



(12)发明专利

(10)授权公告号 CN 106933863 B

(45)授权公告日 2019.04.19

(21)申请号 201511022880.3

(22)申请日 2015.12.30

(65)同一申请的已公布的文献号
申请公布号 CN 106933863 A

(43)申请公布日 2017.07.07

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 蒋瑜

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 杨贝贝 黄健

(51)Int.Cl.
G06F 16/215(2019.01)
G06F 16/22(2019.01)

(56)对比文件

CN 103473373 A, 2013.12.25,
CN 103324617 A, 2013.09.25,
CN 104021160 A, 2014.09.03,
CN 104239304 A, 2014.12.24,

审查员 王晶

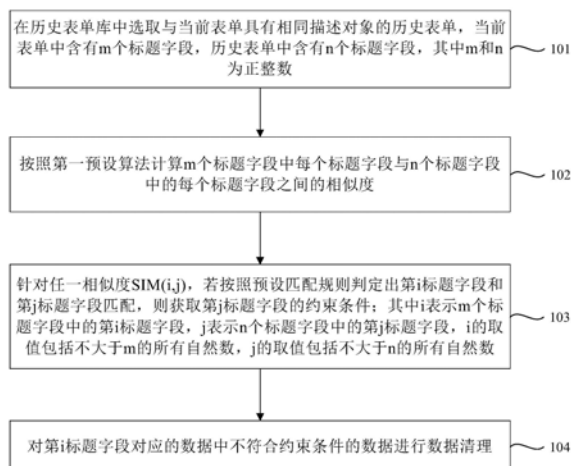
权利要求书4页 说明书16页 附图3页

(54)发明名称

数据清理方法及装置

(57)摘要

本发明提供一种数据清理方法及装置。该方法包括：在历史表单库中选取与当前表单具有相同描述对象的历史表单，当前表单中含有 m 个标题字段，历史表单中含有 n 个标题字段，其中 m 和 n 为正整数；按照第一预设算法计算 m 个标题字段中每个标题字段与 n 个标题字段中的每个标题字段之间的相似度；针对任一相似度 $SIM(i, j)$ ，若按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配，则获取第 j 标题字段的约束条件；其中 i 表示 m 个标题字段中的第 i 标题字段， j 表示 n 个标题字段中的第 j 标题字段， i 的取值包括不大于 m 的所有自然数， j 的取值包括不大于 n 的所有自然数；对第 i 标题字段对应的数据中不符合约束条件的数据进行数据清理。本发明提供的数据清理方法及装置提高了数据清理的效率和准确性，提高了数据源的准确性和可靠性。



1. 一种数据清理方法,其特征在于,所述方法包括:

在历史表单库中选取与当前表单具有相同描述对象的历史表单,所述当前表单中含有 m 个标题字段,所述历史表单中含有 n 个标题字段,其中 m 和 n 为正整数;

按照第一预设算法计算所述 m 个标题字段中每个标题字段与所述 n 个标题字段中的每个标题字段之间的相似度;

针对任一相似度 $SIM(i, j)$,若按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配,则获取所述第 j 标题字段的约束条件;其中 i 表示所述 m 个标题字段中的第 i 标题字段, j 表示所述 n 个标题字段中的第 j 标题字段, i 的取值包括不大于 m 的所有自然数, j 的取值包括不大于 n 的所有自然数;

对所述第 i 标题字段对应的数据中不符合所述约束条件的数据进行数据清理。

2. 根据权利要求1所述的方法,其特征在于,所述针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

若所述相似度 $SIM(i, j)$ 大于第一预设值,则判定出所述 m 个标题字段中的第 i 标题字段和所述 n 个标题字段中的第 j 标题字段匹配。

3. 根据权利要求1所述的方法,其特征在于,所述针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

若所述相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时,依据所述当前表单的预设字段关联关系确定出 k 个标题字段,对所述 k 个标题字段中的第 s 标题字段,在所述第 s 标题字段与所述 n 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据所述当前表单的预设字段关联关系确定出的与所述第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

根据所述相似度 SIM_s 通过第二预设算法对所述 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

若所述 $SIM_0(i, j)$ 大于所述第一预设值,则判定出所述 m 个标题字段中的第 i 标题字段和所述 n 个标题字段中的第 j 标题字段匹配。

4. 根据权利要求1所述的方法,其特征在于,所述针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

若所述相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时,依据所述当前表单的预设字段关联关系确定出 k 个标题字段,对所述 k 个标题字段中的第 s 标题字段,在所述第 s 标题字段与所述 n 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据所述当前表单的预设字段关联关系确定出的与所述第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

根据所述相似度 SIM_s 通过第二预设算法对所述 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

若所述 $SIM_0(i, j)$ 不大于所述第一预设值,则对 $SIM_0(i, j)$ 进行持续修正 Y 次后,得到 $SIM_{0+Y}(i, j)$,若所述 $SIM_{0+Y}(i, j)$ 大于所述第一预设值时,则判定出所述 m 个标题字段中的第 i 标题字段和所述 n 个标题字段中的第 j 标题字段匹配;

其中,在第 y 次修正中,根据所述相似度 SIM_s 通过第二预设算法对 $SIM_{0+y-1}(i, j)$ 进行修正,得到 $SIM_{0+y}(i, j)$, y 的取值包括不大于 Y 的正整数。

5. 根据权利要求3或4所述的方法, 其特征在于, 所述第二预设算法为如下所示的公式一:

$$SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

其中, SIM为待修正的相似度, SIM*为修正后的相似度, a为预设权重系数。

6. 根据权利要求1所述的方法, 其特征在于, 所述按照第一预设算法计算所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度, 包括:

按照所述当前表单中的所述m个标题字段的标题字段名称与所述历史表单中的所述n个标题字段的标题字段名称的重合度, 获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度; 或者

按照所述当前表单中的所述m个标题字段的标题字段名称与所述历史表单中的所述n个标题字段的标题字段名称, 根据第三方知识库获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度; 或者

按照所述当前表单中的所述m个标题字段的字段取值集与所述历史表单中的所述n个标题字段的字段取值集的重合度, 获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度。

7. 根据权利要求6所述的方法, 其特征在于, 所述按照所述当前表单中的所述m个标题字段的字段取值集与所述历史表单中的所述n个标题字段的字段取值集的重合度, 获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度, 包括:

当所述当前表单中的所述m个标题字段的字段取值集为离散型时, 通过如下公式二确定所述相似度:

$$SIM(i, j) = \frac{\emptyset_1 \cap \emptyset_2}{\emptyset_1 \cup \emptyset_2};$$

其中, $\emptyset_1$ 表示第i标题字段的字段取值集, $\emptyset_2$ 表示第j标题字段的字段取值集; 或者

当所述当前表单中的所述m个标题字段的字段取值集为连续型时, 通过如下公式三确定所述相似度:

$$SIM(i, j) = \max\left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0\right);$$

其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。

8. 一种数据清理装置, 其特征在于, 所述装置包括:

历史表单获取模块, 用于在历史表单库中选取与当前表单具有相同描述对象的历史表单, 所述当前表单中含有m个标题字段, 所述历史表单中含有n个标题字段, 其中m和n为正整数;

相似度计算模块, 用于按照第一预设算法计算所述历史表单获取模块获取的所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度;

匹配模块, 用于针对所述相似度计算模块计算得到的任一相似度SIM(i, j), 若按照预

设匹配规则判定出第i标题字段和第j标题字段匹配,则获取所述第j标题字段的约束条件;其中i表示所述m个标题字段中的第i标题字段,j表示所述n个标题字段中的第j标题字段,i的取值包括不大于m的所有自然数,j的取值包括不大于n的所有自然数;

数据清理模块,用于对所述第i标题字段对应的数据中不符合所述匹配模块获取的约束条件的数据进行数据清理。

9. 根据权利要求8所述的装置,其特征在于,所述匹配模块具体用于:

针对所述相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若所述相似度 $SIM(i, j)$ 大于第一预设值,则判定出所述m个标题字段中的第i标题字段和所述n个标题字段中的第j标题字段匹配,则获取所述第j标题字段的约束条件。

10. 根据权利要求8所述的装置,其特征在于,所述匹配模块具体用于:

针对所述相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若所述相似度 $SIM(i, j)$ 不大于第一预设值且m、n大于1时,依据所述当前表单的预设字段关联关系确定出k个标题字段,对所述k个标题字段中的第s标题字段,在所述第s标题字段与所述n个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中s的取值包括不大于k的所有自然数,k为依据所述当前表单的预设字段关联关系确定出的与所述第i标题字段关联的标题字段的总个数,其中k小于m;

根据所述相似度 SIM_s 通过第二预设算法对所述 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

若所述 $SIM_0(i, j)$ 大于所述第一预设值,则判定出所述m个标题字段中的第i标题字段和所述n个标题字段中的第j标题字段匹配。

11. 根据权利要求8所述的装置,其特征在于,所述匹配模块具体用于:

针对所述相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若所述相似度 $SIM(i, j)$ 不大于第一预设值且m、n大于1时,依据所述当前表单的预设字段关联关系确定出k个标题字段,对所述k个标题字段中的第s标题字段,在所述第s标题字段与所述n个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中s的取值包括不大于k的所有自然数,k为依据所述当前表单的预设字段关联关系确定出的与所述第i标题字段关联的标题字段的总个数,其中k小于m;

根据所述相似度 SIM_s 通过第二预设算法对所述 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

若所述 $SIM_0(i, j)$ 不大于所述第一预设值,则对 $SIM_0(i, j)$ 进行持续修正Y次后,得到 $SIM_{0+Y}(i, j)$,若所述 $SIM_{0+Y}(i, j)$ 大于所述第一预设值时,则判定出所述m个标题字段中的第i标题字段和所述n个标题字段中的第j标题字段匹配;

其中,在第y次修正中,根据所述相似度 SIM_s 通过第二预设算法对 $SIM_{0+y-1}(i, j)$ 进行修正,得到 $SIM_{0+y}(i, j)$,y的取值包括不大于Y的正整数。

12. 根据权利要求10或11所述的装置,其特征在于,所述第二预设算法为如下所示的公式一:

$$SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

其中,SIM为待修正的相似度, SIM^* 为修正后的相似度,a为预设权重系数。

13. 根据权利要求8所述的装置,其特征在于,所述相似度计算模块包括:

第一计算单元,用于按照所述当前表单中的所述m个标题字段的标题字段名称与所述历史表单中的所述n个标题字段的标题字段名称的重合度,获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度;或者

第二计算单元,用于按照所述当前表单中的所述m个标题字段的标题字段名称与所述历史表单中的所述n个标题字段的标题字段名称,根据第三方知识库获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度;或者

第三计算单元,用于按照所述当前表单中的所述m个标题字段的字段取值集与所述历史表单中的所述n个标题字段的字段取值集的重合度,获取所述m个标题字段中每个标题字段与所述n个标题字段中的每个标题字段之间的相似度。

14. 根据权利要求13所述的装置,其特征在于,所述第三计算单元具体用于:

当所述当前表单中的所述m个标题字段的字段取值集为离散型时,通过如下公式二确定所述相似度:

$$\text{SIM}(i,j) = \frac{\varnothing_1 \cap \varnothing_2}{\varnothing_1 \cup \varnothing_2};$$

其中, $\varnothing_1$ 表示第i标题字段的字段取值集, $\varnothing_2$ 表示第j标题字段的字段取值集;或者

当所述当前表单中的所述m个标题字段的字段取值集为连续型时,通过如下公式三确定所述相似度:

$$\text{SIM}(i,j) = \max\left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0\right);$$

其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。

15. 一种数据清理装置,其特征在于,包括:存储器、处理器以及总线,所述存储器以及所述处理器分别与所述总线连接,其中:

所述存储器用于存储数据和存储程序代码;

所述处理器,用于读取所述存储器中存储的程序代码,执行如权利要求1至7中任意一项所述的数据清理方法。

数据清理方法及装置

技术领域

[0001] 本发明涉及数据清理技术,尤其涉及一种数据清理方法及装置。

背景技术

[0002] 随着信息技术快速发展,大数据时代到来,各行各业开始建立信息系统并积累大量的数据。而数据的准确性,则是各种数据分析的基本条件。然而,现实中由于采集、传输、存储、处理过程中的各种原因,数据的准确性问题普遍存在。数据清理的目的是检测数据中存在的错误数据,剔除或者改正错误数据,以提高数据的准确性和质量。

[0003] 常见的数据错误包括空值、取值越界等。现有技术中,为了剔除或者改正错误数据,常见的数据清理方法主要为基于领域专用语言的编程数据清理方法,具体地,研发人员每次对表单进行数据清理时,研发人员为该表单制定错误数据的清理规则,然后根据错误数据的清理规则,确定具体的清理算法,再根据清理算法,编写数据清理程序,最后通过该数据清理程序实现数据的自动检测和修正。

[0004] 然而,通过数据清理程序实现数据的自动检测和修正,虽然实现了对数据的自动检测和修正,但是该方法要求研发人员在每次清理数据时,都要编写或修改数据清理程序,不仅对研发人员的要求较高,而且效率低下,使得现有技术的数据清理方法不具有普适性和易用性。

发明内容

[0005] 本发明实施例提供一种数据清理方法及装置,以克服现有数据清理方法效率低,不具有普适性和易用性问题。

[0006] 本发明实施例一方面提供一种数据清理方法,包括:

[0007] 在历史表单库中选取与当前表单具有相同描述对象的历史表单,当前表单中含有m个标题字段,历史表单中含有n个标题字段,其中m和n为正整数;

[0008] 按照第一预设算法计算m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;

[0009] 针对任一相似度 $SIM(i, j)$,若按照预设匹配规则判定出第i标题字段和第j标题字段匹配,则获取第j标题字段的约束条件;其中i表示m个标题字段中的第i标题字段,j表示n个标题字段中的第j标题字段,i的取值包括不大于m的所有自然数,j的取值包括不大于n的所有自然数;

[0010] 对第i标题字段对应的数据中不符合约束条件的数据进行数据清理。

[0011] 通过结合历史表单库中的具有相同描述对象的历史表单,将历史表单的标题字段的约束条件适应性的应用到当前表单的标题字段中,基于该约束条件对当前表单的标题字段对应的数据进行数据清理,无需研发人员在每次进行数据清理时进行清理算法代码程序的编写和维护,降低了用户的使用门槛,具有广泛的适用性,同时减少了人工进行数据清理的工作强度;也实现了数据库中大数据的自动清洗,提高了数据清理的效率和准确性,提高

了数据源的准确性和可靠性。

[0012] 进一步地,针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

[0013] 若相似度 $SIM(i, j)$ 大于第一预设值,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配。

[0014] 进一步地,针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

[0015] 若相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时,依据当前表单的预设字段关联关系确定出 k 个标题字段,对 k 个标题字段中的第 s 标题字段,在第 s 标题字段与 m 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据当前表单的预设字段关联关系确定出的与第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

[0016] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0017] 若 $SIM_0(i, j)$ 大于第一预设值,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配。

[0018] 进一步地,针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配包括:

[0019] 若相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时,依据当前表单的预设字段关联关系确定出 k 个标题字段,对 k 个标题字段中的第 s 标题字段,在第 s 标题字段与 n 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据当前表单的预设字段关联关系确定出的与第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

[0020] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0021] 若 $SIM_0(i, j)$ 不大于第一预设值,则对 $SIM_0(i, j)$ 进行持续修正 Y 次后,得到 $SIM_{0+Y}(i, j)$,若 $SIM_{0+Y}(i, j)$ 大于第一预设值时,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配;

[0022] 其中,在第 y 次修正中,根据相似度 SIM_s 通过第二预设算法对 $SIM_{0+y-1}(i, j)$ 进行修正,得到 $SIM_{0+y}(i, j)$, y 的取值包括不大于 Y 的正整数。

[0023] 进一步地,第二预设算法为如下所示的公式:

$$[0024] \quad SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

[0025] 其中, SIM 为待修正的相似度, SIM^* 为修正后的相似度, a 为预设权重系数。

[0026] 在确定匹配标题字段的过程中,通过利用表单中预设的标题字段间关联关系,对标题字段的相似度进行修正,得到修正后的更为准确的相似度,可确定出更多匹配标题字段,获得更多的约束条件,提高了数据清理的效率。

[0027] 进一步地,按照第一预设算法计算 m 个标题字段中每个标题字段与 n 个标题字段中的每个标题字段之间的相似度,包括:

[0028] 按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;或者

[0029] 按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称,根据第三方知识库获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;或者

[0030] 按照当前表单中的m个标题字段的字段取值集与历史表单中的n个标题字段的字段取值集的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度。

[0031] 进一步地,按照当前表单中的m个标题字段的字段取值集与历史表单中的n个标题字段的字段取值集的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度,包括:

[0032] 当当前表单中的m个标题字段的字段取值集为离散型时,通过如下公式确定相似度:

$$[0033] \quad \text{SIM}(i,j) = \frac{\emptyset_1 \cap \emptyset_2}{\emptyset_1 \cup \emptyset_2};$$

[0034] 其中, $\emptyset_1$ 表示第i标题字段的字段取值集, $\emptyset_2$ 表示第j标题字段的字段取值集;或者

[0035] 当当前表单中的m个标题字段的字段取值集为连续型时,通过如下公式确定相似度:

$$[0036] \quad \text{SIM}(i,j) = \max \left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0 \right);$$

[0037] 其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。

[0038] 进一步地,对当前表单进行数据清理之后,还包括:

[0039] 将当前表单以及当前表单的约束条件存储至历史表单库中。

[0040] 下面介绍本发明实施例提供一种数据清理装置,该装置与方法一一对应,用以实现上述实施例中的数据清理方法,具有相同的技术特征和技术效果,本发明对此不再赘述。

[0041] 本发明实施例另一方面提供一种数据清理装置,包括:

[0042] 历史表单获取模块,用于在历史表单库中选取与当前表单具有相同描述对象的历史表单,当前表单中含有m个标题字段,历史表单中含有n个标题字段,其中m和n为正整数;

[0043] 相似度计算模块,用于按照第一预设算法计算历史表单获取模块获取的m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;

[0044] 匹配模块,用于针对相似度计算模块计算得到的任一相似度SIM(i,j),若按照预设匹配规则判定出第i标题字段和第j标题字段匹配,则获取第j标题字段的约束条件;其中i表示m个标题字段中的第i标题字段,j表示n个标题字段中的第j标题字段,i的取值包括不大于m的所有自然数,j的取值包括不大于n的所有自然数;

[0045] 数据清理模块,用于对第i标题字段对应的数据中不符合匹配模块获取的约束条件的数据进行数据清理。

[0046] 进一步的,匹配模块具体用于:

[0047] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若相似度 $SIM(i, j)$ 大于第一预设值,则判定出m个标题字段中的第i标题字段和n个标题字段中的第j标题字段匹配,则获取第j标题字段的约束条件。

[0048] 进一步的,匹配模块具体用于:

[0049] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若相似度 $SIM(i, j)$ 不大于第一预设值且m、n大于1时,依据当前表单的预设字段关联关系确定出k个标题字段,对k个标题字段中的第s标题字段,在第s标题字段与n个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中s的取值包括不大于k的所有自然数,k为依据当前表单的预设字段关联关系确定出的与第i标题字段关联的标题字段的总个数,其中k小于m;

[0050] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0051] 若 $SIM_0(i, j)$ 大于第一预设值,则判定出m个标题字段中的第i标题字段和n个标题字段中的第j标题字段匹配。

[0052] 进一步的,匹配模块具体用于:

[0053] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若相似度 $SIM(i, j)$ 不大于第一预设值且m、n大于1时,依据当前表单的预设字段关联关系确定出k个标题字段,对k个标题字段中的第s标题字段,在第s标题字段与n个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中s的取值包括不大于k的所有自然数,k为依据当前表单的预设字段关联关系确定出的与第i标题字段关联的标题字段的总个数,其中k小于m;

[0054] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0055] 若 $SIM_0(i, j)$ 不大于第一预设值,则对 $SIM_0(i, j)$ 进行持续修正Y次后,得到 $SIM_{0+Y}(i, j)$,若 $SIM_{0+Y}(i, j)$ 大于第一预设值时,则判定出m个标题字段中的第i标题字段和n个标题字段中的第j标题字段匹配;

[0056] 其中,在第y次修正中,根据相似度 SIM_s 通过第二预设算法对 $SIM_{0+y-1}(i, j)$ 进行修正,得到 $SIM_{0+y}(i, j)$,y的取值包括不大于Y的正整数。

[0057] 进一步地,第二预设算法为如下所示的公式:

$$[0058] \quad SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

[0059] 其中,SIM为待修正的相似度, SIM^* 为修正后的相似度,a为预设权重系数。

[0060] 进一步的,相似度计算模块包括:

[0061] 第一计算单元,用于按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;或者

[0062] 第二计算单元,用于按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称,根据第三方知识库获取m个标题字段中每个标题字段与

n个标题字段中的每个标题字段之间的相似度;或者

[0063] 第三计算单元,用于按照当前表单中的m个标题字段的字段取值集与历史表单中的n个标题字段的字段取值集的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度。

[0064] 进一步地,第三计算单元具体用于:

[0065] 当当前表单中的m个标题字段的字段取值集为离散型时,通过如下公式确定相似度:

$$[0066] \quad \text{SIM}(i,j) = \frac{\varnothing_1 \cap \varnothing_2}{\varnothing_1 \cup \varnothing_2};$$

[0067] 其中, $\varnothing_1$ 表示第i标题字段的字段取值集, $\varnothing_2$ 表示第j标题字段的字段取值集;或者

[0068] 当当前表单中的m个标题字段的字段取值集为连续型时,通过如下公式确定相似度:

$$[0069] \quad \text{SIM}(i,j) = \max\left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0\right);$$

[0070] 其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。

[0071] 进一步地,在上述任一装置实施例的基础上,该装置还包括:

[0072] 存储模块,用于将当前表单以及当前表单的约束条件存储至历史表单库中。

[0073] 本发明实施例另一方面还提供一种数据清理装置,包括:存储器、处理器以及总线,存储器以及处理器分别与总线连接,其中:

[0074] 存储器用于存储数据和存储程序代码;

[0075] 处理器,用于读取存储器中存储的程序代码,执行如上所述的数据清理方法。

附图说明

[0076] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0077] 图1为本发明数据清理方法实施例一的流程示意图;

[0078] 图2为本发明数据清理方法实施例二的表单场景示意图;

[0079] 图3为本发明数据清理装置实施例一的结构示意图;

[0080] 图4为本发明数据清理装置的结构示意图。

具体实施方式

[0081] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中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发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员

在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0082] 在数据存储中，通常有很多统计数据以表单的形式进行存储，如工作种类和内容表单、员工基本信息表单、流动人口信息表单等，随着表单及表单中数据逐渐增多，则表单中可能出现由于采集、传输、存储、处理等过程导致的错误数据，进而影响了后续的基于表单的数据查询及分析。本发明实施例针对可能的错误数据，考虑到表单库中已存储有多个已进行数据清洗过的历史表单，当待清洗的当前表单与历史表单中的内容相一致时，可参考历史表单的数据清洗的约束条件对当前表单进行清洗，提出一种数据清理方法，可广泛应用于以表单形式存储数据的数据清洗中。

[0083] 本发明实施例提供一种数据清理方法，用于自动的为数据库中存储的大量数据进行数据清洗，发现可能存在的错误数据并剔除或者改正错误数据。

[0084] 通常大量的数据在数据库中以表单的形式存储，数据库中存储有多个表单。一个表单中存储了多列数据，一列数据通常由标题字段和该标题字段对应的数据构成。如下所示的表一示出了一种常见的表单。示例性的，表一所示的表单中包括标题字段名称为“姓名”、“ID”、“性别”等多个标题字段。各标题字段依据其对应的数据的不同，具有不同的属性，且各标题字段间存在不同的关联关系。当标题字段对应的数据中存在不符合约束条件的数据时，该不符合约束条件的数据则为错误数据，约束条件包括标题字段的属性或标题字段间的关联关系等。例如，标题字段“ID”中对应的数据具有唯一性，即标题字段“ID”对应的多个数据中不应存在重复数据“4”；标题字段“城市”具有特定取值范围属性，即标题字段“城市”中不应存在“程度”这一不在特定取值范围内的数据；标题字段“城市”与标题字段“区号”间为一一对应的关系，即表一中的标题字段“城市”中的“成都”与标题字段“区号”中的“029”之间的对应关系有误。表一中以下划线的方式示例性的标示出几种可能的错误数据。

[0085] 表一

[0086]

姓名	ID	性别	年龄	区号	城市	学历	职业	月收入	税率
刘一	1	男	<u>350</u>	0571	杭州	博士	程序员	20000	20%
关二	2	<u>male</u>	30	0571	杭州	硕士	程序员	18000	18%
张三	3	男	25	=	杭州	本科	公务员	10000	10%
李四	<u>4</u>	男	27	028	成都	本科	公务员	5000	5%
王五	<u>4</u>	女	34	028	成都	博士	程序员	<u>15000</u>	<u>20%</u>
赵六	6	女	31	<u>029</u>	成都	硕士	程序员	12000	12%
孙七	7	女	32	028	<u>程度</u>	硕士	公务员	8000	8%
周八	8	女	23	028	成都	本科	程序员	=	6%

[0087] 本发明实施例针对如上所述的大量数据中存在错误数据的问题，提出一种数据清理方法，该方法用于对表单进行数据清理，在本实例中，为了便于描述，称待进行数据清理的表单为当前表单。该方法先获取与当前表单相似的历史表单，在历史表单中确定与当前表单的标题字段相匹配的标题字段，然后根据该历史表单中的匹配的标题字段的约束条

件,对当前表单中的对应的匹配的标题字段进行数据清理,清理出不符合该约束条件的数据,该方法旨在解决现有技术中数据清理方法不具有普适性和易用性的问题,以实现数据清理的普适性和易用性。

[0088] 图1为本发明数据清理方法实施例一的流程示意图。本实施例的执行主体为数据清理装置,该装置可以设置在处理器中。如图1所示,本实施例的方法可以包括:

[0089] 步骤101、在历史表单库中选取与当前表单具有相同描述对象的历史表单,当前表单中含有 m 个标题字段,历史表单中含有 n 个标题字段,其中 m 和 n 为正整数;

[0090] 步骤102、按照第一预设算法计算 m 个标题字段中每个标题字段与 n 个标题字段中的每个标题字段之间的相似度;

[0091] 步骤103、针对任一相似度 $SIM(i, j)$,若按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配,则获取第 j 标题字段的约束条件;其中 i 表示 m 个标题字段中的第 i 标题字段, j 表示 n 个标题字段中的第 j 标题字段, i 的取值包括不大于 m 的所有自然数, j 的取值包括不大于 n 的所有自然数;

[0092] 步骤104、对第 i 标题字段对应的数据中不符合约束条件的数据进行数据清理。

[0093] 本实施例为了区分历史表单中的标题字段以及当前表单中的标题字段,将当前表单中的 m 个标题字段中的一个标题字段记为第 i 标题字段,将历史表单中的 n 个标题字段中的一个标题字段记为第 j 标题字段,其中 m 和 n 为正整数, i 的取值包括不大于 m 的所有自然数, j 的取值包括不大于 n 的所有自然数,相似度 $SIM(i, j)$ 则表示第 i 标题字段与第 j 标题字段之间的相似度。

[0094] 具体的,在步骤101中,对于当前表单,首先在历史表单库中存储的多个历史表单中,通过比较或查询获取与当前表单具有相同描述对象的历史表单。

[0095] 具体地,在历史表单库中,可能存在一历史表单与当前表单具有相同描述对象,即包含相同的标题字段,当相同的标题字段个数越多,且不同的标题字段个数越少,则表明当前表单与该历史表单越相似,可参照该相似原则,在历史表单库中筛选出与当前表单具有相同描述对象的历史表单。

[0096] 具体的,在步骤102中,计算筛选出来的当前表单中的 m 个标题字段与历史表单中的 n 个标题字段间的相似度。示例性的,按照第一预设算法遍历当前表单中的所有标题字段,获取各标题字段与历史表单中的每个标题字段的相似度,得到相似度 $SIM(i, j)$,其中 i 表示 m 个标题字段中的第 i 标题字段, j 表示 n 个标题字段中的第 j 标题字段, i 的取值包括不大于 m 的所有自然数, j 的取值包括不大于 n 的所有自然数。

[0097] 在具体实现过程中,获取相似度 $SIM(i, j)$ 的具体实现方式可以为以下实现方式中的任一种或多种相结合。

[0098] 一种可行的实现方式,按照当前表单中的 m 个标题字段的标题字段名称与历史表单中的 n 个标题字段的标题字段名称的重合度,获取 m 个标题字段中每个标题字段与 n 个标题字段中的每个标题字段之间的相似度。

[0099] 示例性的,对于当前表单中的一个标题字段,若其标题字段名称为“姓名”,当历史表单中同样存在标题字段名称为“姓名”的标题字段时,可认为这两个标题字段完全重合,这两个标题字段的相似度为1;当前表单中标题字段名称为“姓名”的标题字段与历史表单中标题字段名称为“月收入”的标题字段的重合度为0,故这两个标题字段的相似度可认为

为0。

[0100] 另一种可行的实现方式,按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称,根据第三方知识库获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度。

[0101] 示例性的,当第三方知识库中存储有字段名“姓名”与“name”为同义词时,可认为当前表单中字段名为“姓名”的字段与历史表单中字段名为“name”的字段的相似度为1。

[0102] 又一种可行的实现方式,按照当前表单中的m个标题字段的字段取值集与历史表单中的n个标题字段的字段取值集的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度。

[0103] 具体的,针对表单中标题字段的字段取值集类型的不同,该可行的实现方式包括以下两种情况:

[0104] 情况一,当标题字段的字段取值集为离散型时,根据如下公式(1),获取当前表单中的离散型标题字段与历史表单中的每个标题字段的相似度;

[0105] 其中,公式(1)为:

$$[0106] \quad \text{SIM}(i,j) = \frac{\emptyset_1 \cap \emptyset_2}{\emptyset_1 \cup \emptyset_2};$$

[0107] 其中, $\emptyset_1$ 表示第i标题字段的字段取值集, $\emptyset_2$ 表示第j标题字段的字段取值集。

[0108] 当采用该方法计算相似度时,当前表单中的离散型标题字段与历史表单中的非离散型标题字段可以认为相似度为0,无需采用公式(1)进行计算。由公式(1)可看出,当两个标题字段的字段取值集中相同的值越多,其相似度越高。

[0109] 示例性的,常见的离散型标题字段有“城市”、“学历”等。例如当一个“工作城市”标题字段的取值集包括:北京、上海、深圳,另一个标题字段“工作地点”的取值集包括:北京、上海、深圳、天津,采用上述公式(1)计算两字段的相似度,可得到相似度 = {北京、上海、深圳} / {北京、上海、深圳、天津} = 75%。

[0110] 情况二,当标题字段的字段取值集为连续型时,根据如下公式(2),获取当前表单中的连续型标题字段与历史表单中的所有标题字段的相似度;

[0111] 其中,公式(2)为:

$$[0112] \quad \text{SIM}(i,j) = \max \left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0 \right);$$

[0113] 其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。示例性的,常见的连续型标题字段有“年龄”、“工资”等。

[0114] 对于采用上述任一实现方式获取到的,当前表单中各标题字段与历史表单中的每个标题字段的相似度,相似度的取值范围为0至1的实数。相似度为0则表明分别来自当前表单和历史表单的两个标题字段没有任何相同的属性或关联关系,可视为无效的相似度。在实际计算相似度的过程中,对于当前表单中的任一标题字段与历史表单中的每个标题字段的相似度,通常最多有1个大于0的有效相似度,其余相似度值均为0。

[0115] 进一步地,还可以根据上述不同实现方式获取的相似度以及各实现方式得到的相

似度的预设权重,获取更为准确的相似度。

[0116] 具体的,在步骤103中,在步骤102获取的所有相似度中,针对任一相似度 $SIM(i, j)$,若按照预设匹配规则判定出第i标题字段和第j标题字段匹配,则获取第j标题字段的约束条件。

[0117] 对于任一表单,根据该表单中每个标题字段的取值集以及多个标题字段相互之间的关系,可为该表单构建约束条件,约束条件通常包括标题字段属性和标题字段间的关联关系。

[0118] 可选的,标题字段属性可以为以下中的一项或多项:可靠性、唯一性、标签、字段近义词、取值范围等。可选的,关联关系也可以为以下中的一项或多项:相关、保序、一一映射等。约束条件主要限定了表单中各标题字段的取值范围,以及多个标题字段对应的数据的相互关系,不符合约束条件的数据可认为是需要清理的错误数据。

[0119] 历史表单库中的历史表单各自均对应有自己的约束条件,当来自历史表单与当前表单的两个标题字段为匹配的标题字段时,可以认为这两个标题字段应该具有相同或相似的标题字段属性和标题字段间关联关系,即两个标题字段对应的数据符合同样的约束条件。因此,可在判定出第i标题字段和第j标题字段匹配后,获取第j标题字段的约束条件。

[0120] 具体的,在步骤104中,由于第i标题字段和第j标题字段匹配,则认为第i标题字段应符合第j标题字段的约束条件,可直接用第j标题字段的约束条件对第i标题字段对应的数据进行排查,确定第i标题字段对应的所有数据中不符合约束条件的数据,对该些数据进行清理。通过判定出尽量多对的标题字段匹配,可对当前表单中的标题字段进行更充分的数据清理。

[0121] 在具体根据约束条件,对当前表单中不符合标题字段属性或标题字段间关系的数值进行清理时,清理过程包括删除错误值并提供修正值。例如,根据约束条件可知标题字段为“性别”时,该标题字段的取值范围可以为“男、女、未知”,而当检测到当前表单中的“性别”标题字段对应的某个数据为“北京”时,则认为该数据为错误数据,需要被清理,删除“北京”后,根据“性别”标题字段的取值范围,可采用“未知”进行修正,从而完成当前表单的数据清理。示例性的,当检测到当前表单中的“性别”标题字段对应的某个数据为“male”时,根据该字段的取值范围以及近义词,可知“male”的近义词为“男”,故可将当前表单中的“male”修正为“男”,完成数据清理。

[0122] 进一步的,在步骤104中进行数据清理时,当进行数据修改和替换时,可选择将需进行修改的源数据与替换数据通过显示屏显示给操作人员,待操作人员确认后再进行修正,或待操作人员否定后不进行修正,通过增加确认步骤可提高数据清理的准确性。

[0123] 在具体实施数据清理的过程中,可调用第三方知识库或专家知识库,以查询标题字段及标题字段对应的数据的同义词,近义词及联想扩展词等。

[0124] 可选的,当步骤101中无法获取到具有相同描述对象的历史表单时,或在步骤101中获取到具有相同描述对象的历史表单的同时,还可直接人工输入标题字段属性和标题字段间关系等约束条件;或由数据清理装置根据专家知识库中已经存储的预置标题字段属性和标题字段间关系,自动进行标题字段匹配,为当前表单中标题字段配置专家知识库中存储的约束条件,并进行数据清理。

[0125] 进一步的,在上述实施例的基础上,本发明实施例提供的数据清理方法还包括:

[0126] 将当前表单以及当前表单的约束条件存储至历史表单库中。

[0127] 通过将每次进行了数据清理后的当前表单,以及当前表单的约束条件存储至历史表单库中,扩充历史表单库,方便后续再次应用本发明实施例提供的数据清理方法进行数据清理。

[0128] 本发明实施例提供的数据清理方法,通过结合历史表单库中的具有相同描述对象的历史表单,将历史表单的标题字段的约束条件适应性的应用到当前表单的标题字段中,基于该约束条件对当前表单的标题字段对应的数据进行数据清理,无需研发人员在每次进行数据清理时进行清理算法代码程序的编写和维护,降低了用户的使用门槛,具有广泛的适用性,同时减少了人工进行数据清理的工作强度;也实现了数据库中大数据的自动清洗,提高了数据清理的效率和准确性,提高了数据源的准确性和可靠性。

[0129] 下面结合具体的实施例,在上述实施例的基础上,对按照预设匹配规则判定两标题字段匹配,进行详细说明。

[0130] 图2为本发明数据清理方法实施例二的表单场景示意图。图2中示意性的示出了应用了本发明数据清理方法的一个当前表单和一个已经筛选出来的具有相同描述对象的历史表单,并标示出当前表单和历史表单中的部分标题字段间的相似度,图2中还示出了当前表单的部分预设的标题字段间关联关系,及历史表单的部分标题字段间关联关系。

[0131] 当获取到当前表单中的各标题字段与历史表单中的每个标题字段的相似度 $SIM(i, j)$ 后,针对任一相似度 $SIM(i, j)$,按照预设匹配规则判定该相似度 $SIM(i, j)$ 对应的第 i 标题字段和第 j 标题字段是否匹配。

[0132] 具体的,预设匹配规则的具体实现方式可以为以下实现方式中的任一种。

[0133] 可行的实现方式一为:

[0134] 若相似度 $SIM(i, j)$ 大于第一预设值,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配。

[0135] 示例性的,结合图2,对于获取到的当前表单中的“职业”标题字段与历史表单中的“职业”标题字段的相似度1,判断该相似度1是否大于第一预设值0.9,由于相似度1大于第一预设值0.9,则可确定第 i 标题字段和第 j 标题字段为相互匹配的标题字段,二者具有相似的标题字段属性和标题字段间的关联关系,即认为当前表单中的“职业”标题字段与历史表单中的“职业”标题字段相互匹配,二者应该具有相同的约束条件。示例性的,当历史表单中的“职业”标题字段具有“工作”这一相似标题字段,则当前表单中的“职业”标题字段同样具有“工作”这一相似标题字段。通过将历史表单的“职业”标题字段的约束条件直接应用到当前表单中的“职业”标题字段上,实现对当前表单的数据清理。其中,第一预设值的取值范围为0至1的实数,可提前预设,也可在匹配过程进行适当修改。

[0136] 可行的实现方式二为:

[0137] 若相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时,依据当前表单的预设字段关联关系确定出 k 个标题字段,对 k 个标题字段中的第 s 标题字段,在第 s 标题字段与 m 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据当前表单的预设字段关联关系确定出的与第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

[0138] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度

$SIM_0(i, j)$;

[0139] 若 $SIM_0(i, j)$ 大于第一预设值,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配。

[0140] 具体的,当确定相似度 $SIM(i, j)$ 小于第一预设值时,则无法直接判定出第 i 标题字段和第 j 标题字段匹配。当 m, n 大于1时,当前表单与历史表单中确实存在多个标题字段,则还可根据与相似度 $SIM(i, j)$ 对应的标题字段具有关联关系的其他标题字段的相似度,对相似度 $SIM(i, j)$ 进行修正,以得到更准确的相似度,当修正后的相似度 $SIM_0(i, j)$ 大于第一预设值,则可同样判定出第 i 标题字段和第 j 标题字段匹配。

[0141] 可选的,第一预设值可为固定值,也可在对相似度进行修正时,适应性的更改为其他值。可选的,在比较相似度 $SIM(i, j)$ 与第一预设值之前,还包括先确定相似度 $SIM(i, j)$ 大于0,对于相似度 $SIM(i, j)$ 为0的第 i 标题字段和第 j 标题字段,则可直接确定第 i 标题字段和第 j 标题字段不匹配,无需进行修正。

[0142] 示例性的,对该相似度 $SIM(i, j)$ 进行修正的方法为:

[0143] 根据当前表单的预设字段关联关系,确定出与第 i 标题字段关联的 k 个标题字段, k 为依据当前表单的预设字段关联关系确定出的与第 i 标题字段关联的标题字段的总个数,其中 k 小于 m 。对 k 个标题字段中的第 s 标题字段,在第 s 标题字段与 m 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数,即根据预设字段关联关系,可确定出 K 个相似度 $SIM_1, SIM_2, \dots, SIM_{k-1}, SIM_k$ 。然后, K 个相似度 $SIM_1, SIM_2, \dots, SIM_{k-1}, SIM_k$ 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$ 。

[0144] 具体的,在修正时,第二预设算法为如下所示的公式(3):

$$[0145] \quad SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

[0146] 其中, SIM 为待修正的相似度, SIM^* 为修正后的相似度, a 为预设权重系数。

[0147] 在对相似对进行修正时,可将所有小于第一预设值的相似度 $SIM(i, j)$ 进行排序,按照相似度从高到低的顺序依次进行修正。通过对所有小于第一预设值的相似度 $SIM(i, j)$ 进行修正,得到修正后的相似度,以提高当前表单与历史表单中的标题字段的匹配率,对当前表单进行更充分的数据清理。

[0148] 结合图2,参考图2中已经计算出的当前表单和历史表单中的多个相似度,其中,当前表单中的“职业”标题字段和历史表单中的“职业”标题字段的相似度为1,大于第一预设值0.9,可直接确定两标题字段相互匹配,从而可根据历史表单中“职业”标题字段的约束条件对当前表单中的“职业”标题字段进行数据清理,示例性的,若当前表单中的“职业”标题字段对应的数据中存在“北京”、“男”、“2000”等数据时,根据历史表单中的“职业”标题字段的约束条件中限定的取值集为“公务员、程序员、无”,可认为数据“北京”、“男”、“2000”为错误数据,需要被修改。当前表单中的“月收入”标题字段和历史表单中的“月薪”标题字段的相似度为0.7,当前表单中的“学历”标题字段和历史表单中的“最高学历”标题字段的相似度为0.8。同时,当前表单中的“月收入”标题字段和“学历”标题字段及“职业”标题字段相关联,当历史单中的“月薪”标题字段和“最高学历”标题字段及“职业”标题字段相关联,标题字段“月收入”和标题字段“月薪”的修正的相似度 $SIM_0(i, j)$,可由如下公式确定:

$$[0149] \quad \text{SIM}_0(i,j) = a \times 0.7 + (1 - a) \times \frac{0.8+1}{2}$$

[0150] 当a取0.4时,可得到修正的相似度为0.817,当第一预设值仍为0.9时,可认为标题字段“月收入”和标题字段“月薪”不匹配,若在修正过程中,将第一预设值修改为0.81时,则可认为标题字段“月收入”和标题字段“月薪”相互匹配,可将历史表单中的标题字段“月薪”相关的约束条件都应用到当前表单中,以供当前表单进行数据清理。

[0151] 可行的实现方式三为:

[0152] 若相似度 $\text{SIM}(i, j)$ 不大于第一预设值且 m, n 大于1时,依据当前表单的预设字段关联关系确定出 k 个标题字段,对 k 个标题字段中的第 s 标题字段,在第 s 标题字段与 n 个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中 s 的取值包括不大于 k 的所有自然数, k 为依据当前表单的预设字段关联关系确定出的与第 i 标题字段关联的标题字段的总个数,其中 k 小于 m ;

[0153] 根据相似度 SIM_s 通过第二预设算法对 $\text{SIM}(i, j)$ 进行修正,得到修正后的相似度 $\text{SIM}_0(i, j)$;

[0154] 若 $\text{SIM}_0(i, j)$ 不大于第一预设值,则对 $\text{SIM}_0(i, j)$ 进行持续修正 Y 次后,得到 $\text{SIM}_{0+Y}(i, j)$,若 $\text{SIM}_{0+Y}(i, j)$ 大于第一预设值时,则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配;

[0155] 其中,在第 y 次修正中,根据相似度 SIM_s 通过第二预设算法对 $\text{SIM}_{0+y-1}(i, j)$ 进行修正,得到 $\text{SIM}_{0+y}(i, j)$, y 的取值包括不大于 Y 的正整数。

[0156] 具体的,与可行的实现方式二中的修正方法相同,可先对任一相似度 $\text{SIM}(i, j)$ 进行修正,得到修正后的相似度 $\text{SIM}_0(i, j)$,具体修正方法本发明对此不再赘述。

[0157] 当修正后的相似度 $\text{SIM}_0(i, j)$ 大于第一预设值时,则判定出第 i 标题字段和第 j 标题字段匹配,当修正后的相似度 $\text{SIM}_0(i, j)$ 仍小于第一预设值时,可对修正后的相似度 $\text{SIM}_0(i, j)$ 进行多次持续修正。可选的,在多次修正过程中,均根据相似度 SIM_s 通过第二预设算法对上一次修正得到的相似度进行迭代修正。具体在迭代过程中,在第 y 次修正中,根据相似度 SIM_s 通过第二预设算法对 $\text{SIM}_{0+y-1}(i, j)$ 进行修正,得到 $\text{SIM}_{0+y}(i, j)$,其中 y 的取值包括不大于 Y 的正整数, Y 为预设值。当 Y 次修正后的 $\text{SIM}_{0+Y}(i, j)$ 大于第一预设值时,则判定出第 i 标题字段和第 j 标题字段匹配。可选的,在迭代修正时,若得到 $\text{SIM}_{0+X}(i, j)$ 大于第一预设值,且 X 小于 Y ,则停止迭代修正。其中第二预设算法与实现方式二中类似,本发明对此不再赘述。

[0158] 可选的,在进行多次持续修正时,还可采用已修正的相似度替换修正前的相似度,以提高修正效率。在具体替换过程中,包括如下两种替换模式。

[0159] 一种可能的替换模式:

[0160] 对于所有相似度,在进行第 y 次修正之前,确定所有大于第一预设值的相似度 $\text{SIM}_{0+y-1}(i, j)$,采用该些相似度值替换修正前的相似度;并确定所有小于第一预设值的相似度 $\text{SIM}_{0+y-1}(i, j)$,组成相似度集合 E ,在相似度集合 E 中,按照相似度从高到低的顺序,进行第 y 次修正,得到修正后的相似度 $\text{SIM}_{0+y}(i, j)$,然后确定所有大于第一预设值的相似度 $\text{SIM}_{0+y}(i, j)$,采用大于第一预设值的相似度 $\text{SIM}_{0+y}(i, j)$ 替换修正前的相似度 $\text{SIM}_{0+y-1}(i, j)$ 。通过在待修正的相似度进行第 y 次修正之前,将修正后的相似度替换修正前的相似度,可提高修

正效率。

[0161] 另一种可能的替换模式：

[0162] 对于所有相似度，在对相似度 $SIM(i, j)$ 进行修正之前，确定所有小于第一预设值的相似度 $SIM(i, j)$ ，组成相似度集合 F ，在相似度集合 F 中，按照相似度从高到低的顺序，进行修正，在每得到一个修正后的相似度 $SIM_0(i, j)$ 时，判断该相似度 $SIM_0(i, j)$ 是否大于第一预设值，若是，则用该修正后的相似度替换修正前的相似度，直至完成相似度集合 F 中的所有相似度的修正和替换，然后，将所有仍小于第一预设值的修正后的相似度 $SIM_0(i, j)$ 组成相似度集合 F^* ，在相似度集合 F^* 中，采用与第一次修正相同的方式进行修正和替换，直至完成 Y 次持续修正。

[0163] 本发明实施例提供的数据清理方法，通过结合历史表单库中的具有相同描述对象的历史表单，将历史表单的标题字段的约束条件适应性的应用到当前表单的标题字段中，基于该约束条件对当前表单的标题字段对应的数据进行数据清理，无需研发人员在每次进行数据清理时进行清理算法代码程序的编写和维护，降低了用户的使用门槛，具有广泛的适用性，同时减少了人工进行数据清理的工作强度；也实现了数据库中大数据的自动清洗，提高了数据清理的效率和准确性，提高了数据源的准确性和可靠性。

[0164] 图3为本发明数据清理装置实施例一的结构示意图。如图3所示，本实施例的装置可以包括：

[0165] 历史表单获取模块301，用于在历史表单库中选取与当前表单具有相同描述对象的历史表单，当前表单中含有 m 个标题字段，历史表单中含有 n 个标题字段，其中 m 和 n 为正整数；

[0166] 相似度计算模块302，用于按照第一预设算法计算历史表单获取模块301获取的 m 个标题字段中每个标题字段与 n 个标题字段中的每个标题字段之间的相似度；

[0167] 匹配模块303，用于针对相似度计算模块302计算得到的任一相似度 $SIM(i, j)$ ，若按照预设匹配规则判定出第 i 标题字段和第 j 标题字段匹配，则获取第 j 标题字段的约束条件；其中 i 表示 m 个标题字段中的第 i 标题字段， j 表示 n 个标题字段中的第 j 标题字段， i 的取值包括不大于 m 的所有自然数， j 的取值包括不大于 n 的所有自然数；

[0168] 数据清理模块304，用于对第 i 标题字段对应的数据中不符合匹配模块303获取的约束条件的数据进行数据清理。

[0169] 本实施例的装置，可以用于执行图1所示方法实施例的技术方案，其实现原理和技术效果类似，此处不再赘述。

[0170] 进一步的，在上述实施例的基础上，匹配模块303具体用于：

[0171] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$ ，若相似度 $SIM(i, j)$ 大于第一预设值，则判定出 m 个标题字段中的第 i 标题字段和 n 个标题字段中的第 j 标题字段匹配，则获取第 j 标题字段的约束条件。

[0172] 进一步的，在上述实施例的基础上，匹配模块303具体用于：

[0173] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$ ，若相似度 $SIM(i, j)$ 不大于第一预设值且 m, n 大于1时，依据当前表单的预设字段关联关系确定出 k 个标题字段，对 k 个标题字段中的第 s 标题字段，在第 s 标题字段与 n 个标题字段中的每个标题字段之间的相似度中，确定最大的相似度 SIM_s ，其中 s 的取值包括不大于 k 的所有自然数， k 为依据当前表单

的预设字段关联关系确定出的与第i标题字段关联的标题字段的总个数,其中k小于m;

[0174] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0175] 若 $SIM_0(i, j)$ 大于第一预设值,则判定出m个标题字段中的第i标题字段和n个标题字段中的第j标题字段匹配。

[0176] 进一步的,在上述实施例的基础上,匹配模块303具体用于:

[0177] 针对相似度计算模块计算得到的任一相似度 $SIM(i, j)$,若相似度 $SIM(i, j)$ 不大于第一预设值且m、n大于1时,依据当前表单的预设字段关联关系确定出k个标题字段,对k个标题字段中的第s标题字段,在第s标题字段与n个标题字段中的每个标题字段之间的相似度中,确定最大的相似度 SIM_s ,其中s的取值包括不大于k的所有自然数,k为依据当前表单的预设字段关联关系确定出的与第i标题字段关联的标题字段的总个数,其中k小于m;

[0178] 根据相似度 SIM_s 通过第二预设算法对 $SIM(i, j)$ 进行修正,得到修正后的相似度 $SIM_0(i, j)$;

[0179] 若 $SIM_0(i, j)$ 不大于第一预设值,则对 $SIM_0(i, j)$ 进行持续修正Y次后,得到 $SIM_{0+Y}(i, j)$,若 $SIM_{0+Y}(i, j)$ 大于第一预设值时,则判定出m个标题字段中的第i标题字段和n个标题字段中的第j标题字段匹配;

[0180] 其中,在第y次修正中,根据相似度 SIM_s 通过第二预设算法对 $SIM_{0+y-1}(i, j)$ 进行修正,得到 $SIM_{0+y}(i, j)$,y的取值包括不大于Y的正整数。

[0181] 进一步地,第二预设算法为如下所示的公式:

$$[0182] \quad SIM^* = a \times SIM + (1 - a) \times \frac{\sum_{s=1}^k SIM_s}{k};$$

[0183] 其中,SIM为待修正的相似度, SIM^* 为修正后的相似度,a为预设权重系数。

[0184] 进一步的,在上述任一装置实施例的基础上,相似度计算模块302包括:

[0185] 第一计算单元,用于按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;或者

[0186] 第二计算单元,用于按照当前表单中的m个标题字段的标题字段名称与历史表单中的n个标题字段的标题字段名称,根据第三方知识库获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度;或者

[0187] 第三计算单元,用于按照当前表单中的m个标题字段的字段取值集与历史表单中的n个标题字段的字段取值集的重合度,获取m个标题字段中每个标题字段与n个标题字段中的每个标题字段之间的相似度。

[0188] 进一步地,在上述实施例的基础上,第三计算单元具体用于:

[0189] 当当前表单中的m个标题字段的字段取值集为离散型时,通过如下公式确定相似度:

$$[0190] \quad SIM(i, j) = \frac{\emptyset_1 \cap \emptyset_2}{\emptyset_1 \cup \emptyset_2};$$

[0191] 其中, $\emptyset_1$ 表示第i标题字段的字段取值集, $\emptyset_2$ 表示第j标题字段的字段取值集;或者

[0192] 当当前表单中的m个标题字段的字段取值集为连续型时,通过如下公式确定相似度:

$$[0193] \quad SIM(i,j) = \max \left(\frac{\min(\max_1, \max_2) - \max(\min_1, \min_2)}{\max(\max_1, \max_2) - \min(\min_1, \min_2)}, 0 \right);$$

[0194] 其中, $\max_1$ 表示第i标题字段的最大值, $\min_1$ 表示第i标题字段的最小值, $\max_2$ 表示第j标题字段的最大值, $\min_2$ 表示第j标题字段的最小值。

[0195] 进一步地,在上述任一装置实施例的基础上,该装置还包括:

[0196] 存储模块,用于将当前表单以及当前表单的约束条件存储至历史表单库中。

[0197] 图4为本发明数据清理装置的结构示意图。该装置可用于执行如上述实施例所述的数据清理方法。如图4所示,该装置包括:处理器401、存储器402以及总线405,处理器401以及存储器402分别与总线405连接,其中:

[0198] 存储器402用于存储数据和存储程序代码;

[0199] 处理器401,用于读取存储器402中存储的程序代码,执行数据清理方法。

[0200] 具体的,存储器402中,存储有大量数据和程序代码,该数据以表单形式存储,处理器401通过实施本发明的数据清理方法对存储器402中可能存在的错误数据进行清理和修正;可选的,用户设备中还包括显示器403,显示器403用于将处理器401的清理和修正结果进行显示,也可用于将处理器401的中间处理过程进行显示;可选的,用户设备中还包括存储器404,存储器404中存储有第三方数据库、历史表单库、专家知识库等预设数据,便于处理器401实施本发明的数据清理方法时调用,可选的,第三方数据库、历史表单库、专家知识库等预设数据也可存储在存储器402中。其中,第三方知识库中包含预设的标题字段的近义词库,历史表单库中包含已经进行过数据清理的历史表单,该些历史表单中的数据准确性高,且该些历史表单对应有各自的约束条件,专家知识库中包含基于专家知识预设的约束条件。图4中上述各装置模块可以为集成在同一计算机中,也可仅通过网络连接。

[0201] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0202] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0203] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、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,所述单元的划分,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,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0204] 本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”“第四”等(如果存在)是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理

解这样使用的数据在适当情况下可以互换,以便这里描述的本发明的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0205] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元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单元来实现本实施例方案的目的。

[0206] 另外,在本发明各个实施例中的各功能单元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。

[0207] 所述功能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,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(Read-Only Memory,简称ROM)、随机存取存储器(Random Access Memory,简称RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0208] 最后应说明的是:以上各实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述各实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分或者全部技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的范围。

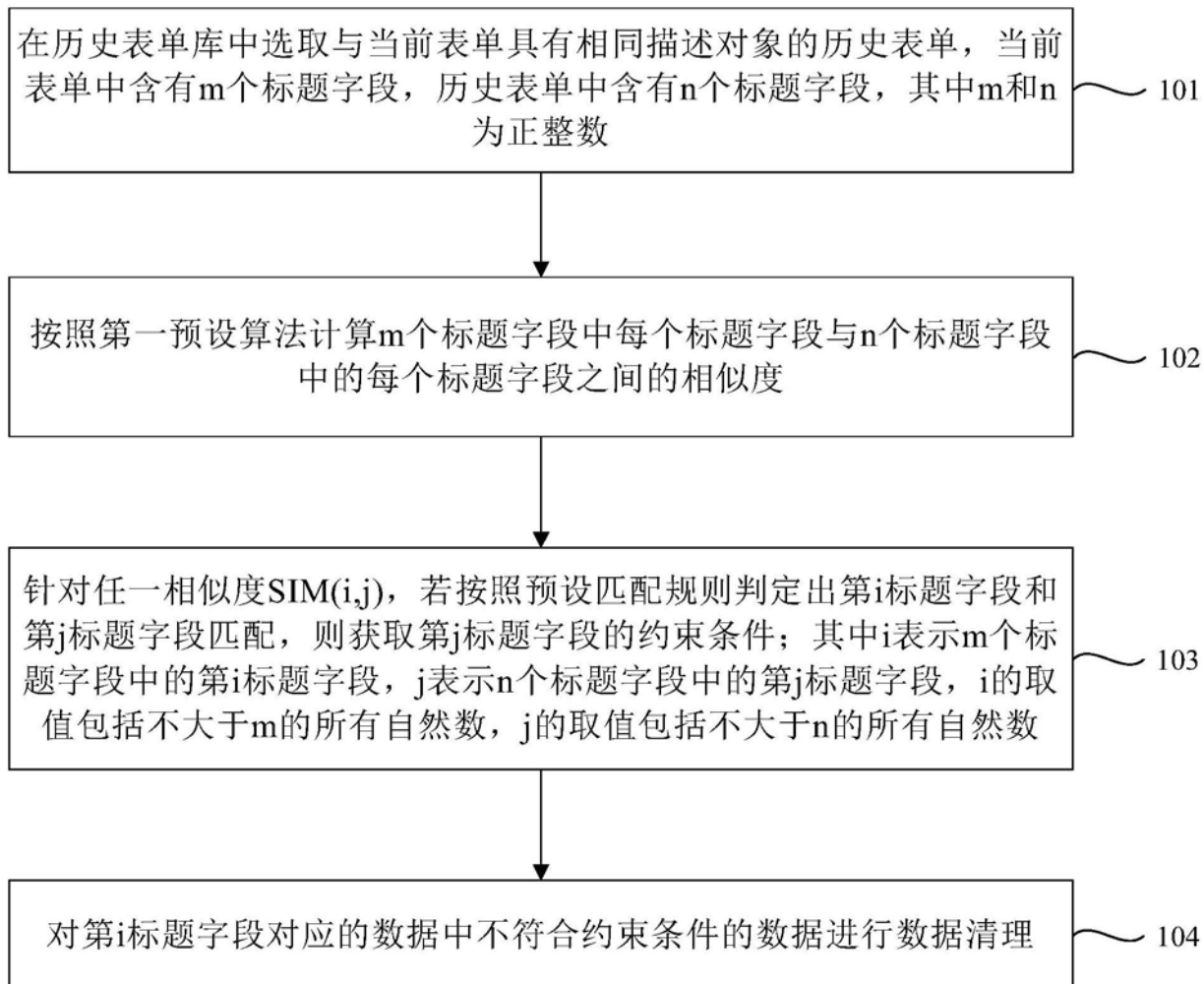


图1

当前表单

ID	性别	年龄	区号	城市	学历	职业	月收入	税率
----	----	----	----	----	----	----	-----	----

历史表单

姓名	gender	年纪	居住地	最高学历	职业	月薪
----	--------	----	-----	------	----	----



图2

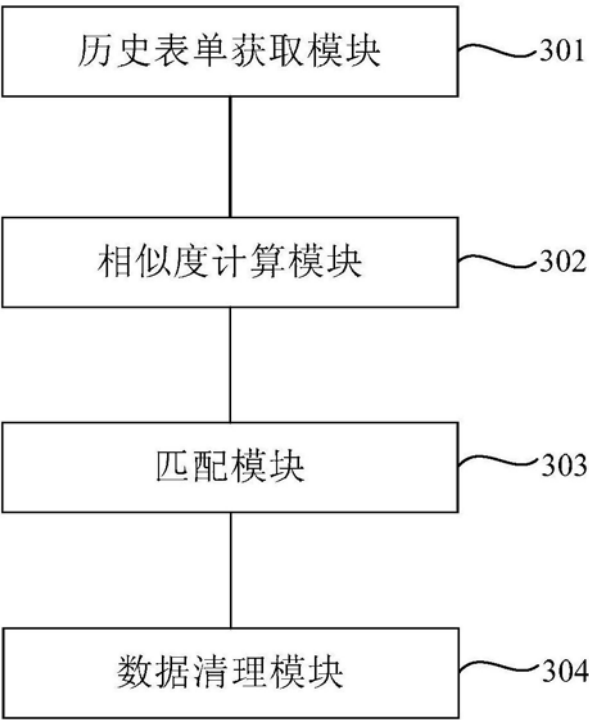


图3

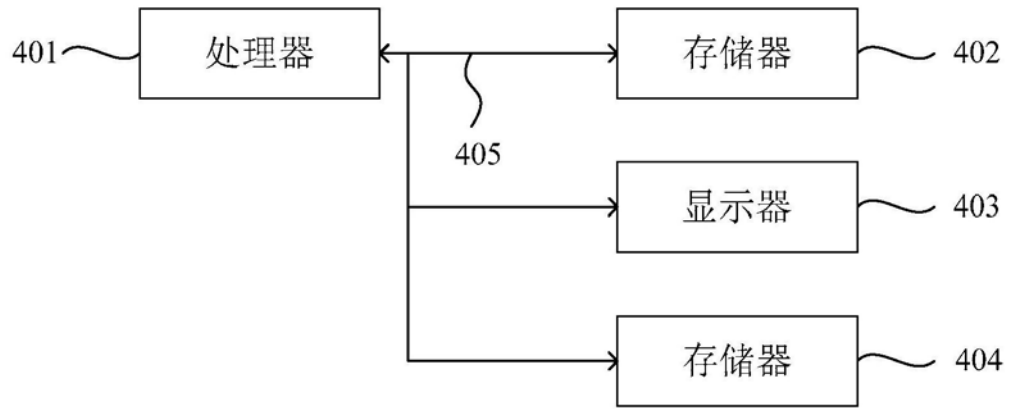


图4