建立边,而是遍历从起始节点到目标节点的所有可能路径,检查这些路径中是否存在至少一条子路径的强度达到第一阈值(即包含强关联片段)。若存在此类候选路径,则从中选取整体概率最高的路径作为“可靠通道”,并沿着这条路径的节点顺序,在知识图谱中依次建立相应的引用关系边,最终形成一条从起始节点指向目标节点的间接引用关系链。

[0080] 在本申请实施例中,当目标引用概率达到第一阈值时直接建立引用边,实现了对高置信度关联的快速、准确构建,极大提升了图谱生成效率,此外,当概率处于中间区间时,通过智能路径分析,仅在存在经强关联子路径验证的可靠间接通道时,才构建引用关系链,有效避免了低质或虚假关联的引入,确保了图谱中每一条关系的可靠性,同时挖掘并固化了那些虽不直接但逻辑严谨的深层知识关联,从而使得最终形成的知识网络兼具高度的准确性、丰富的连接维度和良好的结构深度。

[0081] 根据本公开实施例提出的针对知识文档的知识图谱构建方法,通过自动化自然语言处理提取文档的纯内容文本、语句单元、词句单元及直接引用关系,并协同完成训练的通用大模型与本地领域大模型进行多层级语义处理,有效融合了直接引用关系与深层语义关联度,从而自动化、精准地计算出知识文档间的综合引用关系概率。在此基础上,依据该概率自动构建以知识文档为节点、引用关系为边的知识图谱,显著提升了从海量非结构化文档中挖掘和构建知识关联的效率与准确性,实现了知识体系的结构化、智能化组织。

[0082] 图2是根据一示例性实施例示出的一种针对知识文档的知识图谱构建装置框图。参照图2,该装置包括自然语言处理单元201,语义处理单元202,计算单元203和构建单元204。

[0083] 其中,自然语言处理单元201,用于对知识文档进行自然语言处理,得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系;

语义处理单元202,用于基于完成训练的大模型对词句单元、语句单元进行语义处理,得到第一语义处理结果,以及,将纯内容文本输入本地大模型进行语义处理,得到第二语义处理结果;第一语义处理结果包括词句在其所属语句中的重要程度评分;第二语义处理结果包括知识文档对之间的语义关联度评分;

计算单元203,用于基于直接引用关系、第一语义处理结果及第二语义处理结果,计算任意两个知识文档之间的综合引用关系概率;

构建单元204,用于基于综合引用关系概率,构建以知识文档为节点、以引用关系为边的知识图谱。

[0084] 在本申请一些实施例中,语义处理单元202具体可以用于:

将词句单元和语句单元分别输入至多个已训练完成的通用大模型;

接收各通用大模型返回的语义处理结果,并对语义处理结果进行归一化处理;语义处理结果至少包括对应词句的重要性评价值、大模型检索的总语料数以及命中关联语料数;

采用以下公式对通用大模型返回的重要性评价进行加权融合,得到第一语义处理结果:

$$P\_api\_sentence(word, sentence)=\sum_i \gamma_i * importance(i) * \frac{hit(i)}{num(i)}$$

[0085] 其中,$P_api_sentence(word, sentence)$为第一语义处理结果,$\gamma_i$为第i个大模型的权重,$importance(i)$为第i个大模型输出的重要性评价值,hit(i)为第i个大模型检索的总语料数,num(i)为第i个大模型的命中关联语料数。

[0086] 在本申请一些实施例中,语义处理单元202具体可以用于:

将纯内容文本输入至已完成训练且针对特定领域优化的本地大模型进行语义处理;

获取本地大模型返回的第二语义处理结果;第二语义处理结果还包括专业词库中的词句在纯内容文本中的语义重要度评分。

[0087] 在本申请一些实施例中,装置具体还可以包括:

计算单元203,还用于将词句单元与预设的煤炭开采领域的专业词库中的词句进行匹配,在匹配成功的情况下,计算词句单元在所属语句中的第一重要性概率,以及计算词句在所属文档中的第二重要性概率;

计算单元203,还用于基于词句在其所属语句中的重要程度评分、第一重要性概率和第二重要性概率,采用以下公式计算词句的关键词概率$P_keyword(word,text)$:

$$P\_keyword(word,text)=P'\_sentence*\beta+(1-\beta)*P\_text(word,sentence)$$

[0088] $$P'\_sentence = \frac{\sum_{n} \delta * P\_api\_sentence(word, sentence) + (1-\delta)\, P\_sentence(word, sentence)}{n}$$

[0089] 其中，$P\_api\_sentence$（word，sentence）为第一语义处理结果，P_sentence(word，sentence)为第一重要性概率，$P_text(word, sentence)$为第二重要性概率，β为调节系数；

装置还包括：

构建单元204，还用于基于关键词概率构建知识图谱中的属性信息。

[0090] 在本申请一些实施例中，构建单元204具体可以用于：

在关键词概率大于或者等于1的情况下，将关键词概率更新为1，将词句单元配置以及更新后的关键词概率配置为所属知识文档对应节点的属性信息；

在关键词概率大于或者等于预设关键词概率并且小于1的情况下，将词句单元配置以及关键词概率配置为所属知识文档对应节点的属性信息。

[0091] 在本申请一些实施例中，计算单元203具体可以用于：

采用以下公式计算任意两个知识文档之间的综合引用关系概率P(a，b)：

$$P(a,b) = (\sigma * P\_api(a,b) + (1-\sigma) * P\_pre(a,b)) * (1 - P\_quote(a,b)) + P\_quote(a,b)$$

[0092] 其中，a和b为任意两个知识文档，$P_api(a,b)$为知识文档a与知识文档b之间的语义关联度评分，P_pre(a,b)为知识文档a与知识文档b之间的预处理关联度评分，P_quote(a,b)为知识文档a与知识文档b之间的直接引用关系。

[0093] 在本申请一些实施例中，构建单元204具体可以用于：

以知识文档为节点、以引用关系为边权构建概率关系图；

删除概率关系图中的节点自环；

计算概率关系图中任意两节点x与y间所有连通路径的综合概率均值，得到目标引用概率，其中，单条路径的概率为该路径上所有边的综合引用关系概率的乘积；

基于目标引用概率构建知识图谱。

[0094] 在本申请一些实施例中，构建单元204具体可以用于：

采用以下公式计算概率关系图中任意两节点x与y间所有连通路径的目标引用概率：

$$P\_build(x,y) = \sum \frac{P(x, i \ldots j, y)}{root\ (a, i, \ldots, j, b)}$$

[0095] $$P(x, i \ldots j, y) = P(x,i) * P(i,n) * \ldots * P(m,j) * P(j,y)$$

[0096] 其中，P_build(x，y)为节点x与y之间的目标引用概率，$P(x, i \ldots j, y)$为单条路径的概率，a，i...j，b为遍历路径，root(a，i，...，j，b)为经过节点数。

[0097] 在本申请一些实施例中，构建单元204具体可以用于：

针对任意两个知识文档之间的目标引用概率，若目标引用概率大于或等于第一阈值，则在知识图谱中建立从其中一个知识文档对应的第一节点指向另一个知识文档对应的第二节点的引用关系边；

若目标引用概率小于第一阈值并且大于第二阈值，则遍历概率关系图中从第一节点指向第二节点的所有可能路径，若遍历到存在至少一条子路径的目标引用概率大于或等于第一阈值的至少一个候选路径，则从至少一个候选路径中选取目标引用概率最高的目标

路径,并沿目标路径在知识图谱中建立第一节点指向第二节点的引用关系链。

[0098] 关于上述实施例中的装置,其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述,此处将不做详细阐述说明。

[0099] 根据本公开实施例提出的针对知识文档的知识图谱构建装置,通过自动化自然语言处理提取文档的纯内容文本、语句单元、词句单元及直接引用关系,并协同完成训练的通用大模型与本地领域大模型进行多层级语义处理,有效融合了直接引用关系与深层语义关联度,从而自动化、精准地计算出知识文档间的综合引用关系概率。在此基础上,依据该概率自动构建以知识文档为节点、引用关系为边的知识图谱,显著提升了从海量非结构化文档中挖掘和构建知识关联的效率与准确性,实现了知识体系的结构化、智能化组织。

[0100] 图3是根据一示例性实施例示出的一种用于针对知识文档的知识图谱构建方法的装置的框图。例如,装置300可以是电子设备,例如可以是移动电话,计算机,数字广播终端,消息收发设备,平板设备,个人数字助理等。

[0101] 参照图3,装置300可以包括以下一个或多个组件:处理组件302,存储器304,电力组件306,多媒体组件308,音频组件310,输入/输出(I/ O)的接口312,传感器组件314,以及通信组件316。

[0102] 处理组件302通常控制装置300的整体操作,诸如与显示,电话呼叫,数据通信,相机操作和记录操作相关联的操作。处理组件302可以包括一个或多个处理器320来执行指令,以完成上述的方法的全部或部分步骤。此外,处理组件302可以包括一个或多个模块,便于处理组件302和其他组件之间的交互。例如,处理组件302可以包括多媒体模块,以方便多媒体组件308和处理组件302之间的交互。

[0103] 存储器304被配置为存储各种类型的数据以支持在设备300的操作。这些数据的示例包括用于在装置300上操作的任何应用程序或方法的指令,联系人数据,电话簿数据,消息,图片,视频等。存储器304可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。

[0104] 电力组件306为装置300的各种组件提供电力。电力组件306可以包括电源管理系统,一个或多个电源,及其他与为装置300生成、管理和分配电力相关联的组件。

[0105] 多媒体组件308包括在所述装置300和用户之间的提供一个输出接口的屏幕。在一些实施例中,屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板,屏幕可以被实现为触摸屏,以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界,而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中,多媒体组件308包括一个前置摄像头和/或后置摄像头。当设备300处于操作模式,如拍摄模式或视频模式时,前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0106] 音频组件310被配置为输出和/或输入音频信号。例如,音频组件310包括一个麦克风(MIC),当装置300处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器304或经由通信组

件316发送。在一些实施例中,音频组件310还包括一个扬声器,用于输出音频信号。

[0107] I/ O接口312为处理组件302和外围接口模块之间提供接口,上述外围接口模块可以是键盘,点击轮,按钮等。这些按钮可包括但不限于:主页按钮、音量按钮、启动按钮和锁定按钮。

[0108] 传感器组件314包括一个或多个传感器,用于为装置300提供各个方面的状态评估。例如,传感器组件314可以检测到设备300的打开/关闭状态,组件的相对定位,例如所述组件为装置300的显示器和小键盘,传感器组件314还可以检测装置300或装置300一个组件的位置改变,用户与装置300接触的存在或不存在,装置300方位或加速/减速和装置300的温度变化。传感器组件314可以包括接近传感器,被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件314还可以包括光传感器,如CMOS或CCD图像传感器,用于在成像应用中使用。在一些实施例中,该传感器组件314还可以包括加速度传感器,陀螺仪传感器,磁传感器,压力传感器或温度传感器。

[0109] 通信组件316被配置为便于装置300和其他设备之间有线或无线方式的通信。装置300可以接入基于通信标准的无线网络,如WiFi,2G或3G,或它们的组合。在一个示例性实施例中,通信组件316经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,所述通信组件316还包括近场通信(NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(RFID)技术,红外数据协会(IrDA)技术,超宽带(UWB)技术,蓝牙(BT)技术和其他技术来实现。

[0110] 在示例性实施例中,装置300可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现,用于执行上述方法。

[0111] 在示例性实施例中,还提供了一种包括指令的非临时性计算机可读存储介质,例如包括指令的存储器304,上述指令可由装置300的处理器320执行以完成上述方法。例如,所述非临时性计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0112] 在示例性实施例中,还提供了一种计算机程序产品,包括计算机程序,所述计算机程序在被装置300的处理器320执行时实现上述方法。

[0113] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本发明的其它实施方案。本公开旨在涵盖本发明的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本发明的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本发明的真正范围和精神由下面的权利要求指出。

[0114] 应当理解的是,本发明并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

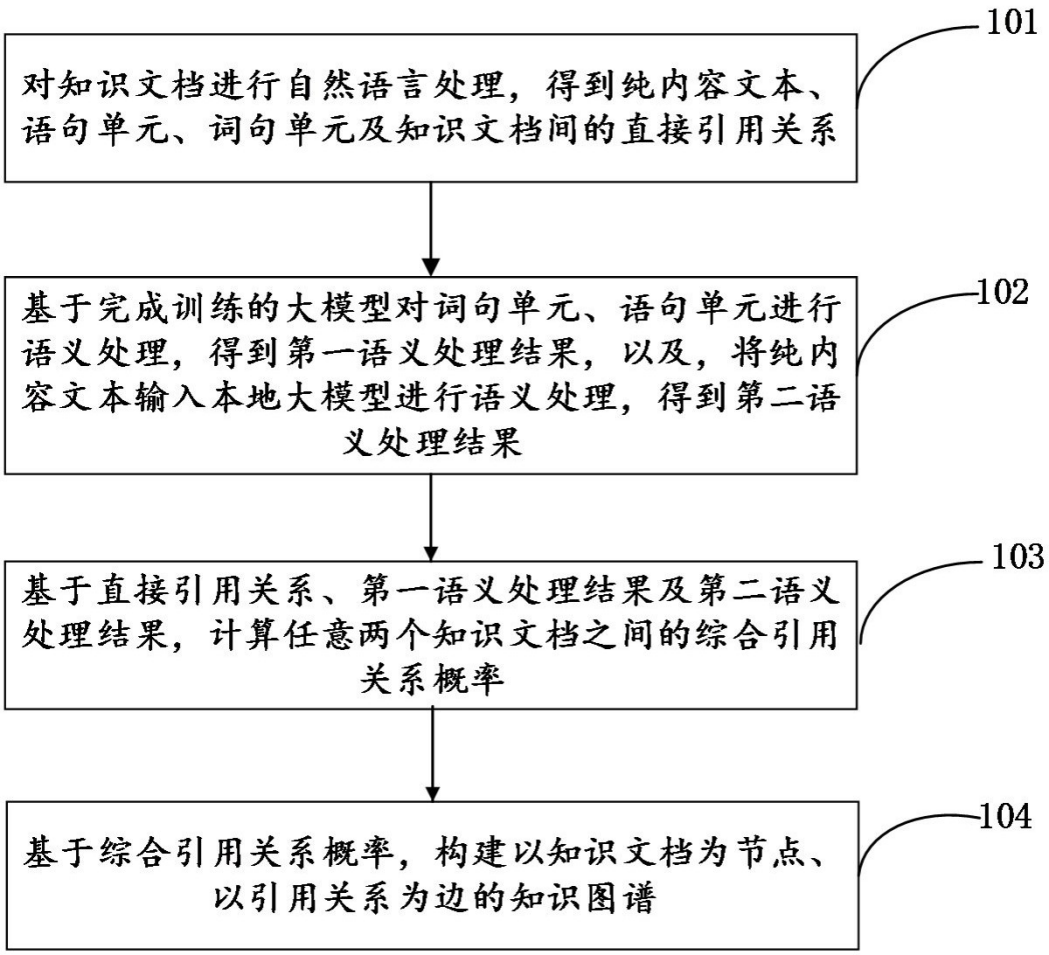

101
对知识文档进行自然语言处理，得到纯内容文本、语句单元、词句单元及知识文档间的直接引用关系
102
基于完成训练的大模型对词句单元、语句单元进行语义处理，得到第一语义处理结果，以及，将纯内容文本输入本地大模型进行语义处理，得到第二语义处理结果
103
基于直接引用关系、第一语义处理结果及第二语义处理结果，计算任意两个知识文档之间的综合引用关系概率
104
基于综合引用关系概率，构建以知识文档为节点、以引用关系为边的知识图谱

图1

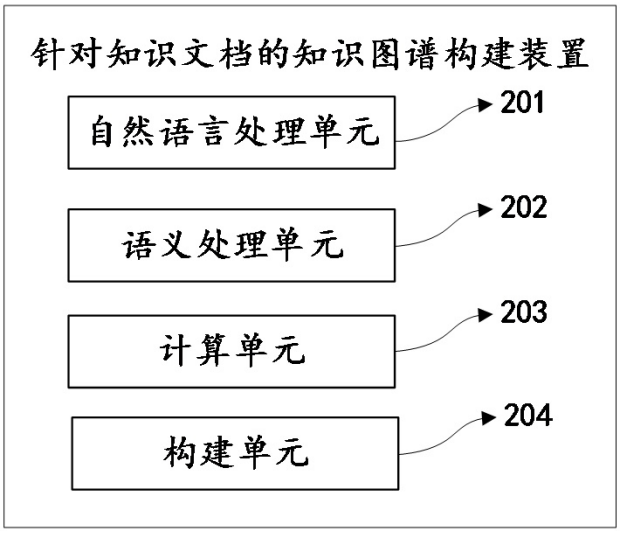

针对知识文档的知识图谱构建装置
自然语言处理单元
201
语义处理单元
202
计算单元
203
构建单元
204

图2

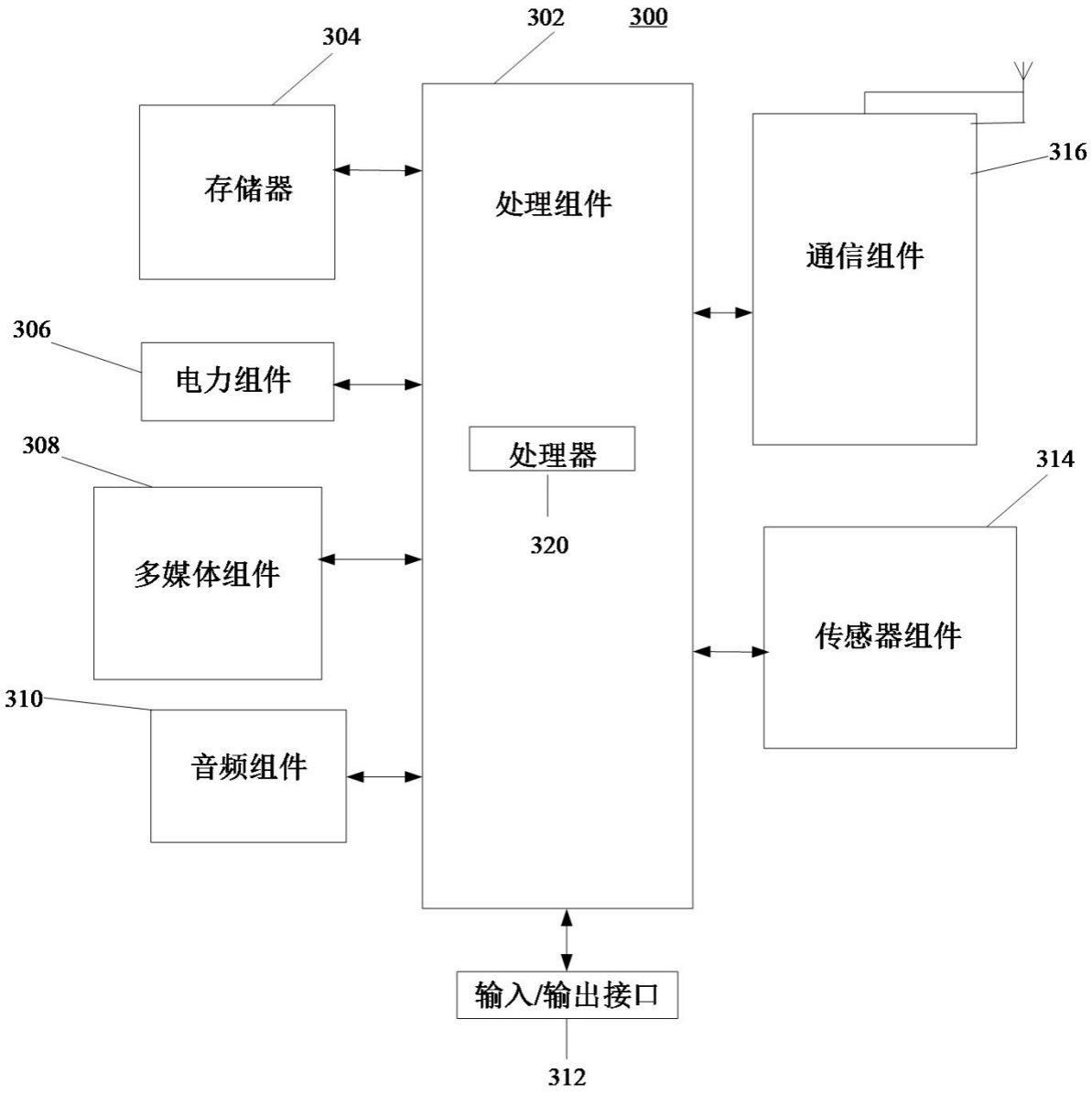

302
300
304
存储器
处理组件
316
通信组件
306
电力组件
处理器
320
308
多媒体组件
314
传感器组件
310
音频组件
输入/输出接口
312

图3