(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113505273 B
(45)授权公告日 2023.08.22

(21)申请号 202110566211.1

(22)申请日 2021.05.24

(65)同一申请的已公布的文献号
申请公布号 CN 113505273 A

(43)申请公布日 2021.10.15

(73)专利权人 平安银行股份有限公司
地址 518000 广东省深圳市罗湖区深南东路5047号

(72)发明人 李珊

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
专利代理师 高杰　于志光

(51)Int.Cl.
G06F 16/901(2019.01)
G06F 16/904(2019.01)
G06F 16/906(2019.01)

(56)对比文件
CN 101477554 A,2009.07.08
CN 111859057 A,2020.10.30
CN 109207606 A,2019.01.15
CN 110046298 A,2019.07.23
CN 110378560 A,2019.10.25
CN 109598307 A,2019.04.09
CN 111008321 A,2020.04.14
US 2016203228 A1,2016.07.14
CN 112328657 A,2021.02.05

审查员 李娇娇

权利要求书2页　说明书12页　附图2页

(54)发明名称

基于重复数据筛选的数据排序方法、装置、设备及介质

(57)摘要

本发明涉及智能决策领域，揭露一种基于重复数据筛选的数据排序方法，包括：根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；对查询结果序列进行标签分类，得到第一分类结果序列；对第一分类结果序列进行相关度分类，得到第二分类结果序列；对第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；将第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列。本发明还涉及一种区块链技术，所述查询结果序列可以存储在区块链节点中。本发明还提出一种基于重复数据筛选的数据排序装置、设备以及介质。本发明可以提高数据排序的效率。

CN 113505273 B

1.一种基于重复数据筛选的数据排序方法，其特征在于，所述方法包括：

根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；

对所述查询结果序列进行标签分类，得到第一分类结果序列；

对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；

将所述目标查询结果序列发送至所述查询请求对应的终端设备；

其中，所述根据所述查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列，包括：提取所述查询请求中的查询字段，将所述查询字段转换为向量，得到查询向量；将所述资源数据集中的每个资源数据转换为向量，得到对应的资源向量；计算所述查询向量及所述资源向量的相关度，得到对应的相关度分数；筛选所述资源数据集中所述相关度分数大于预设相关度的资源数据，得到初始查询结果序列；将所述初始查询结果序列中所有资源数据按照对应的相关度分数大小进行排序，得到所述查询结果序列；

所述对所述第一分类结果序列进行相关度分类，得到第二分类结果序列，包括：根据所述查询结果序列构建分值区间；利用所述分值区间对所述第一分类结果序列进行分类，得到所述第二分类结果序列；

所述根据所述查询结果序列构建分值区间，包括：筛选所述查询结果序列的最大相关度分数，得到第一区间数据；筛选所述查询结果序列的最小相关度分数，得到第二区间数据；将所述第一区间数据及所述第二区间数据进行平均计算，得到第三区间数据；将所述第一区间数据、第二区间数据及所述第三区间数据作为区间端点值构建两个连续区间，得到所述分值区间。

2.如权利要求1所述的基于重复数据筛选的数据排序方法，其特征在于，所述对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列，包括：

利用预设算法对所述第二分类结果序列中每个资源数据进行编码，得到对应的数据编码；

计算所述第二分类结果序列对应的所有数据编码中任意两个数据编码的文本距离；

将小于预设阈值的所述文本距离确定为相似文本距离；

将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类，得到重复数据列表；

将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算，得到所述第三分类结果序列。

3.如权利要求2所述的基于重复数据筛选的数据排序方法，其特征在于，所述将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类，得到重复数据列表，包括：

将所述第二分类结果序列中所有相似文本距离对应的资源数据作为节点进行树状分类，得到分类树；

将所述分类树对应的所有资源数据按照每个资源数据对应相关度分数进行排序，得到所述重复数据列表。

4.如权利要求2所述的基于重复数据筛选的数据排序方法，其特征在于，所述将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算，得到所述第三分类结果序列，包括：

对所述第二分类结果序列对应的重复数据列表中预设排序位置及之后的所有资源数据对应的相关度分数进行指数计算，得到对应的更新后的相关度分数；

利用所述更新后的相关度分数替换对应的所述相关度分数，得到所述第三分类结果序列。

5.一种基于重复数据筛选的数据排序装置，用于实现如权利要求1至4中任一项所述的基于重复数据筛选的数据排序方法，其特征在于，包括：

数据分类模块，用于根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；对所述查询结果序列进行标签分类，得到第一分类结果序列；对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

数据筛选模块，用于对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

数据排序模块，用于将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；将所述目标查询结果序列发送至所述查询请求对应的终端设备。

6.一种电子设备，其特征在于，所述电子设备包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行如权利要求1至4中任一项所述的基于重复数据筛选的数据排序方法。

7.一种计算机可读存储介质，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4中任一项所述的基于重复数据筛选的数据排序方法。

基于重复数据筛选的数据排序方法、装置、设备及介质

技术领域

[0001] 本发明涉及智能决策领域，尤其涉及一种基于重复数据筛选的数据排序方法、装置、电子设备及可读存储介质。

背景技术

[0002] 目前，数据排序在数据检索及数据推荐领域应用的非常广泛。在这种检索和推荐场景下，通常对检索或推荐的数据进行相关度打分，将所有数据按照分值从高到低降序进行排序展示。

[0003] 但是由于检索或推荐的数据通常是非常丰富甚至会出现重复的，目前的数据排序方式会存在将相同或相似的数据堆在一起进行展示的问题，相似内容扎堆出现覆盖占据了大量的显示空间，导致对有效信息的获取变得困难，数据排序的效率低。

发明内容

[0004] 本发明提供一种基于重复数据筛选的数据排序方法、装置、电子设备及计算机可读存储介质，其主要目的在于提高数据排序的效率。

[0005] 为实现上述目的，本发明提供的一种基于重复数据筛选的数据排序方法，包括：

[0006] 根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；

[0007] 对所述查询结果序列进行标签分类，得到第一分类结果序列；

[0008] 对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0009] 对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

[0010] 将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；

[0011] 将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0012] 可选地，所述根据所述查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列，包括：

[0013] 提取所述查询请求中的查询字段，将所述查询字段转换为向量，得到查询向量；

[0014] 将所述资源数据集中的每个资源数据转换为向量，得到对应的资源向量；

[0015] 计算所述查询向量及所述资源向量的相关度，得到对应的相关度分数；

[0016] 筛选所述资源数据集中所述相关度分数大于预设相关度的资源数据，得到所述初始查询结果序列；

[0017] 将所述初始查询结果序列中所有资源数据按照对应的相关度分数大小进行排序，得到所述查询结果序列。

[0018] 可选地，所述对所述第一分类结果序列进行相关度分类，得到第二分类结果序列，包括：

[0019] 根据所述查询结果序列构建分值区间；

[0020] 利用所述分值区间对所述第一分类结果序列进行分类，得到所述第二分类结果序列。

[0021] 可选地，所述根据所述查询结果序列构建分值区间，包括：

[0022] 筛选所述查询结果序列的最大相关度分数，得到第一区间数据；

[0023] 筛选所述查询结果序列的最小相关度分数，得到第二区间数据；

[0024] 将所述第一区间数据及所述第二区间数据进行平均计算，得到第三区间数据；

[0025] 将所述第一区间数据、第二区间数据及所述第三区间数据作为区间端点值构建两个连续区间，得到所述分值区间。

[0026] 可选地，所述对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列，包括：

[0027] 利用预设算法对所述第二分类结果序列中每个资源数据进行编码，得到对应的数据编码；

[0028] 计算所述第二分类结果序列对应的所有数据编码中任意两个数据编码的文本距离；

[0029] 将小于预设阈值的所述文本距离确定为相似文本距离；

[0030] 将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类，得到重复数据列表；

[0031] 将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算，得到所述第三分类结果序列。

[0032] 可选地，所述将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类，得到重复数据列表，包括：

[0033] 将所述第二分类结果序列中所有相似文本距离对应的资源数据作为节点进行树状分类，得到分类树；

[0034] 将所述分类树对应的所有资源数据按照每个资源数据对应相关度分数进行排序，得到所述重复数据列表。

[0035] 可选地，所述将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算，得到所述第三分类结果序列，包括：

[0036] 对所述第二分类结果序列对应的重复数据列表中预设排序位置及之后的所有资源数据对应的相关度分数进行指数计算，得到对应的更新后的相关度分数；

[0037] 利用所述更新后的相关度分数替换对应的所述相关度分数，得到所述第三分类结果序列。

[0038] 为了解决上述问题，本发明还提供一种基于重复数据筛选的数据排序装置，所述装置包括：

[0039] 数据分类模块，用于根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；对所述查询结果序列进行标签分类，得到第一分类结果序列；对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0040] 数据筛选模块，用于对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

[0041] 数据排序模块，用于将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0042] 为了解决上述问题，本发明还提供一种电子设备，所述电子设备包括：

[0043] 存储器，存储至少一个计算机程序；及

[0044] 处理器，执行所述存储器中存储的计算机程序以实现上述所述的基于重复数据筛选的数据排序方法。

[0045] 为了解决上述问题，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个计算机程序，所述至少一个计算机程序被电子设备中的处理器执行以实现上述所述的基于重复数据筛选的数据排序方法。

[0046] 本发明实施例通过根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；对所述查询结果序列进行标签分类，得到第一分类结果序列，将不同标签的数据进行分类，避免同类数据扎堆显示；对所述第一分类结果序列进行相关度分类，得到第二分类结果序列，将每类标签的数据按照高相关度分和地相关度分进行分类，基于重复数据筛选的数据排序更加均衡；对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列，将重复数据的相关度分数进行降分，避免相似数据扎堆，使得排序后的数据显示更加多样，提高了基于重复数据筛选的数据排序的效率更高；将所述目标查询结果序列发送至所述查询请求对应的终端设备。因此本发明实施例提出的基于重复数据筛选的数据排序方法、装置、电子设备及可读存储介质提高了基于重复数据筛选的数据排序的效率。

附图说明

[0047] 图1为本发明一实施例提供的基于重复数据筛选的数据排序方法的流程示意图；

[0048] 图2为本发明一实施例提供的基于重复数据筛选的数据排序装置的模块示意图；

[0049] 图3为本发明一实施例提供的实现基于重复数据筛选的数据排序方法的电子设备的内部结构示意图；

[0050] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0051] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0052] 本发明实施例提供一种基于重复数据筛选的数据排序方法。所述基于重复数据筛选的数据排序方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之，所述基于重复数据筛选的数据排序方法可以由安装在终端设备或服务端设备的软件或硬件来执行，所述软件可以是区块链平台。所述服务端包括但不限于：单台服务器、服务器集群、云端服务器或云端服务器集群等。

[0053] 参照图1所示的本发明一实施例提供的基于重复数据筛选的数据排序方法的流程示意图，在本发明实施例中，所述基于重复数据筛选的数据排序方法包括：

[0054] S1、根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结

果序列；

[0055] 本发明实施例中的，所述查询请求包括：查询字段，所述资源数据集为包含不同资源数据的集合，其中所述资源数据可以为咨询数据、产品数据、活动数据等。

[0056] 详细地，本发明实施例中，提取所述查询请求中的查询字段，将所述查询字段转换为向量，得到查询向量；将所述资源数据集中的每个资源数据转换为向量，得到对应的资源向量；计算所述查询向量及所述资源向量的相关度，得到对应的相关度分数；根据所述相关度分数对所述资源数据集进行数据筛选，得到初始查询结果序列；将所述初始查询结果序列中所有资源数据按照对应的相关度分数大小进行排序，得到所述查询结果序列。

[0057] 可选地，本发明实施例中可利用预设的基于专业领域知识文本(如教材、培训资料)通过迁移学习训练而成的Word2vec模型进行向量转换。进一步地，本发明实施例中筛选所述资源数据集中所述相关度分数大于预设相关度值的资源数据，得到所述初始查询结果序列。

[0058] 可选地，本发明实施例可用如下公式计算所述相关度：

[0059]
$$\mathrm{Sim}=\frac{\sum_{i=1}^{n}X_i*Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2*\sum_{i=1}^{n}(Y_i)^2}}$$

[0060] 其中，X_i表示查询向量X的第i个元素，Y_i为资源向量Y的第i个元素，n表示Sim表示查询向量X和资源向量Y的相关度分数。

[0061] 本发明的另一实施例中，所述查询结果序列可以存储在区块链节点中，利用区块链高吞吐的特性，提高数据的存取效率。

[0062] S2、对所述查询结果序列进行标签分类，得到第一分类结果序列；

[0063] 本发明实施例中，所述第一分类结果序列包含有不同属性类别的资源数据，为了对所述查询结果序列中的资源数据更好的进行排序，对所述查询结果序列进行标签分类，得到第一分类结果序列。

[0064] 详细地，本发明实施中，所述对所述查询结果序列进行标签分类，包括：利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记，得到标签查询结果序列；将所述标签查询结果序列中所有资源数据按照不同的标签进行归类，得到对应的所述第一分类结果序列。

[0065] 本发明实施例中由于所述查询结果序列中不同的资源数据对应的标签不同，每一类标签对应一个第一分类结果序列，因此，对所述查询结果序列进行标签分类，得到的第一分类结果序列有多个，例如：所述查询结果序列中的所有资源数据可以用4类标签进行标记，那么每一类标签对应一个第一分类结果序列，共可以得到4个第一分类结果序列。

[0066] 本发明实施例中，所述标签分类模型可以是一种由Bert网络构建的深度学习模型。

[0067] 详细地，本发明实施例中所述利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记之前，还包括：获取历史资源数据集；对所述历史资源数据集进行预设标签标记，得到训练集，本发明其中一个应用场景中，所述预设标签可以包括：资讯品类标签、信用卡产品品类标签、保险产品品类标签、商城商品品类标签、优惠活动品类标签等；利用所述训练集对预构建的深度学习模型进行迭代训练，得到所述标签分类模型。

[0068] 可选地，本发明另一实施例中，所述第一分类结果序列中每个资源数据本身都包含有对应的标签，而不需要利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记，将所述第一分类结果序列中相同资源类别标签对应的资源数据进行汇总，得到对应的第一分类结果序列。

[0069] S3、对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0070] 本发明实施例中，根据上述的S1可知，所述第一分类结果序列中每个资源数据都有相关度分数，为了保证所述第一分类结果序列中的所有资源数据排序更加均衡，防止后续高相关度分数的资源数据对低相关度分数的资源数据造成干扰，影响数据排序的准确性，本发明实施例进一步对所述第一分类结果序列进行相关度分类，得到第二分类结果序列。

[0071] 详细地，本发明实施例中对所述第一分类结果序列进行相关度分类，包括：根据所述查询结果序列构建分值区间，利用所述分值区间对所述第一分类结果序列进行分类，得到所述第二分类结果序列。

[0072] 进一步地，本发明实施例中，所述根据所述查询结果序列构建分值区间；包括：筛选所述查询结果序列的最大相关度分数及最小相关度分数，将所述最大相关度分数及所述最小相关度分数进行平均计算，得到平均相关度分数；将所述最大相关度分数、最小相关度分数及所述平均相关度分数作为区间端点值构建两个连续区间，得到所述分值区间。例如：所述最大相关度分数为10，最小相关度分数为0，那么平均相关度分数为(10+0)/2＝5，将0、5、10构建两个连续区间，得到分值区间为[0,5]及(5,10]。可选地，本发明另一实施例中，所述分值区间可以根据实践经验或业务需求进行调整。本发明实施例中每个所述第一分类结果序列对应多个第二分类结果序列，每个所述第一分类结果序列对应的第二分类结果序列的个数由所述分值区间包含的区间个数决定，例如：分支区间包含两个区间，那么每个所述第一分类结果序列对应的第二分类结果序列的个数为2个。本发明的另一实施例中，所述对所述第一分类结果序列进行相关度分类，包括：将所述第一分类结果序列中所有资源数据按照对应的相关度分数大小进行先后排序，得到标准第一分类结果序列；将所述标准第一分类结果序列中的数据按照预设的排序百分比进行分类，得到所述第二分类结果序列。如：预设的排序百分比为50％，所述标准第一分类结果序列共有10个资源数据，那么将标准第一分类结果序列中排序在前50％的资源数据分为一类，剩余的资源数据分为一类，得到对应的所述第二分类结果序列。

[0073] S4、对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

[0074] 本发明实施例中为了防止重复或者相似的资源数据扎堆显示，导致数据显示类型狭窄，因此，对所述第二分类结果序列进行重复数据筛选，并将筛选的所述第二分类结果序列中的重复数据进行指数降分计算，得到第三分类结果序列，其中，所述重复数据包括相同或相似的数据。

[0075] 详细地，本发明实施例中对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列，包括：利用预设算法对所述第二分类结果序列中每个资源数据进行编码，得到每个资源数据对应的数据编码；计算所述第二分类结果序列对应的所有数据编码中任意两个数据编码的文本距离；将小于预设阈值的所

述文本距离确定为相似文本距离;将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类,得到重复数据列表;将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算,得到所述第三分类结果序列。

[0076] 可选地,本发明实施例中所述预设算法为simhash算法,

[0077] 详细地,本发明实施例中将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类,得到重复数据列表,包括:将所述第二分类结果序列中所有相似文本距离对应的资源数据作为节点进行树状分类,得到分类树;将所述分类树对应的所有资源数据按照每个资源数据对应相关度分数进行排序,得到对应的重复数据列表。例如:A和B的文本距离为相似距离;A和C为的文本距离为相似距离;B和E的文本距离为相似距离,那么将A作为分类数第一层的节点,B、C作为分类树第二层的节点,将E作为分类树的第三层节点构建得到对应的分类树。

[0078] 进一步地,为了避免重复数据扎堆,本发明实施例中将所述重复数据列表中的资源数据进行指数降分计算,包括:对所述第二分类结果序列对应的重复数据列表中预设排序位置及之后的所有资源数据对应的相关度分数进行指数计算,得到对应的更新后的相关度分数。

[0079] 进一步地,本发明实施例利用所述更新后的相关度分数替换对应的所述相关度分数,得到所述第三分类结果序列。

[0080] 可选地,所述预设排序位置为第二个。

[0081] 可选地,本发明实施例中利用如下公式进行指数计算:

[0082] $N=a^{lgi}*C_i$

[0083] 其中,a为预设的排序参数,较佳地,a为0.5,C_i为所述重复数据列表中第i个资源数据对应的相关度分数,i为所述重复数据列表中的资源数据的排序编号,N为所述重复数据列表中第i个资源数据更新后的相关度分数。

[0084] 可选地,本发明实施例中两个数据编码的文本距离为对应的两个数据编码的海明距离。

[0085] S5、将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序,得到目标查询结果序列;

[0086] 本发明实施例中由上述内容可知所述第二分类结果序列有多个,因此,所述第三分类结果序列也有多个,进一步地,本发明实施例将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序,得到目标查询结果序列,例如:共有两个第三分类结果序列,其中一个第三分类结果序列中包含资源数据A及资源数据B,A的相关度分数为10,B的相关度分数为8,另一个第三分类结果序列中包含资源数据C及资源数据D,C的相关度分数为9,C的相关度分数为7,那么所述第三分类结果序列中所有资源数据为A、B、C、D,根据相关度分数将所述第三分类结果序列中所有资源数据进行排序得到的目标查询结果序列为[A,C,B,D]。

[0087] S6、将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0088] 详细地,本发明实施例中将所述目标查询结果序列发送至所述查询请求对应的终端设备,所述终端设备包括:电脑、平板、手机等智能终端,例如:用户在手机A上发起查询请求,那么就将所述目标查询结果序列发送到手机A,方便用户查看。

[0089] 如图2所示，是本发明基于重复数据筛选的数据排序装置的功能模块图。

[0090] 本发明所述基于重复数据筛选的数据排序装置100可以安装于电子设备中。根据实现的功能，所述基于重复数据筛选的数据排序装置可以包括数据分类模块101、数据筛选模块102、数据排序模块103，本发所述模块也可以称之为单元，是指一种能够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。

[0091] 在本实施例中，关于各模块/单元的功能如下：

[0092] 所述数据分类模块101用于根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；对所述查询结果序列进行标签分类，得到第一分类结果序列；对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0093] 本发明实施例中的，所述查询请求包括：查询字段，所述资源数据集为包含不同资源数据的集合，其中所述资源数据可以为咨询数据、产品数据、活动数据等。

[0094] 详细地，本发明实施例中，所述数据分类模块101提取所述查询请求中的查询字段，将所述查询字段转换为向量，得到查询向量；将所述资源数据集中的每个资源数据转换为向量，得到对应的资源向量；计算所述查询向量及所述资源向量的相关度，得到对应的相关度分数；根据所述相关度分数对所述资源数据集进行数据筛选，得到初始查询结果序列；将所述初始查询结果序列中所有资源数据按照对应的相关度分数大小进行排序，得到所述查询结果序列。

[0095] 可选地，本发明实施例中所述数据分类模块101可利用预设的基于专业领域知识文本(如教材、培训资料)通过迁移学习训练而成的Word2vec模型进行向量转换。进一步地，本发明实施例中所述数据分类模块101筛选所述资源数据集中所述相关度分数大于预设相关度值的资源数据，得到所述初始查询结果序列。

[0096] 可选地，本发明实施例所述数据分类模块101可用如下公式计算所述相关度：

[0097]
$$\mathrm{Sim}=\frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2 * \sum_{i=1}^{n}(Y_i)^2}}$$

[0098] 其中，X_i表示查询向量X的第i个元素，Y_i为资源向量Y的第i个元素，n表示Sim表示查询向量X和资源向量Y的相关度分数。

[0099] 本发明实施例中，所述第一分类结果序列包含有不同属性类别的资源数据，为了对所述查询结果序列中的资源数据更好的进行排序，所述数据分类模块101对所述查询结果序列进行标签分类，得到第一分类结果序列。

[0100] 本发明的另一实施例中，所述查询结果序列可以存储在区块链节点中，利用区块链高吞吐的特性，提高数据的存取效率。

[0101] 详细地，本发明实施中，所述数据分类模块101对所述查询结果序列进行标签分类，包括：利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记，得到标签查询结果序列；将所述标签查询结果序列中所有资源数据按照不同的标签进行归类，得到对应的所述第一分类结果序列。

[0102] 本发明实施例中由于所述查询结果序列中不同的资源数据对应的标签不同，每一类标签对应一个第一分类结果序列，因此，对所述查询结果序列进行标签分类，得到的第一

分类结果序列有多个，例如：所述查询结果序列中的所有资源数据可以用4类标签进行标记，那么每一类标签对应一个第一分类结果序列，共可以得到4个第一分类结果序列。

[0103] 本发明实施例中，所述标签分类模型可以是一种由Bert网络构建的深度学习模型。

[0104] 详细地，本发明实施例中所述数据分类模块101利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记之前，还包括：获取历史资源数据集；对所述历史资源数据集进行预设标签标记，得到训练集，本发明其中一个应用场景中，所述预设标签可以包括：资讯品类标签、信用卡产品品类标签、保险产品品类标签、商城商品品类标签、优惠活动品类标签等；利用所述训练集对预构建的深度学习模型进行迭代训练，得到所述标签分类模型。

[0105] 可选地，本发明另一实施例中，所述第一分类结果序列中每个资源数据本身都包含有对应的标签，而不需要利用预设的标签分类模型对所述查询结果序列中的每个资源数据进行标签标记，所述数据分类模块101将所述第一分类结果序列中相同资源类别标签对应的资源数据进行汇总，得到对应的第一分类结果序列。

[0106] 本发明实施例中，所述第一分类结果序列中每个资源数据都有相关度分数，为了保证所述第一分类结果序列中的所有资源数据排序更加均衡，防止后续高相关度分数的资源数据对低相关度分数的资源数据造成干扰，影响数据排序的准确性，本发明实施例所述数据分类模块101进一步对所述第一分类结果序列进行相关度分类，得到第二分类结果序列。

[0107] 详细地，本发明实施例中所述数据分类模块101对所述第一分类结果序列进行相关度分类，包括：根据所述查询结果序列构建分值区间，利用所述分值区间对所述第一分类结果序列进行分类，得到所述第二分类结果序列。

[0108] 进一步地，本发明实施例中，所述数据分类模块101根据所述查询结果序列构建分值区间；包括：筛选所述查询结果序列的最大相关度分数及最小相关度分数，将所述最大相关度分数及所述最小相关度分数进行平均计算，得到平均相关度分数；将所述最大相关度分数、最小相关度分数及所述平均相关度分数作为区间端点值构建两个连续区间，得到所述分值区间。例如：所述最大相关度分数为10，最小相关度分数为0，那么平均相关度分数为(10+0)/2＝5，将0、5、10构建两个连续区间，得到分值区间为[0,5]及(5,10]。可选地，本发明另一实施例中，所述分值区间可以根据实践经验或业务需求进行调整。本发明实施例中每个所述第一分类结果序列对应多个第二分类结果序列，每个所述第一分类结果序列对应的第二分类结果序列的个数由所述分值区间包含的区间个数决定，例如：分支区间包含两个区间，那么每个所述第一分类结果序列对应的第二分类结果序列的个数为2个。

[0109] 本发明的另一实施例中，所述数据分类模块101对所述第一分类结果序列进行相关度分类，包括：将所述第一分类结果序列中所有资源数据按照对应的相关度分数大小进行先后排序，得到标准第一分类结果序列；将所述标准第一分类结果序列中的数据按照预设的排序百分比进行分类，得到所述第二分类结果序列。如：预设的排序百分比为50％，所述标准第一分类结果序列共有10个资源数据，那么将标准第一分类结果序列中排序在前50％的资源数据分为一类，剩余的资源数据分为一类，得到对应的所述第二分类结果序列。

[0110] 所述数据筛选模块102用于对所述第二分类结果序列进行重复数据筛选，并将筛

选的重复数据进行指数降分计算,得到第三分类结果序列;

[0111] 本发明实施例中为了防止重复或者相似的资源数据扎堆显示,导致数据显示类型狭窄,因此,所述数据筛选模块102对所述第二分类结果序列进行重复数据筛选,并将筛选的所述第二分类结果序列中的重复数据进行指数降分计算,得到第三分类结果序列,其中,所述重复数据包括相同或相似的数据。

[0112] 详细地,本发明实施例中所述数据筛选模块102对所述第二分类结果序列进行重复数据筛选,并将筛选的重复数据进行指数降分计算,得到第三分类结果序列,包括:利用预设算法对所述第二分类结果序列中每个资源数据进行编码,得到每个资源数据对应的数据编码;计算所述第二分类结果序列对应的所有数据编码中任意两个数据编码的文本距离;将小于预设阈值的所述文本距离确定为相似文本距离;将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类,得到重复数据列表;将所述第二分类结果序列对应的重复数据列表中的资源数据进行指数降分计算,得到所述第三分类结果序列。

[0113] 可选地,本发明实施例中所述预设算法为simhash算法,

[0114] 详细地,本发明实施例中所述数据筛选模块102将所述第二分类结果序列中所有相似文本距离对应的资源数据进行关联分类,得到重复数据列表,包括:将所述第二分类结果序列中所有相似文本距离对应的资源数据作为节点进行树状分类,得到分类树;将所述分类树对应的所有资源数据按照每个资源数据对应相关度分数进行排序,得到对应的重复数据列表。例如:A和B的文本距离为相似距离;A和C为的文本距离为相似距离;B和E的文本距离为相似距离,那么将A作为分类数第一层的节点,B、C作为分类树第二层的节点,将E作为分类树的第三层节点构建得到对应的分类树。

[0115] 进一步地,为了避免重复数据扎堆,本发明实施例中所述数据筛选模块102将所述重复数据列表中的资源数据进行指数降分计算,包括:对所述第二分类结果序列对应的重复数据列表中预设排序位置及之后的所有资源数据对应的相关度分数进行指数计算,得到对应的更新后的相关度分数。

[0116] 进一步地,本发明实施例所述数据筛选模块102利用所述更新后的相关度分数替换对应的所述相关度分数,得到所述第三分类结果序列。

[0117] 可选地,所述预设排序位置为第二个。

[0118] 可选地,本发明实施例中所述数据筛选模块102利用如下公式进行指数计算:

[0119] $N=a^{\lg i}*C_i$

[0120] 其中,a为预设的排序参数,较佳地,a为0.5,C_i为所述重复数据列表中第i个资源数据对应的相关度分数,i为所述重复数据列表中的资源数据的排序编号,N为所述重复数据列表中第i个资源数据更新后的相关度分数。

[0121] 可选地,本发明实施例中两个数据编码的文本距离为对应的两个数据编码的海明距离。

[0122] 所述数据排序模块103用于将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序,得到目标查询结果序列;将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0123] 本发明实施例中由上述内容可知所述第二分类结果序列有多个,因此,所述第三分类结果序列也有多个,进一步地,本发明实施例所述数据排序模块103将所述第三分类结

果序列中所有资源数据按照每个资源数据对应相关度分数进行排序,得到目标查询结果序列,例如:共有两个第三分类结果序列,其中一个第三分类结果序列中包含资源数据A及资源数据B,A的相关度分数为10,B的相关度分数为8,另一个第三分类结果序列中包含资源数据C及资源数据D,C的相关度分数为9,C的相关度分数为7,那么所述第三分类结果序列中所有资源数据为A、B、C、D,根据相关度分数将所述第三分类结果序列中所有资源数据进行排序得到的目标查询结果序列为[A,C,B,D]。

[0124] 详细地,本发明实施例中将所述目标查询结果序列发送至所述查询请求对应的终端设备,所述终端设备包括:电脑、平板、手机等智能终端,例如:用户在手机A上发起查询请求,那么就将所述目标查询结果序列发送到手机A,方便用户查看。

[0125] 如图3所示,是本发明实现基于重复数据筛选的数据排序方法的电子设备的结构示意图。

[0126] 所述电子设备可以包括处理器10、存储器11、通信总线12和通信接口13,还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序,如基于重复数据筛选的数据排序程序。

[0127] 其中,所述存储器11至少包括一种类型的可读存储介质,所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如:SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备的内部存储单元,例如该电子设备的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备的外部存储设备,例如电子设备上配备的插接式移动硬盘、智能存储卡(Smart Media Card,SMC)、安全数字(Secure Digital,SD)卡、闪存卡(Flash Card)等。进一步地,所述存储器11还可以既包括电子设备的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备的应用软件及各类数据,例如基于重复数据筛选的数据排序程序的代码等,还可以用于暂时地存储已经输出或者将要输出的数据。

[0128] 所述处理器10在一些实施例中可以由集成电路组成,例如可以由单个封装的集成电路所组成,也可以是由多个相同功能或不同功能封装的集成电路所组成,包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心(Control Unit),利用各种接口和线路连接整个电子设备的各个部件,通过运行或执行存储在所述存储器11内的程序或者模块(例如基于重复数据筛选的数据排序程序等),以及调用存储在所述存储器11内的数据,以执行电子设备的各种功能和处理数据。

[0129] 所述通信总线12可以是外设部件互连标准(perIPheral component interconnect,简称PCI)总线或扩展工业标准结构(extended industry standard architecture,简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。所述通信总线12总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。为便于表示,图中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0130] 图3仅示出了具有部件的电子设备,本领域技术人员可以理解的是,图3示出的结构并不构成对所述电子设备的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0131] 例如,尽管未示出,所述电子设备还可以包括给各个部件供电的电源(比如电池),

优选地，电源可以通过电源管理装置与所述至少一个处理器10逻辑相连，从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备还可以包括多种传感器、蓝牙模块、Wi-Fi模块等，在此不再赘述。

[0132] 可选地，所述通信接口13可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等)，通常用于在该电子设备与其他电子设备之间建立通信连接。

[0133] 可选地，所述通信接口13还可以包括用户接口，用户接口可以是显示器(Display)、输入单元(比如键盘(Keyboard))，可选地，用户接口还可以是标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode，有机发光二极管)触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在电子设备中处理的信息以及用于显示可视化的用户界面。

[0134] 应该了解，所述实施例仅为说明之用，在专利申请范围上并不受此结构的限制。

[0135] 所述电子设备中的所述存储器11存储的基于重复数据筛选的数据排序程序是多个计算机程序的组合，在所述处理器10中运行时，可以实现：

[0136] 根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；

[0137] 对所述查询结果序列进行标签分类，得到第一分类结果序列；

[0138] 对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0139] 对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列；

[0140] 将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；

[0141] 将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0142] 具体地，所述处理器10对上述计算机程序的具体实现方法可参考图1对应实施例中相关步骤的描述，在此不赘述。

[0143] 进一步地，所述电子设备集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。所述计算机可读介质可以是非易失性的，也可以是易失性的。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)。

[0144] 本发明实施例还可以提供一种计算机可读存储介质，所述可读存储介质存储有计算机程序，所述计算机程序在被电子设备的处理器所执行时，可以实现：

[0145] 根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列；

[0146] 对所述查询结果序列进行标签分类，得到第一分类结果序列；

[0147] 对所述第一分类结果序列进行相关度分类，得到第二分类结果序列；

[0148] 对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分

计算，得到第三分类结果序列；

[0149] 将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列；

[0150] 将所述目标查询结果序列发送至所述查询请求对应的终端设备。

[0151] 进一步地，所述计算机可用存储介质可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据区块链节点的使用所创建的数据等。

[0152] 在本发明所提供的几个实施例中，应该理解到，所揭露的设备，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0153] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0154] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0155] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0156] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0157] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0158] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0159] 最后应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。


根据接收的查询请求对预设的资源数据集进行相关性筛选排序，得到查询结果序列
S1
对所述查询结果序列进行标签分类，得到第一分类结果序列
S2
对所述第一分类结果序列进行相关度分类，得到第二分类结果序列
S3
对所述第二分类结果序列进行重复数据筛选，并将筛选的重复数据进行指数降分计算，得到第三分类结果序列
S4
将所述第三分类结果序列中所有资源数据按照每个资源数据对应相关度分数进行排序，得到目标查询结果序列
S5
将所述目标查询结果序列发送至所述查询请求对应的终端设备
S6

图1

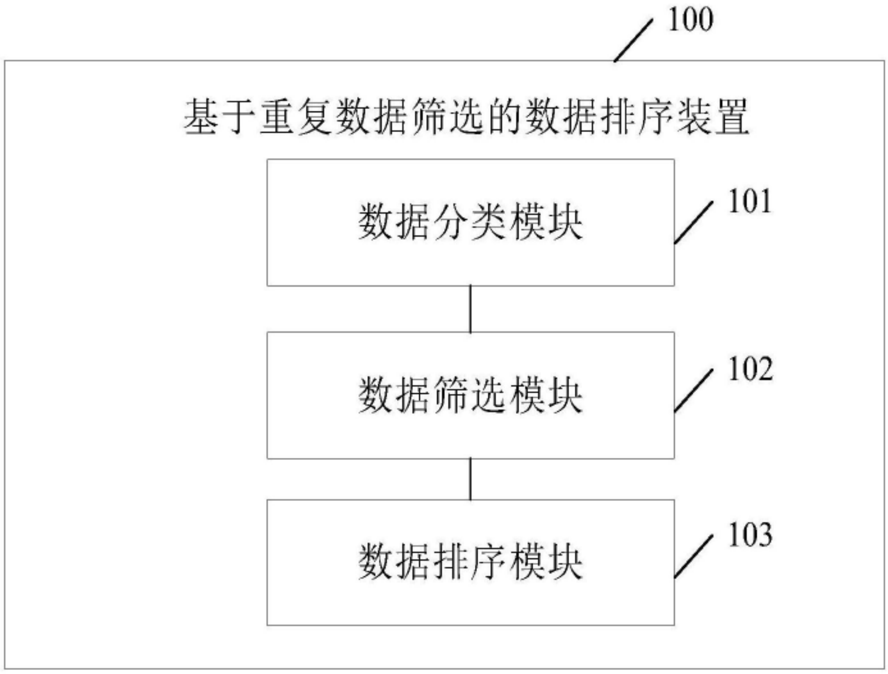
100
基于重复数据筛选的数据排序装置
数据分类模块
101
数据筛选模块
102
数据排序模块
103

图2

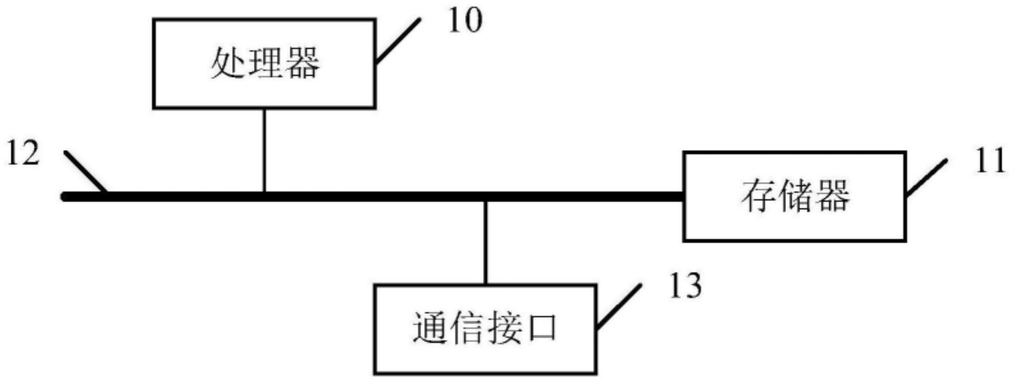
处理器
10
12
存储器
11
通信接口
13

图3