(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108170684 B
(45)授权公告日 2020.06.05

(21)申请号 201810060942.7

(22)申请日 2018.01.22

(65)同一申请的已公布的文献号
申请公布号 CN 108170684 A

(43)申请公布日 2018.06.15

(73)专利权人 京东方科技集团股份有限公司
地址 100015 北京市朝阳区酒仙桥路10号

(72)发明人 张振中

(74)专利代理机构 北京律智知识产权代理有限公司 11438
代理人 袁礼君 王卫忠

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/289(2020.01)
G06F 16/33(2019.01)

(56)对比文件
CN 107436864 A,2017.12.05,
US 2009125805 A1,2009.05.14,
US 2015212993 A1,2015.07.30,
CN 105138647 A,2015.12.09,
陈二静等.文本相似度计算方法研究综述.《数据分析与知识发现》.2017,(第06期),1-11.

审查员 许少滨

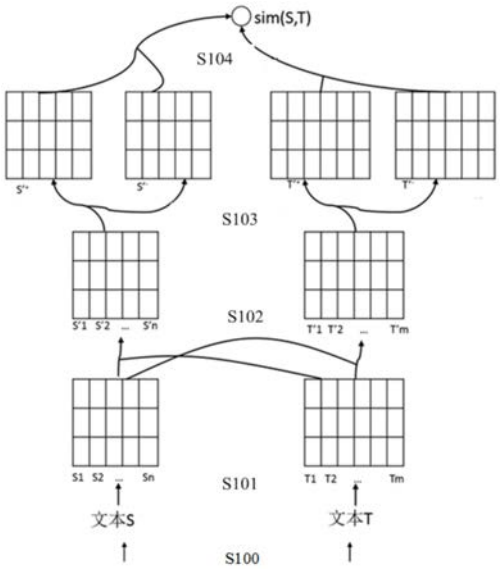
权利要求书2页 说明书10页 附图3页

(54)发明名称

文本相似度计算方法及系统、数据查询系统和计算机产品

(57)摘要

本公开提供一种文本相似度的计算方法及系统、数据查询系统、计算机产品和计算机可读存储介质。该文本相似度的计算方法包括：至少获取第一文本和第二文本；将所述第一文本和第二文本映射为向量；计算所述第一文本与所述第二文本的相似部分和差异部分；利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。根据本公开提供的文本相似度的计算方法和系统，至少部分地考虑了词语语义相似性和差异性对文本相似度的影响，能够从更细的粒度上计算文本之间的语义相似度，进而提高文本匹配的精度，实现高精度的检索或查询等。



CN 108170684 B

1.一种文本相似度的计算方法,包括:

至少获取第一文本和第二文本;

将所述第一文本和第二文本映射为向量;

通过下述公式将第二文本中的词语对应的向量重构所述第一文本的词语对应的向量计算语义覆盖:

$$\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$$ 求解$\alpha_{i,j}$,

其中S_i为第一文本的列向量,T_j为第二文本的列向量,$\alpha_{i,j}$为语义覆盖参数,$\lambda>0$,为事先设定的正实数;i=1,2,…,n;j=1,2,…,m;n和m为正整数;

对所述第一文本与第二文本进行语义分解,得到所述第一文本和第二文本的相似部分和差异部分;

利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

2.根据权利要求1所述的文本相似度的计算方法,将所述第一文本和第二文本映射为向量,还包括:对所述第一文本和第二文本对应的向量进行降维处理。

3.根据权利要求2所述的文本相似度的计算方法,对所述第一文本和第二文本对应的向量进行降维处理,包括采用下述至少一种方法进行降维处理:词向量、句子向量、文章向量。

4.根据权利要求1所述的文本相似度的计算方法,其中,计算所述第一文本的相似部分和差异部分包括:

采用公式

$$S_i'=\sum_{j=1}^{m}A_{i,j}\times T_j$$ 计算相似部分和差异部分,其中$A_{i,j}$为$\alpha_{i,j}$的矩阵,S_i'为所述第一文本的相似部分,S_i-S_i'为所述第一文本的差异部分;

其中i=1,2,…,n;j=1,2,…,m;n和m为正整数。

5.根据权利要求1所述的文本相似度的计算方法,其中,计算所述第二文本的相似部分和差异部分包括:

采用公式

$$T_j'=\sum_{j=1}^{m}A_{i,j}\times S_j$$ 计算相似部分和差异部分,其中$A_{i,j}$为$\alpha_{i,j}$的矩阵,T_j'为所述第二文本的相似部分,T_j-T_j'为所述第二文本的差异部分。

6.根据权利要求1所述的文本相似度的计算方法,其中利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度包括:

输入所述第一文本和所述第二文本的相似部分和差异部分,利用循环神经网络得到所述第一文本和所述第二文本之间的相似度。

7.根据权利要求6所述的文本相似度的计算方法,用循环神经网络得到所述第一文本和所述第二文本之间的相似度,还包括利用样本数据对循环神经网络进行训练的步骤,所述训练数据的格式为(S,T,L),其中S表示第一文本,T表示第二文本,L表示相似度。

8.根据权利要求7所述的文本相似度的计算方法,利用样本数据对所述循环神经网络

进行训练的步骤，还包括预先定义相似程度的粒度，并且将样本数据输入到所述循环神经网络，进行训练。

9.一种文本相似度计算系统，包括：

获取模块，被配置为至少获取第一文本和第二文本的输入；

映射模块，被配置为将所述第一文本和第二文本映射为向量；

文本相似度计算模块，所述文本相似度计算模块包括：

语义匹配模块，被配置为执行下述公式

$$\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$$

计算语义覆盖，

其中S_i为第一文本的列向量，T_j为第二文本的列向量，$\alpha_{i,j}$为语义覆盖参数，$\lambda>0$，为事先设定的正实数；

语义分解模块，被配置为计算所述第一文本和第二文本的相似部分和差异部分；

相似度计算模块，被配置为使用所述相似部分和差异部分计算所述第一文本和第二文本的相似度。

10.如权利要求9所述的文本相似度计算系统，其中所述映射模块包括降维模块，被配置为对所述第一文本和第二文本对应的向量进行降维处理。

11.根据权利要求9的文本相似度计算系统，所述语义分解模块被配置为执行下述公式获得所述第一文本的相似部分和差异部分、所述第二文本的相似部分和差异部分：

$S_i^{'}=\sum_{j=1}^{m}A_{i,j}\times T_j$，其中$A_{i,j}$为$\alpha_{i,j}$的矩阵，$S_i$’为所述第一文本的相似部分，$S_i-S_i$’为所述第一文本的差异部分；

$T_j^{'}=\sum_{j=1}^{m}A_{i,j}\times S_j$，其中，$A_{i,j}$为$\alpha_{i,j}$的矩阵，$T_j$’为所述第二文本的相似部分，$T_j-T_j$’为所述第二文本的差异部分；$i=1,2,\cdots,n$；$j=1,2,\cdots,m$。

12.一种数据查询系统，包括如权利要求9-11中任一所述的文本相似度计算系统。

13.一种计算机产品，包括：

一个或多个处理器，所述处理器被配置为运行计算机指令以执行如权利要求1-8中任一项所述的方法的一个或多个步骤。

14.如权利要求13所述的计算机产品，还包括存储器，连接所述处理器，被配置为存储所述计算机指令。

15.一种计算机可读存储介质，被配置为存储计算机指令，所述计算机指令被处理器运行时执行如权利要求1-8中任一项所述方法的一个或多个步骤。

文本相似度计算方法及系统、数据查询系统和计算机产品

技术领域

[0001] 本申请涉及数据处理领域，尤其涉及一种文本相似度的计算方法及系统、数据查询系统、计算机产品和计算机可读存储介质。

背景技术

[0002] 互联网的快速发展以及大数据时代的到来为人们有效获取各类信息提供了基础。目前人们已经习惯通过网络来获取各种各样的信息。举例来说，在医学领域，医护人员可以通过输入关键词搜索得到所需的文献、书籍或者相关网页等。对于患者来说，可以通过查看医疗网站的社区问答满足自身的信息需求。信息服务系统的基本流程是依据用户输入的查询或者问题，从数据中(文档集、问题集或者知识库等)匹配和查询或者问题最相关的内容返回给用户。

[0003] 但是目前的信息查询系统在满足人们信息需求的同时，也存在着一些不足之处，例如由于文本相似度的计算不够全面，导致匹配精度不高。

发明内容

[0004] 本公开的实施例提供一种文本相似度的计算方法，包括：

[0005] 至少获取第一文本和第二文本；

[0006] 将所述第一文本和第二文本映射为向量；

[0007] 计算所述第一文本与所述第二文本的相似部分和差异部分；

[0008] 利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

[0009] 可选地，将所述第一文本和第二文本映射为向量，还包括：对所述第一文本和第二文本对应的向量进行降维处理。

[0010] 可选地，对所述第一文本和第二文本对应的向量进行降维处理，包括采用下述至少一种方法进行降维处理：词向量、句子向量、文章向量。

[0011] 可选地，计算所述第一文本与所述第二文本之间的相似部分和差异部分，包括：对所述第一文本与第二文本进行语义匹配；对所述第一文本与第二文本进行语义分解，得到所述第一文本和第二文本的相似部分和差异部分。

[0012] 可选地，将所述第一文本与第二文本进行语义匹配，包括：将第二文本中的词语对应的向量重构所述第一文本的词语对应的向量来判断语义覆盖的内容。

[0013] 可选地，通过下述公式将第二文本中的词语对应的向量重构所述第一文本的词语对应的向量计算语义覆盖：

[0014] $$\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$$

[0015] 求解$\alpha_{i,j}$，S_i为第一文本的列向量，T_j为第二文本的列向量，$\alpha_{i,j}$为语义覆盖参数，$\lambda>0$，为事先设定的正实数。

[0016] 可选地，计算所述第一文本的相似部分和差异部分，包括：采用公式

[0017] $S_i^{'}=\sum_{j=1}^{m}A_{i,j}\times T_j$计算相似部分和差异部分，$A_{i,j}$为$\alpha_{i,j}$的矩阵，$S_i$'为所述第一文本的相似部分，$S_i-S_i$'为所述第一文本的差异部分。

[0018] 可选地，计算所述第二文本的相似部分和差异部分，包括：

[0019] 采用公式

[0020] $T_j^{'}=\sum_{j=1}^{m}A_{i,j}\times S_j$计算相似部分和差异部分，$A_{i,j}$为$\alpha_{i,j}$的矩阵，$T_j$'为所述第二文本的相似部分，$T_j-T_j$'为所述第二文本的差异部分。

[0021] 可选地，利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度，包括：输入所述第一文本和所述第二文本的相似部分和差异部分，利用循环神经网络得到所述第一文本和所述第二文本之间的相似度。

[0022] 可选地，用循环神经网络得到所述第一文本和所述第二文本之间的相似度，还包括利用样本数据对所述循环神经网络进行训练的步骤，所述训练数据的格式为(S,T,L)，其中S表示第一文本，T表示第二文本，L表示相似度。

[0023] 可选地，利用样本数据对所述循环神经网络进行训练的步骤，还包括预先定义相似程度的粒度，并将样本数据输入到所述循环神经网络，进行训练。

[0024] 本公开的实施例还提供一种文本相似度计算系统，包括：

[0025] 获取模块，被配置为至少获取第一文本和第二文本的输入；

[0026] 映射模块，被配置为将所述第一文本和第二文本映射为向量；

[0027] 文本相似度计算模块，被配置为计算所述第一文本与第二文本的相似部分和差异部分；利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

[0028] 可选地，所述映射模块包括降维模块，被配置为对所述第一文本和第二文本对应的向量进行降维处理。

[0029] 可选地，所述文本相似度计算模块包括语义匹配模块，被配置将所述第一文本与第二文本进行语义匹配；语义分解模块，被配置为计算所述第一文本和第二文本的相似部分和差异部分；相似度计算模块，被配置为使用所述相似部分和差异部分计算所述第一文本和第二文本的相似度。

[0030] 可选地，所述语义匹配模块，被配置为通过第二文本中词语对应向量重构所述第一文本词语对应向量来判断语义覆盖的内容。

[0031] 可选地，所述语义匹配模块被配置为执行下述公式

[0032] $\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$计算语义覆盖，其中$S_i$为第一文本的列向量，$T_j$为第二文本的列向量，$\alpha_{i,j}$为语义覆盖参数，$\lambda>0$，为事先设定的正实数。

[0033] 可选地，所述语义分解模块被配置为执行下述公式获得所述第一文本的相似部分和差异部分、所述第二文本的相似部分和差异部分：

[0034] $S_i^{'}=\sum_{j=1}^{m}A_{i,j}\times T_j$，其中$A_{i,j}$为$\alpha_{i,j}$的矩阵，$S_i$'为所述第一文本的相似部分，$S_i-S_i$'为所述第一文本的差异部分。

[0035] $T_j' = \sum_{j=1}^{m} A_{i,j} \times S_j$，其中，$A_{i,j}$为$\alpha_{i,j}$的矩阵，$T_j'$为所述第二文本的相似部分，$T_j$-$T_j'$为所述第二文本的差异部分。

[0036] 本公开的实施例还提供一种数据查询系统，包括如前所述的文本相似度计算系统。

[0037] 本公开的实施例还提供一种计算机产品，包括：一个或多个处理器，所述处理器被配置为运行计算机指令以执行如前所述文本相似度的计算方法中的一个或多个步骤。

[0038] 可选地，所述计算机产品还包括存储器，连接所述处理器，被配置为存储所述计算机指令。

[0039] 本公开的实施例提供一种计算机可读存储介质，被配置为存储计算机指令，所述计算机指令被处理器运行时执行如前所述文本相似度的计算方法中的一个或多个步骤。

附图说明

[0040] 图1为根据本公开实施例的文本相似度计算方法的流程图。

[0041] 图2示出了根据本公开一实施例的文本相似度计算系统的结构框图。

[0042] 图3示出了根据本公开另一实施例的文本相似度计算系统的结构框图。

[0043] 图4示出了根据本公开实施例的数据查询系统的结构示意图。

[0044] 图5示出了根据本公开实施例的电子设备的结构框图。

具体实施方式

[0045] 为了使得本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例的附图，对本公开实施例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公开的一部分实施例，而不是全部的实施例。基于所描述的本公开的实施例，本领域普通技术人员在无需创造性劳动的前提下所获得的所有其他实施例，都属于本公开保护的范围。

[0046] 除非另外定义，本公开使用的技术术语或者科学术语应当为本公开所属领域内具有一般技能的人士所理解的通常意义。本公开中使用的“第一”、“第二”以及类似的词语并不表示任何顺序、数量或者重要性，而只是用来区分不同的组成部分。“包括”或者“包含”等类似的词语意指出现该词前面的元件或者物件涵盖出现在该词后面列举的元件或者物件及其等同，而不排除其他元件或者物件。“连接”或者“相连”等类似的词语并非限定于物理的或者机械的连接，而是可以包括电性的连接，不管是直接的还是间接的。“上”、“下”、“左”、“右”等仅用于表示相对位置关系，当被描述对象的绝对位置改变后，则该相对位置关系也可能相应地改变。

[0047] 为了保持本公开实施例的以下说明清楚且简明，本公开省略了已知功能和已知部件的详细说明。在发明人所知的技术中，文本相似度的计算方法通常考虑文本之间的相似性，忽略内容之间的差异性。然而这些差异性能够提供一定的语义信息，这些语义信息有可能有利于实现信息的准确匹配，更好地满足用户的信息需求。

[0048] 例如，在用户进行查询时，用户输入：关于感冒的介绍，在数据库中有：内容1：这篇文章介绍了鼻窦炎；内容2：这篇文章介绍了流行性感冒，而不是普通感冒；内容3：这篇文章

介绍了流行性感冒，而不是肺炎。

[0049] 在计算用户输入的文本与数据库中所具有的文本之间的相似度sim时，很容易得到sim(“输入”,“内容2”)>sim(“输入”,“内容1”)以及sim(“输入”,“内容3”)>sim(“输入”,“内容1”)，其中sim(.,.)表示相似度度量函数，但很难判断sim(“输入”,“内容2”)和sim(“输入”,“内容3”)之间的大小。但如果注意到“内容2”更多的是强调“流行性感冒”和“普通感冒”的差异，即“流行性感冒”的一些独有特性，而非感冒的共性；而“内容3”强调“流行性感冒”和“肺炎”的差异，因为“肺炎”和“感冒”之间得到共性相对较少，所以其包含的感冒共性的内容可能会多一些，从这个意义上说，“内容3”应该比“内容2”更接近“输入”。

[0050] 本公开的实施例提供一种文本相似度的计算方法，包括：

[0051] 至少获取第一文本和第二文本；

[0052] 将所述第一文本和第二文本映射为向量；

[0053] 计算所述第一文本与所述第二文本的相似部分和差异部分；

[0054] 利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

[0055] 现在结合图1具体描述根据本公开实施例提供的文本相似度的计算方法。

[0056] S100，至少获取第一文本和第二文本。

[0057] 在本公开的一些实施例中，至少获取的第一文本和第二文本可以包括多个词语组成的句子，可以包括多个句子组成的段落，可以包括多个段落组成的文章，可以包括海量文本。

[0058] 在本公开的一些实施例中，可以从单一或多个不同的数据源获取至少第一文本和第二文本，例如可以来源于用户在本地设备(手机、平板、电脑等)存储或输出的文本，例如可以来源于通过网络传输的文本。

[0059] S101，将所述第一文本和第二文本映射为向量。

[0060] 例如，将第一文本S映射为S1,S2,…,Sn，第二文本T映射为T1,T2,…,Tm，其中S1-Sn是组成第一文本S的词语序列，也就是第一文本的对应向量，例如S＝“感冒的症状”，则S1＝“感冒”，S2＝“的”，S3＝“症状”，第二文本T1-Tm同理。

[0061] 在本公开的一些实施例中，可通过向量空间模型VSM、文本分布式表示等实现文本的向量化。

[0062] 在本公开的一些实施例中，通过文本分布式表示实现文本的向量化，包括通过潜在语意分析(LSA)、概率潜在语意分析(pLSA)、潜在狄利克雷分配(LDA)等方法，将所获取的文本映射为向量。

[0063] 可选地，将所述第一文本和第二文本映射为向量，还包括：对所述第一文本和第二文本对应的向量进行降维处理，从而将文本表达为比较低维稠密向量形式。

[0064] 在本公开的一些实施例中，通过主成分分析(PCA，将多个变量通过线性变换以选出较少个数重要变量)对本对应的向量进行降维处理。

[0065] 在本公开的一些实施例中，具体的，通过词向量或句子向量或文章向量(Word2Vec或Sentence2Vec或Doc2Vec)的方法对文本对应的向量进行降维处理。

[0066] 在本公开的一些实施例中，具体的，通过词向量实现的词嵌入方法对文本对应的向量进行降维处理。

[0067] 通过降维处理，可以将第一文本、第二文本中的词语S1-Sn和T1-Tm映射成低维度

的向量。例如维度为d，则每个词对应着一个d×1的列向量，文本S对应d×n的矩阵，设为矩阵P；文本T对应d×m的矩阵，设为矩阵Q。

[0068] S102，计算所述第一文本与所述第二文本的相似部分和差异部分。

[0069] 在本公开的一些实施例中，可以通过文本编码的方法获得相似部分与差异部分，例如simhash方法，通过将各文本对应二进制编码，确定文本之间的海明距离(两个文本的simhash进行异或运算)，在异或结果中，1的个数超过3表示不相似(差异部分)，小于等于3表示相似(相似部分)。

[0070] 诸如上述的文本编码方法，还可以使用Shingling、Minhash等。

[0071] 在本公开的一些实施例中，采用S103：基于文本的语义内容的方法获得相似部分和差异部分。

[0072] 例如，对第一文本与第二文本进行语义匹配，以获得第一文本与第二文本的语义覆盖。

[0073] 为了计算文本S和文本T之间的相似部分，需要知道S(T)中每个词语是否可以被T(S)中的词语语义覆盖。例如，上述示例中，“输入”中的词语“感冒”可以语义覆盖内容3中的“流行性感冒”，而不太能够语义覆盖内容1中的“鼻窦炎”，因此sim(“输入”,“内容3”)>sim(“输入”,“内容1”)。

[0074] 在本公开的一些实施例中，对于文本S中的词语S_i，通过文本T中词语向量重构S_i的向量来判断语义覆盖的内容。

[0075] 例如，以下述公式执行语义覆盖内容的判断：

[0076] $$\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2 \qquad (1)$$

[0077] 其中，S_i和T_j是词语对应的d×1列向量(即上述词嵌入步骤中得到的词向量)，$\alpha_{i,j}$为参数，$\lambda>0$，为事先设定的正实数。

[0078] 例如，可通过下述方法求解参数$\alpha_{i,j}$：

[0079] 令 $f_i=\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$ ，因为f_i是凸函数，则令f_i取得最小值的$\alpha_{i,j}$满足

[0080] $$\frac{\partial f_i}{\partial \alpha_{i,j}}=0$$

[0081] 则有，

$$\frac{\partial f_i}{\partial \alpha_{i,j}}=-2(S_i-\sum_{k=1}^{m}\alpha_{i,k}T_k)^T T_j+2\lambda\alpha_{i,j}$$

[0082] $$=-2\left[S_i^T T_j-\sum_{k=1}^{m}\alpha_{i,k}T_k^T T_j-\lambda\alpha_{i,j}\right]=0$$

$$=>S_i^T T_j-\sum_{k=1}^{m}\alpha_{i,k}T_k^T T_j-\lambda\alpha_{i,j}=0$$

[0083] 其中，上标T表示转置。因为上式对于，i＝1,2,…,n和j＝1,2,…,m都成立，这样一

共有n×m个等式，n×m个参数，所以可以解出这n×m个参数的值。

[0084] 为了更清晰地表示求解过程以及最终解。我们将上述式子表示成矩阵形式。设矩阵R为n×m维的矩阵，其中第i行，第j列的元素$R_{i,j}=S_i^T T_j$，则有$R=P^T Q$(P和Q对应着“词嵌入”小节形成的矩阵)。设矩阵U为m×m的矩阵，其中第k行，第j列的元素$U_{k,j}=T_k^T T_j$，则有$U=Q^T Q$。设矩阵A为n×m维的矩阵，其中第i行，第j列的元素为$\alpha_{i,j}$。则上述式子可以表示为，

[0085] $R-AU-\lambda A=0$

[0086] $=>R=AU+\lambda A$

[0087] $=A(U+\lambda I)$

[0088] $=>A=R(U+\lambda I)^{-1}$ (2)

[0089] 其中，I为m×m的单位矩阵。

[0090] 在求解式子(2)的过程中需要保证矩阵(U+λI)满秩，即存在逆矩阵。

[0091] 证明如下：

[0092] 因为$U_{k,j}=T_k^T T_j$，所以U为m×m的实对称矩阵，所以存在对角矩阵∧和U相似，即

[0093] $V^{-1}UV=\Lambda$

[0094] 则有，

[0095] $V^{-1}(U+\lambda I)V=\Lambda+\lambda I$

[0096] 由此可见，只要λ不等于矩阵U的负特征值，矩阵(U+λI)的特征值(即∧+λI的对角线上的元素)就不为0，矩阵就是满秩的。由于U的特征值可以是实数域中的任意实数，设其分布函数为F(x)，由于x的取值是连续的，所以对于给定的实数a，其概率值F(a)=0，因此从概率上说，在没有任何先验知识的情况下，事先确定λ值等于矩阵U的负特征值的概率为0，即U+λI存在逆矩阵。

[0097] 同样的方法，依据上述过程可求解文本T中词语在文本S中的语义覆盖。因此，为了方便叙述，本公开只阐述文本S中词语在文本T中的语义表示。

[0098] 尽管上述描述了一种求解参数$\alpha_{i,j}$的方法，本领域技术人员容易理解，还可以通过多种不同的方法进行求解，例如还可以通过mathematic、maple等软件中提供的数值方法、代数方法等求解。

[0099] 例如，对第一文本与第二文本进行语义分解，以获得第一文本与第二文本的相似部分和差异部分，即第一文本相对第二文本的相似部分和差异部分、第二文本相对第一文本的相似部分和差异部分。

[0100] 在本公开的一些实施例中，基于文本S中每个词语在文本T中的表示矩阵A(参见公式(2))，其中A的第i行表示了词语S_i在文本T中如何表示或覆盖。令$S_i^{'}=\sum_{j=1}^{m}A_{i,j}\times T_j$，即第二文本T可以覆盖第一文本S中词语$S_i$的部分，表示为相似部分，差异部分为$S_i-S_i'$，第一文本S的表示变成两部分：相似部分，由d×n矩阵表示(参见图1中的相似部分S'^{+})；差异部分，也是d×n矩阵表示(参见图1中的差异部分S'^{-})。

[0101] 基于上述方法，文本的表示由d×n矩阵变为d×2n矩阵。采用这样的方法可以计算所述第一文本的相似部分和差异部分。

[0102] 类似地，计算所述第二文本中与所述第一文本的相似部分和差异部分。包括：采用

公式$T_j' = \sum_{j=1}^{m} A_{i,j} \times S_j$，$A_{i,j}$为$\alpha_{i,j}$的矩阵，$T_j'$为所述第二文本中的相似部分，$T_j$-$T_j'$为所述第二文本中的差异部分。第二文本T的表示变成两部分：相似部分，由d×n矩阵表示(参见图1中的相似部分T′⁺)，差异部分，也是d×n矩阵表示(参见图1中的差异部分T′⁻)。

[0103] 采用这样的方法可以计算所述第二文本的相似部分和差异部分。

[0104] S104，基于相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

[0105] 例如，可以是基于文本的编码表示获得Jaccard相似度等。

[0106] 例如，可以是基于文本的语义计算所述第一文本与所述第二文本之间的相似度。

[0107] 在本公开的一些实施例中，通过基于机器学习的方法计算所述第一文本与所述第二文本之间的相似度。

[0108] 例如，使用循环神经网络(RNN)来计算输入文本S和T的相似度。

[0109] 在本公开的一些实施例中，利用样本数据训练RNN，样本数据的格式为(S,T,L)，其中S和T表示文本，L表示相似度。

[0110] 可选地，可以预先定义好相似度的粒度，例如对于2粒度的有L∈{0,1}，0表示不相似，1表示相似。对于3粒度的有L∈{0,0.5,1}，0表示不相似，0.5表示有点相似，1表示很相似。

[0111] 基于样本数据进行训练，先将样本数据输入到循环神经网络中。样本数据中对于第一文本与第二文本的相似部分与差异部分的大小所对应的相似度进行定义。例如对于2粒度，定义第一文本与第二文本的相似部分与差异部分的大小在什么样的范围内表示1相似，在什么样的范围内表示0不相似；对于3粒度，定义第一文本与第二文本的相似部分与差异部分的大小在什么样的范围内表示1相似，在什么样的范围内表示0.5有点相似，在什么样的范围内表示0不相似。

[0112] 采用样本数据对循环神经网络进行训练，使得循环神经网络可以在后续计算相似度。

[0113] 对于完成训练的RNN，以S和T的表示d×2n矩阵，d×2m矩阵作为输入(即输入第一文本和第二文本的相似部分和差异部分)，循环神经网络(RNN)来计算所述输入，输出相似度。

[0114] 本公开实施例提供的文本相似度的计算方法，至少部分地考虑了词语语义相似性和差异性对文本相似度的影响，能够从更细的粒度上计算文本之间的语义相似度，进而提高文本匹配的精度，实现高精度的检索或查询等。

[0115] 本公开的实施例还提供了一种文本相似度计算系统，图2和图3示出了该文本相似度计算系统的框图。该文本相似度计算包括：获取模块200，被配置为获取第一文本和第二文本的输入；映射模块210，被配置为将所述第一文本和第二文本映射为向量；文本相似度计算模块220，计算所述第一文本与第二文本的相似部分和差异部分，利用所述相似部分和差异部分计算所述第一文本与所述第二文本之间的相似度。

[0116] 在本公开的一些实施例中，所述映射模块210包括降维模块211，被配置为对所述第一文本和第二文本对应的向量进行降维处理。

[0117] 在本公开的一些实施例中，所述文本相似度计算模块220包括：语义匹配模块221，

被配置为将所述第一文本与第二文本进行语义匹配;语义分解模块222,计算所述第一文本和第二文本的相似部分和差异部分;相似度计算模块223,被配置为使用所述相似部分和差异部分计算所述第一文本和第二文本的相似度。

[0118] 可选地,所述语义匹配模块221,被配置为通过第二文本中词语对应向量重构所述第一文本词语对应向量来判断语义覆盖的内容。

[0119] 例如,所述语义匹配模块被配置为执行下述公式

[0120] $\min\left\|S_i-\sum_{j=1}^{m}\alpha_{i,j}\times T_j\right\|^2+\lambda\sum_{j=1}^{m}\alpha_{i,j}^2$ 计算语义覆盖,其中S_i为第一文本的列向量,T_j为第二文本的列向量,$\alpha_{i,j}$为语义覆盖参数,$\lambda>0$,为事先设定的正实数。

[0121] 可选地,所述语义分解模块被配置为执行下述公式获得所述第一文本的相似部分和差异部分、所述第二文本的相似部分和差异部分:

[0122] $S_i'=\sum_{j=1}^{m}A_{i,j}\times T_j$,其中$A_{i,j}$为$\alpha_{i,j}$的矩阵,$S_i$'为所述第一文本的相似部分,$S_i-S_i$'为所述第一文本的差异部分。

[0123] $T_j'=\sum_{j=1}^{m}A_{i,j}\times S_j$,其中,$A_{i,j}$为$\alpha_{i,j}$的矩阵,$T_j$'为所述第二文本的相似部分,$T_j-T_j$'为所述第二文本的差异部分。

[0124] 本公开的实施例还提供一种数据查询系统,该数据查询系统包括上述实施例的文本相似度计算系统。例如参考图4,该数据查询系统包括人机交互模块401,例如,对于医学问答系统来说,用于接受患者的问题,以及展示问题的回答。对于医学文献检索系统来说,用户接受用户提供的查询语句,以及展示返回的文献。该数据查询系统还包括内容匹配模块402和数据库403,内容匹配模块402包括上述实施例中的文本相似度计算系统。

[0125] 用户通过人机交互模块401输入待查找的内容,内容匹配模块402从数据库403中获取查询的文本,内容匹配模块402采用上述实施例中的文本相似度计算方法,计算获取查询的文本和用户输入的带查找内容之间的相似度,从而获取用户所需要查询的内容。

[0126] 参考图5,本公开的实施例还提供了一种计算机产品500,以实现上述实施例所描述的文本相似度计算系统。该计算机产品可以包括一个或多个处理器502,处理器502被配置为运行计算机指令以执行如前所述文本相似度的计算方法中的一个或多个步骤。

[0127] 可选地,所述计算机产品500还包括存储器501,连接所述处理器602,被配置为存储所述计算机指令。

[0128] 计算机产品500可以实现为本地计算的计算机产品结构,即计算机产品500在用户侧实现上述实施例所描述的文本相似度计算系统;计算机产品500也可以实现为本地和远端交互的计算机产品结构,即计算机产品500在用户侧的终端实现上述实施例所描述的文本相似度计算系统中的获取至少第一文本和第二文本,在与用户侧终端连接的网络服务器接收至少第一文本和第二文本以执行文本相似度的计算。

[0129] 在本公开的一些实施例中,计算机产品可以包括多个终端设备和与多个终端设备连接的网络服务器。

[0130] 其中,多个终端设备,将各终端设备的至少包括第一文本和第二文本的文本(例如

用户输入或存储的文本)上传至网络服务器；

[0131] 其中，网络服务器，获取各终端设备上传的文本，将所获取的各文本执行上述实施例的文本相似度计算方法。

[0132] 存储器501可以是各种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。

[0133] 处理器502可以是中央处理单元(CPU)或者现场可编程逻辑阵列(FPGA)或者单片机(MCU)或者数字信号处理器(DSP)或者专用集成电路(ASIC)或者图形处理器(GPU)等具有数据处理能力和/或程序执行能力的逻辑运算器件。一个或多个处理器可以被配置为以并行计算的处理器组同时执行上述文本相似度计算方法，或者被配置为以部分处理器执行上述文本相似度计算方法中的部分步骤，部分处理器执行上述文本相似度计算方法中的其它部分步骤等。

[0134] 计算机指令包括了一个或多个由对应于处理器的指令集架构定义的处理器操作，这些计算机指令可以被一个或多个计算机程序在逻辑上包含和表示。

[0135] 该计算机产品500还可以连接各种输入设备(例如用户界面、键盘等)、各种输出设备(例如扬声器等)、以及显示设备等实现计算机产品与其它产品或用户的交互，本文在此不再赘述。

[0136] 连接可以是通过网络连接，例如无线网络、有线网络、和/或无线网络和有线网络的任意组合。网络可以包括局域网、互联网、电信网、基于互联网和/或电信网的物联网(Internet of Things)、和/或以上网络的任意组合等。有线网络例如可以采用双绞线、同轴电缆或光纤传输等方式进行通信，无线网络例如可以采用3G/4G/5G移动通信网络、蓝牙、Zigbee或者Wi-Fi等通信方式。

[0137] 本公开实施例还提供一种计算机可读存储介质，被配置为存储计算机指令，所述计算机指令被处理器运行时执行如前所述文本相似度的计算方法中的一个或多个步骤。

[0138] 应清楚地理解，本公开描述了如何形成和使用特定示例，但本发明的原理不限于这些示例的任何细节。相反，基于本发明公开的内容的教导，这些原理能够应用于许多其它实施方式。

[0139] 本领域技术人员可以理解实现上述实施方式的全部或部分步骤被实现为由处理器执行的计算机程序。在该计算机程序被处理器执行时，执行本发明提供的上述方法所限定的全部或部分功能。所述的计算机程序可以存储于一种计算机可读存储介质中。

[0140] 此外，需要注意的是，上述附图仅是根据本发明示例性实施方式的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

[0141] 需要注意的是，上述附图中所示的框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0142] 通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本发明实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中或网络上，包括若干指令以使得一台计算设备(可以是个人计算机、服务器、移动终端、或者网络设备等)执行根据本发明实施方式的方法。

[0143] 本领域技术人员应当理解，本公开不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本公开的保护范围。因此，虽然通过以上实施例对本公开进行了较为详细的说明，但是本公开不仅仅限于以上实施例，在不脱离本公开构思的情况下，还可以包括更多其他等效实施例，而本公开的范围由所附的权利要求决定。

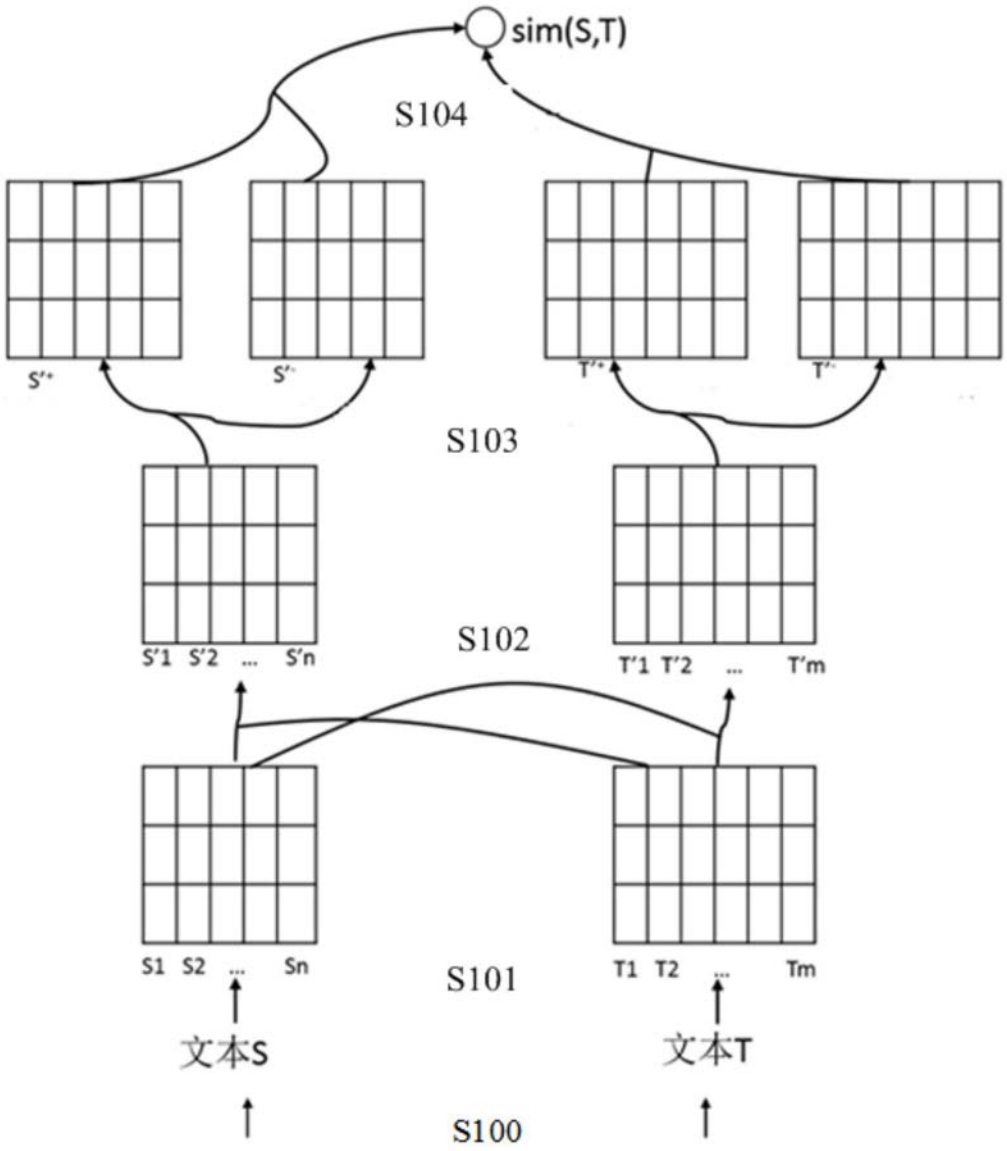

sim(S,T)
S104
S'+
S'-
T'+
T'-
S103
S'1 S'2 ... S'n
T'1 T'2 ... T'm
S102
S1 S2 ... Sn
T1 T2 ... Tm
S101
文本S
文本T
S100

图1

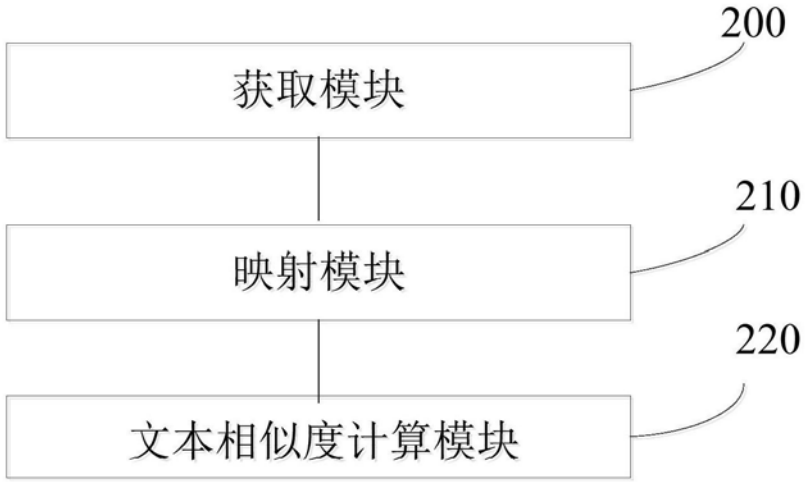

200
获取模块
210
映射模块
220
文本相似度计算模块

图2

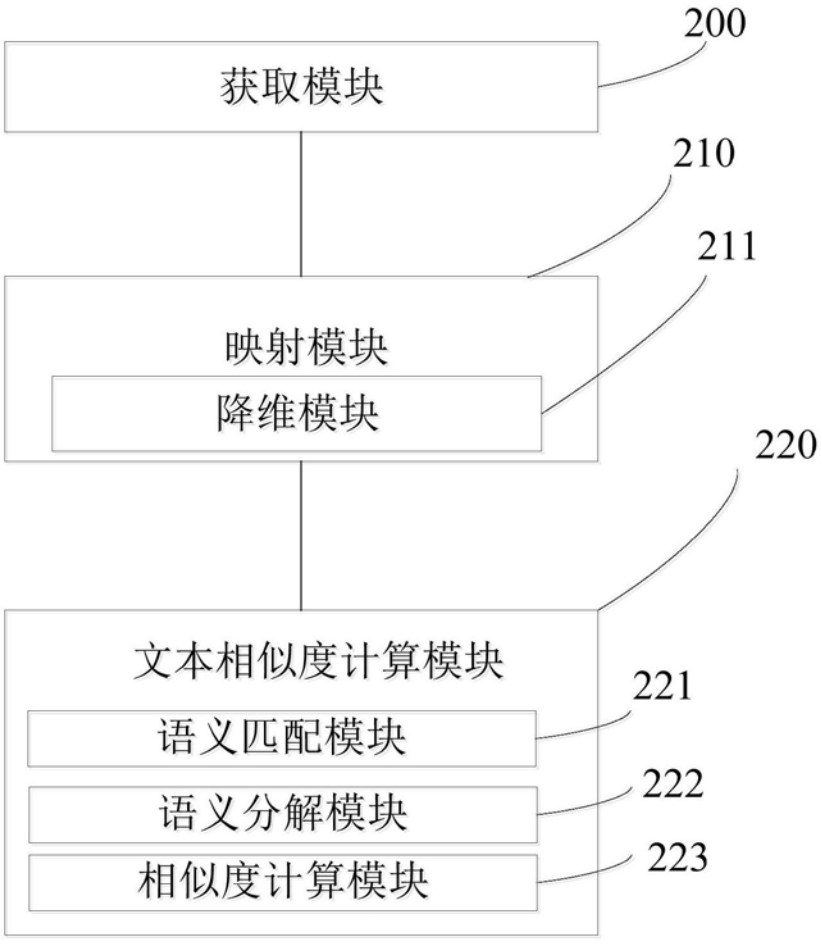

200
获取模块
210
211
映射模块
降维模块
220
文本相似度计算模块
221
语义匹配模块
222
语义分解模块
223
相似度计算模块

图3

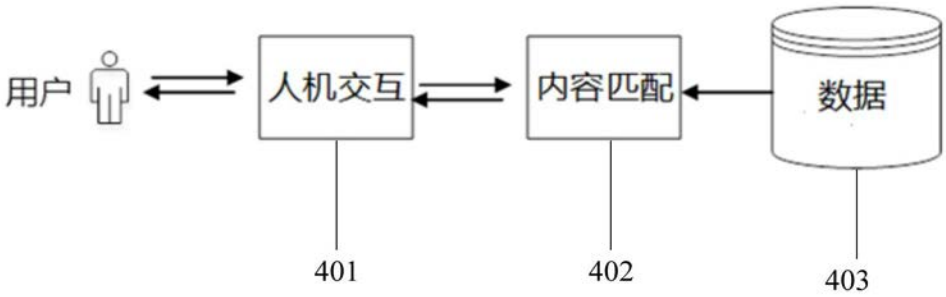


图4

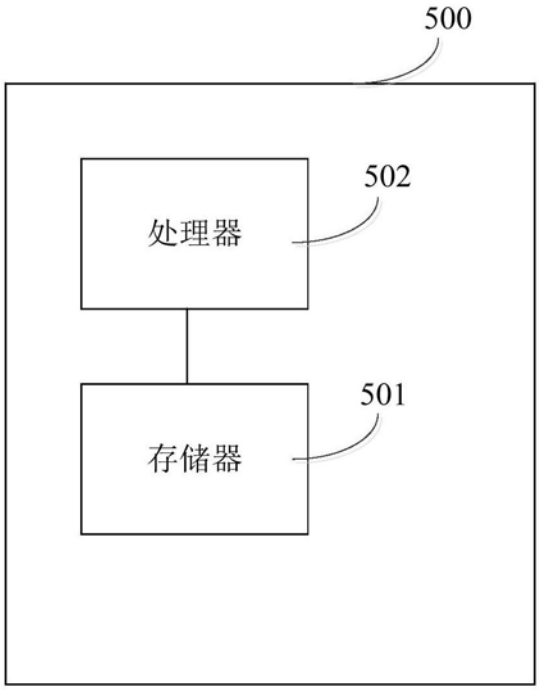


图5