(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112905644 B
(45)授权公告日 2022.08.02

(21)申请号 202110285108.X
(22)申请日 2021.03.17
(65)同一申请的已公布的文献号
申请公布号 CN 112905644 A
(43)申请公布日 2021.06.04
(73)专利权人 杭州电子科技大学
地址 310018 浙江省杭州市钱塘新区白杨街道2号大街
(72)发明人 徐小良 王梦召 吕凌威
(74)专利代理机构 浙江千克知识产权代理有限公司 33246
专利代理师 周希良
(51)Int.Cl.
G06F 16/2455(2019.01)
G06F 16/22(2019.01)
G06F 16/28(2019.01)

(56)对比文件
US 2018032930 A1,2018.02.01
US 2008077570 A1,2008.03.27
US 2019251455 A1,2019.08.15
EP 2836920 A1,2015.02.18
CN 103412925 A,2013.11.27
WO 2017180475 A1,2017.10.19
樊舒等."基于复杂网络的结构化公安情报流程研究".《情报杂志》.2020,86-91.

审查员 朱琦

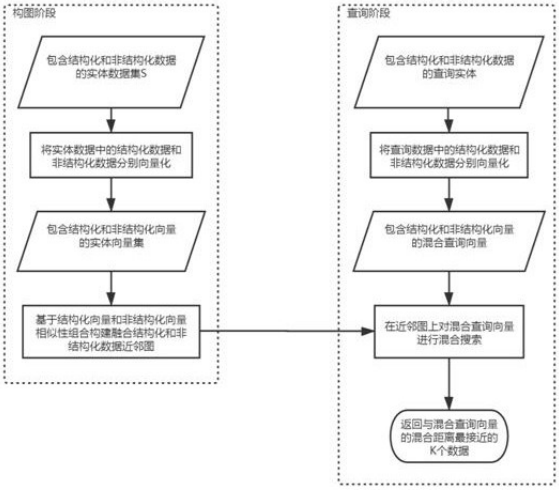
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种融合结构化和非结构化数据的混合搜索方法

(57)摘要

本发明公开了一种融合结构化和非结构化数据的混合搜索方法。该方法首先将数据集中每一个实体所包含的结构化和非结构化数据分别向量化得到包含结构化向量和非结构化向量的实体向量;其次基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图;然后将查询实体所包含的结构化和非结构化数据通过向量化得到包含结构化向量和非结构化向量的混合查询向量;最后混合查询向量在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻。本发明实现了同时对非结构化和结构化数据进行搜索的混合搜索,较之于当前的两种分离的索引系统效率得到较大提升。



CN 112905644 B

1.一种融合结构化和非结构化数据的混合搜索方法,其特征在于包含以下步骤:

(1)将数据集中每一个实体所包含的结构化和非结构化数据分别向量化得到包含结构化向量和非结构化向量的实体向量;

(2)基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图;

(3)将查询实体所包含的结构化和非结构化数据通过与(1)相同的方式向量化得到包含结构化向量和非结构化向量的混合查询向量;

(4)混合查询向量在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻;

其中步骤(1)将数据集S中每一个实体e_i所包含的结构化和非结构化数据分别向量化得到包含非结构化向量α_i和结构化向量β_i的实体向量(α_i,β_i);其中,数据集S表示为:

$S=\{e_i|i=1,2,...,N\}$

其中e_i为数据集中的第i个实体,N为数据集中实体个数;

非结构化向量α_i表示为:

$$\alpha_i = (\alpha_{i_1},\alpha_{i_2},\dots,\alpha_{i_m})$$

其中m为非结构化向量的维数,α_{ij}为非结构化向量α_i在第j维的取值;

结构化向量β_i表示为:

$$\beta_i = (\beta_{i_1},\beta_{i_2},\dots,\beta_{i_n})$$

其中n为结构化向量的维数,β_{ij}结构化向量β_i在第j维的取值;

其中步骤(2)所述的基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图,指通过混合距离计算来评估各个实体向量(α_i,β_i)之间的相似性,从而每个实体向量(α_i,β_i)连接与其混合距离d最近的K个邻居,实体向量(α_1,β_1)与实体向量(α_2,β_2)间的距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$的计算公式为:

$$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))=d_1(\alpha_1,\alpha_2)+w_b\cdot d_2(\beta_1,\beta_2)$$

其中,$d_1(\alpha_1,\alpha_2)$为非结构化向量距离,$d_2(\beta_1,\beta_2)$为结构化向量距离,其中w_b为构建近邻图时结构化向量距离所占的权重,用于调控非结构化向量距离$d_1(\alpha_1,\alpha_2)$和结构化向量距离$d_2(\beta_1,\beta_2)$在混合距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$中的比重;

其中步骤(4)所述的混合查询向量q在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻的过程中采用与以下混合距离计算方式,混合查询向量$q=(q_\alpha,q_\beta)$与实体向量(α_i,β_i)的混合距离d为:

$$d(q,(\alpha_2,\beta_2))=d_1(q_\alpha,\alpha_2)+w_s\cdot d_2(q_\beta,\beta_2)$$

q_α为混合查询向量q的非结构化向量,q_β为混合查询向量q的结构化向量,w_s调节混合距离中非结构化向量距离$d_1(q_\alpha,\alpha_2)$和结构化向量距离$d_2(q_\beta,\beta_2)$所占的比重,通过改变w_s从而调控混合搜索的性能。

一种融合结构化和非结构化数据的混合搜索方法

技术领域

[0001] 本发明涉及近似最近邻搜索领域，具体涉及一种融合结构化和非结构化数据的混合搜索方法。

背景技术

[0002] 各种互联网和智能化应用产生了海量的非结构化数据(图片，视频，语音等)和结构化数据(数字、符号、标签等)，从大规模数据中高效查询获取有用信息是各种人工智能应用的一项核心技术。基于关系型数据库的结构化数据查询已经很成熟并被广泛应用，非结构化数据搜索随着深度学习向量化技术的发展也正在快速应用于各种场景。随着对查询结果一致性越来越高的要求，很多场景都需要同时执行结构化数据和非结构化数据的搜索，即混合搜索。

[0003] 混合搜索方法是目前近似最近邻搜索领域的一个研究热点，在电子商务等平台得到了实际应用。但是，目前的混合搜索系统主要通过对结构化数据和非结构化数据分别执行查询，然后合并它们的查询结果来实现的。这种混合搜索方法存在查询速度慢和查询结果精度低的问题。当前迫切需要一个能够同时执行结构化数据和非结构化数据查询且满足查询精度需求的高效混合搜索解决方案。

发明内容

[0004] 本发明提出了一种融合结构化和非结构化数据的混合搜索方法，这种方法实现了同时对非结构化和结构化数据进行搜索的混合搜索，较之于当前的两种分离的索引系统效率得到较大提升。

[0005] 本发明所提出的一种融合结构化和非结构化数据的混合搜索方法具体内容如下：

[0006] (1)将数据集中每一个实体所包含的结构化和非结构化数据分别向量化得到包含结构化向量和非结构化向量的实体向量；

[0007] (2)基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图；

[0008] (3)将查询实体所包含的结构化和非结构化数据通过与(1)相同的方式向量化得到包含结构化向量和非结构化向量的混合查询向量；

[0009] (4)混合查询向量在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻。

[0010] 其中，步骤(1)将数据集S中每一个实体e_i所包含的结构化和非结构化数据分别向量化得到包含非结构化向量α_i和结构化向量β_i的实体向量(α_i,β_i)。其中，数据集S表示为：

[0011] $S=\{e_i|i=1,2,...,N\}$

[0012] 其中e_i为数据集中的第i个实体，N为数据集中实体个数。

[0013] 非结构化向量α_i表示为：

[0014] $$\alpha_i = (\alpha_{i_1},\alpha_{i_2},\dots,\alpha_{i_m})$$

[0015] 其中m为非结构化向量的维数，α_{ij}为非结构化向量α_i在第j维的取值。

[0016] 结构化向量β_i表示为：

[0017] $$\beta_i = (\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_n})$$

[0018] 其中n为结构化向量的维数，β_{ij}结构化向量β_i在第j维的取值。

[0019] 步骤(2)所述的基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图，指通过混合距离计算来评估各个实体向量(α_i,β_i)之间的相似性，从而每个实体向量(α_i,β_i)连接与其混合距离d最近的K个邻居，实体向量(α_1,β_1)与实体向量(α_2,β_2)间的距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$的计算公式为：

[0020] $d((\alpha_1,\beta_1),(\alpha_2,\beta_2)) = d_1(\alpha_1,\alpha_2) + w_b \cdot d_2(\beta_1,\beta_2)$

[0021] 其中，$d_1(\alpha_1,\alpha_2)$为非结构化向量距离，$d_2(\beta_1,\beta_2)$为结构化向量距离，其中w_b为构建近邻图时结构化向量距离所占的权重，以调控非结构化向量距离$d_1(\alpha_1,\alpha_2)$和结构化向量距离$d_2(\beta_1,\beta_2)$在混合距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$中的比重，进而影响构建的融合结构化和非结构化数据近邻图进行混合搜索的性能。

[0022] 步骤(4)所述的混合查询向量q在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻的过程中采用以下距离计算方式，混合查询向量$q=(q_\alpha,q_\beta)$与实体向量(α_i,β_i)的混合距离d为：

[0023] $d(q,(\alpha_2,\beta_2)) = d_1(q_\alpha,\alpha_2) + w_s \cdot d_2(q_\beta,\beta_2)$

[0024] q_α为混合查询向量q的非结构化向量，q_β为混合查询向量q的结构化向量，w_s调节混合距离中非结构化向量距离$d_1(q_\alpha,\alpha_2)$和结构化向量距离$d_2(q_\beta,\beta_2)$所占的比重，通过改变w_s从而调控混合搜索的性能。

[0025] 本发明的有益效果：本发明提出的基于融合结构化和非结构化数据近邻图的混合搜索方法通过将实体数据中的结构化数据和非结构化数据分别向量化得到实体向量，并基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图，再将查询实体中的结构化数据和非结构化数据分别向量化得到混合查询向量，再用混合查询向量在构建的近邻图上执行贪婪搜索，实现了非结构化和结构化数据的混合搜索，较之于当前的两种分离的索引系统效率得到较大提升。

附图说明

[0026] 图1是本发明的流程示意图。

具体实施方式

[0027] 为了使本发明的技术方案和优点更加明确，下面将结合附图对本发明作近一步的描述说明。

[0028] 图1是本发明的流程示意图，主要包括以下步骤：

[0029] (1)将数据集中每一个实体所包含的结构化和非结构化数据分别向量化得到包含结构化向量和非结构化向量的实体向量；

[0030] 该过程具体为将数据集S中每一个实体e_i所包含的结构化和非结构化数据分别向量化得到包含非结构化向量α_i和结构化向量βi的实体向量(α_i,β_i)。其中，数据集S表示为：

[0031] $S=\{e_i \mid i=1,2,\ldots,N\}$

[0032] 其中e_i为数据集中的第i个实体,N为数据集中实体个数。

[0033] 非结构化向量α_i表示为:

[0034] $$\alpha_i = (\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_m})$$

[0035] 其中m为非结构化向量的维数,α_{i_j}为非结构化向量α_i在第j维的取值。

[0036] 结构化向量β_i表示为:

[0037] $$\beta_i = (\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_n})$$

[0038] 其中n为结构化向量的维数,β_{i_j}结构化向量β_i在第j维的取值。

[0039] (2)基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图;构图过程中通过混合距离计算来评估各个实体向量(α_i,β_i)之间的相似性,从而每个实体向量(α_i,β_i)连接与其混合距离d最近的K个邻居。

[0040] 实体向量(α_1,β_1)与实体向量(α_2,β_2)间的距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$的计算公式为:

[0041] $$d((\alpha_1,\beta_1),(\alpha_2,\beta_2)) = d_1(\alpha_1,\alpha_2) + w_b \cdot d_2(\beta_1,\beta_2)$$

[0042] 其中,$d_1(\alpha_1,\alpha_2)$为非结构化向量距离,$d_2(\beta_1,\beta_2)$为结构化向量距离,其中w_b为构建近邻图时结构化向量距离所占的权重,以调控非结构化向量距离$d_1(\alpha_1,\alpha_2)$和结构化向量距离$d_2(\beta_1,\beta_2)$在混合距离$d((\alpha_1,\beta_1),(\alpha_2,\beta_2))$中的比重,进而影响构建的融合结构化和非结构化数据近邻图进行混合搜索的性能。

[0043] 结构化向量的距离$d_2(\beta_1,\beta_2)$定义如下:

[0044] $$d_2(\beta_1,\beta_2)=\sum_{i=1}^{M} \cdot f(\beta_{1i},\beta_{2i})$$

[0045] 式中,β_{1i}、β_{2i}为结构化向量β_1、β_2的第i维的取值,M为结构化向量的维数,$f(\beta_{1i},\beta_{2i})$为结构化向量第i维的取值的距离计算函数,定义如下:

[0046] $$f(\beta_{1i},\beta_{2i}) = \begin{cases} 0 & (\beta_{1i}=\beta_{2i}) \\ 1 & (\text{其他}) \end{cases}$$

[0047] (3)将查询实体所包含的结构化和非结构化数据通过与(1)相同的方式向量化得到包含结构化向量和非结构化向量的混合查询向量;

[0048] (4)混合查询向量在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻。具体为混合查询向量q在融合结构化和非结构化数据近邻图上通过贪婪算法执行混合搜索得到查询实体的最近邻的过程中采用以下距离计算方式。

[0049] 混合查询向量$q=(q_\alpha,q_\beta)$与实体向量(α_i,β_i)的混合距离d为:

[0050] $$d(q,(\alpha_2,\beta_2)) = d_1(q_\alpha,\alpha_2) + w_s \cdot d_2(q_\beta,\beta_2)$$

[0051] q_α为混合查询向量q的非结构化向量,q_β为混合查询向量q的结构化向量,w_s调节混合距离中非结构化向量距离$d_1(q_\alpha,\alpha_2)$和结构化向量距离$d_2(q_\beta,\beta_2)$所占的比重,通过改变w_s从而调控混合搜索的性能。

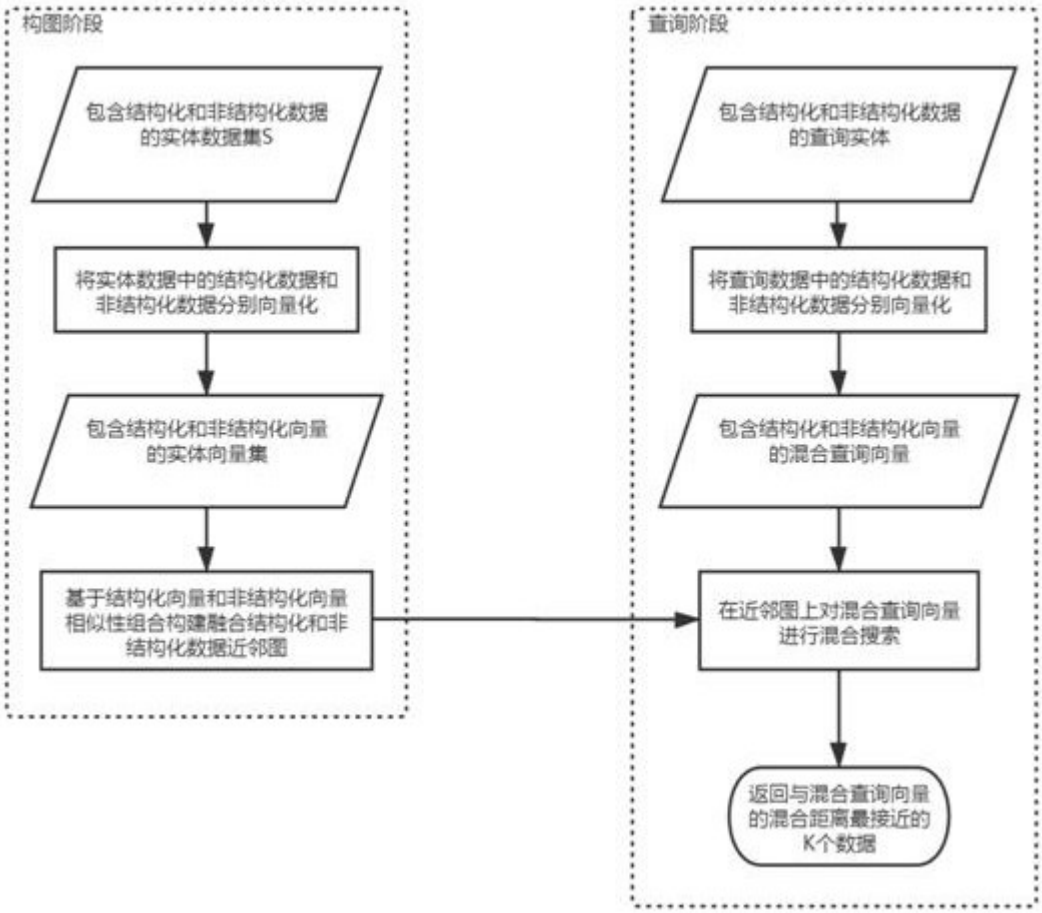

构图阶段
包含结构化和非结构化数据的实体数据集S
将实体数据中的结构化数据和非结构化数据分别向量化
包含结构化和非结构化向量的实体向量集
基于结构化向量和非结构化向量相似性组合构建融合结构化和非结构化数据近邻图
查询阶段
包含结构化和非结构化数据的查询实体
将查询数据中的结构化数据和非结构化数据分别向量化
包含结构化和非结构化向量的混合查询向量
在近邻图上对混合查询向量进行混合搜索
返回与混合查询向量的混合距离最接近的K个数据

图1