(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118797364 A
(43)申请公布日 2024.10.18

(21)申请号 202410969597.4

(22)申请日 2024.07.18

(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 谢丽莎

(51)Int.Cl.
G06F 18/22(2023.01)
G06F 16/33(2019.01)
G06F 40/289(2020.01)
G06N 5/02(2023.01)

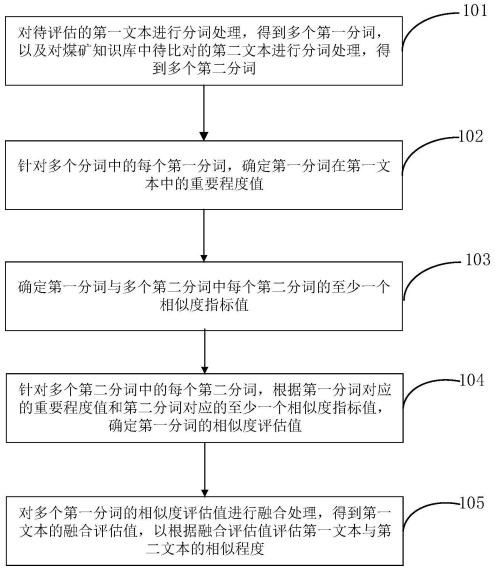
权利要求书2页 说明书11页 附图3页

(54)发明名称

煤矿知识库的文本相似度评估方法、装置及系统

(57)摘要

本公开是关于一种煤矿知识库的文本相似度评估方法、装置及系统,涉及自然语言的计算机处理技术领域。其中,方法包括:通过对待评估的第一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值。本方案对煤矿知识库中相似度较高的文本进行识别,提高文本的真实性和原创性。



CN 118797364 A

1.一种煤矿知识库的文本相似度评估方法,其特征在于,包括:

对待评估的第一文本进行分词处理,得到多个第一分词,以及对所述煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词;

针对所述多个分词中的每个第一分词,确定所述第一分词在所述第一文本中的重要程度值;

确定所述第一分词与所述多个第二分词中每个第二分词的至少一个相似度指标值;

针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值;

对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,以根据所述融合评估值评估所述第一文本与所述第二文本的相似程度。

2.根据权利要求1所述的煤矿知识库的文本相似度评估方法,其特征在于,所述对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,包括:

针对每个第一分词,确定所述第一分词对应的多个相似度评估值的平均值;所述多个相似度评估值与所述多个第二分词一一对应;

对所述多个第一分词各自对应的平均值进行融合处理,得到所述第一文本的融合评估值。

3.根据权利要求2所述的煤矿知识库的文本相似度评估方法,其特征在于,所述对所述多个第一分词各自对应的平均值进行融合处理,得到所述第一文本的融合评估值,包括:

通过以下公式对所述多个第一分词各自对应的平均值进行融合处理,得到所述融合评估值Value:

$$Value=\sqrt[x]{\prod_{1}^{x}Target[x]}$$

其中,x为任意一个第一分词,Target[x]为包括所述第一文本中全部第一分词的数组。

4.根据权利要求1所述的煤矿知识库的文本相似度评估方法,其特征在于,所述针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值,包括:

所述针对所述多个第二分词中的每个第二分词,对所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值进行加权求和,得到所述第一分词的相似度评估值。

5.根据权利要求1所述的煤矿知识库的文本相似度评估方法,其特征在于,所述根据所述融合评估值评估所述第一文本与所述第二文本的相似程度,包括:

根据预设的多个融合评估阈值区间,确定所述融合评估值的所属融合评估阈值区间;

确定所述所属融合评估阈值区间对应的预设级别;

在所述预设级别满足预设要条件的情况下,确定所述第一文本与所述第二文本相似。

6.根据权利要求4所述的煤矿知识库的文本相似度评估方法,其特征在于,在所述对所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值进行加权求和,得到所述第一分词的相似度评估值之前,方法还包括:

分别对每个第一分词和所述第一分词对应的重要程度值以键值对的形式存储至第一

哈希表中;

针对所述至少一个相似度指标值中的每个相似度指标值,将所述相似度指标值和所述相似度指标值对应的第一分词以键值对的形式存储至相应的第二哈希表中;所述第二哈希表存储有相同类型的相似度指标值;

所述根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值,包括:

从所述第一哈希表中查找与所述第一分词对应的所述重要程度值;

从至少一个第二哈希表中的每个第二哈希表中查找与所述第一分词对应的相似度指标值;

根据所述重要程度值和每个第二哈希表中与所述第一分词对应的相似度指标值,确定所述第一分词的相似度评估值。

7.一种煤矿知识库的文本相似度评估装置,其特征在于,包括:

分词单元,用于对待评估的第一文本进行分词处理,得到多个第一分词,以及对待比对的第二文本进行分词处理,得到多个第二分词;所述第二文本用于与所述第一文本进行相似度比对;

第一确定单元,用于针对所述多个分词中的每个第一分词,确定所述第一分词在所述第一文本中的重要程度值;

第二确定单元,用于确定所述第一分词与所述多个第二分词中每个第二分词的至少一个相似度指标值;

第三确定单元,用于针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值;

评估单元,用于对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,以根据所述融合评估值评估所述第一文本与所述第二文本的相似程度。

8.一种电子设备,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6中任一项所述的方法。

9.一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6中任一项所述的方法。

10.一种计算机程序产品,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6中任一项所述的方法。

煤矿知识库的文本相似度评估方法、装置及系统

技术领域

[0001] 本公开涉及自然语言的计算机处理技术领域,尤其涉及一种煤矿知识库的文本相似度评估方法、装置及系统。

背景技术

[0002] 相关技术中,随着信息化技术的迅速发展,知识库应用和管理系统成为了学术研究和技术开发的重要工具。这些系统存储了大量的研究报告、技术文档和数据资料,为科研人员提供了便捷的资源共享和知识管理平台。然而,随着知识库的传播和共享,知识库中的文件也面临着高相似度的问题,导致知识库中信息质量低,难以保证数据的真实性和原创性。

发明内容

[0003] 为克服相关技术中存在的问题,本公开提供一种煤矿知识库的文本相似度评估方法、装置及系统。

[0004] 根据本公开实施例的第一方面,提供一种煤矿知识库的文本相似度评估方法,包括:

[0005] 对待评估的第一文本进行分词处理,得到多个第一分词,以及对所述煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词;

[0006] 针对所述多个分词中的每个第一分词,确定所述第一分词在所述第一文本中的重要程度值;

[0007] 确定所述第一分词与所述多个第二分词中每个第二分词的至少一个相似度指标值;

[0008] 针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值;

[0009] 对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,以根据所述融合评估值评估所述第一文本与所述第二文本的相似程度。

[0010] 在本公开一些实施例中,所述对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,包括:

[0011] 针对每个第一分词,确定所述第一分词对应的多个相似度评估值的平均值;所述多个相似度评估值与所述多个第二分词一一对应;

[0012] 对所述多个第一分词各自对应的平均值进行融合处理,得到所述第一文本的融合评估值。

[0013] 在本公开一些实施例中,所述对所述多个第一分词各自对应的平均值进行融合处理,得到所述第一文本的融合评估值,包括:

[0014] 通过以下公式对所述多个第一分词各自对应的平均值进行融合处理,得到所述融合评估值Value:

[0015] $$Value=\sqrt[x]{\prod_{1}^{x} Target[x]}$$

[0016] 其中,x为任意一个第一分词,Target[x]为包括所述第一文本中全部第一分词的数组。

[0017] 在本公开一些实施例中,所述针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值,包括:

[0018] 所述针对所述多个第二分词中的每个第二分词,对所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值进行加权求和,得到所述第一分词的相似度评估值。

[0019] 在本公开一些实施例中,所述根据所述融合评估值评估所述第一文本与所述第二文本的相似程度,包括:

[0020] 根据预设的多个融合评估阈值区间,确定所述融合评估值的所属融合评估阈值区间;

[0021] 确定所述所属融合评估阈值区间对应的预设级别;

[0022] 在所述预设级别满足预设要条件的情况下,确定所述第一文本与所述第二文本相似。

[0023] 在本公开一些实施例中,在所述对所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值进行加权求和,得到所述第一分词的相似度评估值之前,方法还包括:

[0024] 分别对每个第一分词和所述第一分词对应的重要程度值以键值对的形式存储至第一哈希表中;

[0025] 针对所述至少一个相似度指标值中的每个相似度指标值,将所述相似度指标值和所述相似度指标值对应的第一分词以键值对的形式存储至相应的第二哈希表中;所述第二哈希表存储有相同类型的相似度指标值;

[0026] 所述根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值,包括:

[0027] 从所述第一哈希表中查找与所述第一分词对应的所述重要程度值;

[0028] 从至少一个第二哈希表中的每个第二哈希表中查找与所述第一分词对应的相似度指标值;

[0029] 根据所述重要程度值和每个第二哈希表中与所述第一分词对应的相似度指标值,确定所述第一分词的相似度评估值。

[0030] 根据本公开实施例的第二方面,提供一种煤矿知识库的文本相似度评估装置,包括:

[0031] 分词单元,用于对待评估的第一文本进行分词处理,得到多个第一分词,以及对待比对的第二文本进行分词处理,得到多个第二分词;所述第二文本用于与所述第一文本进行相似度比对;

[0032] 第一确定单元,用于针对所述多个分词中的每个第一分词,确定所述第一分词在所述第一文本中的重要程度值;

[0033] 第二确定单元,用于确定所述第一分词与所述多个第二分词中每个第二分词的至少一个相似度指标值;

[0034] 第三确定单元,用于针对所述多个第二分词中的每个第二分词,根据所述第一分词对应的所述重要程度值和所述第二分词对应的至少一个相似度指标值,确定所述第一分词的相似度评估值;

[0035] 评估单元,用于对所述多个第一分词的相似度评估值进行融合处理,得到所述第一文本的融合评估值,以根据所述融合评估值评估所述第一文本与所述第二文本的相似程度。

[0036] 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。

[0037] 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。

[0038] 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。

[0039] 本公开的实施例提供的技术方案可以包括以下有益效果:通过对待评估的第一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值,对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,从而根据融合评估值评估第一文本与第二文本的相似程度,进而能够对煤矿知识库中相似度较高的文本进行识别,提高文本的真实性和原创性。

[0040] 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

[0041] 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。

[0042] 图1是根据一示例性实施例示出的一种煤矿知识库的文本相似度评估方法的流程图。

[0043] 图2是根据一示例性实施例示出的一种煤矿知识库的文本相似度评估装置的框图。

[0044] 图3是根据一示例性实施例示出的一种用于煤矿知识库的文本相似度评估的装置的框图。

具体实施方式

[0045] 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。

[0046] 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。

[0047] 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”。

[0048] 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。

[0049] 相关技术中,随着信息化技术的迅速发展,知识库应用和管理系统成为了学术研究和技术开发的重要工具。这些系统存储了大量的研究报告、技术文档和数据资料,为科研人员提供了便捷的资源共享和知识管理平台。然而,随着知识库的传播和共享,知识库中的文件也面临着高相似度的问题,导致知识库中信息质量低,难以保证数据的真实性和原创性。

[0050] 为了解决上述问题,本公开提供了一种煤矿知识库的文本相似度评估方法、装置及系统,通过对待评估的第一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值,对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,从而根据融合评估值评估第一文本与第二文本的相似程度,进而能够对煤矿知识库中相似度较高的文本进行识别,提高文本的真实性和原创性。

[0051] 图1是根据一示例性实施例示出的一种煤矿知识库的文本相似度评估方法、装置及系统方法的流程图,如图1所示,需要说明的是,本公开实施例的煤矿知识库的文本相似度评估方法、装置及系统方法应用于煤矿知识库的文本相似度评估方法、装置及系统装置中。如图1所示,该方法可以包括以下步骤:

[0052] 步骤101,对待评估的第一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词。

[0053] 可以理解的是,第一文本为需要进行相似度评估的文本,第二文本为第一文本比对的对象。

[0054] 在一个实施例中,第一文本可以是煤矿知识库中的文本,也可以是其他文本。

[0055] 在本公开实施例中,对待评估的第一文本以及待比对的第二文本进行分词处理,得到第一文本对应的多个第一分词,以及第二文本对应的多个第二分词。

[0056] 需要说明的是,第一分词和第二分词均可以是单词,也可以是词组。

[0057] 在本公开一些实施例中,步骤101具体可以包括以下步骤:

[0058] 步骤a1,对待评估的第一文本进行分词处理,得到第一分词结果。

[0059] 步骤a2,从第一分词结果中去除词性为虚词的分词,将除词性为虚词的分词以外的其他分词确定为第一分词。

[0060] 在本公开实施例中,在得到第一分词结果后,可以对第一分词结果中的每个分词进行词性标注,从而区分动词、名词、定语和虚词。根据标注的磁性,从第一分词结果中去除性为虚词的分词,如“的”、“是”等,将除词性为虚词的分词以外的其他分词确定为第一分词,从而降低干扰,提高文本信息率。

[0061] 步骤a3,对待评估的第二文本进行分词处理,得到第二分词结果。

[0062] 步骤a4,从第二分词结果中去除词性为虚词的分词,将除词性为虚词的分词以外的其他分词确定为第二分词。

[0063] 在本公开实施例中,在得到第二分词结果后,可以对第二分词结果中的每个分词进行词性标注,从而区分动词、名词、定语和虚词。根据标注的磁性,从第二分词结果中去除性为虚词的分词,如“的”、“是”等,将除词性为虚词的分词以外的其他分词确定为第二分词,从而降低干扰,提高文本信息率。

[0064] 步骤102,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值。

[0065] 在本公开一些实施例中,可以采用TF-IDF(Term Frequency-Inverse Document Frequency,术语频率-反向文档频率)计算第一分词在第一文本中的重要程度值。

[0066] 需要说明的是,TF-IDF是一种用于资讯检索与文本挖掘的常用加权技术。TF-IDF是一种统计方法,用以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。

[0067] 步骤103,确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值。

[0068] 在本公开一些实施例中,至少一个相似度指标值包括以下一种或者多种:

[0069] 编辑距离指标值,余弦相似度指标值,杰卡德Jaccard相似系数。

[0070] 在一个实施例中,编辑距离指标值可以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。例如,"kitten"转换为"sitting"至少需要3次编辑操作。

[0071] 在一个实施例中,余弦相似度是将文本表示为向量,计算两个向量之间的余弦相似度。余弦相似度(Cosine Similarity)是一种用于衡量两个非零向量间夹角余弦值的方法。它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。

[0072] 需要说明的是,第一文本包括多个第一分词,可以将每个第一分词均与全部第二分词分别进行相似度比较,也就是说,任意一个第一分词均与全部第二分词分别进行相似度比较,比较相似度的方式可以是计算编辑距离指标值,余弦相似度指标值,杰卡德Jaccard相似系数中的任意一种或者多种。

[0073] 步骤104,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值。

[0074] 在本公开一些实施例中,步骤104具体可以包括以下步骤:针对多个第二分词中的每个第二分词,对第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值进行加权求和,得到第一分词的相似度评估值。

[0075] 在一个实施例中,可以通过以下公式计算第一分词的相似度评估值V(n):

[0076] V(n)=(tf_hash(Test[n])*P1+d_hash(Test[n])*P2+cos_hash(Test[n])*P3+jac_hash(Test[n])*P4)/(P1+P2+P3+P4)

[0077] 其中,tf_hash(Test[n])为第一哈希表中第二分词n对应的重要程度值,d_hash(Test[n])为编辑距离指标值对应的第二哈希表中的第一相似度评估值,cos_hash(Test[n])为余弦相似度指标值对应的第二哈希表中的第二相似度评估值,jac_hash(Test[n])为Jaccard相似系数对应的第二哈希表中的第三相似度评估值,P1为重要程度值对应的权重值,P2为第一相似度评估值对应的权重值,P3为第二相似度评估值对应的权重值,P4为第三相似度评估值对应的权重值。

[0078] 在本公开一些实施例中,在步骤104之前,方法还可以包括以下步骤:

[0079] 分别对每个第一分词和第一分词对应的重要程度值以键值对的形式存储至第一哈希表中;

[0080] 针对至少一个相似度指标值中的每个相似度指标值,将相似度指标值和相似度指标值对应的第一分词以键值对的形式存储至相应的第二哈希表中;第二哈希表存储有相同类型的相似度指标值;

[0081] 根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值,包括:

[0082] 从第一哈希表中查找与第一分词对应的重要程度值;

[0083] 从至少一个第二哈希表中的每个第二哈希表中查找与第一分词对应的相似度指标值;

[0084] 根据重要程度值和每个第二哈希表中与第一分词对应的相似度指标值,确定第一分词的相似度评估值。

[0085] 可以理解的是,由于第一文本中包括多个第一分词,因此每个第一分词对应多个重要程度值,另外,由于第二文本中包括多个第二分词,因此每个第一分词对应的第二分词均对应有多个相似度指标值。

[0086] 为了在计算相似度评估值时能够高效的查找到当前第一分词对应的重要程度值和相似度指标值,可以将分别对每个第一分词和第一分词对应的重要程度值以键值对的形式存储至第一哈希表中针对至少一个相似度指标值中的每个相似度指标值,另外,将相似度指标值和相似度指标值对应的第一分词以键值对的形式存储至相应的第二哈希表中。其中,第二哈希表存储有相同类型的相似度指标值;

[0087] 在一个实施例中,在需要计算第一分词的相似度评估值时,可以从第一哈希表中查找与第一分词对应的重要程度值,从至少一个第二哈希表中的每个第二哈希表中查找与第一分词对应的相似度指标值,根据重要程度值和每个第二哈希表中与第一分词对应的相似度指标值,确定第一分词的相似度评估值。

[0088] 步骤105,对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,以根据融合评估值评估第一文本与第二文本的相似程度。

[0089] 可以理解的是,由于第一文本中包括多个第一分词,因此,在得到每个第一分词的相似度评估值后,可以将多个相似度评估值进行融合处理,得到第一文本的融合评估值,以根据融合评估值评估第一文本与第二文本的相似程度。

[0090] 在本公开一些实施例中,步骤105中的对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,具体可以包括以下步骤:

[0091] 步骤a1,针对每个第一分词,确定第一分词对应的多个相似度评估值的平均值。

[0092] 多个相似度评估值与多个第二分词一一对应。

[0093] 在一个实施例中,可以通过以下公式计算上述评估值R:

[0094] $$R=\frac{\sum V(n)}{\mathrm{sum}(V(n))}$$

[0095] 其中,V(n)为第一分词的相似度评估值。

[0096] 步骤a2,对多个第一分词各自对应的平均值进行融合处理,得到第一文本的融合评估值。

[0097] 在本公开一些实施例中,步骤a2具体可以包括以下步骤:

[0098] 通过以下公式对多个第一分词各自对应的平均值进行融合处理,得到融合评估值Value:

[0099] $$Value=\sqrt[x]{\prod_{1}^{x}\mathrm{Target}[x]}$$

[0100] 其中,x为任意一个第一分词,Target[x]为包括第一文本中全部第一分词的数组。

[0101] 在本公开一些实施例中,步骤105中的根据融合评估值评估第一文本与第二文本的相似程度,具体可以包括以下步骤:

[0102] 步骤b1,根据预设的多个融合评估阈值区间,确定融合评估值的所属融合评估阈值区间。

[0103] 步骤b2,确定所属融合评估阈值区间对应的预设级别。

[0104] 步骤b3,在预设级别满足预设要条件的情况下,确定第一文本与第二文本相似。

[0105] 可以理解的是,可以预先根据实际需求划分融合评估阈值区间,按照融合评估阈值区间确定融合评估值的级别,根据级别大小来判断第一文本是否与第二文本相似,从而能够进一步提高判断效率。

[0106] 举例来说,可以预先设定评估等级如表1所示:

[0107] 表1相似度等级评估表

[0108]

相似度等级	1相似度低	2相似度较低	3相似度较高	4相似度高
测试结果	[0,0.2)	[0.2,0.45)	[0.45,0.75)	[0.75,1.0]

[0109] 当对比结果测定相似度等级为3时,输出针对该文本的警告信息,提示疑似高相似度声明,当对比结果测定相似度等级为4时,知识库判定该第一文本与第二文本高度相似,确定第一文本为非原创文本。

[0110] 根据本公开实施例提出的煤矿知识库的文本相似度评估方法,通过对待评估的第

一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值,对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,从而根据融合评估值评估第一文本与第二文本的相似程度,进而能够对煤矿知识库中相似度较高的文本进行识别,提高文本的真实性和原创性。

[0111] 图2是根据一示例性实施例示出的一种煤矿知识库的文本相似度评估装置框图。参照图2,该装置包括分词单元201,第一确定单元202,第二确定单元203,第三确定单元204和评估单元205。

[0112] 其中,分词单元201,用于对待评估的第一文本进行分词处理,得到多个第一分词,以及对待比对的第二文本进行分词处理,得到多个第二分词;第二文本用于与第一文本进行相似度比对;

[0113] 第一确定单元202,用于针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;

[0114] 第二确定单元203,用于确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值;

[0115] 第三确定单元204,用于针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值;

[0116] 评估单元205,用于对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,以根据融合评估值评估第一文本与第二文本的相似程度。

[0117] 在本公开一些实施例中,评估单元205具体可以用于:

[0118] 针对每个第一分词,确定第一分词对应的多个相似度评估值的平均值;多个相似度评估值与多个第二分词一一对应;

[0119] 对多个第一分词各自对应的平均值进行融合处理,得到第一文本的融合评估值。

[0120] 在本公开一些实施例中,评估单元205具体可以用于:

[0121] 通过以下公式对多个第一分词各自对应的平均值进行融合处理,得到融合评估值Value:

[0122] $$Value=\sqrt[x]{\prod_{1}^{x} Target[x]}$$

[0123] 其中,x为任意一个第一分词,Target[x]为包括第一文本中全部第一分词的数组。

[0124] 在本公开一些实施例中,第三确定单元204具体可以用于:针对多个第二分词中的每个第二分词,对第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值进行加权求和,得到第一分词的相似度评估值。

[0125] 在本公开一些实施例中,评估单元205具体可以用于:

[0126] 根据预设的多个融合评估阈值区间,确定融合评估值的所属融合评估阈值区间;

[0127] 确定所属融合评估阈值区间对应的预设级别;

[0128] 在预设级别满足预设要条件的情况下,确定第一文本与第二文本相似。

[0129] 在本公开一些实施例中,装置还可以包括:

[0130] 第一存储单元,用于分别对每个第一分词和第一分词对应的重要程度值以键值对的形式存储至第一哈希表中;

[0131] 第二存储单元,用于针对至少一个相似度指标值中的每个相似度指标值,将相似度指标值和相似度指标值对应的第一分词以键值对的形式存储至相应的第二哈希表中;第二哈希表存储有相同类型的相似度指标值;

[0132] 第三确定单元204具体可以用于:

[0133] 从第一哈希表中查找与第一分词对应的重要程度值;

[0134] 从至少一个第二哈希表中的每个第二哈希表中查找与第一分词对应的相似度指标值;

[0135] 根据重要程度值和每个第二哈希表中与第一分词对应的相似度指标值,确定第一分词的相似度评估值。

[0136] 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。

[0137] 根据本公开实施例提出的煤矿知识库的文本相似度评估方法,通过对待评估的第一文本进行分词处理,得到多个第一分词,以及对煤矿知识库中待比对的第二文本进行分词处理,得到多个第二分词,针对多个分词中的每个第一分词,确定第一分词在第一文本中的重要程度值;确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值,针对多个第二分词中的每个第二分词,根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值,确定第一分词的相似度评估值,对多个第一分词的相似度评估值进行融合处理,得到第一文本的融合评估值,从而根据融合评估值评估第一文本与第二文本的相似程度,进而能够对煤矿知识库中相似度较高的文本进行识别,提高文本的真实性和原创性。

[0138] 图3是根据一示例性实施例示出的一种用于煤矿知识库的文本相似度评估的装置的框图。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。

[0139] 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。

[0140] 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。

[0141] 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。

[0142] 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。

[0143] 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。

[0144] 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。

[0145] 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。

[0146] 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。

[0147] 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。

[0148] 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。

[0149] 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。

[0150] 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。

[0151] 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。

[0152] 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。

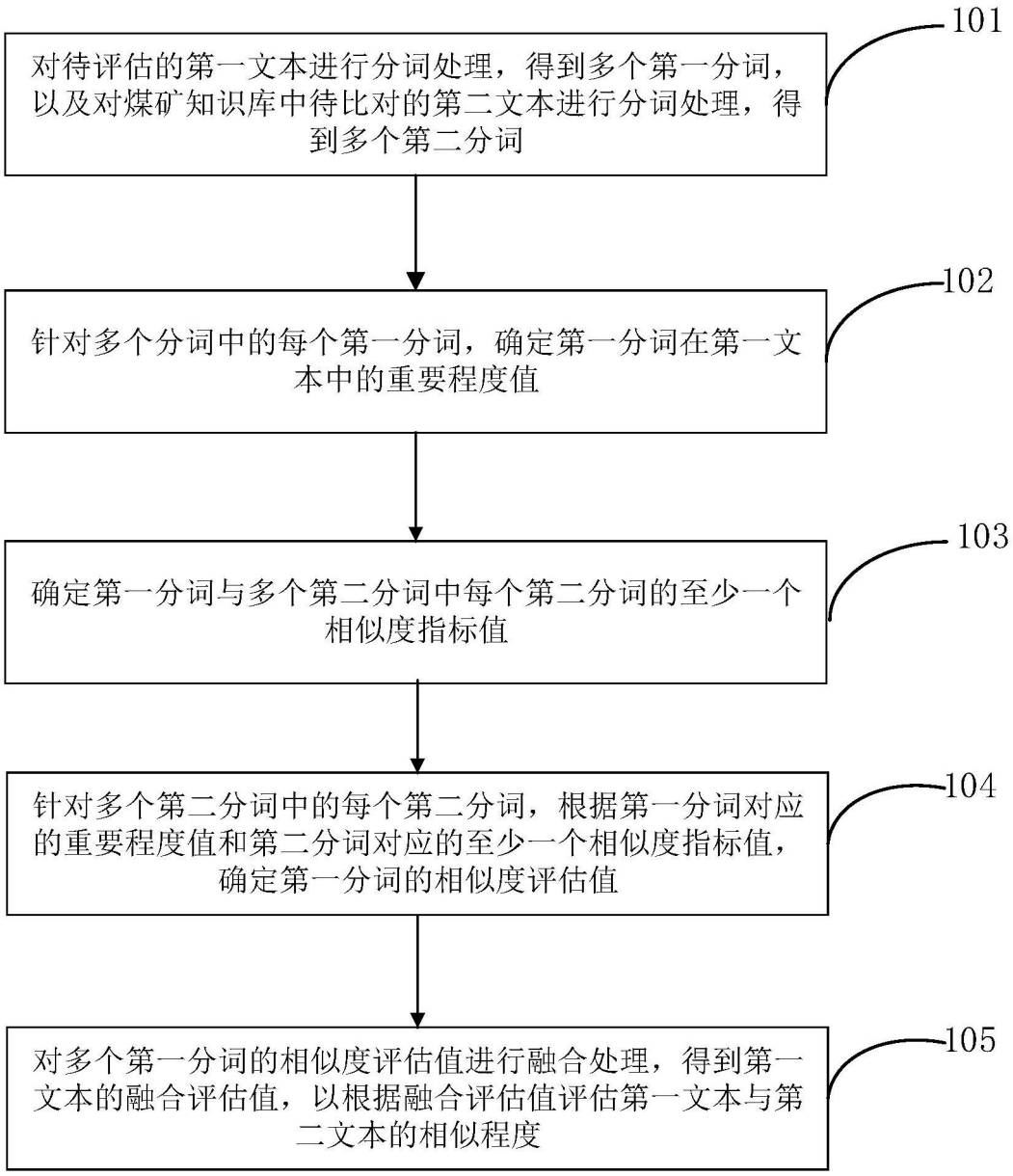

101
对待评估的第一文本进行分词处理，得到多个第一分词，以及对煤矿知识库中待比对的第二文本进行分词处理，得到多个第二分词
102
针对多个分词中的每个第一分词，确定第一分词在第一文本中的重要程度值
103
确定第一分词与多个第二分词中每个第二分词的至少一个相似度指标值
104
针对多个第二分词中的每个第二分词，根据第一分词对应的重要程度值和第二分词对应的至少一个相似度指标值，确定第一分词的相似度评估值
105
对多个第一分词的相似度评估值进行融合处理，得到第一文本的融合评估值，以根据融合评估值评估第一文本与第二文本的相似程度

图1

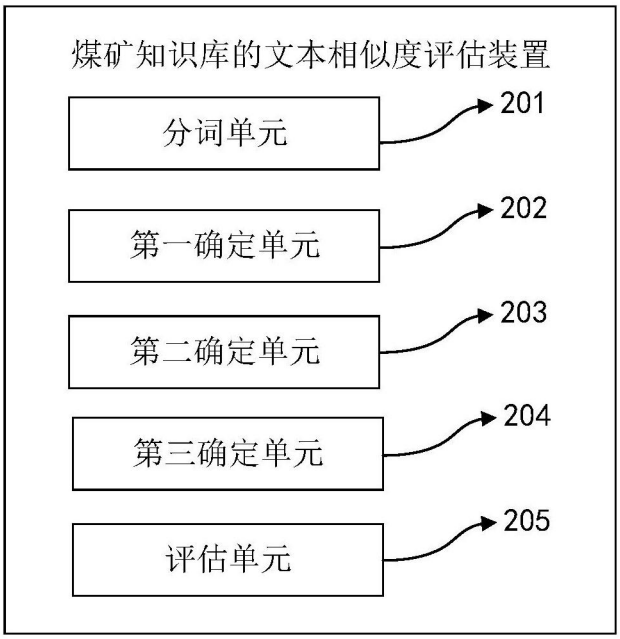

煤矿知识库的文本相似度评估装置
分词单元
201
第一确定单元
202
第二确定单元
203
第三确定单元
204
评估单元
205

图2

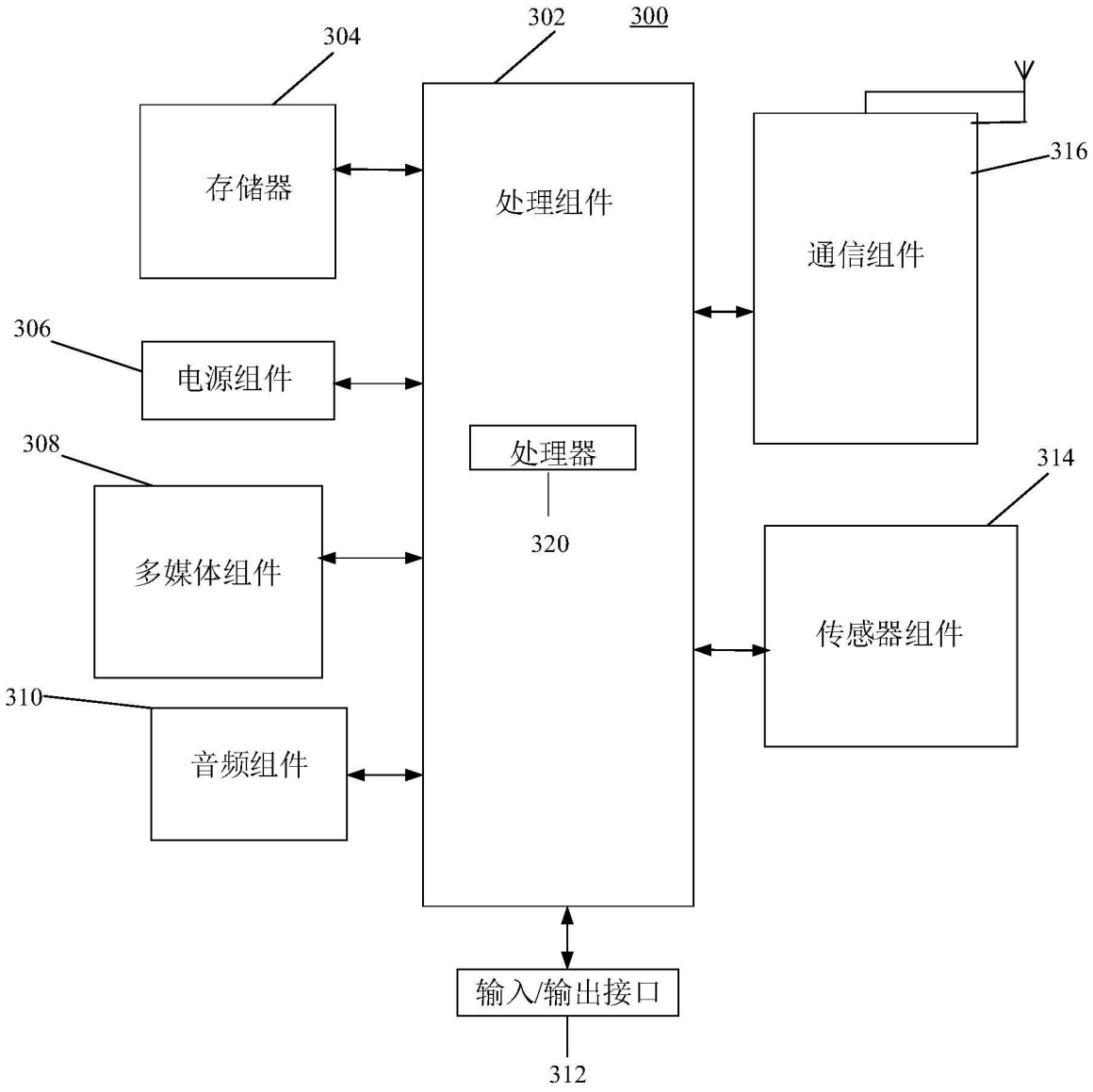

302
300
304
存储器
处理组件
316
通信组件
306
电源组件
处理器
320
308
多媒体组件
314
传感器组件
310
音频组件
输入/输出接口
312

图3