(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112507688 A
(43)申请公布日 2021.03.16

(21)申请号 202011488930.8

(22)申请日 2020.12.16

(71)申请人 咪咕数字传媒有限公司
地址 310000 浙江省杭州市西湖区文二西路820号2幢102室
申请人 咪咕文化科技有限公司
中国移动通信集团有限公司

(72)发明人 徐欣辰

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 黄灿 尹倩

(51)Int.Cl.
G06F 40/194(2020.01)
G06F 40/253(2020.01)
G06F 40/279(2020.01)
G06F 40/30(2020.01)

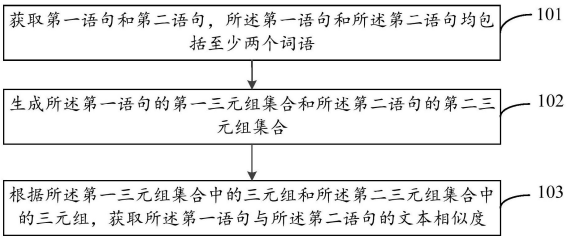
权利要求书2页 说明书16页 附图2页

(54)发明名称

文本相似度分析方法、装置、电子设备及可读存储介质

(57)摘要

本申请提供一种文本相似度分析方法、装置、电子设备及可读存储介质，涉及信息分析技术领域。所述文本相似度分析方法包括：获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。本申请提供的技术方案能够解决现有技术中对语句相似度的分析结果准确性较低的问题。



CN 112507688 A

1.一种文本相似度分析方法,其特征在于,包括:

获取第一语句和第二语句,所述第一语句和所述第二语句均包括至少两个词语;

生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合,其中,所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组,所述三元组包括两个词语及所述两个词语之间的语法关系;

根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组,获取所述第一语句与所述第二语句的文本相似度。

2.根据权利要求1所述的方法,其特征在于,所述第一三元组集合中的三元组为第一三元组,所述第二三元组集合中的三元组为第二三元组;

所述根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组,获取所述第一语句与所述第二语句的文本相似度,包括:

将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合,获得多个配对三元组;所述配对三元组包括一个第一三元组和一个第二三元组;

获取每个配对三元组的相似度值;

基于所述每个配对三元组的相似度值,获取所述第一语句与所述第二语句的文本相似度。

3.根据权利要求2所述的方法,其特征在于,所述获取每个配对三元组的相似度值,包括:

基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语,获取所述每个配对三元组的词语匹配得分;

基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系,获取所述每个配对三元组的语法关系匹配得分;

基于所述词语匹配得分和所述语法关系匹配得分,计算所述每个配对三元组的相似度值。

4.根据权利要求3所述的方法,其特征在于,所述每个配对三元组均包括第一配对词语和第二配对词语,所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个,所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个;

所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语,获取所述每个配对三元组的词语匹配得分,包括:

基于词向量的余弦相似度算法,获取每个配对三元组中第一配对词语的第一分值和第二配对词语的第二分值;

对所述第一分值和所述第二分值进行加权求和计算,获取所述每个配对三元组的词语匹配得分。

5.根据权利要求3所述的方法,其特征在于,所述每个配对三元组均包括第三配对词语,所述第三配对词语包括第一短语和第二短语,所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语,所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语;

所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分，包括：

基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

6.根据权利要求2所述的方法，其特征在于，所述基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度，包括：

获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

确定每一个所述第一三元组对应的目标配对三元组；

基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

7.根据权利要求6所述的方法，其特征在于，所述基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度，包括：

获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

8.一种文本相似度分析装置，其特征在于，包括：

第一获取模块，用于获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

生成模块，用于生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

第二获取模块，用于根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

9.一种电子设备，其特征在于，包括处理器、存储器及存储在所述存储器上并可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求1至7中任一项所述的文本相似度分析方法的步骤。

10.一种可读存储介质，其特征在于，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如权利要求1至7中任一项所述的文本相似度分析方法的步骤。

文本相似度分析方法、装置、电子设备及可读存储介质

技术领域

[0001] 本申请涉及信息分析技术领域，具体涉及一种文本相似度分析方法、装置、电子设备及可读存储介质。

背景技术

[0002] 句子作为在词语之上、段落之下的结构形式在语言处理的各项工作中都扮演着重要角色，而对于句子的相似性分析也逐渐成为文本研究的重要方向之一。目前通常基于词的层面来分析两个句子之间是否相似，具体方式为，寻找句子中每个词在另一个句子中语义相近的词，并基于这些语义相近的词来计算两个句子之间的相似度，用以判断两者是否相似。但由于句子语义的复杂性，这种基于词的层面对于两个句子是否相似的分析结果，通常准确性较低。

发明内容

[0003] 本申请实施例提供一种文本相似度分析方法、装置、电子设备及可读存储介质，能够解决现有技术中对语句相似度的分析结果准确性较低的问题。

[0004] 为了解决上述技术问题，本申请是这样实现的：

[0005] 第一方面，本申请实施例提供了一种文本相似度分析方法，包括：

[0006] 获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

[0007] 生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

[0008] 根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0009] 可选地，所述第一三元组集合中的三元组为第一三元组，所述第二三元组集合中的三元组为第二三元组；

[0010] 所述根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度，包括：

[0011] 将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合，获得多个配对三元组；所述配对三元组包括一个第一三元组和一个第二三元组；

[0012] 获取每个配对三元组的相似度值；

[0013] 基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度。

[0014] 可选地，所述获取每个配对三元组的相似度值，包括：

[0015] 基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分；

[0016] 基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系，获取所述每个配对三元组的语法关系匹配得分；

[0017] 基于所述词语匹配得分和所述语法关系匹配得分，计算所述每个配对三元组的相似度值。

[0018] 可选地，所述每个配对三元组均包括第一配对词语和第二配对词语，所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个，所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个；

[0019] 所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分，包括：

[0020] 基于词向量的余弦相似度算法，获取每个配对三元组中第一配对词语的第一分值和第二配对词语的第二分值；

[0021] 对所述第一分值和所述第二分值进行加权求和计算，获取所述每个配对三元组的词语匹配得分。

[0022] 可选地，所述每个配对三元组均包括第三配对词语，所述第三配对词语包括第一短语和第二短语，所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语，所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语；

[0023] 所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分，包括：

[0024] 基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

[0025] 可选地，所述基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度，包括：

[0026] 获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

[0027] 确定每一个所述第一三元组对应的目标配对三元组；

[0028] 基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

[0029] 基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

[0030] 可选地，所述基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度，包括：

[0031] 获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

[0032] 基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

[0033] 基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

[0034] 第二方面，本申请实施例提供了一种文本相似度分析装置，包括：

[0035] 第一获取模块，用于获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

[0036] 生成模块，用于生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

[0037] 第二获取模块，用于根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0038] 第三方面，本申请实施例提供了一种电子设备，包括处理器、存储器及存储在所述存储器上并可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第一方面中所述的文本相似度分析方法的步骤。

[0039] 第四方面，本申请实施例提供了一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如第一方面所述的文本相似度分析方法的步骤。

[0040] 第五方面，本申请实施例提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如第一方面所述的文本相似度分析方法。

[0041] 本申请实施例中，通过生成第一语句的第一三元组集合和第二语句的第二三元组集合，所述三元组包括两个词语及所述两个词语之间的语法关系，因而在根据所述第一三元组集合和所述第二三元组集合获取第一语句和第二语句的文本相似度时，不仅只是获取第一语句中词语的相似度，还包括获取词语之间的语法关系的相似度，基于语法关系能够更好地考虑到句子的语义，这样也就更进一步提高了对两个语句之间相似度分析的准确性。

附图说明

[0042] 图1是本申请实施例提供的一种文本相似度分析方法的流程图；

[0043] 图1a是本申请实施例提供的另一种文本相似度分析方法的流程图；

[0044] 图2是本申请实施例提供的一种文本相似度分析装置的结构图；

[0045] 图3是本申请实施例提供的一种电子设备的结构图。

具体实施方式

[0046] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0047] 本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”等所区分的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”，一般表示前后关联对象是一种“或”的关系。

[0048] 下面结合附图，通过具体的实施例及其应用场景对本申请实施例提供的文本相似

度分析方法、装置及电子设备进行详细地说明。

[0049] 请参见图1,图1是本申请实施例提供的一种文本相似度分析方法的流程图,所述文本相似度分析方法可以是应用于计算机、平板电脑、手机等电子设备。

[0050] 如图1所示,所述文本相似度分析方法包括以下步骤:

[0051] 步骤101、获取第一语句和第二语句,所述第一语句和所述第二语句均包括至少两个词语。

[0052] 本申请实施例中,所述第一语句和所述第二语句可以是同一文本中的两个语句,或者也可以是不同文本中的两个语句。例如,电子设备在显示某小说文本的情况下,若电子设备接收到用户的文本相似度分析指令,则可以是随机选取当前显示的小说文本中的两个语句分别作为第一语句和第二语句。或者,也可以是用户选择第一语句和第二语句输入特定的文本相似度分析程序中,进而电子设备获取到第一语句和第二语句。当然,所述第一语句和第二语句的获取还可以是其他方式,本实施例不做具体限定。

[0053] 其中,所述第一语句和第二语句均包括至少两个词语,例如,所述第一语句为“第三中学排球队击败了第四中学排球队”,则第一语句可以是进行分词处理,拆分成包括“击败”、“第三中学排球队”、“第四中学排球队”等词语。

[0054] 步骤102、生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合。

[0055] 其中,第一三元组集合中包括至少一个三元组,后续将第一三元组集合中的三元组称之为第一三元组;并且第二三元组集合中也包括至少一个三元组,后续将第二三元组集合中的三元组称之为第二三元组。其中,每个三元组中均包括两个词语及这两个词语之间的语法关系。也就是说,每一个第一三元组包括两个词语及这两个词语之间的语法关系,每一个第二三元组也包括两个词语及这两个词语之间的语法关系。

[0056] 在生成第一语句的第一三元组集合时,可以先将该第一语句拆分成多个词语,然后将这些词语各自之间进行两两组合,并添加两两组合的词语之间的语法关系,从而得到多个第一三元组,然后通过这些第一三元组得到该第一三元组集合。同理,通过相同的方式,可以得到第二语句相应的第二三元组集合。

[0057] 例如,第一语句为“第三中学排球队击败第四中学排球队”,可以将该第一语句拆分为“第三中学排球队”、“击败”、“第四中学排球队”三个词语,通过这三个词语之间的两两组合,则可以生成多个第一三元组,例如第一个第一三元组R1=(第三中学排球队,主谓关系,击败),第二个第一三元组R2=(第四中学排球队,动宾关系,击败),第三个第一三元组R3=(第三中学排球队,主语和宾语,第四中学排球),第一三元组集合也就至少包括R1、R2、R3。

[0058] 当然,对于每个三元组中所包括的两个词语以及语法关系,这里并不对它们之间的排列顺序进行限定。比如上述的R1还可以为R1=(第三中学排球队,击败,主谓关系),或者为R1=(击败,第三中学排球队,主谓关系)。

[0059] 假设第二语句为“第四中学排球队击败第三中学排球队”,同样也可以是拆分成包括“击败”、“第三中学排球队”、“第四中学排球队”等词语,基于这些词语生成多个第二三元组,例如第一个第二三元组R1’=(第三中学排球队,动宾关系,击败),第二个第二三元组R2’=(第四中学排球队,主谓关系,击败),第三个第二三元组R3’=(第四中学排球队,主语

和宾语，第三中学排球)，第二三元组集合也就至少包括R1’、R2’、R3’。

[0060] 这样，通过获取第一语句和第二语句，对第一语句和第二语句分别进行分词处理，并对分词处理后得到的词语进行组合，也就能够获得包括至少一个第一三元组的第一三元组集合，及包括至少一个第二三元组的第二三元组集合。

[0061] 步骤103、根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0062] 可选地，将所述第一三元组集合所包括的第一三元组与第二三元组集合所包括的第二三元组进行文本相似度分析，以获得第一语句与第二语句的文本相似度。

[0063] 例如，可以是将第一三元组集合中的第一个第一三元组与第二三元组集合中的每一个第二三元组一一比对，进行文本相似度分析，获取相似度值；而后将第一三元组集合中的第二个第一三元组与第二三元组集合中的每一个第二三元组一一比对，进行文本相似度分析，获取相似度值……依此获得每一个第一三元组与每一个第二三元组的相似度值，进而对获得的所有相似度值就可以是通过加权平均算法来计算第一语句和第二语句的相似度值，进而以获得第一语句和第二语句的文本相似度。

[0064] 本申请实施例中，通过生成第一语句的第一三元组集合和第二语句的第二三元组集合，所述三元组包括两个词语及所述两个词语之间的语法关系，因而在对所述第一三元组集合和所述第二三元组集合进行文本相似度分析时，不仅只是对第一语句中词语的相似度进行分析，还包括对词语之间的语法关系的相似度的分析，基于语法关系能够更好地考虑到句子的语义，这样也就更进一步提高了对两个语句之间相似度分析的准确度。

[0065] 可选地，所述步骤103可以包括：

[0066] 将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合，获得多个配对三元组；所述配对三元组包括一个第一三元组和一个第二三元组；

[0067] 获取每个配对三元组的相似度值；

[0068] 基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度。

[0069] 例如，所述第一三元组集合包括三个第一三元组R1、R2、R3，第二三元组集合包括三个第二三元组R1’、R2’、R3’，则将上述三个第一三元组与三个第二三元组一一配对，得到多个配对三元组：R1R1’、R1R2’、R1R3’、R2R1’、R2R2’、R2R3’、R3R1’、R3R2’、R3R3’，这样也就使得获得的配对三元组更为全面，更有利于对第一语句和第二语句相似度的分析。进一步地，获取每一个配对三元组的相似度值，基于每个配对三元组的相似度值来获取第一语句和第二语句的文本相似度。

[0070] 可以理解地，每一个配对三元组包括第一语句的一个第一三元组和第二语句的第一第二三元组，而每一个三元组包括两个词语和这两个词语之间的语法关系，进而一个配对三元组的相似度值也就能够在一定程度上反映第一语句和第二语句之间的文本相似度。本申请实施例中，多个配对三元组是基于将第一三元组集合中的每一个第一三元组与第二三元组集合中的每一个第二三元组进行组合而获得，这样的方式获得的多个配对三元组也就更为全面地覆盖了第一语句的词语和第二语句的词语之间的组合方式，使得对第一语句和第二语句文本相似度的分析更为精确，能够有效提高第一语句与第二语句之间文本相似

度分析的准确性。

[0071] 可选地，所述获取每个配对三元组的相似度值，包括：

[0072] 基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分；

[0073] 基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系，获取所述每个配对三元组的语法关系匹配得分；

[0074] 基于所述词语匹配得分和所述语法关系匹配得分，计算所述每个配对三元组的相似度值。

[0075] 可以理解地，每个配对三元组包括一个第一三元组和一个第二三元组，第一三元组包括两个词语和这两个词语之间的语法关系，第二三元组也包括两个词语和两个词语之间的语法关系，则可以是分别计算两个三元组中的词语匹配得分以及语法关系匹配得分，基于所述词语匹配得分和语法关系匹配得分，也就能够获得一个配对三元组的相似度值。

[0076] 需要说明的是，第一三元组中包括两个词语，第二三元组也包括两个词语，一个配对三元组也就包括四个词语，则可以是通过不同的组合方式来对这四个词语进行匹配，以计算第一三元组和第二三元组之间的词语匹配得分。

[0077] 可选地，在一种实施方式中，所述每个配对三元组均包括第一配对词语和第二配对词语，所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个，所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个；

[0078] 所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分，包括：

[0079] 基于词向量的余弦相似度算法，获取每个配对三元组中第一配对词语的第一分值和第二配对词语的第二分值；

[0080] 对所述第一分值和所述第二分值进行加权求和计算，获取所述每个配对三元组的词语匹配得分。

[0081] 例如，第一语句为“第三中学排球队击败第四中学排球队”，第一三元组为R1＝(第三中学排球队，主谓关系，击败)；第二语句为“第四中学排球队击败第三中学排球队”，第二三元组R1’＝(第三中学排球队，动宾关系，击败)，得到一个配对三元组R1R1’；那么所述第一配对词语slot1＝(第三中学排球队，第三中学排球队)，第二配对词语slot2＝(击败，击败)；或者slot1＝(第三中学排球队，击败)，slot2＝(击败，第三中学排球队)；基于词向量的余弦相似度算法，计算第一配对词语slot1的第一分值和第二配对词语slot2的第二分值，对第一分值和第二分值进行加权求和计算，进而获取到该配对三元组的词语匹配得分。

[0082] 具体地，基于上述配对三元组R1R1’中，第一配对词语slot1＝(第三中学排球队，第三中学排球队)，第二配对词语slot2＝(击败，击败)，计算第一配对词语slot1的第一分值$score_{slot1}$：

[0083] $score_{slot1}=sim(slot_{R11},slot_{R11'})$，

[0084] 其中，$slot_{R11}$为第一三元组中的第一词语(例如上述slot1中的第三中学排球队)，$slot_{R11'}$为第二三元组中的第三词语或第四词语(例如上述slot1中的第三中学排球队)；

[0085] 计算第二配对词语slot2的第二分值$score_{slot2}$：

[0086] $score_{slot2}=sim(slot_{R12},slot_{R12}')$，

[0087] 其中，$slot_{R12}$为第一三元组中的第二词语(例如上述slot2中的击败)，$slot_{R12'}$为第二三元组中的第三词语或第四词语(例如上述slot2中的击败)；

[0088] 对第一分值和第二分值进行加权求和计算，获得配对三元组的词语匹配得分$score_{word}$：

[0089] $score_{word}=a\times score_{slot1}+(1-a)\times score_{slot2}$；

[0090] 其中，a为预设加权值，0<a<1。可选地，所述预设加权值可以是用户预先设定。

[0091] 这样，通过上述方式也就能够计算出每一个配对三元组的词语匹配得分。

[0092] 在该实施方式中，通过将第一三元组中的一个词语与第二三元组中的一个词语进行组合得到第一配对词语和第二配对词语，通过词向量的余弦相似度算法分别计算两个配对词语的分值，也就能够得到两个配对词语的相似度得分，这样也就能够获得配对三元组所包括的第一三元组和第二三元组两组词语的相似度得分，进而也就使得对于第一语句和第二语句中词语的相似度计算更为精细，能够提高对于两个语句文本相似度的准确度。

[0093] 或者，在另一种实施方式中，所述每个配对三元组均包括第三配对词语，所述第三配对词语包括第一短语和第二短语，所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语，所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语；

[0094] 所述基于每个配对三元组中所述第一三元组中的两个词语和所述第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分，包括：

[0095] 基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

[0096] 例如，第一语句为“第三中学排球队击败第四中学排球队”，第一三元组为R1＝(第三中学排球队，主谓关系，击败)；第二语句为“第四中学排球队击败第三中学排球队”，第二三元组R1’＝(第三中学排球队，动宾关系，击败)，得到一个配对三元组R1R1’；那么得到第一短语$slotpair_{R1}$＝(第三中学排球队，击败)，第二短语$slotpair_{R1'}$＝(第三中学排球队，击败)，第三配对词语也就包括第一短语和第二短语；进一步地，基于词向量的余弦相似度算法，获取第三配对词语的第三分值，所述第三分值也就为对应的配对三元组的词语匹配得分。

[0097] 具体地，在上述配对三元组R1R1’中，第三配对词语包括$slotpair_{R1}$和$slotpair_{R1'}$，该配对三元组的词语匹配得分$score_{word}$也即对第三配对词语进行计算获得的第三分值$score_{pair}$，所述第三分值$score_{pair}$通过如下方式计算：

[0098] $score_{word}=score_{pair}=sim(slotpair_{R1},slotpair_{R1'})$。

[0099] 这样，通过上述方式也就能够计算出每一个配对三元组的词语匹配得分。

[0100] 在该实施方式中，通过将第一三元组中的两个词语组合成第一短语，将第二三元组中的两个词语组合成第二短语，而后计算第一短语和第二短语之间的相似度来获得配对三元组的词语匹配得分。这样也就能够通过另一种词语组合方式，来对第一三元组和第二三元组的词语进行组合，以获得配对三元组的词语匹配得分，以提高对两个语句文本相似度分析的准确度。

[0101] 可选地，本申请实施例中，可以采用word2vec的词向量训练方式用于获得计算词

语匹配的词向量。word2vec由Cbow和skip-gram两种训练模型组成，其训练方式都在于通过统计句子中前后词语的共现概率来获得低维的词向量。word2vec的训练模型都有输入层、隐藏层和输出层组成，每一个词通过其前后出现的词语来预测其出现的概率，假定一个词语序列$w_1\cdots\cdots w_t$，每个词语的词向量通过其近邻词来训练器出现的最大log概率获得，公式如下：

[0102] $$\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in nb(t)}\log p(w_i|w_t)$$

[0103] 其中，nb(t)为$|w_t$的近邻词集合，$p(w_i|w_t)$为计算联系此向量w_i和$|w_t$的隐藏层softmax函数，具体的计算原理可以是参照相关技术，本实施例不做赘述。

[0104] 可以理解地，所述配对三元组还包括第一三元组中的语法关系，以及第二三元组中的语法关系，进而还需要获得所述配对三元组的语法关系得分。

[0105] 例如，第一语句为“第三中学排球队击败第四中学排球队”，第一三元组为R1＝(第三中学排球队，主谓关系，击败)；第二语句为“第四中学排球队击败第三中学排球队”，第二三元组R1’＝(第三中学排球队，动宾关系，击败)，得到一个配对三元组R1R1’，虽然第一三元组中的两个词语和第二三元组中的的两个词语是相同的，但是这两个三元组的语法关系不同，表达的意思也就不同，进而也就需要计算该配对三元组的语法关系得分。容易理解地，该配对三元组中R1R1’的语法关系包括主谓关系和动宾关系，本申请实施例中，可以是基于stanford parser的依存关系来计算配对三元组的语法关系得分，具体如下：

[0106] $score_{rel}$＝match(rel_{R1},$rel_{R1'}$)；

[0107] 其中，$score_{rel}$为配对三元组的语法关系得分，rel_{R1}为配对三元组中第一三元组的语法关系(例如R1R1’中第一三元组包括的主谓关系)，$rel_{R1'}$为配对三元组中第二三元组的语法关系(例如R1R1’中第二三元组包括的动宾关系)。

[0108] 可以理解地，在获得配对三元组的词语匹配得分和语法关系匹配得分后，进一步计算该配对三元组的相似度度值，具体如下：

[0109] $score_{dep}$＝$score_{word}$×$score_{rel}$；

[0110] 这样，也就可以基于上述方式获取第一语句和第二语句得到的每个配对三元组的相似度值，进而基于每个配对三元组的相似度值，计算第一语句和第二语句的文本相似度。

[0111] 可选地，所述基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度，包括：

[0112] 获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

[0113] 确定每一个所述第一三元组对应的目标配对三元组；

[0114] 基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

[0115] 基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

[0116] 可以理解地，第一语句能够生成第一三元组集合，第一三元组集合包括至少一个第一三元组，所述目标第一三元组为所述第一三元组集合中的任一个第一三元组。本申请实施例中，将第一三元组集合中的每一个第一三元组和第二三元组集合中的每一个第二三元组进行组合后，得到多个配对三元组，那么目标第一三元组也就会和每一个第二三元组

进行组合配对，获得配对三元组，并计算每个配对三元组的相似度值。可以理解地，目标第一三元组和每一个第二三元组不会都很相似，基于目标第一三元组得到的配对三元组的相似度值也就会不同，则将其中相似度值最高的目标三元组确定为目标配对三元组，并基于同样的方法得到每一个第一三元组对应的目标配对三元组。当然，也可以是以第二三元组集合中的某一个作为目标第二三元组，确定目标配对三元组，并基于相同的方式获得每一个第二三元组对应的目标配对三元组。这样，通过逐一比较相似度值，能够更为准确地确定出每一个第一三元组最为相似的第二三元组。

[0117] 本申请实施例中，可以预先根据配对三元组中包括的词语或语法关系在第一语句或第二语句中的重要程度，分别设定每个配对三元组对应的权重值，进而也就能够得到预设的语句权重值表。所述预设的语句权重值表也就包括与每一个配对三元组对应的权重值，进而也就能够获取与每一个目标配对三元组对应的权重值，基于所述目标配对三元组的相似度值及对应的所述权重值，计算所述第一语句与所述第二语句的文本相似度。

[0118] 可选地，第一语句与第二语句的文本相似度$score_{sent}$计算公式如下：

[0119] $$score_{sent}=\frac{\sum_{i}^{sentl} edgeweight_i * score_{highdep\_i}}{\sum_{i}^{sentl} edgeweight_i};$$

[0120] 其中，$score_{highdep}$为目标配对三元组的相似度值；edgeweight为与目标配对三元组对应的权重值；sentl表示第一语句和第二语句中较长的一个语句，可以理解地，语句之间的蕴含意思相似分析是具有方向性的，长句子更可能包含短句子所蕴含的意思，例如“张三朝医院走动”和“张三在走动”这两个句子中，长句子包含了短句子中的意向，而短句子不能涵盖长句子中所蕴含的所有意思。

[0121] 这样，也就通过对第一语句和第二语句所生成的所有目标配对三元组与对应的权重值进行加权计算，进而以获得第一语句和第二语句的文本相似度。并且，本申请实施例中，目标配对三元组中不仅包括第一语句的两个和第二语句的两个词语，还包括每个语句两个词语之间的语法关系，进而目标配对三元组的相似度值也就不仅只是对词语的相似度进行分析，还包括对词语之间语法关系的相似度进行分析，更进一步提高了对两个语句之间相似度分析的准确度。

[0122] 可选地，所述基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度，包括：

[0123] 获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

[0124] 基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

[0125] 基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

[0126] 其中，所述预设词语可以是用户预先设置的词语，例如可以是将动词作设置为预设词语。本申请实施例中，所述预设词语可以是语句中的核心词，比如核心词可以是动词、修饰词等，例如“第三中学排球队击败第四中学排球队”中的核心词为“击败”，“小红今天穿了一件漂亮的衣服”中的核心词为“漂亮”。

[0127] 可选地，在设定预设词语后，可以是基于目标配对三元组中所包括的预设词语的数量来确定目标配对三元组的权重衰减系数，进而基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减。例如，若所述目标配对三元组所包括的第一三元组的两个词语或第二三元组中的两个词语都为预设词语，则所述目标配对三元组对应的权重值不进行衰减；若所述目标配对三元组所包括的第一三元组或第二三元组中只包括一个预设词语，则所述目标配对三元组对应的权重值进行二分之一的衰减；所述目标配对三元组所包括的第一三元组和第二三元组中都不包括预设词语，则所述目标配对三元组对应的权重值进行四分之一的衰减。进一步地，基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度，计算公式如下：

[0128] $$score_{sent}=\frac{\sum_{i}^{sentl} reduce(edgeweight_i)*score_{highdep\_i}}{\sum_{i}^{sentl} reduce(edgeweight_i)}$$；

[0129] 其中，reduce(edgeweight)为目标配对三元组衰减后的权重值，其他参数请参照上述公式中的描述。可选地，还可以进一步采用皮尔逊相关系数计算语句之间的相关程度，相关计算方法可以是参考相关技术，本实施例不做赘述。

[0130] 这样，通过目标配对三元组中包括的预设词语的数量，来对目标配对三元组对应的权重值进行衰减，进而使得语句中包括的词语的不同或者词语的数量能够对语句的文本相似度分析产生影响，使得语句之间的文本相似度分析与所包括的词语相关性更强，进一步提高了语句之间文本相似度分析的准确度。

[0131] 为更好地理解本申请实施例提供的方案，请参照图1a，图1a是本申请实施例提供的另一种文本相似度分析方法的流程图。如图1a所示，在获取到第一语句和第二语句的情况下，生成第一语句的三元组集合Rsent和第二语句的三元组集合Rsent'，其中，第一语句的三元组集合Rsent包括多个第一三元组R1、R2、R3……Rn，第二语句的三元组集合Rsent'包括多个第一三元组R1'、R2'、R3'……Rn'，三元组及三元组集合的生成方式可以是参数上述图1所述实施例中的具体描述，本实施例不再举例赘述。

[0132] 可选地，可以是将多个第一三元组与多个第二三元组任意组合，以获得配对三元组，如图1a中所示，将R1与R2'进行组合，将R2与R3'进行组合，将R3与R1'进行组合，将Rn与Rn'进行组合等，对组合后的配对三元组Rn Rn'进行文本相似度分析。例如，将第一三元组Rn的词语$slot_{n1}$与第二三元组Rn'中的词语$slot_{n',1}$计算第一词语匹配得分，将第一三元组Rn的词语$slot_{n2}$与第二三元组Rn'中的词语$slot_{n',2}$计算第一词语匹配得分，将第一三元组Rn的语法关系rel_n与第二三元组Rn'中的语法关系$rel_{n'}$计算语法关系匹配得分，基于第一词语匹配得分、第二词语匹配得分和语法关系匹配得分获取配对三元组Rn Rn'的相似度值。基于同样的方式，也就能够获得每一个配对三元组的相似度值，将获取到的每一个配对三元组的相似度值进行加权均值计算，以获取第一语句和第二语句的相似度，其中，具体的计算方式可以是参照上述图1所述实施例中的具体描述，本实施例不再赘述；本申请实施例提供的文本相似度分析方法，通过对两个语句中词语的相似度及语法关系的相似度进行分析计算，提高了文本相似度分析的准确性。

[0133] 需要说明的是，本申请实施例提供的文本相似度分析方法可以是应用于机器翻

译、文本挖掘、文本分析、文本数据采集等领域，能够满足用户对于文本相似度分析的需求。

[0134] 需要说明的是，本申请实施例提供的文本相似度分析方法，执行主体可以为文本相似度分析装置，或者该文本相似度分析装置中的用于执行加载文本相似度分析方法的控制模块。本申请实施例中以文本相似度分析装置执行加载文本相似度分析方法为例，说明本申请实施例提供的文本相似度分析装置。

[0135] 请参见图2，图2是本申请实施例提供的一种文本相似度分析装置的结构图。如图2所示，所述文本相似度分析装置200包括：

[0136] 第一获取模块201，用于获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

[0137] 生成模块202，用于生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

[0138] 第二获取模块203，用于用于根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0139] 可选地，所述第一三元组集合中的三元组为第一三元组，所述第二三元组集合中的三元组为第二三元组；

[0140] 所述第二获取模块203包括：

[0141] 配对子模块，用于将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合，获得多个配对三元组；所述配对三元组包括一个第一三元组和一个第二三元组；

[0142] 获取子模块，用于获取每个配对三元组的相似度值；

[0143] 分析子模块，用于基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度。

[0144] 可选地，所述获取子模块还用于：

[0145] 基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分；

[0146] 基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系，获取所述每个配对三元组的语法关系匹配得分；

[0147] 基于所述词语匹配得分和所述语法关系匹配得分，计算所述每个配对三元组的相似度值。

[0148] 可选地，所述每个配对三元组均包括第一配对词语和第二配对词语，所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个，所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个；

[0149] 所述获取子模块还用于：

[0150] 基于词向量的余弦相似度算法，获取每个配对三元组中第一配对词语的第一分值和第二配对词语的第二分值；

[0151] 对所述第一分值和所述第二分值进行加权求和计算，获取所述每个配对三元组的词语匹配得分。

[0152] 可选地，所述每个配对三元组均包括第三配对词语，所述第三配对词语包括第一短语和第二短语，所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语，所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语；

[0153] 所述获取子模块还用于：

[0154] 基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

[0155] 可选地，所述分析子模块还用于：

[0156] 获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

[0157] 确定每一个所述第一三元组对应的目标配对三元组；

[0158] 基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

[0159] 基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

[0160] 可选地，所述分析子模块还用于：

[0161] 获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

[0162] 基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

[0163] 基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

[0164] 本申请实施例提供的文本相似度分析装置200，通过生成第一语句的第一三元组集合和第二语句的第二三元组集合，所述三元组包括两个词语及所述两个词语之间的语法关系，因而在根据所述第一三元组集合和所述第二三元组集合获取第一语句和第二语句的文本相似度时，不仅只是对第一语句中词语的相似度进行分析，还包括对词语之间的语法关系的相似度的分析，基于语法关系能够更好地考虑到句子的语义，这样也就更进一步提高了对两个语句之间相似度分析的准确度。

[0165] 本申请实施例中的文本相似度分析装置200可以是装置，也可以是终端中的部件、集成电路、或芯片。该装置可以是移动电子设备，也可以为非移动电子设备。示例性的，移动电子设备可以为手机、平板电脑、笔记本电脑、掌上电脑、车载电子设备、可穿戴设备、超级移动个人计算机(ultra-mobile personal computer,UMPC)、上网本或者个人数字助理(personal digital assistant,PDA)等，非移动电子设备可以为服务器、网络附属存储器(Network Attached Storage,NAS)、个人计算机(personal computer,PC)、电视机(television,TV)、柜员机或者自助机等，本申请实施例不作具体限定。

[0166] 本申请实施例中的文本相似度分析装置200可以为具有操作系统的装置。该操作系统可以为安卓(Android)操作系统，可以为ios操作系统，还可以为其他可能的操作系统，本申请实施例不作具体限定。

[0167] 本申请实施例提供的文本相似度分析装置200能够实现图1所述方法实施例实现的各个过程，为避免重复，这里不再赘述。

[0168] 请参见图3，图3是本申请实施例提供的一种电子设备的结构图，如图3所示，所述电子设备包括：处理器300、存储器320及存储在所述存储器320上并可在所述处理器300上

运行的程序或指令，处理器300，用于读取存储器320中的程序或指令；所述电子设备还包括总线接口和收发机310。

[0169] 收发机310，用于在处理器300的控制下接收和发送数据。

[0170] 其中，在图3中，总线架构可以包括任意数量的互联的总线和桥，具体由处理器300代表的一个或多个处理器和存储器320代表的存储器的各种电路链接在一起。总线架构还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路链接在一起，这些都是本领域所公知的，因此，本文不再对其进行进一步描述。总线接口提供接口。收发机310可以是多个元件，即包括发送机和收发机，提供用于在传输介质上与各种其他装置通信的单元。处理器300负责管理总线架构和通常的处理，存储器320可以存储处理器300在执行操作时所使用的数据。

[0171] 在本申请实施例的一种实施方式中，处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0172] 获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

[0173] 生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

[0174] 根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0175] 可选地，所述第一三元组集合中的三元组为第一三元组，所述第二三元组集合中的三元组为第二三元组；处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0176] 将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合，获得多个配对三元组；所述配对三元组包括一个第一三元组和一个第二三元组；

[0177] 获取每个配对三元组的相似度值；

[0178] 基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度。

[0179] 可选的，处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0180] 基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分；

[0181] 基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系，获取所述每个配对三元组的语法关系匹配得分；

[0182] 基于所述词语匹配得分和所述语法关系匹配得分，计算所述每个配对三元组的相似度值。

[0183] 可选的，所述每个配对三元组均包括第一配对词语和第二配对词语，所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个，所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个；

[0184] 处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0185] 基于词向量的余弦相似度算法，获取每个配对三元组中第一配对词语的第一分值

和第二配对词语的第二分值；

[0186] 对所述第一分值和所述第二分值进行加权求和计算，获取所述每个配对三元组的词语匹配得分。

[0187] 可选的，所述每个配对三元组均包括第三配对词语，所述第三配对词语包括第一短语和第二短语，所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语，所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语；

[0188] 处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0189] 基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

[0190] 可选的，处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0191] 获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

[0192] 确定每一个所述第一三元组对应的目标配对三元组；

[0193] 基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

[0194] 基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

[0195] 可选的，处理器300，用于读取存储器320中的程序或指令，执行如下步骤：

[0196] 获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

[0197] 基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

[0198] 基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

[0199] 本实施方式中，电子设备能够执行上述图1所述文本信息搜索方法实施例的全部技术特征，能够提高电子设备对于文本相似度分析的准确度，其实现原理和技术效果类似，本实施例此处不再赘述。

[0200] 本发明实施例还提供一种可读存储介质，可读存储介质上存储有计算机程序。

[0201] 其中，所述计算机程序被处理器执行时实现如下步骤：

[0202] 获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语；

[0203] 生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合，其中，所述第一三元组集合和所述第二三元组集合中均包括至少一个三元组，所述三元组包括两个词语及所述两个词语之间的语法关系；

[0204] 根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度。

[0205] 可选的，所述第一三元组集合中的三元组为第一三元组，所述第二三元组集合中的三元组为第二三元组；所述计算机程序被处理器执行时还能实现如下步骤：

[0206] 将所述第一三元组集合中的每一个第一三元组与所述第二三元组集合中的每一个第二三元组进行组合，获得多个配对三元组；所述配对三元组包括一个第一三元组和一个第二三元组；

[0207] 获取每个配对三元组的相似度值；

[0208] 基于所述每个配对三元组的相似度值，获取所述第一语句与所述第二语句的文本相似度。

[0209] 可选的，所述计算机程序被处理器执行时还能实现如下步骤：

[0210] 基于每个配对三元组中的第一三元组中的两个词语和第二三元组中的两个词语，获取所述每个配对三元组的词语匹配得分；

[0211] 基于所述每个配对三元组中所述第一三元组中的语法关系和所述第二三元组中的语法关系，获取所述每个配对三元组的语法关系匹配得分；

[0212] 基于所述词语匹配得分和所述语法关系匹配得分，计算所述每个配对三元组的相似度值。

[0213] 可选的，所述每个配对三元组均包括第一配对词语和第二配对词语，所述第一配对词语为组成所述配对三元组的第一三元组中的第一词语和第二三元组中的第三词语和第四词语中的一个，所述第二配对词语为组成所述配对三元组的第一三元组中的第二词语和第二三元组中的所述第三词语和所述第四词语中的另一个；所述计算机程序被处理器执行时还能实现如下步骤：

[0214] 基于词向量的余弦相似度算法，获取每个配对三元组中第一配对词语的第一分值和第二配对词语的第二分值；

[0215] 对所述第一分值和所述第二分值进行加权求和计算，获取所述每个配对三元组的词语匹配得分。

[0216] 可选的，所述每个配对三元组均包括第三配对词语，所述第三配对词语包括第一短语和第二短语，所述第一短语为组成所述配对三元组的第一三元组中的第一词语和第二词语，所述第二短语为组成所述配对三元组的第二三元组中的第三词语和第四词语；所述计算机程序被处理器执行时还能实现如下步骤：

[0217] 基于词向量的余弦相似度算法，获取每个配对三元组中第三配对词语的第三分值，所述第三分值为对应的配对三元组的词语匹配得分。

[0218] 可选的，所述计算机程序被处理器执行时还能实现如下步骤：

[0219] 获取目标第一三元组与每一个第二三元组形成的配对三元组中，相似度值最高的目标配对三元组；

[0220] 确定每一个所述第一三元组对应的目标配对三元组；

[0221] 基于预设的语句权重值表，获取每一个目标配对三元组对应的权重值；

[0222] 基于所述目标配对三元组的相似度值及对应的所述权重值，获取所述第一语句与所述第二语句的文本相似度。

[0223] 可选的，所述计算机程序被处理器执行时还能实现如下步骤：

[0224] 获取所述目标配对三元组中包括的预设词语的数量，确定所述目标配对三元组的权重衰减系数；

[0225] 基于所述权重衰减系数对所述目标配对三元组对应的权重值进行衰减；

[0226] 基于所述目标配对三元组的相似度值及衰减后的权重值，获取所述第一语句与所述第二语句的文本相似度。

[0227] 在该实施方式中，可读存储介质能够执行上述图1所述文本相似度分析方法实施例的全部技术特征，其实现原理和技术效果类似，本实施例此处不再赘述。

[0228] 其中,所述的可读存储介质,如只读存储器(Read-Only Memory,简称ROM)、随机存取存储器(Random Access Memory,简称RAM)、磁碟或者光盘等。

[0229] 本申请实施例另提供了一种芯片,所述芯片包括处理器和通信接口,所述通信接口和所述处理器耦合,所述处理器用于运行程序或指令,实现上述图1所述文本相似度分析方法实施例的各个过程,且能达到相同的技术效果,为避免重复,这里不再赘述。

[0230] 应理解,本申请实施例提到的芯片还可以称为系统级芯片、系统芯片、芯片系统或片上系统芯片等。

[0231] 需要说明的是,在本文中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0232] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本发明各个实施例所述的方法。

[0233] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

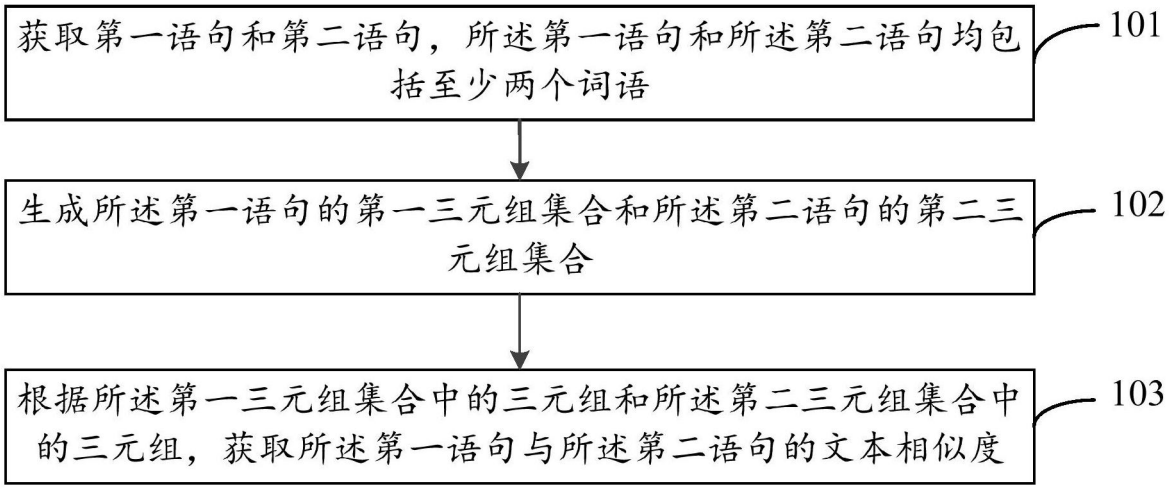

获取第一语句和第二语句，所述第一语句和所述第二语句均包括至少两个词语
101
生成所述第一语句的第一三元组集合和所述第二语句的第二三元组集合
102
根据所述第一三元组集合中的三元组和所述第二三元组集合中的三元组，获取所述第一语句与所述第二语句的文本相似度
103

图1

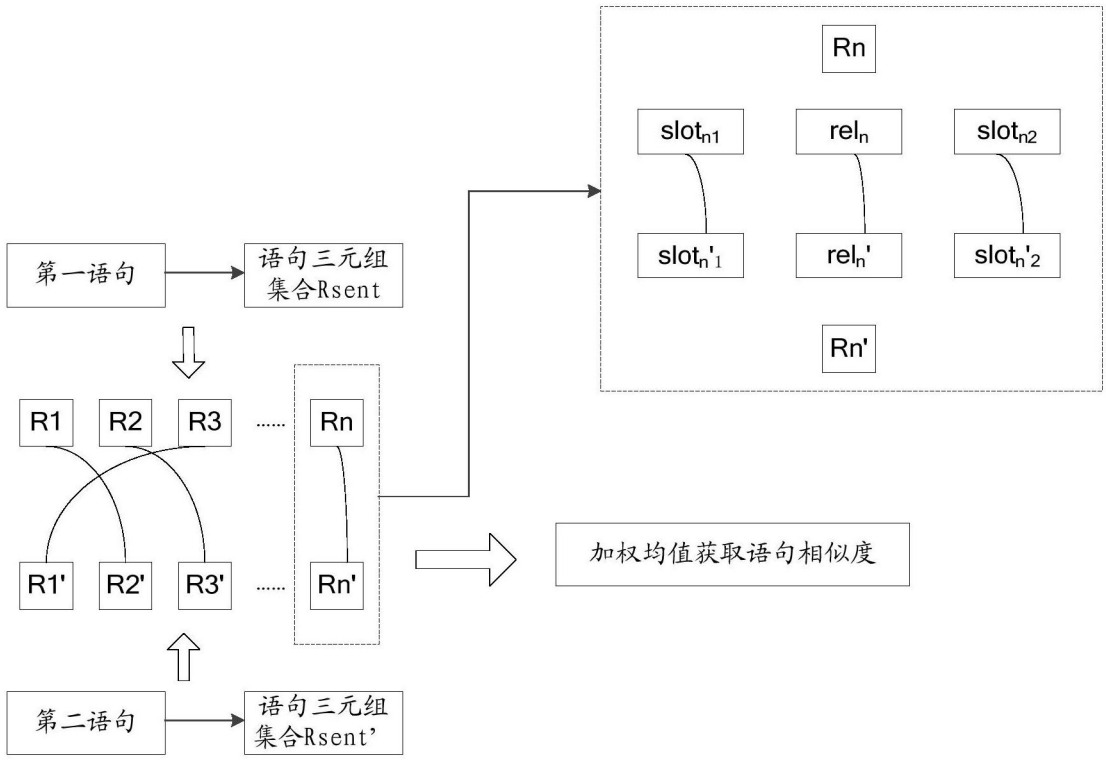

第一语句
语句三元组集合Rsent
R1
R2
R3
.....
Rn
R1'
R2'
R3'
.....
Rn'
第二语句
语句三元组集合Rsent'
Rn
slot_n1
rel_n
slot_n2
slot_n'1
rel_n'
slot_n'2
Rn'
加权均值获取语句相似度

图1a

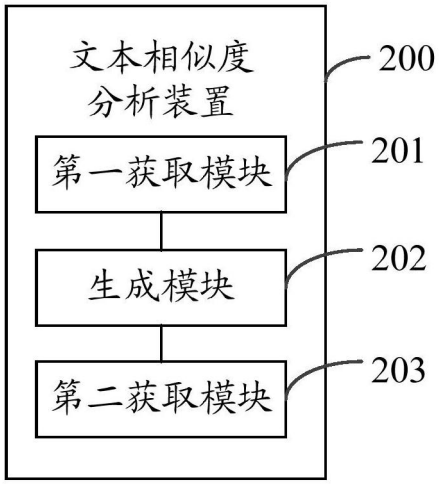

文本相似度
分析装置
200
第一获取模块
201
生成模块
202
第二获取模块
203

图2

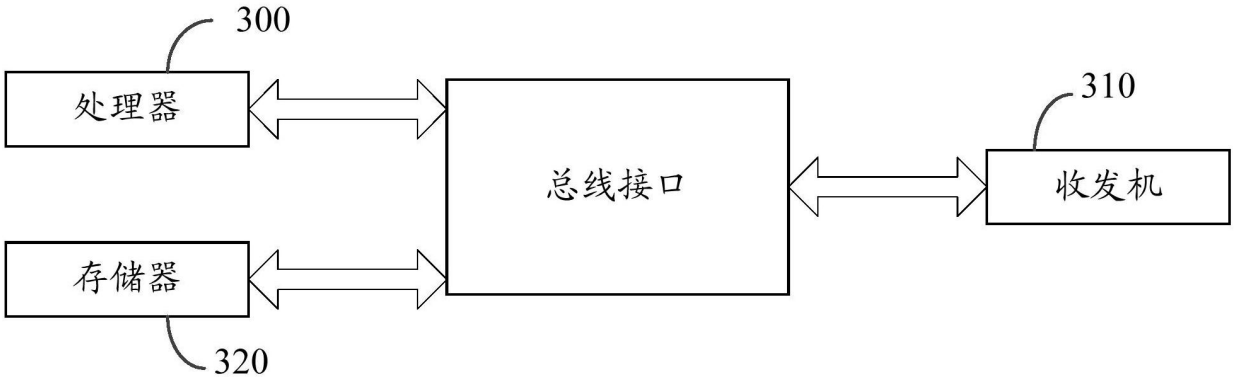

300
处理器
存储器
320
总线接口
310
收发机

图3