(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 112417154 A
(43)申请公布日 2021.02.26

(21)申请号 202011336796.X

(22)申请日 2020.11.25

(71)申请人 上海创米科技有限公司
地址 200241 上海市闵行区紫星路588号1幢11层001A室

(72)发明人 秦泓杰

(74)专利代理机构 北京市一法律师事务所 11654
代理人 刘荣娟

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/335(2019.01)

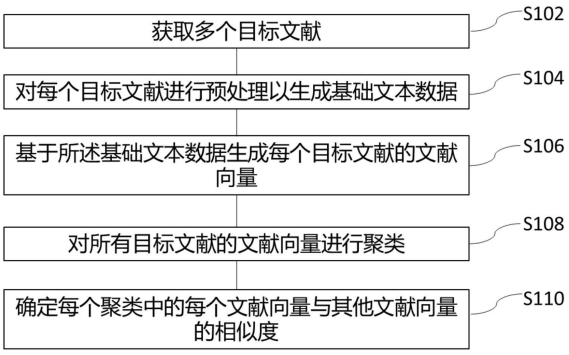
权利要求书1页 说明书10页 附图3页

(54)发明名称

确定文献相似度的方法和装置

(57)摘要

本申请公开了确定文献相似度的方法和装置。该方法包括:获取多个目标文献;对每个目标文献进行预处理以生成基础文本数据;基于所述基础文本数据生成每个目标文献的文献向量;对所有目标文献的文献向量进行聚类;以及确定每个聚类中的每个文献向量与其他文献向量的相似度。该方法和装置采用无监督模型,将词及文献分步映射到同一语义空间,通过聚类分析,缩小了相似文献的备选集,从而同时提高了性能及准确度。



CN 112417154 A

1.一种确定文献相似度的方法,其特征在于,包括:

获取多个目标文献;

对每个目标文献进行预处理以生成基础文本数据;

基于所述基础文本数据生成每个目标文献的文献向量;

对所有目标文献的文献向量进行聚类;以及

确定每个聚类中的每个文献向量与其他文献向量的相似度。

2.如权利要求1所述的方法,其特征在于,基于所述基础文本数据生成每个目标文献的文献向量包括:

基于所述基础文本数据生成词向量文本数据和文献向量文本数据;

基于所述词向量文本数据,通过词向量模型生成词向量;以及

基于所述文献向量文本数据和所述词向量,通过文献向量模型生成每个目标文献的文献向量。

3.如权利要求2所述的方法,其特征在于,所述基础文本数据包括以下中的至少一个:文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。

4.如权利要求2所述的方法,其特征在于,所述词向量文本数据包括文献标题和文献摘要,所述文献向量文本数据包括文献标题和文献关键字。

5.如权利要求2所述的方法,其特征在于,所述词向量模型为连续词袋模型,所述文献向量模型为跳字模型。

6.一种确定相似文献的装置,其特征在于,包括:

文献获取单元,被配置为获取多个目标文献;

预处理单元,被配置为对每个目标文献进行预处理以生成基础文本数据;

文献向量生成单元,被配置为基于所述基础文本数据生成每个目标文献的文献向量;

聚类单元,被配置为对所有目标文献的文献向量进行聚类;以及

相似度确定单元,被配置为确定每个聚类中的每个文献向量与其他文献向量的相似度。

7.如权利要求6所述的装置,其特征在于,基于所述基础文本数据生成每个目标文献的文献向量包括:

基于所述基础文本数据生成词向量文本数据和文献向量文本数据;

基于所述词向量文本数据,通过词向量模型生成词向量;以及

基于所述文献向量文本数据和所述词向量,通过文献向量模型生成每个目标文献的文献向量。

8.如权利要求7所述的装置,其特征在于,所述基础文本数据包括以下中的至少一个:文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。

9.如权利要求7所述的装置,其特征在于,所述词向量文本数据包括文献标题和文献摘要,所述文献向量文本数据包括文献标题和文献关键字。

10.如权利要求7所述的装置,其特征在于,所述词向量模型为连续词袋模型,所述文献向量模型为跳字模型。

确定文献相似度的方法和装置

技术领域

[0001] 本公开涉及大数据信息处理技术领域，尤其涉及确定文献相似度的方法和装置。

背景技术

[0002] 随着科学技术的飞速发展，电子、机械、计算机、生化、医药等领域的研究成果发布周期越来越短，各学科文献的数目极速增长。通过查阅相关科技文献，可以了解当前研究领域内的主要研究成果、同行研究动态、该领域内已解决的问题及有待于改进和完善的问题等，从而进一步明确研究课题的科学价值，找准研究的真正起点。

[0003] 文献数量的激增，一方面表明文献信息资源的丰富，但同时也产生了"文献信息污染"，给人们选择、利用文献造成了障碍。因此，面对日益增长的文献资源，如何快捷准确地获取感兴趣的文献，已成为人们关注的热点问题。故而，搜索和推荐相似文献，在学术上起着举足轻重的作用。

[0004] 因此，需要一种确定文献相似度的方法和装置。

发明内容

[0005] 本公开的目的在于提出一种基于无监督模型确定文献相似度的方法和装置，以解决现有在计算文献相似度时，方法复杂、数据庞大、性能低下、准确度低的问题。

[0006] 为达上述目的，本公开的一个方面提供了一种确定文献相似度的方法，其包括：获取多个目标文献；对每个目标文献进行预处理以生成基础文本数据；基于所述基础文本数据生成每个目标文献的文献向量；对所有目标文献的文献向量进行聚类；以及确定每个聚类中的每个文献向量与其他文献向量的相似度。

[0007] 可选地，基于所述基础文本数据生成每个目标文献的文献向量包括：基于所述基础文本数据生成词向量文本数据和文献向量文本数据；基于所述词向量文本数据，通过词向量模型生成词向量；以及基于所述文献向量文本数据和所述词向量，通过文献向量模型生成每个目标文献的文献向量。

[0008] 可选地，所述基础文本数据包括以下中的至少一个：文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。

[0009] 可选地，所述词向量文本数据包括文献标题和文献摘要，所述文献向量文本数据包括文献标题和文献关键字。

[0010] 可选地，所述词向量模型为连续词袋模型，所述文献向量模型为跳字模型。

[0011] 本公开的另一个方面提供了一种确定相似文献的装置，其包括：文献获取单元，被配置为获取多个目标文献；预处理单元，被配置为对每个目标文献进行预处理以生成基础文本数据；文献向量生成单元，被配置为基于所述基础文本数据生成每个目标文献的文献向量；聚类单元，被配置为对所有目标文献的文献向量进行聚类；以及相似度确定单元，被配置为确定每个聚类中的每个文献向量与其他文献向量的相似度。

[0012] 可选地，基于所述基础文本数据生成每个目标文献的文献向量包括：基于所述基

础文本数据生成词向量文本数据和文献向量文本数据;基于所述词向量文本数据,通过词向量模型生成词向量;以及基于所述文献向量文本数据和所述词向量,通过文献向量模型生成每个目标文献的文献向量。

[0013] 可选地,所述基础文本数据包括以下中的至少一个:文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。

[0014] 可选地,所述词向量文本数据包括文献标题和文献摘要,所述文献向量文本数据包括文献标题和文献关键字。

[0015] 可选地,所述词向量模型为连续词袋模型,所述文献向量模型为跳字模型。

[0016] 本公开的又一个方面提供了一种计算设备,其包括:至少一个存储介质,存储有至少一组指令;以及至少一个处理器,同所述至少一个存储介质通讯连接,其中,当所述至少一个处理器运行所述至少一组指令时,所述至少一个处理器执行前述方法。

[0017] 本公开的一个或多个实施例提出的确定文献相似度的方法和装置采用无监督模型,将词及文献分步映射到同一语义空间,通过聚类分析,缩小了相似文献的备选集,从而同时提高了性能及准确度。

[0018] 另外,本公开的一个或多个实施例提出的确定文献相似度的方法和装置仅依赖于文献向量,而文献向量又只依赖于训练好的词向量模型,与文献数量无关,所以即使面对海量文献,也依然适用。

附图说明

[0019] 以下附图详细描述了本公开中披露的示例性实施例。其中相同的附图标记在附图的若干视图中表示类似的结构。本领域的一般技术人员将理解这些实施例是非限制性的、示例性的实施例,附图仅用于说明和描述的目的,并不旨在限制本公开的范围,其他方式的实施例也可能同样的完成本公开中的构思意图。应当理解,附图未按比例绘制。其中:

[0020] 图1为根据本公开一个或多个实施例的确定文献相似度的方法的流程图;

[0021] 图2为根据本公开一个或多个实施例的基于基础文本数据生成文献向量的流程图;

[0022] 图3为根据本公开一个或多个实施例的确定文献相似度的装置的示意图;

[0023] 图4为根据本公开一个或多个实施例的计算设备的示意图。

具体实施方式

[0024] 以下描述提供了本公开的特定应用场景和要求,目的是使本领域技术人员能够制造和使用本公开中的内容。对于本领域技术人员来说,对所公开的实施例的各种局部修改是显而易见的,并且在不脱离本公开的精神和范围的情况下,可以将这里定义的一般原理应用于其他实施例和应用。因此,本公开不限于所示的实施例,而是具有与权利要求一致的最宽范围。

[0025] 本领域技术人员将理解,本公开中使用的术语仅用于描述特定示例实施例的目的,而不是限制性的。比如,除非上下文另有明确说明,这里所使用的,单数形式"一"、"一个"、"该"和"所述"也可以包括复数形式。当在本公开中使用时,术语"包括"、"包含"、"具有"、"含有"、"配备有"和/或"设置有"意思是指所关联的整数、步骤、操作、元素、组件和/或

组的存在，但不排除一个或多个其他特征、整数、步骤、操作、元素、组件和/或组的存在，或在该系统/方法中可以添加其他特征、整数、步骤、操作、元素、组件和/或组。

[0026] 本领域技术人员将理解，特定术语已被用于描述本公开的实施例。例如，"实施例"、"一个实施例"、"一些实施例"、"多个实施例"和/或"若干实施例"意味着结合该实施例描述的特定特征、结构或特性可以包括在本公开的至少一个实施例中。因此，可以强调并且应当理解，在本公开的各个部分中对"实施例"或"替代实施例"的两个或更多个引用不一定都指代相同的实施例。此外，特定特征、结构或特性可以在本公开的一个或多个实施例中适当地组合。

[0027] 本领域技术人员将理解，除非另外指定，序数形容词"第一"、"第二"、"第三"等用于描述普通对象仅指示被提及的相像对象的不同实例，而不旨在暗示这样描述的对象必须在时间上、空间上、按排名或以任意其他方式按给定顺序。

[0028] 本领域技术人员将理解，本公开的方面可以在许多可获得专利的类别或内容中的任何一个中示出和描述，这些类别或内容包括任何新的和有用的过程、机器、制造或物质的组合物，或其任何新的和有用的改进。因此，本公开的各方面可以完全由硬件(电路、芯片、逻辑器件等)，完全由软件(包括固件、常驻软件、微代码等)或软硬件组合来实现，这些实现在本文中通常都称为"块"、"模块"、"引擎"、"单元"、"组件"、或"系统"。此外，本公开的各方面可以采取体现在一个或多个计算机可读介质中的计算机程序产品的形式，该计算机可读介质包含在其上具现化的计算机可读程序代码。

[0029] 本领域技术人员将理解，本公开中的算法通常被认为是通向期望结果的自相一致的一系列动作或操作。这些动作或操作包括物理量的物理操纵。通常，但不是必要的，这些量采取能够被存储、转移、组合、比较且以其他方式操纵的电或磁信号的形式。主要是由于共用的原因，有时已经证明便利的是将这些信号称为位、值、元素、标记、字符、术语、数字诸如此类。然而，应理解，所有这些和类似术语与适当的物理量关联，并且仅是应用于这些量的方便标记。

[0030] 本领域技术人员将理解，本公开中关于"处理"、"运算"、"计算"、"确定"、"创建"、"分析"、"检查"等的讨论可以指计算机、计算平台、计算系统或其他电子计算设备的操作和/或处理，这些设备将被表示为计算机的寄存器和/或存储器内的物理(例如电子)量的数据操纵和/或变换成被类似地表示为计算机的寄存器和/或存储器或可以存储执行操作和/或处理的指令的其他信息存储介质内的物理量的其他数据。

[0031] 传统的文本检索系统或者搜索引擎都是基于用户输入的关键字进行搜索，以关键字与文献间的匹配程度为依据返回结果。这种搜索方式虽然在一定程度上能够解决文献相似度计算问题，但当用户对信息精度的需求较高时，基于关键字匹配的计算方式就无法满足用户需求，相似度计算结果也不够准确。文献相似度计算是指：对输入的每对文献之间的相似程度进行量化(例如，给出一个分值)。在此基础上，可以进一步应用于搜索引擎、推荐系统、问答系统、筛选定位等领域中。

[0032] 图1为根据本公开一个或多个实施例的确定文献相似度的方法的流程图。

[0033] 如图1所示，确定文献相似度的方法可包括步骤S102、步骤S104、步骤S106、步骤S108和步骤S110。

[0034] 步骤S102：获取多个目标文献。

[0035] 例如，所述目标文献可以是任何类型的记录知识的载体。所述目标文献可以来自于图书、报刊、网络、数据库等。

[0036] 步骤S104：对每个目标文献进行预处理以生成基础文本数据。

[0037] 所述预处理可包括可用字段提取、数据清洗、数据分词、停用词剔除等。

[0038] 所述可用字段提取指的是提取所述目标文献记录的可用字段数据(又可称为有效字段数据)。所述可用字段数据是具有实际意义或者能够从中提取特征的文本。所述可用字段数据例如可以包括：文献标题、文献摘要、文献关键词、文献正文、文献作者、文献分类号等。

[0039] 所述数据清洗是从所述可用字段数据中检测和纠正(或删除)损坏或不准确的记录的过程。例如，可通过所述数据清洗识别所述可用字段数据中的不完整(例如，缺词少句)、不正确(例如，错别字)、不准确或不相关(例如，脚本)的部分，然后替换、修改、或删除脏数据或粗数据。

[0040] 所述数据分词是将所述可用字段数据中的句子、段落、文章分解为以字词为单位的数据结构，以便后续的处理分析。

[0041] 所述停用词剔除根据预先设定的停用词表将所述可用字段数据中的某些字或词过滤或剔除，以节省存储空间和提高后续语言处理的效率。

[0042] 步骤S106：基于所述基础文本数据生成每个目标文献的文献向量。

[0043] 图2为根据本公开一个或多个实施例的基于基础文本数据生成文献向量的流程图。

[0044] 如图2所示，步骤S106可进一步包括子步骤S1062、子步骤S1064和子步骤S1066。

[0045] 子步骤S1062：基于所述基础文本数据生成词向量文本数据和文献向量文本数据。

[0046] 例如，可根据应用场景需求，重新组织所述基础文本数据中的文本数据，以生成词向量模型文本数据和文献向量文本数据，二者又可称为语料库。

[0047] 在一些实施例中，可基于所生成的词向量模型文本数据生成用于词向量模型的训练数据集。假设所述词向量模型文本数据如表1所示。

[0048] 表1

[0049]

我	家	兔子	爱	吃	胡萝卜
你	家	兔子	爱	吃	胡萝卜
谁	家	兔子	爱	吃	胡萝卜

[0050] 在一些实施例中，生成所述训练数据集可包括如下步骤：

[0051] (1)统计所述词向量模型文本数据中的单词的词频，基于所述词频构建词汇表，所述词汇表包括每个单词的词频以及每个单词的索引；

[0052] (2)定义上下文单词；

[0053] (3)设置低频单词过滤；

[0054] (4)逐行遍历所述词向量模型文本数据，以生成单词对(又称为训练数据正样本)；以及

[0055] (5)通过算法(例如，洗牌算法)将单词对随机打乱。

[0056] 在一些实施例中，根据表1中单词的词频构造的词汇表可如表2所示。

[0057] 表2

[0058]

单词	词频	索引
兔子	3	0
爱	3	1
吃	3	2
胡萝卜	3	3
家	3	4
我	1	5
你	1	6
谁	1	7

[0059] 在一些实施例中,定义单词的上下文包括确定与中心单词相邻间隔不超过预定数量的单词的单词集合。例如,当所述预定数量为2个时,对于"我/家/兔子/爱/吃/胡萝卜","爱"的上下文单词就是"家"、"兔子"、"吃"、"胡萝卜"。

[0060] 在一些实施例中,部分单词及其索引可如表3所示。

[0061] 表3

[0062]

索引对	单词对
(4,0)	(家,兔子)
(4,1)	(家,爱)
(0,4)	(兔子,家)
(0,1)	(兔子,爱)
(0,2)	(兔子,吃)
(1,4)	(爱,家)
(1,2)	(爱,吃)
(1,3)	(爱,胡萝卜)
(2,3)	(吃,胡萝卜)

[0063] 所述词向量模型文本数据可包括以下中的一个或多个:文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。在本实施例中,所述词向量文本数据可包括文献标题和文献摘要。

[0064] 子步骤S1064:基于所述词向量文本数据,通过词向量模型生成词向量。

[0065] 例如,可基于所述词向量文本数据的训练数据集对词向量模型进行训练。具体地,可通过使所述词向量模型遍历所述训练数据集,来训练词汇表中的每个单词的词向量。对于所述词向量模型中的每个词向量,还可通过对每个正样本和随机采样得到的N个负样本进行随机梯度下降(SGD)来实现更新迭代。

[0066] 在一些实施例中,可基于词向量文本数据构建词向量训练数据集。例如,所述词向量模型的训练数据集可以包含所述可用文本数据,也可以是所述可用文本数据的一部分。在本实施例中,所述词向量文本数据包括文献标题和文献摘要。所述词向量模型可以是无监督词向量模型,例如,连续词袋(CBOW)模型或跳字(Skip-gram)模型。在本实施例中,所述词向量模型是连续词袋模型。

[0067] 所述连续词袋模型可包括:输入层、投影层和输出层。所述连续词袋模型的输入是词向量训练数据集。在本实施例中,所述词向量训练数据集为一系列的单词索引对。在不考

虑内存、硬盘开销及训练时长的情况下，所述连续词袋模型输入也可以是单词对。例如，对于文本"我/家/兔子/爱/吃/胡萝卜"，输入词向量模型的单词对可以是(我，家)、(家，我)、(家，兔子)、(兔子，家)、(我，兔子)、(兔子，爱)、(兔子、吃)、(爱、吃)、(爱、胡萝卜)、(吃、胡萝卜)、(爱，兔子)、(吃，兔子)、(吃，爱)、(胡萝卜，爱)、(胡萝卜，吃)，所述词向量模型的输出可以是各中心单词"兔子"、"爱"、"吃"、"胡萝卜"所对应的迭代更新后的词向量。所述词向量模型的训练目标是使得相邻(或在一定单词间隔内)的两单词，在向量空间内要靠的比较"近"。

[0068] 所述连续词袋模型的输出是迭代更新后的词向量，也即所述连续词袋模型本身，这是一个自迭代的过程。仍以"我/家/兔子/爱/吃/胡萝卜"为例，结合词汇表2，训练流程片段可包括如下步骤：

[0069] (1)从训练数据集中读取一个(正)样本，例如(4，0)，表示当前中心单词为4(家)，上下文单词为0(兔子)；

[0070] (2)根据词汇表中各单词出现的频率，对中心单词(例如，家)随机采样得到N个负样本，例如，(4，3)、(4，1)等；以及

[0071] (3)基于每个样本整数对，迭代更新输入单词索引对应的词向量。

[0072] 所述跳字模型可包括：输入层、投影层和输出层。所述跳字模型输入的是训练数据集，输出的是迭代更新后的中心单词的上下文单词对应的向量。

[0073] 例如，在训练生成词向量的过程中，可首先对所述词向量模型进行随机初始化，随后反复遍历所述词向量训练数据集，以迭代更新所述词向量模型。

[0074] 子步骤S1066：基于所述文献向量文本数据和所述词向量，通过文献向量模型生成每个目标文献的文献向量。

[0075] 例如，可基于所述文献向量文本数据和所述词向量对文献向量模型进行训练，以生成文献向量。在一些实施例中，可基于所生成的文献向量模型文本数据生成用于文献向量模型的训练数据集。与词向量模型的训练类似，文献向量模型的训练过程包括：首先对文献向量模型进行随机初始化，随后从训练数据集中，逐个读取样本，更新对应的向量。在一些实施例中，训练数据集可以仅包括正样本。在一些实施例中，训练数据集可以包括正样本和负样本。例如，在硬件资源丰富条件下，可以提前采样好负样本，随正样本一起打乱构成训练数据集。

[0076] 在一些实施例中，可基于所述文献向量文本数据构建文献向量训练数据集。例如，所述文献向量模型的训练数据集可以包含所述可用文本数据，也可以是所述可用文本数据的一部分。在本实施例中，所述文献向量文本数据包括文献标题和文献关键字。所述文献向量模型可以是无监督词向量模型，例如，连续词袋模型或跳字模型。在本实施例中，所述文献向量模型是跳字模型。

[0077] 在一些实施例中，可以向所述文献向量模型输入单词对。例如，对于文本数据"兔子/爱/吃/胡萝卜"，输入所述文献向量模型的单词对可以是(兔子，所述文本数据的虚拟单词)、(爱，所述文本数据的虚拟单词)、(吃，所述文本数据的虚拟单词)，(胡萝卜，所述文本数据的虚拟单词)，所述词向量模型的输出可以是所述文本数据的虚拟单词向量，其中"兔子"、"爱"、"吃"、"胡萝卜"的词向量均为所述词向量模型之前(例如，在子步骤S1064中)生成的相应词向量，虚拟单词相当于浓缩了该文本数据中所有单词语义信息的一个超级单

词。虚拟单词向量为所述文本数据的向量化表示，即代表该文本数据的向量。所述文献向量模型的训练目标是使得该篇文献的超级单词和代表该篇文献的所有单词(例如，在本实施例中选取的标题和关键词)，在向量空间中靠的都比较"近"。通过这种方式，能够在前面得到的词向量空间里面找到一个合适的点来描述/表示一篇文献。对于一篇文献来说，其标题和关键词本身往往已经浓缩了该篇文献的语义，本身就已经有部分"超级单词"属性了，虽然它们在文献中出现的频次，并不一定和它们的重要性成正比。

[0078] 在一个或多个实施例中，可通过优化如下目标函数L，分别对词向量和文献向量进行更新迭代：

[0079] $$L=\begin{cases}\sum_{w\in C}\log\prod_{u\in\{w\}\cup NEG(w)}p(u\mid Context(w)), & CBOW\\ \sum_{d\in D}\log\prod_{w\in Content(d)}\prod_{u\in\{w\}\cup NEG(w)}p(u\mid d,w), & Skip-gram\end{cases}$$

[0080] 其中，第一行表达式对应词向量模型所采用的CBOW模型，第二行表达式对应文献向量模型所采用的Skip-gram模型，C表示词向量训练数据集或词向量文本数据，D表示文献向量训练数据集或文献向量文本数据，Context(w)表示单词w的上下文单词集合，Content(d)表示代表文献d的单词集合，{w}表示由单词w构成的集合，NEG(w)表示针对单词w采样得到的负样本单词集合，u表示{w}与负样本单词集合的并集中的元素，P()为概率函数。

[0081] 所述目标函数L中的条件概率为：

[0082] $$\begin{cases}p(u\mid Context(w))=\begin{cases}\sigma(X_w^Tv_u), & u=w;\\ 1-\sigma(X_w^Tv_u), & u\neq w;\end{cases}\\ p(u\mid d,w)=\begin{cases}\sigma(X_d^Tv_u), & u=w;\\ 1-\sigma(X_d^Tv_u), & u\neq w;\end{cases}\end{cases}$$

[0083] 其中，vx和v_d分别表示单词w和文献d对应的向量，X_w和X_d分别表示单词w和文献d对应的投影向量，v_u为单词u的向量表示，σ()为预设的激活函数，例如Sigmoid函数。

[0084] 投影层函数可以为求和函数、均值函数或恒等函数等。本实施例中，连续词袋模型的投影层函数可采用均值函数，跳字模型的投影层函数可采用恒等函数。x_w和X_d具体可以为：

[0085] $$\begin{cases} X_w = \dfrac{1}{\left|Context(w)\right|} \sum_{\tilde{w} \in Context(w)} v_{\tilde{w}}, \\ X_d = v_d, \end{cases}$$

[0086] 其中，|Context(w)|表示单词w的上下文单词集合中单词的总数，$\tilde{w}$表示单词w的上下文单词集合中的单词。

[0087] 连续词袋模型和跳字模型中的每个词向量都可通过对每个正样本及随机采样得到的N个负样本进行随机梯度下降(SGD)来实现更新迭代。不同之处在于，在使用跳字模型作为文献向量模型时，固定了输入的词向量，迭代更新时只更新文献向量。

[0088] 例如，在训练生成文献向量的过程中，可首先对所述文献向量模型进行随机初始化，随后反复遍历所述文献向量训练数据集，以迭代更新所述文献向量模型。

[0089] 步骤S108：对所有目标文献的文献向量进行聚类。

[0090] 例如，可根据预设聚类算法，对所有文献向量进行聚类分析，并按聚类结果分类输出。在本实施例中，聚类算法可以为K-Means算法。

[0091] 步骤S110：确定每个聚类中的每个文献向量与其他文献向量的相似度。

[0092] 相似度的计算方法可以为余弦度量法(cosine measure)：

[0093] $$\mathrm{sim}_{\cos ine}(x, y) = \frac{\vec{x}^T \vec{y}}{\left|\vec{x}\right|\left|\vec{y}\right|},$$

[0094] 其中，$|\vec{x}|$表示向量$\vec{x}$的长度(模)，$|\vec{y}|$表示向量$\vec{y}$的长度(模)。

[0095] 由于每个文献向量代表一篇文献，通过比较文献向量之间的相似度，可以得知文献之间的相似度。例如，在每个聚类中，计算每篇文献的前K篇最相似文献，K为预置的正整数。

[0096] 本公开的一个或多个实施例通过先训练单词向量，再训练文献向量，将单词及文献分步映射到同一语义空间。由于训练文献向量时，已经固定了词向量，从而有效减少了迭代更新所需的计算量。另一方面，由于文献向量仅依赖于词向量，故可以逐个训练，大大减少参数及内存开销。

[0097] 本公开的一个或多个实施例通过文献向量的夹角余弦来度量文献相似度。在所有文献向量单位化后，仅涉及浮点数加法和乘法，所以可以利用多线程技术或GPU计算进一步加速计算过程。

[0098] 图3为根据本公开一个或多个实施例的确定文献相似度的装置的示意图。如图3所示，确定文献相似度的装置300可包括文献获取单元310、预处理单元320、文献向量生成单元330、聚类单元340和相似度确定单元350。

[0099] 文献获取单元310可被配置为获取多个目标文献。例如，文献获取单元310可从图书、报刊、网络、数据库等获取、采集或下载中多个目标文献。

[0100] 预处理单元320可被配置为对每个目标文献进行预处理以生成基础文本数据。所述预处理可包括可用字段提取、数据清洗、数据分词、停用词剔除等。所述基础文本数据可包括以下中的至少一个:文献标题、文献摘要、文献关键字、文献正文、文献分类号和文献作者。

[0101] 文献向量生成单元330可被配置为基于所述基础文本数据生成每个目标文献的文献向量。基于所述基础文本数据生成每个目标文献的文献向量可包括:基于所述基础文本数据生成词向量文本数据和文献向量文本数据;基于所述词向量文本数据,通过词向量模型生成词向量;以及基于所述文献向量文本数据和所述词向量,通过文献向量模型生成每个目标文献的文献向量。所述词向量文本数据可包括文献标题和文献摘要,所述文献向量文本数据可包括文献标题和文献关键字。所述词向量模型可为连续词袋模型,所述文献向量模型可为跳字模型。

[0102] 聚类单元340可被配置为对所有目标文献的文献向量进行聚类。聚类算法可以为K-Means算法。

[0103] 相似度确定单元350可被配置为确定每个聚类中的每个文献向量与其他文献向量的相似度。相似度的计算方法可以为余弦度量法。

[0104] 本公开的一个或多个实施例通过先训练单词向量,再训练文献向量,将单词及文献分步映射到同一语义空间。由于训练文献向量时,已经固定了词向量,从而有效减少了迭代更新所需的计算量。另一方面,由于文献向量仅依赖于词向量,故可以逐个训练,大大减少参数及内存开销。

[0105] 本公开的一个或多个实施例通过文献向量的夹角余弦来度量文献相似度。在所有文献向量单位化后,仅涉及浮点数加法和乘法,所以可以利用多线程技术或GPU计算进一步加速计算过程。

[0106] 图4为根据本公开一个或多个实施例的计算设备的示意图。计算设备100可包括至少一个存储介质,存储有至少一组指令;以及至少一个处理器,同所述至少一个存储介质通讯连接。当所述至少一个处理器运行所述至少一组指令时,所述至少一个处理器执行前述方法。

[0107] 计算设备100可以使用合适的硬件部件和/或软件部件(例如,处理器、控制器、存储单元、储存单元、输入单元、输出单元、通信单元、操作系统、应用、诸如此类)来实施。

[0108] 在一些示例性实施方式中,计算设备100例如可以包括计算设备、移动电话、智能电话、蜂窝电话、笔记本、移动计算机、膝上型计算机、笔记本计算机、台式计算机、手持计算机、手持设备、PDA设备、手持PDA设备、无线通信设备、合并无线通信设备的PDA设备、诸如此类。

[0109] 在一些示例性实施方式中,计算设备100例如可以包括处理器191、输入单元192、输出单元193、存储单元194、和/或储存单元195中的一个或多个。计算设备100可选地可以包括其他合适的硬件部件和/或软件部件。在一些示例性实施方式中,计算设备100的一个或多个的一些或全部部件可以被围绕在公共壳体或包装中,并且可以使用一个或多个有线或无线链路互连或可操作地关联。在其他实施方式中,计算设备100的一个或多个的部件可以分布在多个或单独的设备中。

[0110] 在一些示例性实施方式中,处理器191例如可以包括中央处理单元(CPU)、数字信

号处理器(DSP)、一个或多个处理器核心、单核处理器、双核处理器、多核处理器、微处理器、主处理器、控制器、多个处理器或控制器、芯片、微芯片、一个或多个电路、电路系统、逻辑单元、集成电路(IC)、专用IC(ASIC)或任意其他合适的多功能或专用处理器或控制器。处理器191例如可以执行计算设备100的操作系统(OS)和/或一个或多个合适应用的指令。

[0111] 在一些示例性实施方式中,输入单元192例如可以包括键盘、小键盘、鼠标、触摸屏、触摸板、跟踪球、触针、麦克风或其他合适的指向设备或输入设备。输出单元193例如可以包括监视器、屏幕、触摸屏、平板显示器、发光二极管(LED)显示单元、液晶显示器(LCD)显示单元、等离子体显示单元、一个或多个扬声器或耳机、或其他合适的输出设备。

[0112] 在一些示例性实施方式中,存储介质194例如可以包括随机存取存储器(RAM)、只读存储器(ROM)、动态RAM(DRAM)、同步DRAM(SD-RAM)、闪速存储器、易失性存储器、非易失性存储器、高速缓冲存储器、缓冲器、短期存储单元、长期存储单元、硬盘驱动器、软盘驱动器、压缩盘(CD)驱动器、CD-ROM驱动器、DVD驱动器、或其他合适的可移动或不可移动储存单元。存储介质194例如可以存储由计算设备100处理的数据。

[0113] 在一些示例性实施方式中,存储介质194可储存逻辑195,逻辑195可以包括指令、数据、和/或代码,这些指令、数据、和/或代码在由机器执行时,可以使得机器执行如这里所描述的方法、处理和/或操作。机器例如可以包括任意合适的处理平台、计算平台、计算设备、处理设备、计算系统、处理系统、计算机、处理器、诸如此类,并且可以使用硬件、软件、固件、诸如此类的任意合适组合来实施。逻辑195可以包括或可以被实施为软件、软件模块、应用、程序、子例程、指令、指令集、计算代码、词、值、标记、诸如此类。指令可以包括任意合适类型的代码(诸如源代码、编译代码、翻译代码、可执行代码、静态代码、动态代码、诸如此类)。指令可以根据预定义计算机语言、方式或语法来实施,用于指示处理器执行特定功能。指令可以使用任意合适的高级的、低级的、面向对象的、视觉的、编译的和/或翻译的编程语言(诸如C、C++、Java、BASIC、Python、Matlab、Pascal、Visual BASIC、汇编语言、机器代码、诸如此类)来实施。

[0114] 在一些示例性实施方式中,计算设备100可以被配置为经由无线和/或有线网络与一个或多个其他设备通信。所述网络可以包括有线网络、局域网(LAN)、无线LAN(WLAN)网络、无线电网络、蜂窝网络、无线保真(WiFi)网络、IR网络、蓝牙(BT)网络、诸如此类。

[0115] 在一些示例性实施方式中,计算设备100可以允许一个或多个用户与例如如这里所述的计算设备100的一个或多个处理、应用和/或模块交互。

[0116] 在一些示例性实施方式中,计算设备100可以被配置为执行和/或实行一个或多个操作、模块、处理、过程和/或诸如此类。

[0117] 综上所述,在阅读本详细公开内容之后,本领域技术人员可以明白,前述详细公开内容可以仅以示例的方式呈现,并且可以不是限制性的。尽管这里没有明确说明,本领域技术人员可以理解本公开意图囊括对实施例的各种合理改变、改进和修改。这些改变、改进和修改旨在由本公开提出,并且在本公开的示例性实施例的精神和范围内。

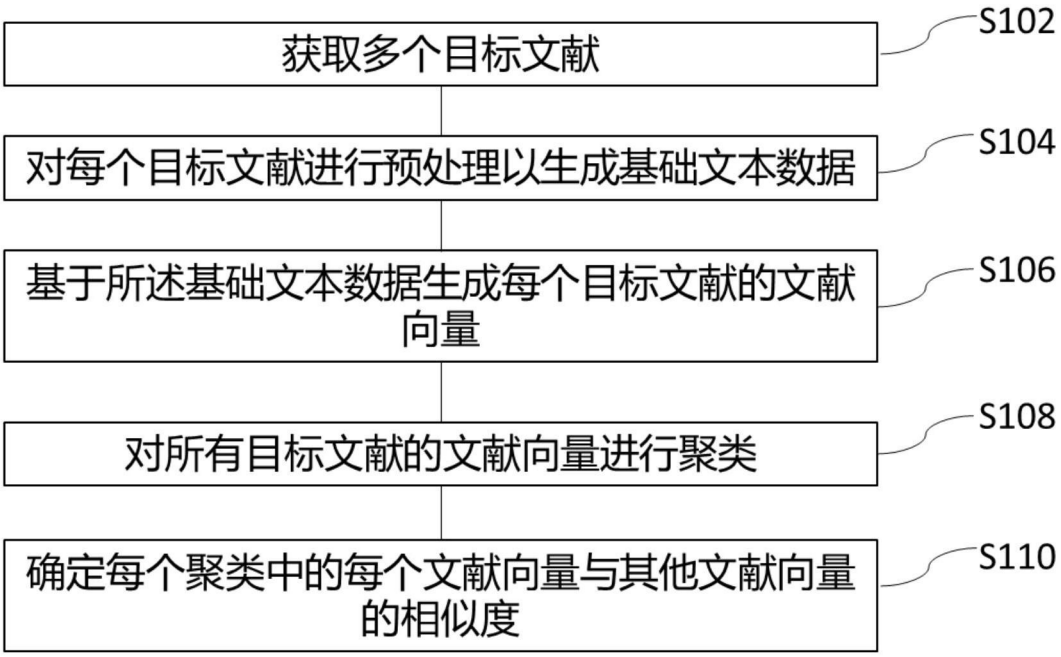

获取多个目标文献
S102
对每个目标文献进行预处理以生成基础文本数据
S104
基于所述基础文本数据生成每个目标文献的文献向量
S106
对所有目标文献的文献向量进行聚类
S108
确定每个聚类中的每个文献向量与其他文献向量的相似度
S110

图1

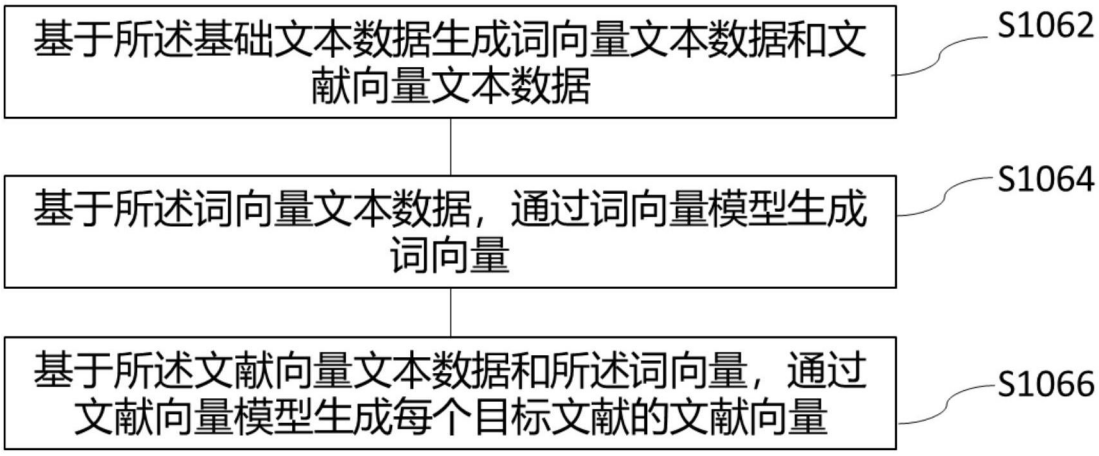

基于所述基础文本数据生成词向量文本数据和文献向量文本数据
S1062
基于所述词向量文本数据，通过词向量模型生成词向量
S1064
基于所述文献向量文本数据和所述词向量，通过文献向量模型生成每个目标文献的文献向量
S1066

图2

图3

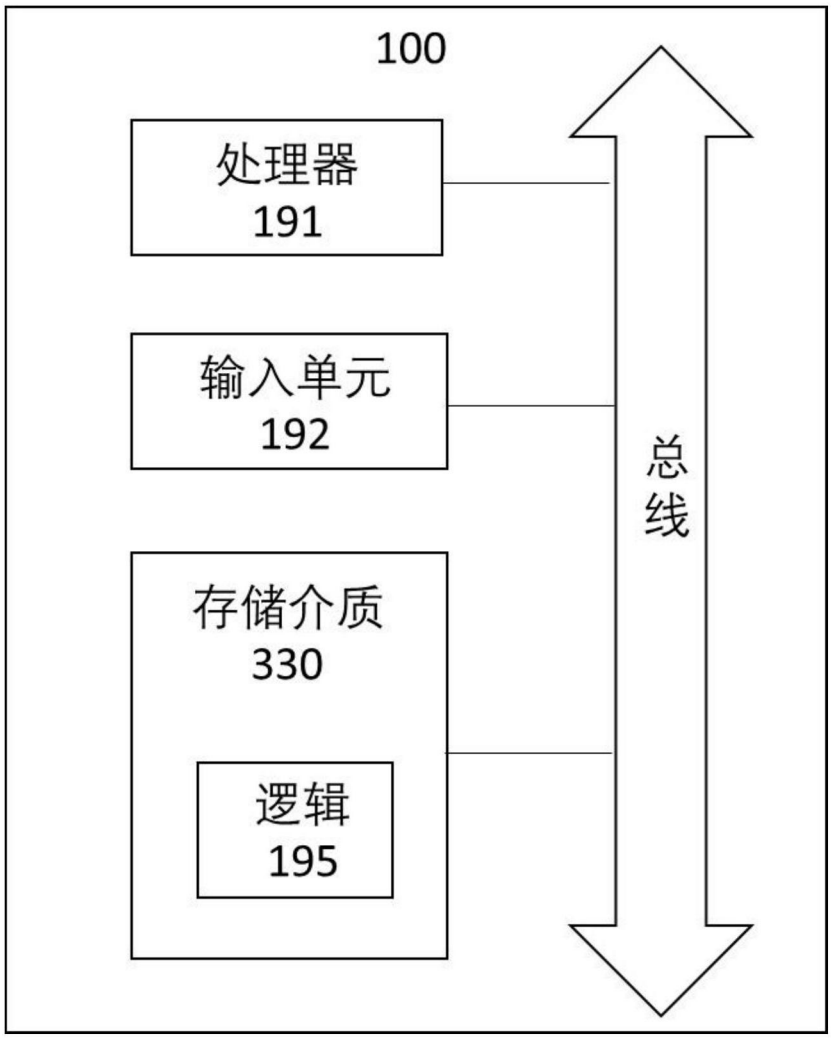


图4