(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 119646215 B
(45)授权公告日 2026.01.02

(21)申请号 202411679015.5
(22)申请日 2024.11.21
(65)同一申请的已公布的文献号
申请公布号 CN 119646215 A
(43)申请公布日 2025.03.18
(73)专利权人 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 杜月
(51)Int.Cl.
G06F 16/35(2025.01)
G06F 40/289(2020.01)
G06F 40/247(2020.01)
G06F 40/216(2020.01)
G06F 18/22(2023.01)
(56)对比文件
CN 114357990 A,2022.04.15
CN 116796723 A,2023.09.22
审查员 徐晓孜

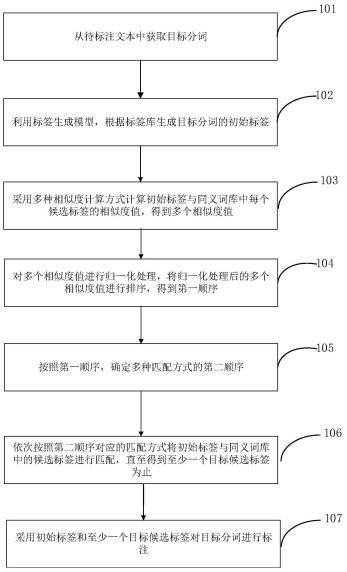
权利要求书3页 说明书13页 附图3页

(54)发明名称

一种文本标签的标注方法、装置及系统

(57)摘要

本公开是关于一种文本标签的标注方法、装置及系统。其中,方法包括:从待标注文本中获取目标分词,利用标签生成模型,根据标签库生成目标分词的初始标签,采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值,对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序,按照第一顺序,确定多种匹配方式的第二顺序,依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,采用初始标签和至少一个目标候选标签对目标分词进行标注。本方案提升了文本标签的全面性。



CN 119646215 B

1.一种文本标签的标注方法,其特征在于,包括:

从待标注文本中获取目标分词;

利用标签生成模型,根据标签库生成所述目标分词的初始标签;所述标签库为与所述待标注文本所属领域关联的标签库;

采用多种相似度计算方式计算所述初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值;

对所述多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序;

按照所述第一顺序,确定多种匹配方式的第二顺序;

依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止;

采用所述初始标签和所述至少一个目标候选标签对所述目标分词进行标注;

其中,所述按照所述第一顺序,确定多种匹配方式的第二顺序,包括:

获取所述多种相似度计算方式与所述多种匹配方式的映射关系;

按照所述映射关系和所述第一顺序,确定所述多种匹配方式的第二顺序;其中,所述多种相似度计算方式包括TF-IDF算法、编辑距离算法、余弦相似度算法和杰卡德Jaccard相似度算法;所述匹配方式包括同义词匹配、精确匹配、模糊匹配和语义匹配;所述映射关系为TF-IDF算法对应同义词匹配,编辑距离算法对应精确匹配,余弦相似度算法对应模糊匹配,Jaccard相似度算法对应语义匹配;

其中,所述精确匹配是比较初始标签与同义词库中的每个同义词,如果完全相同,则匹配成功;所述语义匹配是利用自然语言处理技术捕捉初始标签和同义词之间的语义关系,如果初始标签和同义词在语义上相近或相关,则匹配成功。

2.根据权利要求1所述的文本标签的标注方法,其特征在于,所述对所述多个相似度值进行归一化处理,包括:

在所述多种相似度计算方式包括第一方式的情况下,采用以下公式对所述第一方式对应的相似度值进行归一化处理:

$$a=\frac{tf\_idf\_h(Test[n])}{\log(|n|)}$$

其中,$tf_idf_h(Test[n]$为采用第一方式对所述同义词库中第n个候选标签进行相似度计算得到的相似度值,a为第一方式对应的归一化处理后的相似度值;所述第一方式为采用词频-逆向文件频率TF-IDF算法计算所述相似度值;

在所述多种相似度计算方式包括第二方式的情况下,采用以下公式对所述第一方式对应的相似度值进行归一化处理:

$$b=1-\frac{levenshtein\_hash(Test[n])}{length(n)}$$

其中,$levenshtein_hash(Test[n])$为采用第二方式对所述同义词库中第n个候选标签进行相似度计算得到的相似度值,b为第二方式对应的归一化处理后的相似度值,length(n)为第n个候选标签的长度;所述第二方式为采用编辑距离算法计算所述相似度值。

3.根据权利要求1所述的文本标签的标注方法,其特征在于,所述依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,包括:

按照所述第二顺序从所述多种匹配方式中确定当前匹配方式;

按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果;所述匹配结果包括是否匹配成功以及匹配到的候选标签;

在所述匹配结果为没有匹配成功,并且所述当前匹配方式不是所述第二顺序中的最后一个匹配方式的情况下,按照所述第二顺序将下一个匹配方式确定为当前匹配方式,返回执行所述按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果的步骤;

在所述匹配结果为匹配成功的情况下,将所述匹配到的候选标签确定为所述目标候选标签;

在所述匹配结果为没有匹配成功,并且所述当前匹配方式是所述第二顺序中的最后一个匹配方式的情况下,确定匹配失败。

4.根据权利要求3所述的文本标签的标注方法,其特征在于,所述多种匹配方式中包括模糊匹配,所述按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果,包括:

在所述当前匹配方式为模糊匹配的情况下,针对同义词库中每个候选标签,通过以下公式计算所述初始标签与所述候选标签的综合相似度值:

$$Target = \frac{a * \lambda_1 + b * \lambda_2 + c * \lambda_3 + d * \lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}$$

其中,Target为综合相似度值,a为采用第一方式计算得到的相似度值,b为采用第二方式计算得到的相似度值,c为采用第三方式计算得到的相似度值,d为采用第四方式计算得到的相似度值,λ_1为第一方式对应的权重值,λ_2为第二方式对应的权重值,λ_3为第三方式对应的权重值,λ_4为第四方式对应的权重值;

在综合相似度值满足第一预设条件的情况下,将所述候选标签确定为目标候选标签。

5.根据权利要求4所述的文本标签的标注方法,其特征在于,所述在综合相似度值满足第一预设条件的情况下,将所述候选标签确定为目标候选标签包括:

通过以下公式计算相似度均值x:

$$x = \frac{\sum_{1}^{n} a\,\lambda_1 + \sum_{1}^{n} b\,\lambda_2 + \sum_{1}^{n} c\,\lambda_3 + \sum_{1}^{n} d\,\lambda_4}{0.5n(n+1)(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)}$$

其中,n为所述同义词库中候选标签的数量;

在Target<=x的情况下,采用多个第一区间确定所述候选标签的相似度等级;

在Target>x的情况下,采用多个第二区间确定所述候选标签的相似度等级;

在所述相似度等级满足第二预设条件的情况下,将所述候选标签确定为目标候选标签;其中,相似度等级满足所述第二预设条件的第一区间对应的区间范围大于相似度等级满足所述第二预设条件的第二区间对应的区间范围。

6.一种文本标签的标注装置,其特征在于,应用于权利要求1-5任一项所述的方法,包

括：

获取单元，用于从待标注文本中获取目标分词；

生成单元，用于利用标签生成模型，根据标签库生成所述目标分词的初始标签；所述标签库为与所述待标注文本所属领域关联的标签库；

计算单元，用于采用多种相似度计算方式计算所述初始标签与同义词库中每个候选标签的相似度值，得到多个相似度值；

排序单元，用于对所述多个相似度值进行归一化处理，将归一化处理后的多个相似度值进行排序，得到第一顺序；

确定单元，用于按照所述第一顺序，确定多种匹配方式的第二顺序；

匹配单元，用于依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配，直至得到至少一个目标候选标签为止；

标注单元，用于采用所述初始标签和所述至少一个目标候选标签对所述目标分词进行标注。

7.一种电子设备，其特征在于，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，实现如权利要求1至5中任一项所述的方法。

8.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至5中任一项所述的方法。

9.一种计算机程序产品，包括计算机程序，其特征在于，所述计算机程序在被处理器执行时实现如权利要求1至5中任一项所述的方法。

一种文本标签的标注方法、装置及系统

技术领域

[0001] 本公开涉及数据处理技术领域,尤其涉及一种文本标签的标注方法、装置及系统。

背景技术

[0002] 相关技术中,对于存储的科研报告或技术资料等知识资源需要进行标签分类,以便于对其进行查询和检索,进而使知识资源得到有效的利用和共享,然而,传统的人工标签分类方式不仅耗时耗力,还难以保证标签的一致性和全面性,从而限制了知识资源的有效利用和共享。

发明内容

[0003] 为克服相关技术中存在的问题,本公开提供一种文本标签的标注方法、装置及系统。

[0004] 根据本公开实施例的第一方面,提供一种文本标签的标注方法,包括:

[0005] 从待标注文本中获取目标分词;

[0006] 利用标签生成模型,根据标签库生成所述目标分词的初始标签;所述标签库为与所述待标注文本所属领域关联的标签库;

[0007] 采用多种相似度计算方式计算所述初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值;

[0008] 对所述多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序;

[0009] 按照所述第一顺序,确定多种匹配方式的第二顺序;

[0010] 依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止;

[0011] 采用所述初始标签和所述至少一个目标候选标签对所述目标分词进行标注。

[0012] 在本公开一些实施例中,所述对所述多个相似度值进行归一化处理,包括:

[0013] 在所述多种相似度计算方式包括第一方式的情况下,采用以下公式对所述第一方式对应的相似度值进行归一化处理:

[0014] $$a=\frac{tf\_idf\_h(Test[n])}{\log(|n|)}$$

[0015] 其中,tf_idf_h(Test[n]为采用第一方式对所述同义词库中第n个候选标签进行相似度计算得到的相似度值,a为第一方式对应的归一化处理后的相似度值;所述第一方式为采用词频-逆向文件频率TF-IDF算法计算所述相似度值;

[0016] 在所述多种相似度计算方式包括第二方式的情况下,采用以下公式对所述第一方式对应的相似度值进行归一化处理:

[0017] $$b=1-\frac{levenshtein\_hash(Test[n])}{length(n)}$$

[0018] 其中,levenshtein_hash(Test[n])为采用第二方式对所述同义词库中第n个候选标签进行相似度计算得到的相似度值,b为第二方式对应的归一化处理后的相似度值,length(n)为第n个候选标签的长度;所述第二方式为采用编辑距离算法计算所述相似度值。

[0019] 在本公开一些实施例中,所述按照所述第一顺序,确定多种匹配方式的第二顺序,包括:

[0020] 获取所述多种相似度计算方式与所述多种匹配方式的映射关系;

[0021] 按照所述映射关系和所述第一顺序,确定所述多种匹配方式的第二顺序;其中,所述多种相似度计算方式包括TF-IDF算法、编辑距离算法、余弦相似度算法和杰卡德Jaccard相似度算法;所述匹配方式包括同义词匹配、精确匹配、模糊匹配和语义匹配;所述映射关系为TF-IDF算法对应同义词匹配,编辑距离算法对应精确匹配,余弦相似度算法对应模糊匹配,Jaccard相似度算法对应语义匹配。

[0022] 在本公开一些实施例中,所述依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,包括:

[0023] 按照所述第二顺序从所述多个匹配方式中确定当前匹配方式;

[0024] 按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果;所述匹配结果包括是否匹配成功以及匹配到的候选标签;

[0025] 在所述匹配结果为没有匹配成功,并且所述当前匹配方式不是所述第二顺序中的最后一个匹配方式的情况下,按照所述第二顺序将下一个匹配方式确定为当前匹配方式,返回执行所述按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果的步骤;

[0026] 在所述匹配结果为匹配成功的情况下,将所述匹配到的候选标签确定为所述目标候选标签;

[0027] 在所述匹配结果为没有匹配成功,并且所述当前匹配方式是所述第二顺序中的最后一个匹配方式的情况下,确定匹配失败。

[0028] 在本公开一些实施例中,所述多种匹配方式中包括模糊匹配,所述按照所述当前匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配,得到匹配结果,包括:

[0029] 在所述当前匹配方式为模糊匹配的情况下,针对同义词库中每个候选标签,通过以下公式计算所述初始标签与所述候选标签的综合相似度值:

[0030] $$Target=\frac{a*\lambda_1+b*\lambda_2+c*\lambda_3+d*\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}$$

[0031] 其中,Target为综合相似度值,a为采用第一方式计算得到的相似度值,b为采用第二方式计算得到的相似度值,c为采用第三方式计算得到的相似度值,d为采用第四方式计算得到的相似度值,λ_1为第一方式对应的权重值,λ_2为第二方式对应的权重值,λ_3为第三方式对应的权重值,λ_4为第四方式对应的权重值;

[0032] 在综合相似度值满足第一预设条件的情况下,将所述候选标签确定为目标候选标签。

[0033] 在本公开一些实施例中,所述在综合相似度值满足第一预设条件的情况下,将所述候选标签确定为目标候选标签包括:

[0034] 通过以下公式计算相似度均值x：

[0035] $$x=\frac{\sum_{1}^{n}a\,\lambda_1+\sum_{1}^{n}b\,\lambda_2+\sum_{1}^{n}c\,\lambda_3+\sum_{1}^{n}d\,\lambda_4}{0.5n(n+1)(\lambda_1+\lambda_2+\lambda_3+\lambda_4)}$$

[0036] 其中，n为所述同义词库中候选标签的数量；

[0037] 在Target<＝x的情况下，采用多个第一区间确定所述候选标签的相似度等级；

[0038] 在Target>x的情况下，采用多个第二区间确定所述候选标签的相似度等级；

[0039] 在所述相似度等级满足第二预设条件的情况下，将所述候选标签确定为目标候选标签；其中，相似度等级满足所述第二预设条件的第一区间对应的区间范围大于相似度等级满足所述第二预设条件的第二区间对应的区间范围。

[0040] 根据本公开实施例的第二方面，提供一种文本标签的标注装置，包括：

[0041] 获取单元，用于从待标注文本中获取目标分词；

[0042] 生成单元，用于利用标签生成模型，根据标签库生成所述目标分词的初始标签；所述标签库为与所述待标注文本所属领域关联的标签库；

[0043] 计算单元，用于采用多种相似度计算方式计算所述初始标签与同义词库中每个候选标签的相似度值，得到多个相似度值；

[0044] 排序单元，用于对所述多个相似度值进行归一化处理，将归一化处理后的多个相似度值进行排序，得到第一顺序；

[0045] 确定单元，用于按照所述第一顺序，确定多种匹配方式的第二顺序；

[0046] 匹配单元，用于依次按照所述第二顺序对应的匹配方式将所述初始标签与所述同义词库中的候选标签进行匹配，直至得到至少一个目标候选标签为止；

[0047] 标注单元，用于采用所述初始标签和所述至少一个目标候选标签对所述目标分词进行标注。

[0048] 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。

[0049] 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。

[0050] 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。

[0051] 本公开的实施例提供的技术方案可以包括以下有益效果：通过从待标注文本中获取目标分词，利用标签生成模型，根据标签库生成目标分词的初始标签，采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值，得到多个相似度值，对多个相似度值进行归一化处理，将归一化处理后的多个相似度值进行排序，得到第一顺序，按照第一顺序，确定多种匹配方式的第二顺序，依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配，直至得到至少一个目标候选标签为止，采用初始标签和至少一个目标候选标签对目标分词进行标注。通过二次生成标识的方式对文本的标签进行扩充，从而在提高对文本标注标签的效率的同时，提高了文本标签的全面性，进而能够提升检索或查询文本的准确性，使知识资源得到充分利用。

[0052] 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

[0053] 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。

[0054] 图1是根据一示例性实施例示出的一种文本标签的标注方法的流程图。

[0055] 图2是根据一示例性实施例示出的一种文本标签的标注装置的框图。

[0056] 图3是根据一示例性实施例示出的一种用于文本标签的标注方法的装置的框图。

具体实施方式

[0057] 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。

[0058] 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。

[0059] 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”。

[0060] 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。

[0061] 相关技术中,对于存储的科研报告或技术资料等知识资源需要进行标签分类,以便于对其进行查询和检索,进而使知识资源得到有效的利用和共享,然而,传统的人工标签分类方式不仅耗时耗力,还难以保证标签的一致性和全面性,从而限制了知识资源的有效利用和共享。

[0062] 为了解决上述问题,本公开提供了一种文本标签的标注方法、装置及系统,通过从待标注文本中获取目标分词,利用标签生成模型,根据标签库生成目标分词的初始标签,采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值,对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序,按照第一顺序,确定多种匹配方式的第二顺序,依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,采用初始标签和至少一个目标候选标签对目标分词进行标注。通过二次生成标识的方式对文本的标签进行扩充,从而在提高对文本标注标签的效率的同时,提高了文本标签的全面性,进而能够提升检索或查询文本的准确性,使知识资源得到充分利用。

[0063] 图1是根据一示例性实施例示出的一种文本标签的标注方法的流程图,如图1所示,需要说明的是,本公开实施例的文本标签的标注方法应用于文本标签的标注装置中。如图1所示,该方法可以包括以下步骤:

[0064] 步骤101,从待标注文本中获取目标分词。

[0065] 在本公开一些实施例中,可以预先将需要标注标签的知识资源转换为纯文本格式,并进行去噪、分词、停用词过滤等预处理操作,为后续处理提供高质量的文本数据。利用自然语言处理(NLP)技术,如TF-IDF、Word2Vec或BERT等模型,对预处理后的文本进行特征提取,捕捉文本中的关键信息,包括关键词、短语、句子级别的语义特征等,得到多个分词。

[0066] 在一个实施例中,可以采用自然语言处理NLP技术进行文本预处理,对文档进行清洗,去除无用的特殊字符,并进行分词(针对中文或其他非空格分隔语言)和停用词(是指在文本中频繁出现但通常没有太多有意义的词语)过滤,以降低后续处理的复杂度。采用词嵌入(如Word2Vec、BERT)将文本转换为数值向量,便于计算机处理。此外,还可以通过TF-IDF、TextRank等算法提取关键词或短语作为文档的特征表示。

[0067] 可以理解的是,上述多个分词中的每个分词均可作为上述目标分词,以执行本公开提出的文本标签的标注方法,以完成整个代表著文本的标签标注。

[0068] 步骤102,利用标签生成模型,根据标签库生成目标分词的初始标签。

[0069] 其中,标签库为与待标注文本所属领域关联的标签库。

[0070] 在一个实施例中,可以预先根据代表著文本所属领域的专业知识构建标签库,还可以在标签库中构建标签之间的关联网络,例如构建基于标签的知识图谱,从而提升标签标注的全面性和逻辑性。

[0071] 作为一种示例,可以采用预训练的机器学习或深度学习算法(如K-means聚类、LDA主题模型、序列标注模型等)作为标签生成模型,自动生成一组初始标签。由于是根据包括领域专业知识的标签库生成的,因此初始标签既覆盖了文本的主要内容,又体现了领域特色。

[0072] 在本公开一些实施例中,由于自动生成的初始标签可能包含冗余、不相关或错误的标签,因此可以对初始标签进行优化与选择来进一步提升标签的质量。如对初始标签进行去重与筛选,移除重复的标签,基于统计或规则过滤掉明显不相关的标签。

[0073] 步骤103,采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值。

[0074] 在一些实施例中,上述多种相似度计算方式可以包括以下任意多种:

[0075] TF-IDF算法、编辑距离算法、余弦相似度算法和杰卡德Jaccard相似度算法。

[0076] 步骤104,对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序。

[0077] 在一个实施例中,可以将归一化处理后的多个相似度值按照从大到小的顺序进行排序。

[0078] 在本公开一些实施例中,步骤104中的对多个相似度值进行归一化处理,可以包括以下步骤:

[0079] 步骤a1,在多种相似度计算方式包括第一方式的情况下,采用以下公式对第一方式对应的相似度值进行归一化处理:

[0080] $$a=\frac{tf\_idf\_h(Test[n])}{\log(|n|)}$$

[0081] 其中,tf_idf_h(Test[n]为采用第一方式对同义词库中第n个候选标签进行相似度计算得到的相似度值,a为第一方式对应的归一化处理后的相似度值;第一方式为采用词频-逆向文件频率TF-IDF算法计算相似度值;

[0082] 步骤a2,在多种相似度计算方式包括第二方式的情况下,采用以下公式对第一方式对应的相似度值进行归一化处理:

[0083] $$b=1-\frac{levenshtein\_hash(Test[n])}{length(n)}$$

[0084] 其中,levenshtein_hash(Test[n])为采用第二方式对同义词库中第n个候选标签进行相似度计算得到的相似度值,b为第二方式对应的归一化处理后的相似度值,length(n)为第n个候选标签的长度;第二方式为采用编辑距离算法计算相似度值。

[0085] 步骤105,按照第一顺序,确定多种匹配方式的第二顺序。

[0086] 可以理解的是,相似度值越大,对应的匹配方式优先应用。

[0087] 在本公开一些实施例中,步骤105具体可以包括以下步骤:

[0088] 获取多种相似度计算方式与多种匹配方式的映射关系;

[0089] 按照映射关系和第一顺序,确定多种匹配方式的第二顺序。

[0090] 其中,多种相似度计算方式包括TF-IDF算法、编辑距离算法、余弦相似度算法和杰卡德Jaccard相似度算法;匹配方式包括同义词匹配、精确匹配、模糊匹配和语义匹配;映射关系为TF-IDF算法对应同义词匹配,编辑距离算法对应精确匹配,余弦相似度算法对应模糊匹配,Jaccard相似度算法对应语义匹配。

[0091] 需要说明的是,精确匹配是指直接比较生成的初始标签与同义词库中的每个同义词,如果完全相同,则视为匹配成功;模糊匹配是指为了能够考虑到语言的多样性和灵活性,采用模糊匹配算法(如编辑距离、Jaccard相似度等)来比较初始标签和同义词之间的相似度,如果相似度超过一定的阈值,则视为匹配成功;语义匹配是指利用自然语言处理技术(如词嵌入、语义角色标注等)来捕捉初始标签和同义词之间的语义关系,如果初始标签和同义词在语义上相近或相关,则视为匹配成功。

[0092] 可以理解的是,TF-IDF(Term Frequency-Inverse Document Frequency)算法用于评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜索引擎应用,作为文件与用户查询之间相关程度的度量或评级。基于此特性,TF-IDF数值更容易检测同义词。

[0093] 此外,编辑距离算法是计算两个文本之间的编辑距离,衡量其相似性,用于计算两个字符串之间的最小编辑操作次数,包括插入、删除和替换。基于此特性,更容易检测精确匹配。

[0094] 另外,余弦相似度算法是将文本标签词组表示为向量,计算两个向量之间的余弦相似度。余弦相似度是一种用于衡量两个非零向量间夹角余弦值的方法,通常用于比较文档相似度。余弦相似度的值介于-1和1之间,其中1表示完全相同,0表示不相似,-1表示完全相反。基于此特性,更适合采用模糊匹配的方式进行匹配。

[0095] 另外,Jaccard相似度算法是基于两个文本词组的词汇集合,计算其交集和并集的比率,用于衡量两个集合之间相似度的指标,其定义为两个集合的交集大小除以并集大小。Jaccard相似度值在0和1之间,1表示两个集合完全相同,0表示完全不同。基于此特性,更适合采用语义匹配的方式进行匹配。

[0096] 在一个实施例中,可以将利用TF-IDF算法计算得到的相似度值以键值对的形式存入Hash表tf_idf_hash中;可以将利用编辑距离算法计算得到的相似度值以键值对的形式存入Hash表levenshteind_hash中;可以将利用余弦相似度算法计算得到的相似度值以键值对的形式存入Hash表cosine_similaritycos_hash中;可以将利用Jaccard相似度算法计算得到的相似度值以键值对的形式存入Hash表Jaccardjac_hash中。

[0097] 在一个实施例中,可以对a、b、c、d四个值进行排续,得到上述第一排序。值大的优先选择对应方法进行词与词的匹配,a为采用第一方式计算得到的相似度值,b为采用第二方式计算得到的相似度值,c为采用第三方式计算得到的相似度值,d为采用第四方式计算得到的相似度值,c=consine_similarity_hash(Test[n]),d=jaccard_hash(Test[n])。

[0098] 步骤106,依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止。

[0099] 在一个实施例中,可以预先创建同义词库,同义词库中可以包含一词多义、多词同义的情况,从而增强搜索的准确性和丰富性。将初始标签与标签库中的同义词(也即候选标签)进行匹配,以选择最合适的标签作为目标候选标签。

[0100] 可以理解的是,为了提高标签配效率,可以按照第二顺序,选取最适合的匹配方式进行匹配,如果没有匹配到目标候选标签,可以再采用下一个匹配方式进行匹配,直至当前匹配方式能够匹配到目标候选标签为止。

[0101] 需要说明的是,目标候选标签可以是一个,也可以是多个。此外,初始标签可以是一个,也可以是多个。在上述目标分词有多个初始标签的情况下,可以分别为每个初始标签匹配至少一个目标候选标签。

[0102] 在本公开一些实施例中,步骤106具体可以包括以下步骤:

[0103] 步骤b1,按照第二顺序从多个匹配方式中确定当前匹配方式。

[0104] 步骤b2,按照当前匹配方式将初始标签与同义词库中的候选标签进行匹配,得到匹配结果。

[0105] 其中,匹配结果包括是否匹配成功以及匹配到的候选标签。

[0106] 步骤b3,在匹配结果为没有匹配成功,并且当前匹配方式不是第二顺序中的最后一个匹配方式的情况下,按照第二顺序将下一个匹配方式确定为当前匹配方式,返回执行步骤b2。

[0107] 步骤b4,在匹配结果为匹配成功的情况下,将匹配到的候选标签确定为目标候选标签。

[0108] 步骤b5,在匹配结果为没有匹配成功,并且当前匹配方式是第二顺序中的最后一个匹配方式的情况下,确定匹配失败。

[0109] 在一个实施例中,如果确定匹配失败,可以将初始标签存储至同义词库中,以便未来使用。

[0110] 在本公开一些实施例中,步骤b2具体可以包括以下步骤:

[0111] 步骤b21,在当前匹配方式为模糊匹配的情况下,针对同义词库中每个候选标签,通过以下公式计算初始标签与候选标签的综合相似度值:

[0112] $$Target=\frac{a*\lambda_1+b*\lambda_2+c*\lambda_3+d*\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}$$

[0113] 其中,Target为综合相似度值,a为采用第一方式计算得到的相似度值,b为采用第二方式计算得到的相似度值,c为采用第三方式计算得到的相似度值,d为采用第四方式计算得到的相似度值,λ_1为第一方式对应的权重值,λ_2为第二方式对应的权重值,λ_3为第三方式对应的权重值,λ_4为第四方式对应的权重值;

[0114] 步骤b22,在综合相似度值满足第一预设条件的情况下,将候选标签确定为目标候选标签。

[0115] 在本公开一些实施例中,步骤b22具体可以包括以下步骤:

[0116] 通过以下公式计算相似度均值x:

[0117] $$x=\frac{\sum_1^n a\,\lambda_1+\sum_1^n b\,\lambda_2+\sum_1^n c\,\lambda_3+\sum_1^n d\,\lambda_4}{0.5n(n+1)(\lambda_1+\lambda_2+\lambda_3+\lambda_4)}$$

[0118] 其中,n为同义词库中候选标签的数量;

[0119] 在Target<=x的情况下,采用多个第一区间确定候选标签的相似度等级;

[0120] 在Target>x的情况下,采用多个第二区间确定候选标签的相似度等级;

[0121] 在相似度等级满足第二预设条件的情况下,将候选标签确定为目标候选标签。

[0122] 其中,相似度等级满足第二预设条件的第一区间对应的区间范围大于相似度等级满足第二预设条件的第二区间对应的区间范围。

[0123] 在一个实施例中,上述第二预设条件可以是相似度等级大于或者等于预设等级。

[0124] 以理解的是,x越大说明初始标签与同义词库的整体相似性较高,较为容易匹配到目标候选标签,x越小则说明初始标签与同义词库的整体相似性较低,匹配目标候选标签较为困难。因此,在初始标签与候选标签的综合相似度值Target<=x的情况下,说明匹配较为困难,可以放宽筛选标准,也即采用多个第一区间确定候选标签的相似度等级,以利用相似度等级定是否能够将候选标签作为目标候选标签;在初始标签与候选标签的综合相似度值Target>x的情况下,说明匹配较为容易,可以收紧筛选标准,从而在确保能够匹配成功的前提下提升匹配精度,也即采用多个第二区间确定候选标签的相似度等级,以利用相似度等级定是否能够将候选标签作为目标候选标签。

[0125] 举例来说,当Target<=x时,采用的多个第一区间如表1所示:

[0126] 表1第一区间与相似度等级的映射关系表

[0127]

额定值等级	1相似度低	2相似度较低	3相似度较高	4相似度高
第一区间	[0,0.3)	[0.3,0.5)	[0.5,0.7)	[0.7,1.0]

[0128] 当Target>x时,采用的多个第一区间如表2所示:

[0129] 表2第一区间与相似度等级的映射关系表

[0130]

额定值等级	1相似度低	2相似度较低	3相似度较高	4相似度高
第二区间	[0,0.4)	[0.4,0.6)	[0.6,0.85)	[0.85,1.0]

[0131] 当确定相似度等级为3时,标签相似度较高,当确定相似度等级为4时,相似度高。可以设定阈值为a,0.6<=a<=1,在此范围内由知识库系统设定,如果Value>=a,则视为模糊匹配成功。

[0132] 步骤107,采用初始标签和至少一个目标候选标签对目标分词进行标注。

[0133] 在本公开一些实施例中,在步骤107之后,还可以将初始标签和至少一个目标候选标签显示在终端设备上,用户可实时进行反馈,对标签的反馈包含“正确”、“错误”、“有用”、“无关”等。

[0134] 根据本公开实施例提出的文本标签的标注方法,通过从待标注文本中获取目标分词,利用标签生成模型,根据标签库生成目标分词的初始标签,采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值,对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序,按照第一顺序,确定多种匹配方式的第二顺序,依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,采用初始标签和至少一个目标候选标签对目标分词进行标注。通过二次生成标识的方式对文本的标签进行扩充,从而在提高对文本标注标签的效率的同时,提高了文本标签的全面性,进而能够提升检索或查询文本的准确性,使知识资源得到充分利用。

[0135] 图2是根据一示例性实施例示出的一种文本标签的标注装置框图。参照图2,该装置包括获取单元201,生成单元202,计算单元203,排序单元204,确定单元205,匹配单元206和标注单元207。

[0136] 其中,获取单元201,用于从待标注文本中获取目标分词;

[0137] 生成单元202,用于利用标签生成模型,根据标签库生成目标分词的初始标签;标签库为与待标注文本所属领域关联的标签库;

[0138] 计算单元203,用于采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值;

[0139] 排序单元204,用于对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序;

[0140] 确定单元205,用于按照第一顺序,确定多种匹配方式的第二顺序;

[0141] 匹配单元206,用于依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止;

[0142] 标注单元207,用于采用初始标签和至少一个目标候选标签对目标分词进行标注。

[0143] 在本公开一些实施例中,排序单元204具体可以用于:

[0144] 在多种相似度计算方式包括第一方式的情况下,采用以下公式对第一方式对应的相似度值进行归一化处理:

[0145] $$a=\frac{tf\_idf\_h(Test[n])}{\log(|n|)}$$

[0146] 其中,tf_idf_h(Test[n]为采用第一方式对同义词库中第n个候选标签进行相似度计算得到的相似度值,a为第一方式对应的归一化处理后的相似度值;第一方式为采用词频-逆向文件频率TF-IDF算法计算相似度值;

[0147] 在多种相似度计算方式包括第二方式的情况下,采用以下公式对第一方式对应的

相似度值进行归一化处理：

[0148] $$b = 1 - \frac{\text{levenshtein}\_hash(Test[n])}{\text{length}(n)}$$

[0149] 其中，levenshtein_hash(Test[n])为采用第二方式对同义词库中第n个候选标签进行相似度计算得到的相似度值，b为第二方式对应的归一化处理后的相似度值，length(n)为第n个候选标签的长度；第二方式为采用编辑距离算法计算相似度值。

[0150] 在本公开一些实施例中，确定单元205具体可以用于：

[0151] 获取多种相似度计算方式与多种匹配方式的映射关系；

[0152] 按照映射关系和第一顺序，确定多种匹配方式的第二顺序；其中，多种相似度计算方式包括TF-IDF算法、编辑距离算法、余弦相似度算法和杰卡德Jaccard相似度算法；匹配方式包括同义词匹配、精确匹配、模糊匹配和语义匹配；映射关系为TF-IDF算法对应同义词匹配，编辑距离算法对应精确匹配，余弦相似度算法对应模糊匹配，Jaccard相似度算法对应语义匹配。

[0153] 在本公开一些实施例中，匹配单元206具体可以用于：

[0154] 按照第二顺序从多个匹配方式中确定当前匹配方式；

[0155] 按照当前匹配方式将初始标签与同义词库中的候选标签进行匹配，得到匹配结果；匹配结果包括是否匹配成功以及匹配到的候选标签；

[0156] 在匹配结果为没有匹配成功，并且当前匹配方式不是第二顺序中的最后一个匹配方式的情况下，按照第二顺序将下一个匹配方式确定为当前匹配方式，返回执行按照当前匹配方式将初始标签与同义词库中的候选标签进行匹配，得到匹配结果的步骤；

[0157] 在匹配结果为匹配成功的情况下，将匹配到的候选标签确定为目标候选标签；

[0158] 在匹配结果为没有匹配成功，并且当前匹配方式是第二顺序中的最后一个匹配方式的情况下，确定匹配失败。

[0159] 在本公开一些实施例中，匹配单元206具体可以用于：

[0160] 在当前匹配方式为模糊匹配的情况下，针对同义词库中每个候选标签，通过以下公式计算初始标签与候选标签的综合相似度值：

[0161] $$\text{Target} = \frac{a*\lambda_1 + b*\lambda_2 + c*\lambda_3 + d*\lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}$$

[0162] 其中，Target为综合相似度值，a为采用第一方式计算得到的相似度值，b为采用第二方式计算得到的相似度值，c为采用第三方式计算得到的相似度值，d为采用第四方式计算得到的相似度值，λ_1为第一方式对应的权重值，λ_2为第二方式对应的权重值，λ_3为第三方式对应的权重值，λ_4为第四方式对应的权重值；

[0163] 在综合相似度值满足第一预设条件的情况下，将候选标签确定为目标候选标签。

[0164] 在本公开一些实施例中，匹配单元206具体可以用于：

[0165] 通过以下公式计算相似度均值x：

[0166] $$x = \frac{\sum_1^n a\,\lambda_1 + \sum_1^n b\,\lambda_2 + \sum_1^n c\,\lambda_3 + \sum_1^n d\,\lambda_4}{0.5n(n+1)(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)}$$

[0167] 其中，n为同义词库中候选标签的数量；

[0168] 在Target<=x的情况下,采用多个第一区间确定候选标签的相似度等级;

[0169] 在Target>x的情况下,采用多个第二区间确定候选标签的相似度等级;

[0170] 在相似度等级满足第二预设条件的情况下,将候选标签确定为目标候选标签;其中,相似度等级满足第二预设条件的第一区间对应的区间范围大于相似度等级满足第二预设条件的第二区间对应的区间范围。

[0171] 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。

[0172] 根据本公开实施例提出的文本标签的标注装置,通过从待标注文本中获取目标分词,利用标签生成模型,根据标签库生成目标分词的初始标签,采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值,得到多个相似度值,对多个相似度值进行归一化处理,将归一化处理后的多个相似度值进行排序,得到第一顺序,按照第一顺序,确定多种匹配方式的第二顺序,依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配,直至得到至少一个目标候选标签为止,采用初始标签和至少一个目标候选标签对目标分词进行标注。通过二次生成标识的方式对文本的标签进行扩充,从而在提高对文本标注标签的效率的同时,提高了文本标签的全面性,进而能够提升检索或查询文本的准确性,使知识资源得到充分利用。

[0173] 图3是根据一示例性实施例示出的一种用于文本标签的标注方法的装置的框图。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,游戏控制台,平板设备,医疗设备,健身设备,个人数字助理等。

[0174] 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。

[0175] 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。

[0176] 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。

[0177] 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。

[0178] 多媒体组件308包括在装置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。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。

[0179] 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。

[0180] 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。

[0181] 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。

[0182] 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。

[0183] 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。

[0184] 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。

[0185] 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。

[0186] 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。

[0187] 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。

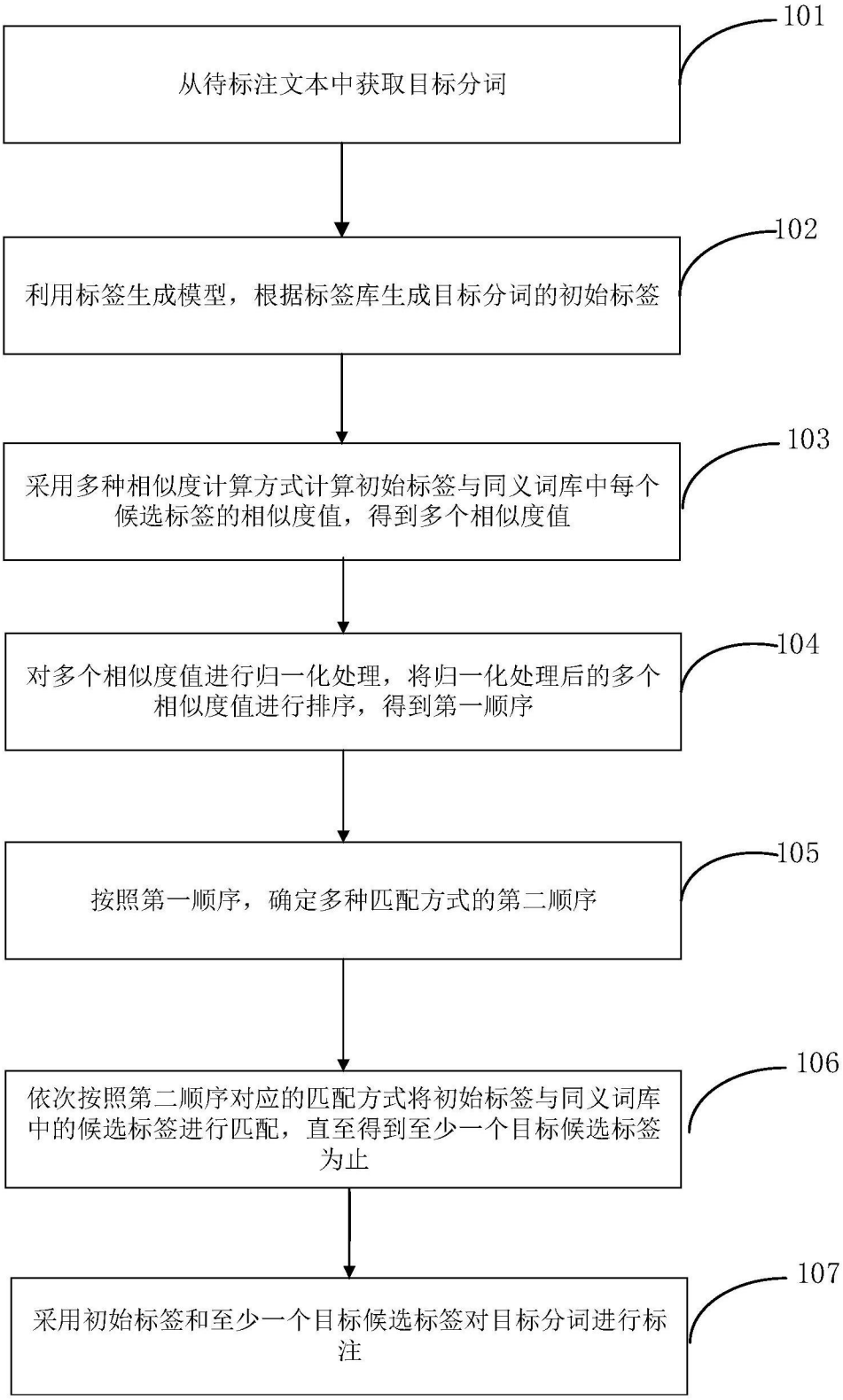

101
从待标注文本中获取目标分词
102
利用标签生成模型，根据标签库生成目标分词的初始标签
103
采用多种相似度计算方式计算初始标签与同义词库中每个候选标签的相似度值，得到多个相似度值
104
对多个相似度值进行归一化处理，将归一化处理后的多个相似度值进行排序，得到第一顺序
105
按照第一顺序，确定多种匹配方式的第二顺序
106
依次按照第二顺序对应的匹配方式将初始标签与同义词库中的候选标签进行匹配，直至得到至少一个目标候选标签为止
107
采用初始标签和至少一个目标候选标签对目标分词进行标注

图1

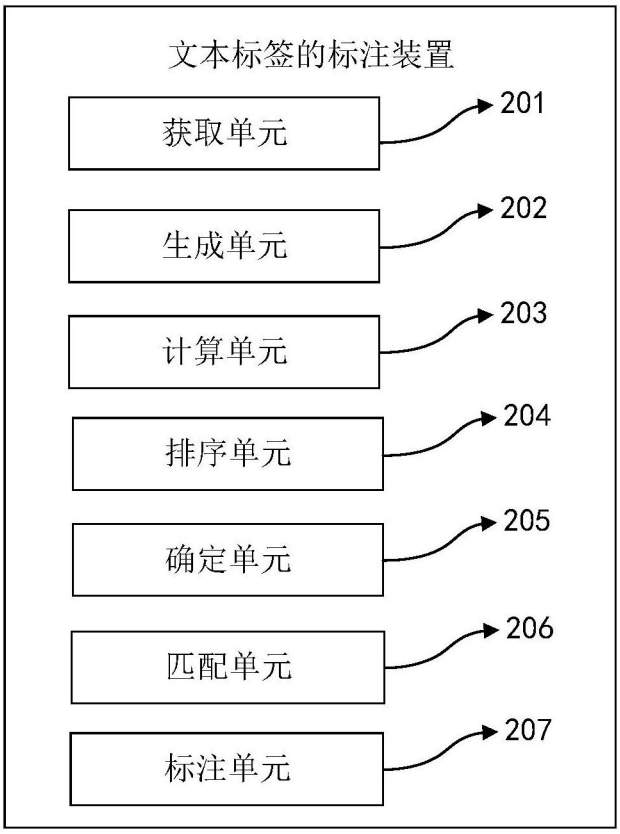

文本标签的标注装置
获取单元
201
生成单元
202
计算单元
203
排序单元
204
确定单元
205
匹配单元
206
标注单元
207

图2

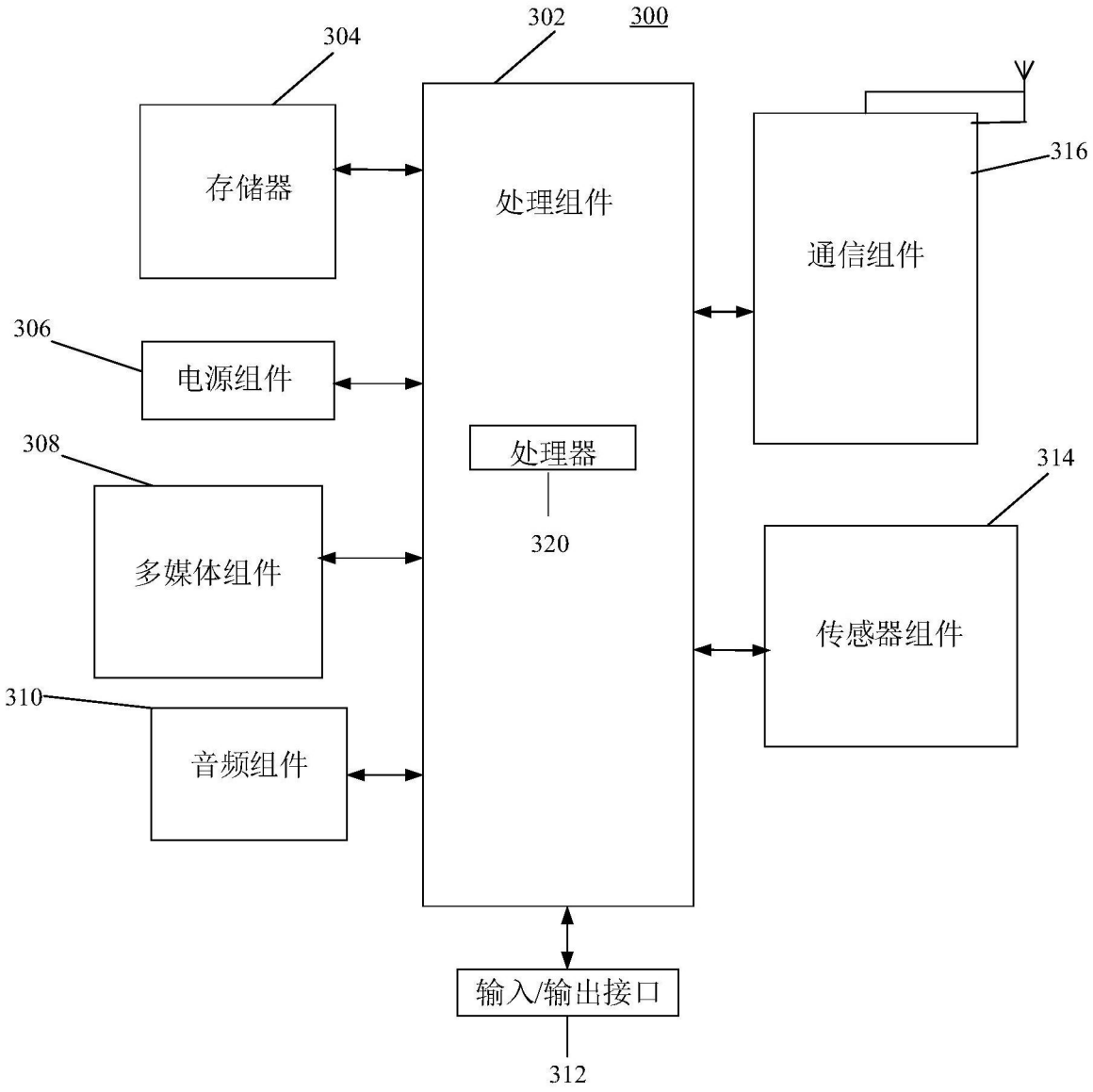

300
302
304
存储器
处理组件
316
通信组件
306
电源组件
处理器
320
308
多媒体组件
314
传感器组件
310
音频组件
输入/输出接口
312

图3