(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114186552 B
(45)授权公告日 2023.04.07

(21)申请号 202111521511.4

(22)申请日 2021.12.13

(65)同一申请的已公布的文献号
申请公布号 CN 114186552 A

(43)申请公布日 2022.03.15

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 夏琦 黄昉 史亚冰 蒋烨 柴春光 朱勇

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 杨泽 刘芳

(51)Int.Cl.
G06F 40/253(2020.01)
G06F 40/284(2020.01)
G06F 40/268(2020.01)
G06F 16/35(2019.01)
G06F 18/22(2023.01)
G06N 20/00(2019.01)
G06N 7/02(2006.01)
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 16/338(2019.01)

(56)对比文件
CN 109241538 A,2019.01.18
CN 112232074 A,2021.01.15
CN 112527981 A,2021.03.19
CN 113536770 A,2021.10.22
US 2012209606 A1,2012.08.16
WO 2021134524 A1,2021.07.08
CN 106777275 A,2017.05.31
CN 111984778 A,2020.11.24
US 2021150140 A1,2021.05.20
陈珂 等.基于最短依存路径和BERT的关系抽取算法研究.《西南师范大学学报(自然科学版)》.2021,第46卷(第46期),56-66.

审查员 刘莲花

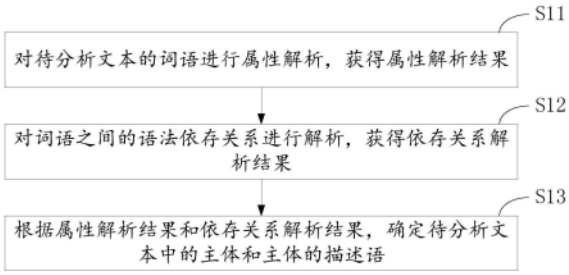
权利要求书3页 说明书11页 附图6页

(54)发明名称

文本分析方法、装置、设备及计算机存储介质

(57)摘要

本公开提供了文本分析方法、装置、设备及计算机存储介质，计算机技术领域，尤其涉及大数据、NLP、智能搜索、知识图谱、深度学习等人工智能领域。具体实现方案为：对待分析文本的词语进行属性解析，获得属性解析结果；对所述词语之间的语法依存关系进行解析，获得依存关系解析结果；根据所述属性解析结果和所述依存关系解析结果，确定所述待分析文本中的主体和所述主体的描述语。本公开实施例能够提高文本分析的准确性。



CN 114186552 B

1.一种文本分析方法,包括:

对待分析文本的词语进行属性解析,获得属性解析结果;

对所述词语之间的语法依存关系进行解析,获得依存关系解析结果;所述依存关系解析结果包括:至少一个主体候选项和至少一个主体的描述语候选项,组成的主体与主体的描述语组合项;

根据所述属性解析结果和所述依存关系解析结果,确定所述待分析文本中的主体和所述主体的描述语;

在所述待分析文本中存在实体词的情况下,所述主体候选项与主体的描述语候选项的获得方式包括:

将所述实体词作为主体候选项;

根据所述主体候选项和设定模式,确定主体的描述语候选项,所述设定模式包括主体、主体的描述语和其它设定词语,以及主体、主体的描述语和其它设定词语之间的相对顺序;

在所述待分析文本中包括设定关键词的情况下,所述根据所述语法依存关系,获得主体候选项和主体的描述语候选项,包括:

在所述待分析文本中,确定与设定关键词存在预设先后顺序的候选词语,所述预设先后顺序包括相邻的先后顺序、存在间隔的先后顺序;

在所述候选词语满足预设条件时,确定所述设定关键词为所述主体候选项,其中,所述预设条件为所述候选词语为预设词语、所述候选词语的属性为设定属性或者所述候选词语为设定类别的词语;

所述组成的主体与主体的描述语组合项,包括:

将所述主体候选项的集合与所述主体的描述语候选项的集合进行组合,得到所有可能的组合;

将所述组合项作为所述主体与主体的描述语组合项。

2.根据权利要求1所述的方法,其中,所述对待分析文本的词语进行属性解析,获得属性解析结果,包括:

确定每个所述词语的属性;

针对每个所述词语,确定所述词语在所述属性下的子分类;

将所有所述词语的属性和子分类,作为所述属性解析结果。

3.根据权利要求1或2所述的方法,其中,在所述待分析文本中存在由至少两个设定词性的词语按照预设顺序组合成的词组的情况下,所述根据所述语法依存关系,获得主体候选项和主体的描述语候选项,包括:

将所述词组拆分,获得拆分词语;

根据拆分词语,确定所述主体候选项和主体的描述语候选项中的至少一个。

4.根据权利要求3所述的方法,其中,所述至少两个设定词性的词语包括设定词性的起始词、和设定词性的终止词,所述起始词和所述终止词在所述待分析文本中的字数距离或词数距离处于设定范围。

5.根据权利要求1所述的方法,其中,所述将所述实体词作为所述主体候选项,包括:

在所述待分析文本中包括两个以上顺序衔接的同类实体的情况下,将所述两个以上顺序衔接的同类实体合并为所述实体词。

6.一种文本分析装置,包括:

属性解析结果获得模块,用于对待分析文本的词语进行属性解析,获得属性解析结果;

依存关系解析结果获得模块,用于对所述词语之间的语法依存关系进行解析,获得依存关系解析结果;

分析结果模块,用于根据所述属性解析结果和所述依存关系解析结果,确定所述待分析文本中的主体和所述主体的描述语;

所述依存关系解析结果包括:至少一个主体候选项和至少一个主体的描述语候选项,组成的主体与主体的描述语组合项;在所述待分析文本中存在实体词的情况下,所述依存关系解析结果获得模块包括:

候选项获得单元,用于将所述实体词作为所述主体候选项;根据所述主体候选项和设定模式,确定所述主体的描述语候选项,所述设定模式包括主体、主体的描述语和其它设定词语,以及所述主体、主体的描述语和其它设定词语之间的相对顺序;

在所述待分析文本中包括设定关键词的情况下,所述候选项获得单元还用于:

在所述待分析文本中,确定与设定关键词存在预设先后顺序的候选词语,所述预设先后顺序包括相邻的先后顺序、存在间隔的先后顺序;

在所述候选词语满足预设条件时,确定所述设定关键词为所述主体候选项,其中,所述预设条件为所述候选词语为预设词语、所述候选词语的属性为设定属性或者所述候选词语为设定类别的词语;

组合项组成单元用于:

将所述主体候选项的集合与所述主体的描述语候选项的集合进行组合,得到所有可能的组合;

将所述组合项作为所述主体与主体的描述语组合项。

7.根据权利要求6所述的装置,其中,所述属性解析结果获得模块包括:

属性确定单元,用于确定每个所述词语的属性;

子分类确定单元,用于针对每个所述词语,确定所述词语在所述属性下的子分类;

结果单元,用于将所有所述词语的属性和子分类,作为所述属性解析结果。

8.根据权利要求6或7所述的装置,其中,在所述待分析文本中存在由至少两个设定词性的词语按照预设顺序组合成的词组的情况下,所述候选项获得单元还用于:

将所述词组拆分,获得拆分词语;

根据拆分词语,确定所述主体候选项和主体的描述语候选项中的至少一个。

9.根据权利要求8所述的装置,其中,所述至少两个设定词性的词语包括设定词性的起始词、和设定词性的终止词,所述起始词和所述终止词在所述待分析文本中的字数距离或词数距离处于设定范围。

10.根据权利要求6所述的装置,其中,所述候选项获得单元还用于:

在所述待分析文本中包括两个以上顺序衔接的同类实体的情况下,将所述两个以上顺序衔接的同类实体合并为所述实体词。

11. 一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-5中任一项所述的方法。

12.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使计算机执行根据权利要求1-5中任一项所述的方法。

文本分析方法、装置、设备及计算机存储介质

技术领域

[0001] 本公开涉及计算机技术领域，尤其涉及大数据、NLP(Natural Language Processing，自然语言处理)、智能搜索、知识图谱、深度学习等人工智能领域。

背景技术

[0002] 随着计算机技术的发展，计算机技术对数据、信息的处理效果也显著提升，不仅处理速度加快，且灵活程度提高，在人工智能等领域，计算机生成的图像、语言等处理结果，也能够逐渐减少机械化的模板感，在保证正确率的情况下，达到越来越贴合实际生活场景的效果。

[0003] 比如，机器可以对一部分文本进行分析，实现信息的提取等目的。机器对文本的分析可应用于搜索、对话等多种场景，由于这些场景中的文本，与用户的使用习惯、普通群体的表达习惯息息相关，因此，需要对文本分析技术进行改进，以更好地适应用户群体在使用产品过程中的一般习惯。

发明内容

[0004] 本公开提供了一种文本分析方法、装置、设备及计算机存储介质。

[0005] 根据本公开的一方面，提供了一种文本分析方法，包括：对待分析文本的词语进行属性解析，获得属性解析结果；

[0006] 对词语之间的语法依存关系进行解析，获得依存关系解析结果；

[0007] 根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语。

[0008] 根据本公开的另一方面，提供了一种文本分析装置，包括：

[0009] 属性解析结果获得模块，用于对待分析文本的词语进行属性解析，获得属性解析结果；

[0010] 依存关系解析结果获得模块，用于对词语之间的语法依存关系进行解析，获得依存关系解析结果；

[0011] 分析结果模块，用于根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语。

[0012] 根据本公开的另一方面，提供了一种电子设备，包括：

[0013] 至少一个处理器；以及

[0014] 与该至少一个处理器通信连接的存储器；其中，

[0015] 该存储器存储有可被该至少一个处理器执行的指令，该指令被该至少一个处理器执行，以使该至少一个处理器能够执行本公开任一实施例中的方法。

[0016] 根据本公开的另一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，该计算机指令用于使计算机执行本公开任一实施例中的方法。

[0017] 根据本公开的另一方面，提供了一种计算机程序产品，包括计算机程序/指令，该

计算机程序/指令被处理器执行时实现本公开任一实施例中的方法。

[0018] 根据本公开的技术，能够根据词语的属性和待分析文本的句法依存信息，确定待分析文本中的主体和对主体的描述语，从而有助于对待分析文本进行理解，以从待分析文本中提取出关键的重点信息。

[0019] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0020] 附图用于更好地理解本方案，不构成对本公开的限定。其中：

[0021] 图1是根据本公开一实施例的文本分析方法流程示意图；

[0022] 图2是根据本公开另一实施例的文本分析方法流程示意图；

[0023] 图3是根据本公开又一实施例的文本分析方法流程示意图；

[0024] 图4是根据本公开又一实施例的文本分析方法流程示意图；

[0025] 图5是根据本公开一示例的文本分析方法示意图；

[0026] 图6是根据本公开一实施例的文本分析装置流程示意图；

[0027] 图7是根据本公开另一实施例的文本分析装置流程示意图；

[0028] 图8是根据本公开又一实施例的文本分析装置流程示意图；

[0029] 图9是根据本公开又一实施例的文本分析装置流程示意图；

[0030] 图10是用来实现本公开实施例的文本分析方法的电子设备的框图。

具体实施方式

[0031] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0032] 根据本公开的实施例，提供了一种文本分析方法，图1是根据本公开实施例的基于文本分析方法的流程示意图，该方法可以应用于文本分析装置，例如，该装置可以部署于终端或服务器或其它处理设备执行的情况下，可以执行待分析文本的获取、待分析文本的分析等步骤。其中，终端可以为用户设备(UE,User Equipment)、移动设备、蜂窝电话、无绳电话、个人数字处理(PDA,Personal Digital Assistant)、手持设备、计算设备、车载设备、可穿戴设备等。在一些可能的实现方式中，该方法还可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。如图1所示，文本分析方法包括：

[0033] 步骤S11：对待分析文本的词语进行属性解析，获得属性解析结果；

[0034] 步骤S12：对词语之间的语法依存关系进行解析，获得依存关系解析结果；

[0035] 步骤S13：根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语。

[0036] 本实施例中，待分析文本可以是一段文字或一句文字，还可以是词语的组合。

[0037] 本公开实施例可应用于多种需要对文本进行分析的场景，比如机器阅读、搜索等。

[0038] 在应用于搜索场景的情况下，待分析文本可以是用于搜索查询的文本，可以至少

包括一个或一个以上的词语。对待分析文本的词语进行属性解析，可以包括对待分析的文本进行词语的提取，对提取的词语进行属性解析。本公开实施例中的词语，可以至少包括一个文字最小单位。比如，在待分析文本为中文的情况下，待分析文本中的词语至少包括一个汉字、数字或特殊符号。

[0039] 对待分析文本的词语进行属性解析，可以包括对待分析文本的词语进行词语的某种性质的分析，比如，确定词语的词性或词类。

[0040] 在待分析文本为中文的情况下，确定词语的词性，可以包括确定词语为实词或虚词。确定词语的词性。确定词语的词性，还可以包括确定词语具体为名词、动词、形容词、代词、数词、量词、区别词、副词、介词、连词、拟声词、助词、叹词等中的至少一种。

[0041] 在搜索场景下，待分析文本中还可能包括其它属性，比如，数学符号、阿拉伯数字、希腊字母(α、β等)、其它具有含义的特殊字符、字母或字母组合(比如物品英文首字母缩写、人名首字母缩写)、常见英文单词、常见其它语种单词等。

[0042] 在另一种可能的实现方式中，待分析文本的词语的属性，与待分析文本的语种有关，包括与待分析文本的语种对应的属性。这种情况下，可先确定待分析文本的语种，比如，待分析文本的语种默认为中文，但是如果待分析文本中包含超过设定长度的外文，或者待分析文本中包含了专业性程度较高的外文，则可调整待分析文本的默认语种，进而根据调整后的语种确定待分析文本可能具有的属性。

[0043] 词语之间的语法依存关系，可以是采用句法分析的方法，或者是语言需要遵循语法规则，如“主谓宾”这种句式等句子对应的句法结构。对待分析文本进行分析所获得的词语之间的依存关系结果，即词语构成待分析文本相互之间在语句结构上的关系。具体可以包括：主谓关系、定状语关系、动宾关系、宾语关系等。

[0044] 在一种可能的实现方式中，确定语法依存关系所使用的句法分析方法主要可以包括两方面的内容，一是确定语言的语法体系，即对语言中合法的句子的语法结构给与形式化的定义；另一方面是句法分析方式，即根据给定的语法体系，自动推导出句子的句法结构，分析句子所包含的句法单位和这些句法单位之间的关系。

[0045] 根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语，具体可以包括，根据属性解析结果和依存关系解析结果，对待分析文本中可能是主体的词语进行可能性打分，对待分析文本中可能是主体的描述语的词语进行可能性打分，根据打分，确定主体和主体的描述语。

[0046] 本实施例中，主体可以是待分析文本主要描述的对象。主体的描述语可以是对主体的属性、相关信息等的限定。比如，待分析文本为关于名词A的解释，则待分析文本的主体为A。再如，待分析文本为关于名词A的问句，则A为主体，对A的提问点为主体的描述语。比如，“A的父母”这一语句中，A可以为主体，“父母”可以为针对主体A的描述语。

[0047] 本公开实施例提供的文本分析方法，能够根据词语的属性和待分析文本的句法依存信息，确定待分析文本中的主体和对主体的描述语，从而有助于对待分析文本进行理解，以从待分析文本中提取出关键的重点信息。

[0048] 在一种实施方式中，对待分析文本的词语进行属性解析，获得属性解析结果，如图2所示，包括：

[0049] 步骤S21：确定每个词语的属性；

[0050] 步骤S22:针对每个词语,确定词语在属性下的子分类;

[0051] 步骤S23:将所有词语的属性和子分类,作为属性解析结果。

[0052] 本实施例中,词语在属性下的子分类,可以是在属性基础上的分类。比如,在属性为词性的情况下,若一个具体的词语为主语,子分类可以为主语的具体类别,可以是人物、动物、植物、有机物、无机物、金属等等。若一个具体的词语为谓语,子分类可以为谓语的具体类别,可以是各分支学科的术语、动词、形容词等。

[0053] 属性解析结果可以包括每个词语可能的属性、在可能属性下的至少一个子分类,以及属性的概率、属性下至少一个子分类的概率等。

[0054] 本实施例能够获取词语的属性和子分类,从而有助于根据属性和子分类确定待分析文本中的主体和主体的描述语。

[0055] 在一种实施方式中,对词语之间的语法依存关系进行解析,获得依存关系解析结果,如图3所示,包括:

[0056] 步骤S31:根据待分析文本,获得词语之间的语法依存关系;

[0057] 步骤S32:根据语法依存关系,获得主体候选项和主体的描述语候选项;

[0058] 步骤S33:选择至少一个主体候选项和至少一个主体的描述语候选项,组成主体与主体的描述语组合项;

[0059] 步骤S34:将主体与主体的描述语组合项作为依存关系解析结果。

[0060] 主体候选项可以是待分析文本中可能为主体的候选词。选择至少一个主体候选项和至少一个主体的描述语候选项,组成主体与主体的描述语组合项可以是根据一定的规则,选择一个主体候选项和一个主体的描述语候选项,进行组合。

[0061] 比如,待分析文本包含词语ABCDE,主体候选项为A、B;主体的描述语候选项为B、C、D、E。选择主体候选项A、主体的描述语候选项C,组成主体与主体的描述语组合项:AC。

[0062] 本实施例中,能够确定主体候选项、主体描述语候选项,并将二者组合为主体与主体的描述语组合项,从而在文本分析过程中,不仅能够提供词语之间的句法依存信息,还能够提供词语之间的关系信息,从而能够更为准确地获得主体和主体的描述语句。

[0063] 在一种实施方式中,选择至少一个主体候选项和至少一个主体的描述语候选项,组成主体与主体的描述语组合项,包括:

[0064] 获得由所有主体候选项,结合主体的描述语候选项所组成的组合;

[0065] 将组合项作为主体与主体的描述语组合项。

[0066] 本实施例中,获得由所有主体候选项,结合主体的描述语候选项所组成的组合,可以是将每一个主体候选项与任意一个主体的描述语候选项进行结合,将所有的组合作为依存关系解析结果。比如,主体候选项包括A、B,主题的描述语候选项包括C、D、E,则根据主体候选项的集合和主体的描述语候选项的集合进行组合,获得所有可能的组合:AC、AD、AE、BC、BD、BE,将所有可能的组合作为依存关系解析结果。

[0067] 本是实施中,将主体候选项和主体的描述语候选项的所有组合作为依存关系解析结果,从而有助于得到更为准确的主体和主体的候选项。

[0068] 在一种实施方式中,在待分析文本中包括设定关键词的情况下,根据语法依存关系,获得主体候选项和主体的描述语候选项,包括:

[0069] 在待分析文本中,确定与设定关键词存在预设先后顺序的候选词语;

[0070] 根据候选词语，确定主体候选项或主体的描述语候选项中的至少一个。

[0071] 本实施例中，设定关键词可以是具体的词语，比如词语A为设定关键词。

[0072] 预设先后顺序，可以包括相邻的先后顺序、存在间隔的先后顺序等。比如，若关键词为A，预设先后顺序为A前间隔出现的词语，则除了A前相邻的词语，其余在待分析文本中排列在A前的词语，均为A的预设先后顺序的候选词语。

[0073] 根据候选词语，确定主体候选项或主体的描述语候选项中的至少一个，可以根据候选词语是否为具体某个词语、候选词的属性是否为设定属性或者候选词是否为设定类别的词语等信息，对确定关键词是否为主体或主体的描述语。

[0074] 比如，设定关键词为A，设定顺序为A前相邻词语，判断条件为A前相邻词语为动词的情况下，确定A为主体，则若待分析语句中存在“CA”词语组合，且C为动词，则可认为A为主体。

[0075] 本实施例通过关键词结合设定模式的方式，对主体进行判断，从而能够提高主体判断的准确性。

[0076] 在一种实施方式中，在待分析文本中存在由至少两个设定词性的词语按照预设顺序组合成的词组的情况下，根据语法依存关系，获得主体候选项和主体的描述语候选项，包括：

[0077] 将词组拆分，获得拆分词语；

[0078] 根据拆分词语，确定主体候选项和主体的描述语候选项中的至少一个。

[0079] 本实施例中，至少两个设定词性的词语按照预设顺序组合成的词组，可以是比如动词+名词组合成的词组、名词+动词组合成的词组、名词+形容词组合成的词组、动词+代词组合成的词组、名词+动词+动词组合成的词组等。

[0080] 根据拆分词语，确定主体候选项和主体的描述语候选项中的至少一个，比如可以是根据动词+代词组合成的词组，将拆分后的动词作为主体候选项，或者将拆分后的代词作为主体候选项，或者将拆分后的动词作为主体的描述语候选项，或者将拆分后的代词作为主体的描述语候选项。

[0081] 本实施例中，能够根据设定的词性组成的词组，确定主语候选项或者主语的描述语候选项，从而能够根据语言使用习惯，确定待分析文本的主体和主体的描述语。

[0082] 在一种实施方式中，至少两个设定词性的词语包括设定词性的起始词、和设定词性的终止词，起始词和终止词在待分析文本中的字数距离或词数距离处于设定范围。

[0083] 具体比如，设定词性的起始词为动词、设定词性的终止词为名词，起始词和终止词之间至少包括一个词，则待分析文本中任意位置出现的“起始词+X+终止词”的组合，将可被识别为至少两个设定词性的词语，其中，X代表任意词语。从而，可根据预设的分配规则，将组合中的起始词或终止词判定为主体候选项，或者判定为主体的描述语候选项。

[0084] 本实施例中，通过设定起始词的词性、设定终止词的词性，能够根据用户的常用表达习惯进行主体分析，提高分析的准确性和命中率。

[0085] 在一种实施方式中，在待分析文本中存在实体词的情况下，根据语法依存关系，获得主体候选项和主体的描述语候选项，包括：

[0086] 将实体词作为主体候选项；

[0087] 根据主体候选项和设定模式，确定主体的描述语候选项，设定模式包括主体、主体

的描述语和其它设定词语，以及主体、主体的描述语和其它设定词语之间的相对顺序。

[0088] 本实施例中，其它设定词语可以是其它设定词性的词语，或者是其它具体词语。比如，设定模式可以为“实体+动词+实体的描述语”。则在待分析文本中出现实体词的情况下，将实体词确定为主体候选项，若主体后相邻位置处为动词，动词后相邻位置处的词确定为实体的描述语候选项。

[0089] 再如，设定模式可以为“A+实体+实体的描述语”，则在待分析文本中出现实体词的情况下，将实体词确定为主体候选项，若主体前相邻位置处的词语为A，则将主体后相邻位置的词语确定为主体的描述语候选项。

[0090] 本实施例中，能够根据实体以及与实体词有关的设定模式，确定待分析文本中的主体候选项、主体的描述语候选项，能够借助一般用户语言表达的官场习惯进行主体分析，提高分析的准确性。

[0091] 在一种实施方式中，将实体词作为主体候选项，包括：

[0092] 在待分析文本中包括两个以上顺序衔接的同类实体的情况下，将两个以上顺序衔接的同类实体合并为实体词。

[0093] 本实施例中，两个以上顺序衔接的同类实体，可以包括两个以上实体词分别相邻排列，且两个以上实体词属于同类实体的情况。比如，待分析文本中存在两个以上的实体词均为地名，则可将两个以上地名合并为一个实体词，即在后续文本分析的过程中作为一个实体词看待。

[0094] 本实施例中，通过将相邻分布的两个以上的同类实体词作为一个实体词处理的方式，能够借助一般用户使用语言表达的习惯，提高主体判断的准确性。

[0095] 本公开一种示例中，文本分析方法可应用于搜索场景。提供一套适用于待分析文本的信息抽取的抽取系统，具体可以基于依存句法和词性模板的方法，用于在无标注数据的情况下，将用户输入的Query(查询语句)作为待分析文本，从Query中抽取S(Subject，主语，相当于前述实施例中的主体)和P(Predicate，谓语，相当于前述实施例中的主体的描述语句)的问题。

[0096] 具体而言，针对用户输入的查询语句，确定待分析文本，完成Query(查询语句)抽取任务，从用户搜索时给定的Query中抽取出S和P。例如，图4中所示的搜索页面中，用户在搜索框中输入“F国的首都是什么”。针对这个Query，确定待分析语句为Query本身，通过本公开实施例提供的文本分析方法，从中提取主体为F国，主体的描述语为F国的首都。再如，若用户输入“G明星多高”，通过本公开实施例提供的文本分析方法，从中提取出主体为G明星，主体的描述语为多高。

[0097] 在本公开一种示例中，将文本分析方法应用于搜索场景，流程如图5所示，包括：

[0098] 步骤S51：过滤特定类别的Query。

[0099] 根据Query对应的行业标签、禁用词语标签，和预先定义的要过滤的行业标签列表，对Query进行过滤，过滤非法Query词条。

[0100] 步骤S52：关键词识别以及二分类过滤。

[0101] 在本步骤中，可将过滤任务看成是一个分类任务，即Query包含和不包含S、P。使用预训练分类语言模型对Query进行分类，将预测结果为“不包含S和P的Query”过滤掉。该分类器的训练数据来自于人工标注。由于标注任务简单，只需要标注Query是否包含S和P，这

种非正即负的简单二分类，标注速度极快，可以视作是极低成本的解决了无标注数据的困窘，并且只需要少量标注即可。此外，该模块提供了白名单(比如，白名单可以包含“演员表”之类的词)的功能，即对于包含特定关键词的Query提供了强制性召回策略，只要Query包含白名单中预定义关键词，即传入后续的S、P抽取流程。

[0102] 步骤S53：基于设置的Query-Tag标签对Query进行过滤。

[0103] 本步骤中，可使用序列标注模型，对Query进行解析，得到词性、类别标签和对应的类别标签概率，通过类别标签或者类别标签的组合进行过滤。

[0104] 步骤S54：基于依存句法的抽取。

[0105] 该模型是基于自然语言处理领域广泛使用的依存句法工具抽取获得。使用依存句法工具对Query进行解析，依据内置的词典，获得S和P候选列表，并对其两两组合进行分类，获得最优的SP候选对。

[0106] 其中，内置的词典可以包括可能为主体的所有词、可能为主体的描述语的所有词。根据内置的词典的查询结果，若一词语存在与内置的词典中，则可将该词语直接确定为S候选项，加入S候选列表，或者直接确定为P候选项，加入P候选列表。

[0107] 步骤S55：基于关键词对Query进行抽取。

[0108] 该模块将Query看成是“S+关键词+P”的模式。根据预定义的关键词对Query进行分界，如果Query中存在预设的特定关键词，则在输出结果时，将关键词划归到P中。

[0109] 步骤S56：基于Pattern(设定模板)的抽取。

[0110] 基于Pattern的方法，比如，可以查询Query中是否包含“S+分界词+P”的模式。分界词由设定属性的起始词和设定属性的终止词组成，起始词和终止词中的至少一个可以是特定词性类别的词汇。并且，可设置起始词和终止词之间的步幅限制，起始词和结束词之间允许不超过N个词汇。N就是预定义的步幅。通过步幅对S和P的长度进行模糊处理，增强了泛化性。

[0111] 步骤S57：基于短语块的抽取S。

[0112] 考虑到实际的Query可能涉及复杂短语类的P抽取问题。在Query为中文的情况下，由于中文句式复杂，词汇的语序要求不严格，传统的抽取方式很难解决此类抽取。因此，根据前述步骤得到每个词语对应的类别标签，得到词语所属实体的类别。将类别相同的词语合并，并依据类别过滤出实体，即得到S。

[0113] 本示例中，步骤S54-S57中，任何步骤得出候选S或候选P的情况下，都可以不再采用其它步骤重复执行获取候选S或候选P的步骤，而是直接进入最后的步骤S58。

[0114] 步骤S58：合并抽取结果。可以执行合并标签过滤结果和基于依存句法的抽取结果。在确定S之后，可基于S确定P。

[0115] 在本步骤中，可以使用预先定义的词性模板，对Query的词性序列进行解析，进而得到P。词性模版，借鉴正则匹配的思想，使用模糊匹配的方式，将多个能归类到统一范式的模板合并成一个，从而增强模板的泛化性，并减少模板的数量。

[0116] 本公开示例提供的文本分析方法，采用了基于词性的序列标注方案，并且对模板方法进行改进，引入模糊匹配的方法，简单且非常有效。

[0117] 同时，本公开示例提供的文本分析方法的人工成本低。成本低体现在两个方面：一，本文摒弃了构建词典或深度学习的序列标注这种耗费人力的求解方式，而是将问题的

求解思路转化为Query分类和模板匹配问题。在运用预训练分类模型时，并不需要大量训练数据，仅需人工标注少量数据，且是标注是否的这种二维简单任务，极大减轻人力成本。二，使用模糊匹配的方法，在保证抽取效果的同时，能有效减少模板数量，减少维护成本。

[0118] 此外，本公开示例提供的文本分析方法的泛化能力强。泛化性强体现在两个方面：一，对不包含S和P的Query的过滤能力。用户输入的内容是无限的，传统的解决思路在实际应用时，会存在极大概率的错误抽取。本公开示例通过引入大规模预训练语言模型，将非法Query过滤，简化抽取难度。二，对复杂中文句式的超强抽取能力。中文句式比较随意，传统的词典式抽取，需要维护极大数量的词典集合，泛化性通过数量来堆叠。词汇是无穷变化的，但词性却是有固定范式的。本公开示例跳脱于词汇的思路，从词性这种更高维度求解，通过设计简单的词性模板，有效解决复杂中文句式的抽取问题。

[0119] 本公开实施例还提供一种文本分析装置，如图6所示，包括：

[0120] 属性解析结果获得模块61，用于对待分析文本的词语进行属性解析，获得属性解析结果；

[0121] 依存关系解析结果获得模块62，用于对词语之间的语法依存关系进行解析，获得依存关系解析结果；

[0122] 分析结果模块63，用于根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语。

[0123] 在一种实施方式中，如图7所示，属性解析结果获得模块包括：

[0124] 属性确定单元71，用于确定每个词语的属性；

[0125] 子分类确定单元72，用于针对每个词语，确定词语在属性下的子分类；

[0126] 结果单元73，用于将所有词语的属性和子分类，作为属性解析结果。

[0127] 在一种实施方式中，如图8所示，依存关系解析结果获得模块包括：

[0128] 语法依存关系获得单元81，用于根据待分析文本，获得词语之间的语法依存关系；

[0129] 候选项获得单元82，用于根据语法依存关系，获得主体候选项和主体的描述语候选项；

[0130] 组合项组成单元83，用于选择至少一个主体候选项和至少一个主体的描述语候选项，组成主体与主体的描述语组合项；

[0131] 依存关系解析结果单元84，用于将主体与主体的描述语组合项作为依存关系解析结果。

[0132] 在一种实施方式中，组合项组成单元还用于：

[0133] 获得由所有主体候选项，结合主体的描述语候选项所组成的组合；

[0134] 将组合项作为主体与主体的描述语组合项。

[0135] 在一种实施方式中，在待分析文本中包括设定关键词的情况下，候选项获得单元还用于：

[0136] 在待分析文本中，确定与设定关键词存在预设先后顺序的候选词语；

[0137] 根据候选词语，确定主体候选项或主体的描述语候选项中的至少一个。

[0138] 在一种实施方式中，在待分析文本中存在由至少两个设定词性的词语按照预设顺序组合成的词组的情况下，候选项获得单元还用于：

[0139] 将词组拆分，获得拆分词语；

[0140] 根据拆分词语，确定主体候选项和主体的描述语候选项中的至少一个。

[0141] 在一种实施方式中，至少两个设定词性的词语包括设定词性的起始词、和设定词性的终止词，起始词和终止词在待分析文本中的字数距离或词数距离处于设定范围。

[0142] 在一种实施方式中，在待分析文本中存在实体词的情况下，候选项获得单元还用于：

[0143] 将实体词作为主体候选项；

[0144] 根据主体候选项和设定模式，确定主体的描述语候选项，设定模式包括主体、主体的描述语和其它设定词语，以及主体、主体的描述语和其它设定词语之间的相对顺序。

[0145] 在一种实施方式中，候选项获得单元还用于：

[0146] 在待分析文本中包括两个以上顺序衔接的同类实体的情况下，将两个以上顺序衔接的同类实体合并为实体词。

[0147] 在本公开一种具体示例中，如图9所示，文本分析的整个系统架构包括：

[0148] 数据过滤模块91：用于过滤非法数据，如禁用词条、特定行业(比如风险较高的医疗行业等)的Query；

[0149] 基于依存句法的抽取模块92：使用依存句法工具对Query进行解析，从中提取S和P的候选，并且对S和P的候选对进行分类判定，优选出最佳S和P词条；

[0150] 基于词性的抽取模块93：使用词性标注工具对Query进行标注，获得Query的词性，通过词性模板，获得S和P；

[0151] 抽取结果合并94：将基于依存句法的抽取结果与基于词性的抽取结果进行合并，输出得到最终的抽取结果。

[0152] 本公开的技术方案中，所涉及的用户个人信息的获取，存储和应用等，均符合相关法律法规的规定，且不违背公序良俗。

[0153] 根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0154] 图10示出了可以用来实施本公开的实施例的示例电子设备1000的示意性框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0155] 如10图10所示，设备1000包括计算单元1001，其可以根据存储在只读存储器(ROM)1002中的计算机程序或者从存储单元1010加载到随机访问存储器(RAM)1003中的计算机程序，来执行各种适当的动作和处理。在RAM 1003中，还可存储设备1000操作所需的各种程序和数据。计算单元1001、ROM 1002以及RAM 1003通过总线1004彼此相连。输入/输出(I/O)接口1005也连接至总线1004。

[0156] 设备1000中的多个部件连接至I/O接口1005，包括：输入单元1006，例如键盘、鼠标等；输出单元1007，例如各种类型的显示器、扬声器等；存储单元10010，例如磁盘、光盘等；以及通信单元1009，例如网卡、调制解调器、无线通信收发机等。通信单元1009允许设备1000通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0157] 计算单元1001可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元1001的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元1001执行上文所描述的各个方法和处理,例如文本分析方法。例如,在一些实施例中,文本分析方法可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元10010。在一些实施例中,计算机程序的部分或者全部可以经由ROM1002和/或通信单元1009而被载入和/或安装到设备1000上。当计算机程序加载到RAM 1003并由计算单元1001执行时,可以执行上文描述的文本分析方法的一个或多个步骤。备选地,在其他实施例中,计算单元1001可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行文本分析方法。

[0158] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0159] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器,使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0160] 在本公开的上下文中,机器可读介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0161] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0162] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算

系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0163] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0164] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发公开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。

[0165] 上述具体实施方式,并不构成对本公开保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等,均应包含在本公开保护范围之内。

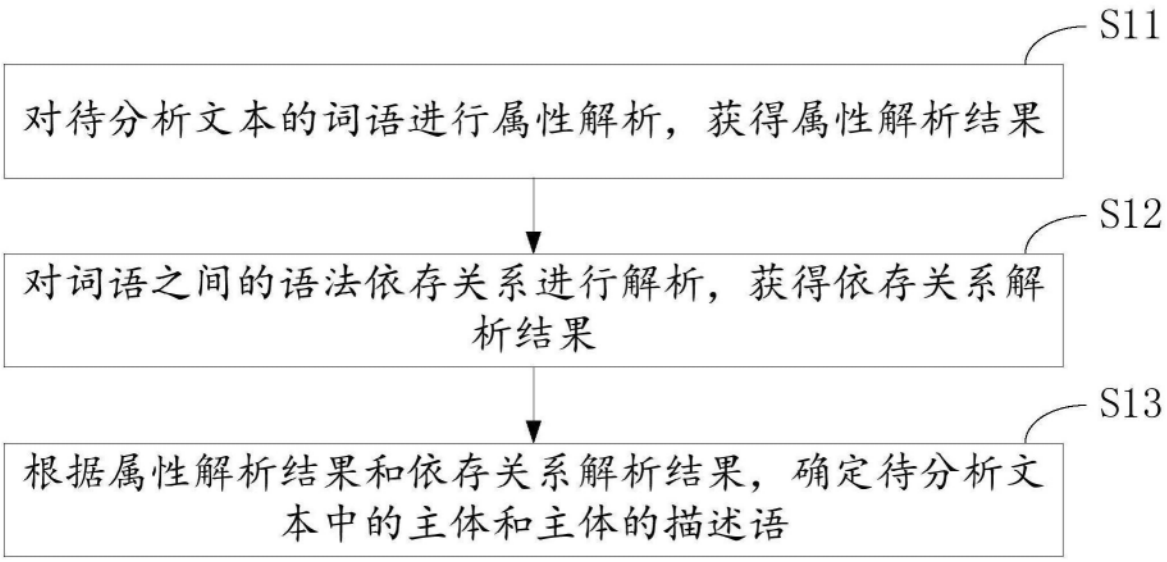

S11
对待分析文本的词语进行属性解析，获得属性解析结果
S12
对词语之间的语法依存关系进行解析，获得依存关系解析结果
S13
根据属性解析结果和依存关系解析结果，确定待分析文本中的主体和主体的描述语

图1

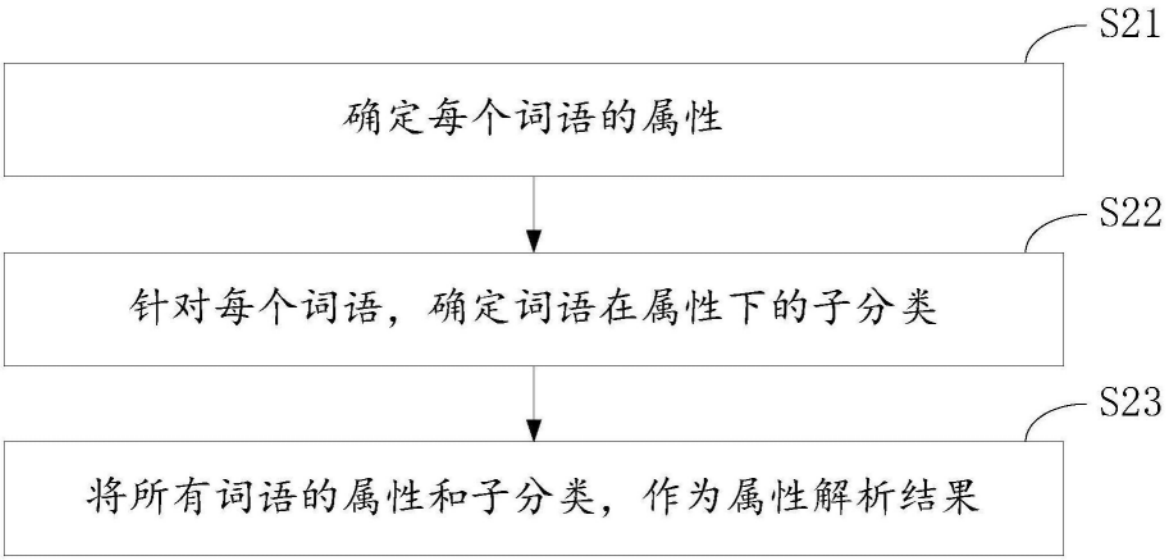

S21
确定每个词语的属性
S22
针对每个词语，确定词语在属性下的子分类
S23
将所有词语的属性和子分类，作为属性解析结果

图2

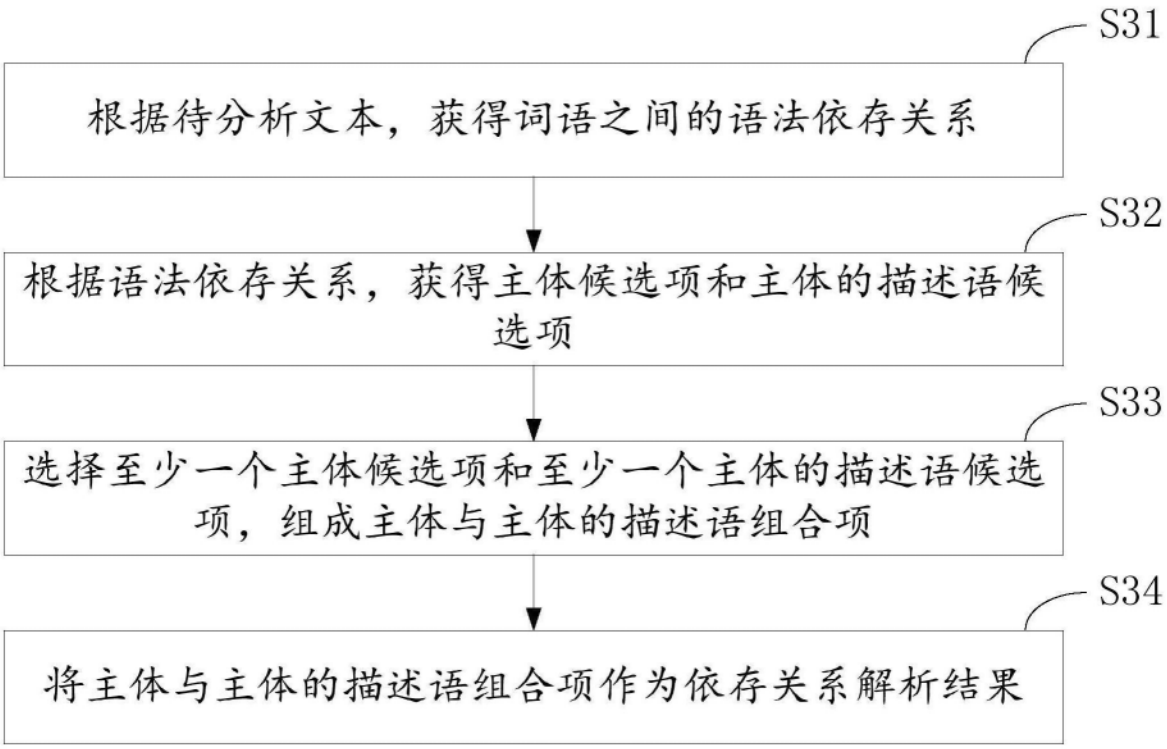

S31
根据待分析文本，获得词语之间的语法依存关系
S32
根据语法依存关系，获得主体候选项和主体的描述语候选项
S33
选择至少一个主体候选项和至少一个主体的描述语候选项，组成主体与主体的描述语组合项
S34
将主体与主体的描述语组合项作为依存关系解析结果

图3

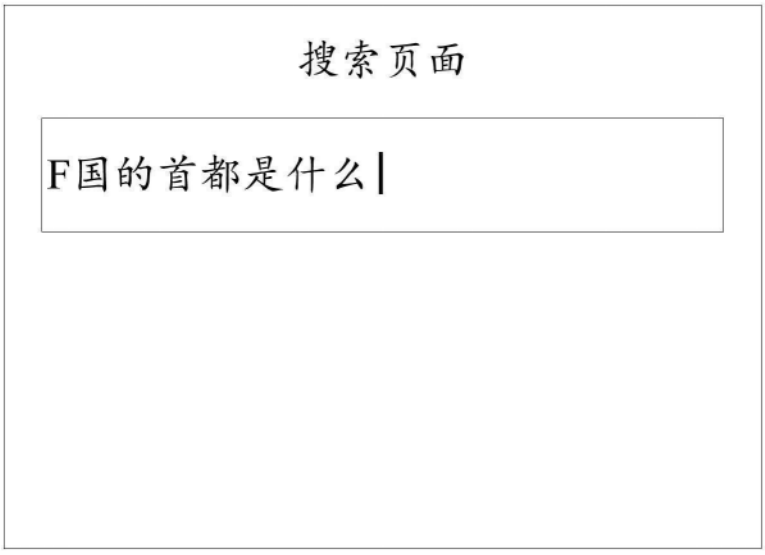

搜索页面
F国的首都是什么

图4

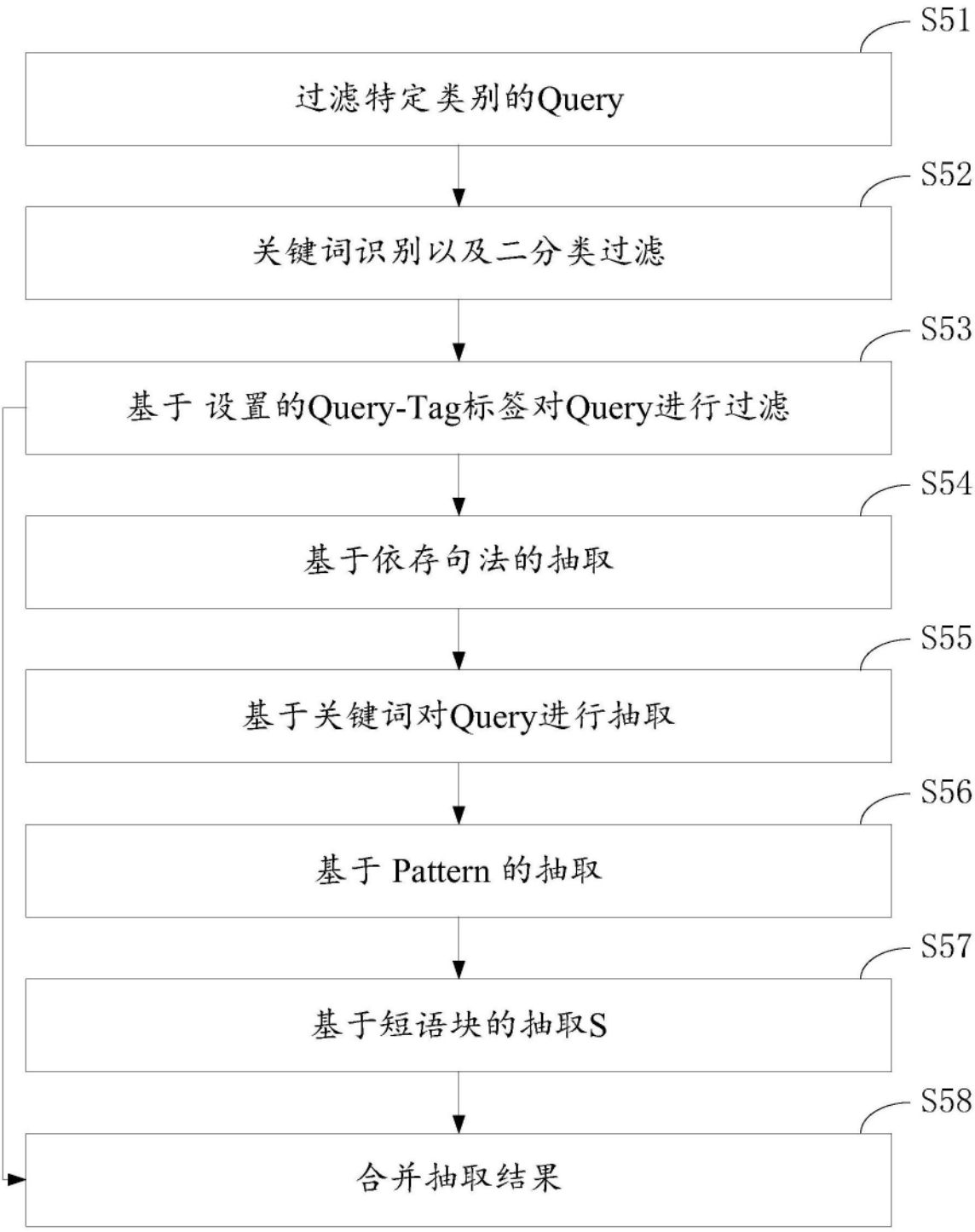

S51
过滤特定类别的Query
S52
关键词识别以及二分类过滤
S53
基于 设置的Query-Tag标签对Query进行过滤
S54
基于依存句法的抽取
S55
基于关键词对Query进行抽取
S56
基于 Pattern 的抽取
S57
基于短语块的抽取S
S58
合并抽取结果

图5

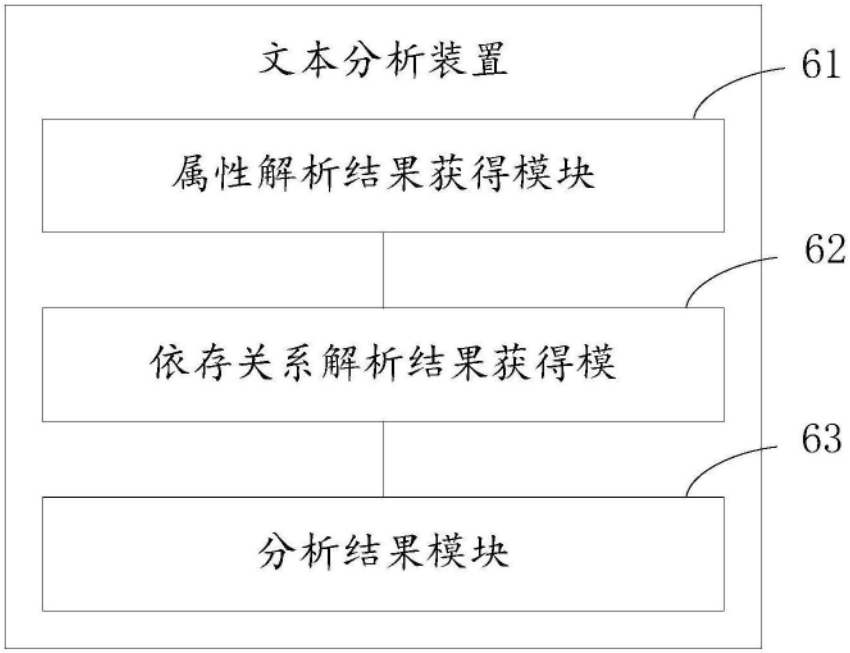

文本分析装置
61
属性解析结果获得模块
62
依存关系解析结果获得模
63
分析结果模块

图6

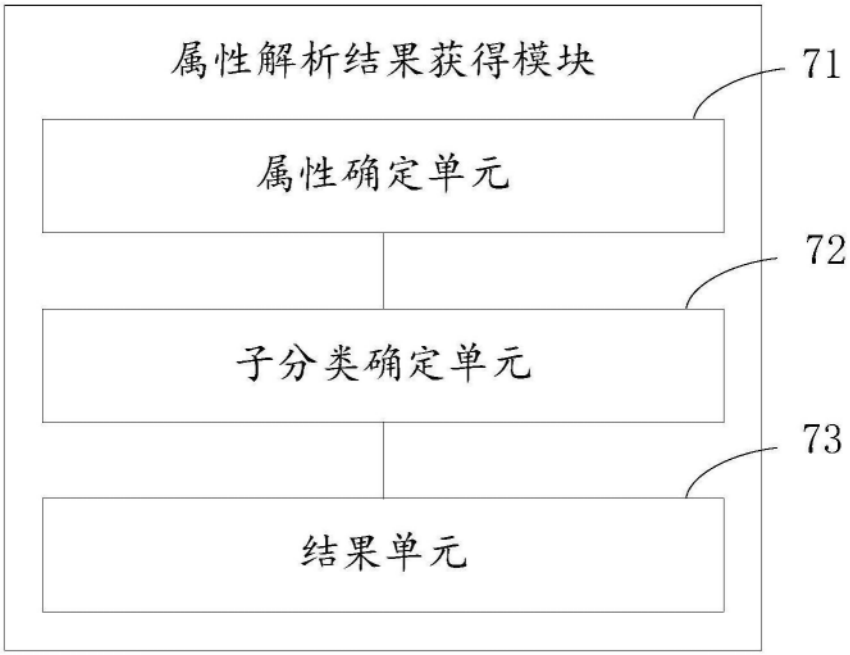

属性解析结果获得模块
71
属性确定单元
72
子分类确定单元
73
结果单元

图7

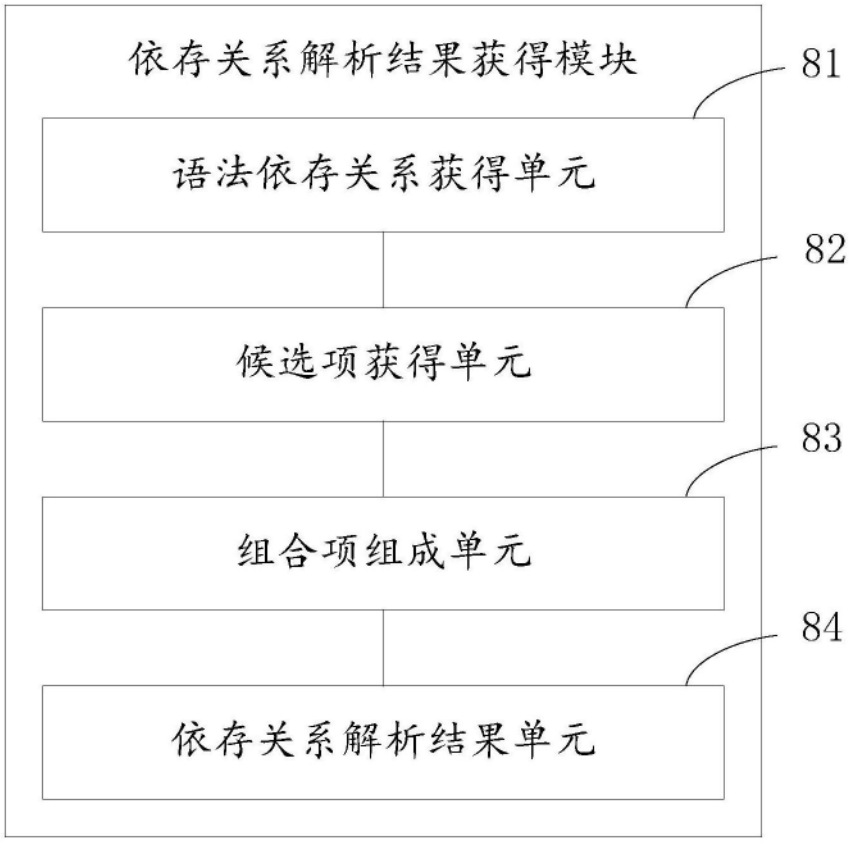

依存关系解析结果获得模块
81
语法依存关系获得单元
82
候选项获得单元
83
组合项组成单元
84
依存关系解析结果单元

图8

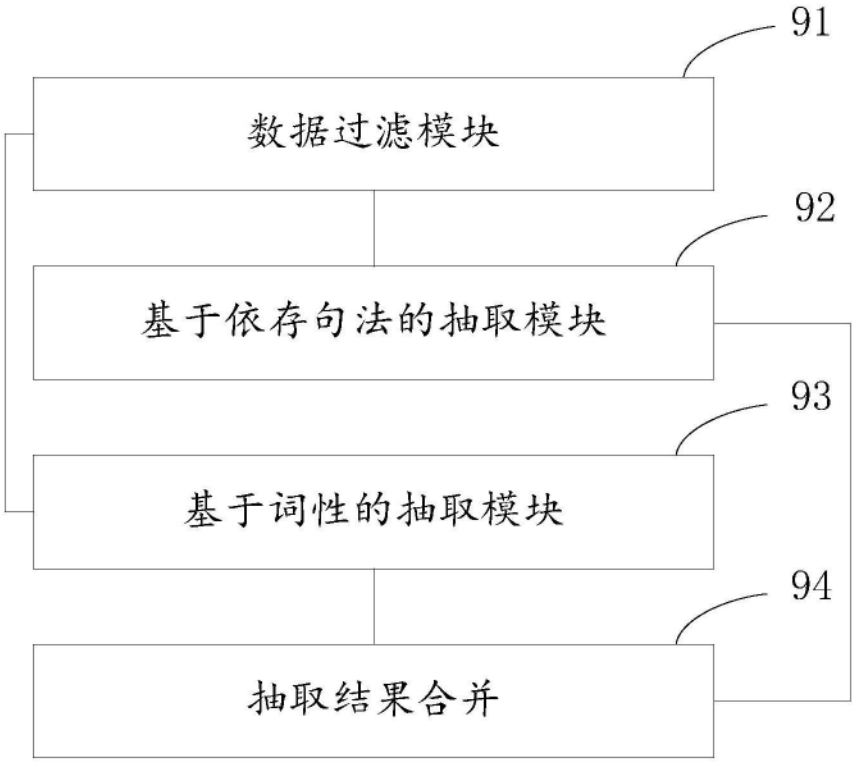

91
数据过滤模块
92
基于依存句法的抽取模块
93
基于词性的抽取模块
94
抽取结果合并

图9

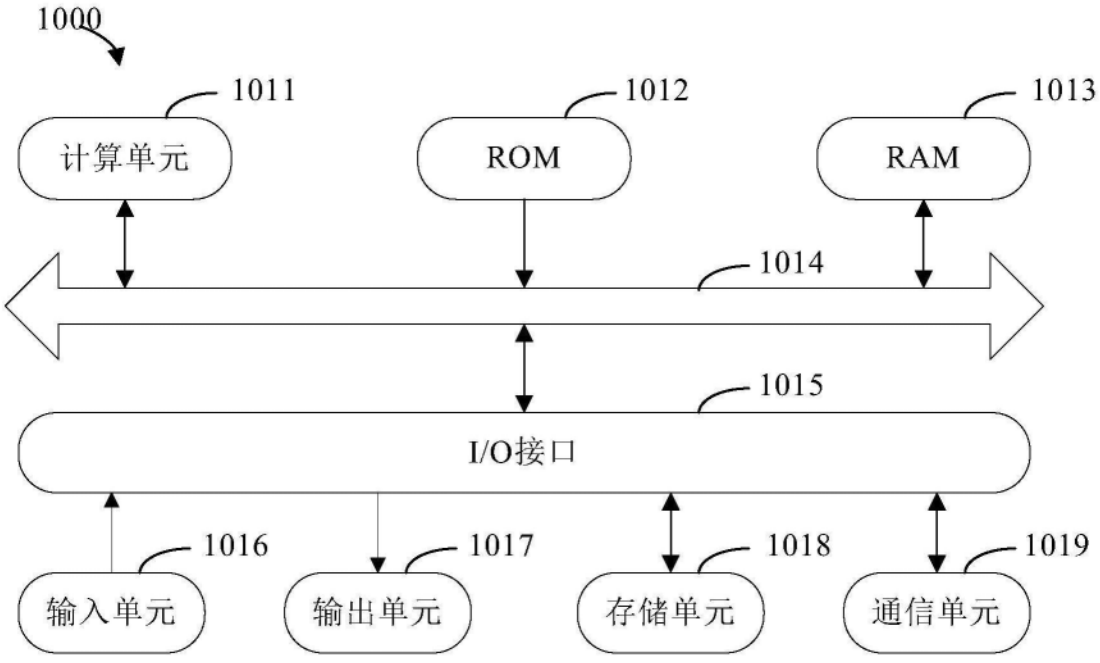

1000
1011
计算单元
1012
ROM
1013
RAM
1014
1015
I/O接口
1016
输入单元
1017
输出单元
1018
存储单元
1019
通信单元

图10