(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111177115 B
(45)授权公告日 2023.07.28

(21)申请号 201911268767.1

(22)申请日 2019.12.11

(65)同一申请的已公布的文献号
申请公布号 CN 111177115 A

(43)申请公布日 2020.05.19

(73)专利权人 中电普信(北京)科技发展有限公司
地址 100000 北京市海淀区东北旺村南1号楼6层609室

(72)发明人 王运春 杨晓勇 孟炎杰 石武军 王占果

(74)专利代理机构 北京冠和权律师事务所 11399
专利代理师 张楠楠

(51)Int.Cl.
G06F 16/21(2019.01)

(56)对比文件
CN 106886578 A,2017.06.23

审查员 饶锐

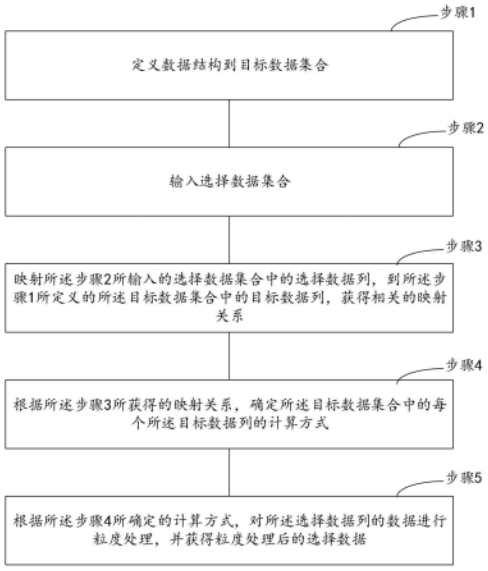
权利要求书4页 说明书8页 附图2页

(54)发明名称

一种数据预处理通用流程方法及系统

(57)摘要

本发明提供了一种数据预处理通用流程方法,包括:步骤1:定义数据结构到目标数据集合;步骤2:输入选择数据集合;步骤3:映射步骤2所输入的选择数据集合中的选择数据列,到步骤1所定义的目标数据集合中的目标数据列,获得相关的映射关系;步骤4:根据步骤3所获得的映射关系,确定目标数据集合中的每个目标数据列的计算方式;步骤5:根据步骤4所确定的计算方式,对选择数据列的数据进行粒度处理,并获得粒度处理后的选择数据。通过定义数据结构和相关映射,来提高处理流程的扩张性和灵活性。



CN 111177115 B

1.一种数据预处理通用流程方法,其特征在于,包括:

步骤1:定义数据结构到目标数据集合;

步骤2:输入选择数据集合;

步骤3:映射所述步骤2所输入的选择数据集合中的选择数据列,到所述步骤1所定义的所述目标数据集合中的目标数据列,获得相关的映射关系;

步骤4:根据所述步骤3所获得的映射关系,确定所述目标数据集合中的每个所述目标数据列的计算方式;

步骤5:根据所述步骤4所确定的计算方式,对所述选择数据列的数据进行粒度处理,并获得粒度处理后的选择数据;

其中,所述每个目标数据列的计算方式是基于预先存储的预设数目个预设算子组合得到的,且所述预设算子之间映射连接,其中所述预设算子是一些基础算子;

所述步骤3中,映射所述步骤2所输入的选择数据集合中的选择数据列,到所述步骤1所定义的所述目标数据集合中的目标数据列的具体过程包括:

对所述选择数据集合中的选择数据列中的数据元素进行第一标记;

对所述目标数据集合中的目标数据列中的数据元素进行第二标记;

获取所述选择数据集合中的选择数据列对应的选择映射区域、及所述目标数据集合中的目标数据列对应的目标映射区域;

其中,所述选择映射区域包括:至少一个进行第一标记的选择特征点;

所述目标映射区域包括:至少一个进行第二标记的目标特征点;

根据所述进行第一标记的选择特征点对应的选择特征点信息,在所述目标映射区域中查找进行第二标记的目标特征点;

同时,判断所获取的选择特征点和目标特征点是否满足预先存储的映射数据库中的预设映射规则;

若是,将所述选择特征点映射到所述目标特征点上,获得两者之间的点映射关系,并将所述点映射关系存储到相应的待扩展数据库中,实现选择数据列到目标数据列的映射;

否则,将所述选择映射区域映射到所述目标映射区域上,获得两者之间的区域映射关系,并将所述区域映射关系存储到相应的待扩展数据库中,实现选择数据列到目标数据列的映射;

所述步骤5中,根据所述步骤4所确定的计算方式,对所述选择数据列的数据进行粒度处理之前,还包括:筛选出目标数据列的最优计算方式,对所述选择数据列的数据进行粒度处理;

其中,筛选出目标数据列的最优计算方式具体包括:

确定所述步骤3所获得的映射关系中,所述选择数据列与所述目标数据列的映射关系是否为一对一映射关系或多对一映射关系,若是,确定所述目标数据列的计算方式为最优计算方式;

否则,筛选出所述目标数据列的最优计算方式,其步骤包括:

步骤S1:确定当前选择数据列所映射的N个目标数据列;

步骤S2:根据公式(1)获取所述N个目标数据列中第j个目标数据列的占比值D_j;

$$D_j=\frac{\sqrt{\frac{1}{\mathrm{m}-1}\sum_{i=1}^{m-1}[P_{\mathrm{ji}}\log\frac{P_{\mathrm{ji}}}{P_{\mathrm{ji+1}}}+(1-P_{\mathrm{ji}})\log\frac{1-P_{\mathrm{ji}}}{1-P_{\mathrm{ji+1}}}]}}{\frac{1}{m}\sum_{i=1}^{m}P_{\mathrm{ji}}}\times\log\left(\frac{N}{m}+0.1\right)\quad(1);$$

其中，m表示第j个目标数据列中包括有m个元素；P_{ji}表示第j个目标数据列中的第i个元素的相似值；P_{ji+1}表示第j个目标数据列中第i+1个元素的相似值；且0＜j≤N；

步骤S3：根据公式(2)对所述步骤S2获取到的占比值D_j进行修正处理，得到修成处理值F_j；

$$F_j=D_{\mathrm{j}}-D_{\mathrm{j}}\frac{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})(D_{\mathrm{j}}-\overline{D})}{\sqrt{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})^2\sum_{\mathrm{i=1}}^{\mathrm{m}}(D_{\mathrm{j}}-\overline{D})^2}}\times e^{\sum_{\mathrm{i=1}}^{m}(1-\frac{\mathrm{d_{ji}}}{d_{ji}+\eta_{ji}})\frac{2\mathrm{d}_{ji}}{\mathrm{d_{ji}}+\delta_{ji}}}\quad(2);$$

其中，$\overline{P}$表示第j个目标数据列中的m个元素的平均相似值；D_j表示第N个目标数据列的平均占比值，e表示自然常数；d_{ji}表示第j个目标数据列中得第i个元素占用第j个目标数据列总存储空间的存储值；η_{ji}表示第j个目标数据列中的第一修正参数；δ_{ji}表示第j个目标数据列中中第二修正参数；

步骤S4：根据公式(3)对修正处理值F_j进行从高到低的排序，确定最高修正处理值A对应的目标数据列为最优目标数据列，并获得最优计算方式；

$A=\max(F_j,j=1、2、3\cdots N)$ (3)；

其中，max()表示获取F_j中的最高修正处理值A。

2.如权利要求1所述的方法，其特征在于，在执行完所述步骤5后，还包括：

步骤6：保存与数据处理通用流程相关的所述步骤1-5。

3.如权利要求1所述的方法，其特征在于，在执行所述步骤1之前，还包括：

步骤01：获取原始数据；

步骤02：对所述步骤01所获取的原始数据进行预设处理，获得目标数据集合。

4.如权利要求1所述的方法，其特征在于，在执行完所述步骤1之后，且未执行所述步骤2之前，还包括：

步骤21：选择预先输入的数据源中的数据列；

步骤22：将所述步骤21所选择的数据列进行组合处理，构成选择数据集合。

5.如权利要求1所述的方法，其特征在于，

所述选择数据集合中的选择数据列为输入数据；

所述目标数据集合中的目标数据列为输出数据。

6.如权利要求1所述的方法，其特征在于，

所述点映射关系，可以是一对一、一对多、多对一、多对多的点映射关系；

所述区域映射关系，可以是一对一、一对多、多对一、多对多的区域映射关系。

7.一种数据预处理通用流程系统，其特征在于，包括：

定义模块,用于定义数据结构到目标数据集合;

输入模块,用于接收输入的选择数据集合;

映射模块,用于映射所述输入模块所输入的选择数据集合中的选择数据列,到所述定义模块所定义的所述目标数据集合中的目标数据列,获得相关的映射关系;

确定模块,用于根据所述映射模块所获得的映射关系,确定所述目标数据集合中的每个所述目标数据列的计算方式;

获得模块,用于根据所述确定模块所确定的计算方式,对所述选择数据列的数据进行粒度处理,并获得粒度处理后的选择数据;

其中,所述每个目标数据列的计算方式是基于预先存储的预设数目个预设算子组合得到的,且所述预设算子之间映射连接,其中所述预设算子是一些基础算子;

所述映射模块,映映射所述输入模块所输入的选择数据集合中的选择数据列,到所述定义模块所定义的所述目标数据集合中的目标数据列的具体过程包括:

对所述选择数据集合中的选择数据列中的数据元素进行第一标记;

对所述目标数据集合中的目标数据列中的数据元素进行第二标记;

获取所述选择数据集合中的选择数据列对应的选择映射区域、及所述目标数据集合中的目标数据列对应的目标映射区域;

其中,所述选择映射区域包括:至少一个进行第一标记的选择特征点;

所述目标映射区域包括:至少一个进行第二标记的目标特征点;

根据所述进行第一标记的选择特征点对应的选择特征点信息,在所述目标映射区域中查找进行第二标记的目标特征点;

同时,判断所获取的选择特征点和目标特征点是否满足预先存储的映射数据库中的预设映射规则;

若是,将所述选择特征点映射到所述目标特征点上,获得两者之间的点映射关系,并将所述点映射关系存储到相应的待扩展数据库中,实现选择数据列到目标数据列的映射;

否则,将所述选择映射区域映射到所述目标映射区域上,获得两者之间的区域映射关系,并将所述区域映射关系存储到相应的待扩展数据库中,实现选择数据列到目标数据列的映射;

所述获得模块,还用于根据所述确定模块所确定的计算方式,对所述选择数据列的数据进行粒度处理之前,筛选出目标数据列的最优计算方式,对所述选择数据列的数据进行粒度处理;

其中,筛选出目标数据列的最优计算方式具体包括:

确定在所述映射模块所获得的映射关系中,所述选择数据列与所述目标数据列的映射关系是否为一对一映射关系或多对一映射关系,若是,确定所述目标数据列的计算方式为最优计算方式;

否则,筛选出所述目标数据列的最优计算方式,其步骤包括:

步骤S1:确定当前选择数据列所映射的N个目标数据列;

步骤S2:根据公式(1)获取所述N个目标数据列中第j个目标数据列的占比值D_j;

$$D_j=\frac{\sqrt{\frac{1}{\mathrm{m}-1}\sum_{i=1}^{m-1}[P_{\mathrm{ji}}\log\frac{P_{\mathrm{ji}}}{P_{\mathrm{ji+1}}}+(1-P_{\mathrm{ji}})\log\frac{1-P_{\mathrm{ji}}}{1-P_{\mathrm{ji+1}}}]}}{\frac{1}{m}\sum_{i=1}^{m}P_{\mathrm{ji}}}\times\log\left(\frac{N}{m}+0.1\right) \quad (1);$$

其中，m表示第j个目标数据列中包括有m个元素；P_{ji}表示第j个目标数据列中的第i个元素的相似值；P_{ji+1}表示第j个目标数据列中第i+1个元素的相似值；且0＜j≤N；

步骤S3：根据公式(2)对所述步骤S2获取到的占比值D_j进行修正处理，得到修成处理值F_j；

$$F_j=D_{\mathrm{j}}-D_{\mathrm{j}}\frac{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})(D_{\mathrm{j}}-\overline{D})}{\sqrt{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})^2\sum_{\mathrm{i=1}}^{\mathrm{m}}(D_{\mathrm{j}}-\overline{D})^2}}\times e^{\sum_{\mathrm{i=1}}^{m}(1-\frac{\mathrm{d_{ji}}}{d_{ji}+\eta_{ji}})\frac{2\mathrm{d}_{ji}}{\mathrm{d_{ji}}+\delta_{ji}}} \quad (2);$$

其中，$\overline{P}$表示第j个目标数据列中的m个元素的平均相似值；D_j表示第N个目标数据列的平均占比值，e表示自然常数；d_{ji}表示第j个目标数据列中得第i个元素占用第j个目标数据列总存储空间的存储值；η_{ji}表示第j个目标数据列中的第一修正参数；δ_{ji}表示第j个目标数据列中中第二修正参数；

步骤S4：根据公式(3)对修正处理值F_j进行从高到低的排序，确定最高修正处理值A对应的目标数据列为最优目标数据列，并获得最优计算方式；

$$A=\max(F_j,j=1、2、3...N) \quad (3)；$$

其中，max()表示获取F_j中的最高修正处理值A。

一种数据预处理通用流程方法及系统

技术领域

[0001] 本发明涉及数据处理技术领域,特别涉及一种数据预处理通用流程方法及系统。

背景技术

[0002] 目前原始仿真数据在进入分析系统之前的数据清洗和预处理工作主要由操作员手动进行处理,操作员使用的工具主要有excel、sql数据库、python脚本语言等,但这些方式都有各自的缺点,具体如下:

[0003] 1.使用excel处理数据需要原始数据是excel格式或者是excel软件支持的格式,如果仿真系统将数据保存为其它格式,例如普通文本,数据库,等格式则excel无法处理,且excel需要操作员手动进行整个过程,效率较低,其处理流程无法保存,对结构相同但具体值不同的不同批次数据需要每次重复手工处理流程,造成大量的重复劳动;

[0004] 2.利用sql数据库对原始数据进行处理时可以保存处理流程,但其处理的建立需要操作人员懂得sql语句,sql语句的编写和调试均需要专业人员才能完成,这种方式对操作人员的要求较高,不是特别通用,且只能处理数据库中的数据,对其它来源的则数据无能为力;

[0005] 3.利用python等脚本语言进行数据处理时灵活性较高,其支持的格式很多,处理流程也能持久化,但同样需要懂得相应脚本的专业人员才能使用,对操作员的要求较高。

发明内容

[0006] 本发明提供一种数据预处理通用流程方法,用以通过定义数据结构和相关映射,来提高处理流程的扩张性和灵活性。

[0007] 本发明实施例提供一种数据预处理通用流程方法,包括:

[0008] 步骤1:定义数据结构到目标数据集合;

[0009] 步骤2:输入选择数据集合;

[0010] 步骤3:映射所述步骤2所输入的选择数据集合中的选择数据列,到所述步骤1所定义的所述目标数据集合中的目标数据列,获得相关的映射关系;

[0011] 步骤4:根据所述步骤3所获得的映射关系,确定所述目标数据集合中的每个所述目标数据列的计算方式;

[0012] 步骤5:根据所述步骤4所确定的计算方式,对所述选择数据列的数据进行粒度处理,并获得粒度处理后的选择数据。

[0013] 在一种可能实现的方式中,在执行完所述步骤5后,还包括:

[0014] 步骤6:保存与所述数据处理通用流程相关的所述步骤1-5。

[0015] 在一种可能实现的方式中,在执行所述步骤1之前,还包括:

[0016] 步骤01:获取原始数据;

[0017] 步骤02:对所述步骤01所获取的原始数据进行预设处理,获得目标数据集合。

[0018] 在一种可能实现的方式中,在执行完所述步骤1之后,且未执行所述步骤2之前,还

包括：

[0019] 步骤21：选择预先输入的数据源中的数据列；

[0020] 步骤22：将所述步骤21所选择的数据列进行组合处理，构成选择数据集合。

[0021] 在一种可能实现的方式中，

[0022] 所述选择数据集合中的选择数据列为输入数据；

[0023] 所述目标数据集合中的目标数据列为输出数据。

[0024] 在一种可能实现的方式中，

[0025] 所述步骤1中，所述每个目标数据列的计算方式是基于预先存储的预设数目个预设算子组合得到的，且所述预设算子之间映射连接。

[0026] 在一种可能实现的方式中，

[0027] 所述步骤3中，映射所述步骤2所输入的选择数据集合中的选择数据列，到所述步骤1所定义的所述目标数据集合中的目标数据列的具体过程包括：

[0028] 对所述选择数据集合中的选择数据列中的数据元素进行第一标记；

[0029] 对所述目标数据集合中的目标数据列中的数据元素进行第二标记；

[0030] 获取所述选择数据集合中的选择数据列对应的选择映射区域、及所述目标数据集合中的目标数据列对应的目标映射区域；

[0031] 其中，所述选择映射区域包括：至少一个进行第一标记的选择特征点；

[0032] 所述目标映射区域包括：至少一个进行第二标记的目标特征点；

[0033] 根据所述进行第一标记的选择特征点对应的选择特征点信息，在所述目标映射区域中查找进行第二标记的目标特征点；

[0034] 同时，判断所获取的选择特征点和目标特征点是否满足预先存储的映射数据库中的预设映射规则；

[0035] 若是，将所述选择特征点映射到所述目标特征点上，获得两者之间的点映射关系，并将所述点映射关系存储到相应的待扩展数据库中，实现选择数据列到目标数据列的映射；

[0036] 否则，将所述选择映射区域映射到所述目标映射区域上，获得两者之间的区域映射关系，并将所述区域映射关系存储到相应的待扩展数据库中，实现选择数据列到目标数据列的映射。

[0037] 在一种可能实现的方式中，

[0038] 所述点映射关系，可以是一对一、一对多、多对一、多对多的点映射关系；

[0039] 所述区域映射关系，可以是一对一、一对多、多对一、多对多的区域映射关系。

[0040] 本发明实施例提供一种数据预处理通用流程系统，包括：

[0041] 定义模块，用于定义数据结构到目标数据集合；

[0042] 输入模块，用于接收输入的选择数据集合；

[0043] 映射模块，用于映射所述输入模块所输入的选择数据集合中的选择数据列，到所述定义模块所定义的所述目标数据集合中的目标数据列，获得相关的映射关系；

[0044] 确定模块，用于根据所述映射模块所获得的映射关系，确定所述目标数据集合中的每个所述目标数据列的计算方式；

[0045] 获得模块，用于根据所述确定模块所确定的计算方式，对所述选择数据列的数据

进行粒度处理，并获得粒度处理后的选择数据。

[0046] 在一种可能实现的方式中，

[0047] 所述步骤5中，根据所述步骤4所确定的计算方式，对所述选择数据列的数据进行粒度处理之前，还包括：筛选出目标数据列的最优计算方式，对所述选择数据列的数据进行粒度处理；

[0048] 其中，筛选出目标数据列的最优计算方式具体包括：

[0049] 确定所述步骤3所获得的映射关系中，所述选择数据列与所述目标数据列的映射关系是否为一对一映射关系或多对一映射关系，若是，确定所述目标数据列的计算方式为最优计算方式；

[0050] 否则，筛选出所述目标数据列的最优计算方式，其步骤包括：

[0051] 步骤S1：确定当前选择数据列所映射的N个目标数据列；

[0052] 步骤S2：根据公式(1)获取所述N个目标数据列中第j个目标数据列的占比值D_j；

[0053] $$D_j=\frac{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m-1}[P_{ji}\log\frac{P_{ji}}{P_{ji+1}}+(1-P_{ji})\log\frac{1-P_{ji}}{1-P_{ji+1}}]}}{\frac{1}{m}\sum_{i=1}^{m}P_{ji}}\times\log\left(\frac{N}{m}+0.1\right) \quad (1);$$

[0054] 其中，m表示第j个目标数据列中包括有m个元素；P_{ji}表示第j个目标数据列中的第i个元素的相似值；P_{ji+1}表示第j个目标数据列中第i+1个元素的相似值；且0＜j≤N；

[0055] 步骤S3：根据公式(2)对所述步骤S2获取到的占比值D_j进行修正处理，得到修成处理值F_j；

[0056] $$F_j=D_j-D_j\frac{\sum_{i=1}^{m}(P_{ji}-\overline{P})(D_j-\overline{D})}{\sqrt{\sum_{i=1}^{m}(P_{ji}-\overline{P})^2\sum_{i=1}^{m}(D_j-\overline{D})^2}}\times e^{\sum_{i=1}^{m}(1-\frac{d_{ji}}{d_{ji}+\eta_{ji}})\frac{2d_{ji}}{d_{ji}+\delta_{ji}}} \quad (2);$$

[0057] 其中，$\overline{P}$表示第j个目标数据列中的m个元素的平均相似值；D_j表示第N个目标数据列的平均占比值，e表示自然常数；d_{ji}表示第j个目标数据列中得第i个元素占用第j个目标数据列总存储空间的存储值；η_{ji}表示第j个目标数据列中的第一修正参数；δ_{ji}表示第j个目标数据列中中第二修正参数；

[0058] 步骤S4：根据公式(3)对修正处理值F_j进行从高到低的排序，确定最高修正处理值A对应的目标数据列为最优目标数据列，并获得最优计算方式；

[0059] $A=\max(F_j, j=1、2、3\cdots N)$ (3)；

[0060] 其中，max()表示获取F_j中的最高修正处理值A。

[0061] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在所写的说明书、权利要求书、以及附图中所特别指出的结构来实现和获得。

[0062] 下面通过附图和实施例，对本发明的技术方案做进一步的详细描述。

附图说明

[0063] 附图用来提供对本发明的进一步理解，并且构成说明书的一部分，与本发明的实施例一起用于解释本发明，并不构成对本发明的限制。在附图中：

[0064] 图1为本发明实施例中一种数据预处理通用流程方法的流程图；

[0065] 图2为本发明实施例中一种数据预处理通用流程系统的结构示意图。

具体实施方式

[0066] 以下结合附图对本发明的优选实施例进行说明，应当理解，此处所描述的优选实施例仅用于说明和解释本发明，并不用于限定本发明。

[0067] 本发明实施例提供一种数据预处理通用流程方法，如图1所示，包括：

[0068] 步骤1：定义数据结构到目标数据集合；

[0069] 步骤2：输入选择数据集合；

[0070] 步骤3：映射所述步骤2所输入的选择数据集合中的选择数据列，到所述步骤1所定义的所述目标数据集合中的目标数据列，获得相关的映射关系；

[0071] 步骤4：根据所述步骤3所获得的映射关系，确定所述目标数据集合中的每个所述目标数据列的计算方式；

[0072] 步骤5：根据所述步骤4所确定的计算方式，对所述选择数据列的数据进行粒度处理，并获得粒度处理后的选择数据。

[0073] 上述目标数据集合通常为一个二维数据表(也可以是一维)；

[0074] 且在对上述目标数据集合进行数据结构的定义，一般是对目标数据定义列数、及每列名称、每列的数据类型如默认为double类型等结构。

[0075] 上述输入选择数据集合，一般是通过选择数据源，再从数据源中选择数据列，且选择的数据源可以是数据库，文件等不同类型的数据源，而且数据源一般是一个二维数据集合；

[0076] 在选择数据源的过程中，可以是同时选择多个数据源，且数据源的类型可以不相同；每个数据源的选择可以具体到数据列(例如某个数据源中有十列数据，可以只选择其中的某一列或者某几列参与处理)；

[0077] 最终将选择的各个数据源的数据列组合成一个输入数据集合，即选择数据集。

[0078] 上述选择数据集合的选择数据列和目标数据集合的目标数据列的列数可以不同，也可以相同，其两者的映射关系也可以不是一一对应的；

[0079] 还可以允许用户自定义选择数据列与目标数据列的映射关系，列与列之间可以随意映射，如一对一，一对多，多对多，多对一等映射关系；

[0080] 其中，选择数据集合的选择数据列为输入数据，目标数据集合的目标数据列为输出数据。

[0081] 上述计算方式，是一些基础算子的自由组合，如spark基础算子等，其目的是为了方便扩展。

[0082] 上述对选择数据列的数据进行粒度处理，并获得粒度处理后的选择数据，其对应的基本处理单位可以具体到每个数据源的数据列，通过组合算子可以对选择数据列的数据进行更细粒度的处理。

[0083] 上述技术方案的有益效果是:通过定义数据结构和相关映射,来提高处理流程的扩张性和灵活性。

[0084] 本发明实施例提供一种数据预处理通用流程方法,在执行完所述步骤5后,还包括:

[0085] 步骤6:保存与所述数据处理通用流程相关的所述步骤1-5。

[0086] 通过使用上述步骤1-5,不仅方便处理普通文本保存的数据,也可以处理数据库中的数据,且支持扩展,可方便的扩展至网络、内存等其它来源的数据。

[0087] 上述技术方案的有益效果是:通过保存,便于进行下次使用,提高下次数据处理流程的速度,节省时间。

[0088] 本发明实施例提供一种数据预处理通用流程方法,在执行所述步骤1之前,还包括:

[0089] 步骤01:获取原始数据;

[0090] 步骤02:对所述步骤01所获取的原始数据进行预设处理,获得目标数据集合。

[0091] 上述原始数据,可以是数据库中的数据或文档文件中的相关数据;

[0092] 上述对原始数据进行预设处理,例如可以是对原始数据进行的数据格式的转换,方便非专业人员对其进行操作的可能性。

[0093] 例如当上述原始数据为sql语句相关的一维阵列数据时,将其进行预设处理后,得到二维阵列的目标数据集合。

[0094] 上述技术方案的有益效果是:通过进行预处理,方便进行统一,同时,便于后续操作。

[0095] 本发明实施例提供一种数据预处理通用流程方法,在执行完所述步骤1之后,且未执行所述步骤2之前,还包括:

[0096] 步骤21:选择预先输入的数据源中的数据列;

[0097] 步骤22:将所述步骤21所选择的数据列进行组合处理,构成选择数据集合。

[0098] 上述数据源可以是多种不同种类的数据源,其选择的数据列可以是选择的不同数据源中的不同数据列,来组合形成选择数据集合。

[0099] 上述技术方案的中的有益效果是:便于构成不同的选择数据集合,提高其的多样性。

[0100] 本发明实施例提供一种数据预处理通用流程方法,

[0101] 所述步骤1中,所述每个目标数据列的计算方式是基于预先存储的预设数目个预设算子组合得到的,且所述预设算子之间映射连接。

[0102] 上述预设数目的预设算子组合得到的,例如有2个预设数目的预设算子,其组合的计算方式,可以是2种,其好处是,提高计算方式的多样性。

[0103] 上述预设算子之间的映射连接,例如存在算子1、算子2、算子3;

[0104] 其中,算子1与算子2和算子3映射连接;

[0105] 算子2与算子1和算子3映射连接;

[0106] 算子3与算子1和算子2映射连接。

[0107] 上述技术方案的有益效果是:便于通过基本算子的组合来定义计算方式,且由于其算子具有方便扩展性,便于获得多种计算方式。

[0108] 本发明实施例提供一种数据预处理通用流程方法，

[0109] 所述步骤3中，映射所述步骤2所输入的选择数据集合中的选择数据列，到所述步骤1所定义的所述目标数据集合中的目标数据列的具体过程包括：

[0110] 对所述选择数据集合中的选择数据列中的数据元素进行第一标记；

[0111] 对所述目标数据集合中的目标数据列中的数据元素进行第二标记；

[0112] 获取所述选择数据集合中的选择数据列对应的选择映射区域、及所述目标数据集合中的目标数据列对应的目标映射区域；

[0113] 其中，所述选择映射区域包括：至少一个进行第一标记的选择特征点；

[0114] 所述目标映射区域包括：至少一个进行第二标记的目标特征点；

[0115] 根据所述进行第一标记的选择特征点对应的选择特征点信息，在所述目标映射区域中查找进行第二标记的目标特征点；

[0116] 同时，判断所获取的选择特征点和目标特征点是否满足预先存储的映射数据库中的预设映射规则；

[0117] 若是，将所述选择特征点映射到所述目标特征点上，获得两者之间的点映射关系，并将所述点映射关系存储到相应的待扩展数据库中，实现选择数据列到目标数据列的映射；

[0118] 否则，将所述选择映射区域映射到所述目标映射区域上，获得两者之间的区域映射关系，并将所述区域映射关系存储到相应的待扩展数据库中，实现选择数据列到目标数据列的映射。

[0119] 上述对选择数据列中的数据元素进行第一标记，是为了方便后续的映射；

[0120] 上述对目标数据列中的数据元素进行第二标记，也是为了方便后续的映射；

[0121] 上述映射关系，是选择数据列映射到目标数据列中，其映射关系可以是一对一、一对多、多对一、多对多的映射关系；

[0122] 上述选择数据列对应的选择映射区域，例如选择数据列对应的地址是从首地址000到地址100，其首地址000到地址100中是包含其选择数据列对应的数据的，例如可以从首地址000到地址100，选择地址050到地址080作为选择映射区域，可以有效节省映射时间，提高映射效率，其获取目标数据集合中的目标数据列对应的目标映射区域，与上述效果类似。

[0123] 上述选择特征点，可以是进行第一标记的数据元素中的一个或多个元素；

[0124] 上述目标特征点，可以是进行第二标记的数据元素中的一个或多个元素；

[0125] 上述判断所获取的选择特征点和目标特征点是否满足预先存储的映射数据库中的预设映射规则，其中例如，预设映射规则为，选择数据集合中的数据元素2可以映射到目标数据集合中的数据元素2、3和4，若此时对应的选择数据集合中的数据元素2为选择特征点，且设此时的目标特征点为5，此时，就不符合预设映射规则，若设此时的目标特征点位2、3，此时，就符合预设映射规则；

[0126] 并且，对应的点映射关系2(映射)2、2(映射)3；

[0127] 上述存储到待扩展数据库，是为了将符合的点映射关系进行存储，方便后续使用时，节省使用的时间；

[0128] 上述例如，选择映射区域中包括元素2、3；则将2、3映射到目标映射区域上，将其存

储到待扩展数据库，是为了进一步丰富映射的样本，进一步节省时间。

[0129] 上述技术方的有益效果是：通过设定预设映射规则，来判断选择特征点和目标特征点是否符合其规则，来提高选择数据集合映射的效率和精度。

[0130] 本发明实施例提供一种数据预处理通用流程方法，

[0131] 所述点映射关系，可以是一对一、一对多、多对一、多对多的点映射关系；

[0132] 所述区域映射关系，可以是一对一、一对多、多对一、多对多的区域映射关系。

[0133] 上述技术方案的有益效果是：便于提供映射的多样性。

[0134] 本发明实施例提供一种数据预处理通用流程系方法，

[0135] 10、所述步骤5中，根据所述步骤4所确定的计算方式，对所述选择数据列的数据进行粒度处理之前，还包括：筛选出目标数据列的最优计算方式，对所述选择数据列的数据进行粒度处理；

[0136] 其中，筛选出目标数据列的最优计算方式具体包括：

[0137] 确定所述步骤3所获得的映射关系中，所述选择数据列与所述目标数据列的映射关系是否为一对一映射关系或多对一映射关系，若是，确定所述目标数据列的计算方式为最优计算方式；

[0138] 否则，筛选出所述目标数据列的最优计算方式，其步骤包括：

[0139] 步骤S1：确定当前选择数据列所映射的N个目标数据列；

[0140] 步骤S2：根据公式(1)获取所述N个目标数据列中第j个目标数据列的占比值D_j；

[0141] $$D_j=\frac{\sqrt{\frac{1}{\mathrm{m}-1}\sum_{i=1}^{m-1}[P_{\mathrm{ji}}\log\frac{P_{\mathrm{ji}}}{P_{\mathrm{ji+1}}}+(1-P_{\mathrm{ji}})\log\frac{1-P_{\mathrm{ji}}}{1-P_{\mathrm{ji+1}}}]}}{\frac{1}{m}\sum_{i=1}^{m}P_{\mathrm{ji}}}\times\log\left(\frac{N}{m}+0.1\right) \quad (1);$$

[0142] 其中，m表示第j个目标数据列中包括有m个元素；P_{ji}表示第j个目标数据列中的第i个元素的相似值；P_{ji+1}表示第j个目标数据列中第i+1个元素的相似值；且0＜j≤N；

[0143] 步骤S3：根据公式(2)对所述步骤S2获取到的占比值D_j进行修正处理，得到修成处理值F_j；

[0144] $$F_j=D_{\mathrm{j}}-D_{\mathrm{j}}\frac{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})(D_{\mathrm{j}}-\overline{D})}{\sqrt{\sum_{\mathrm{i=1}}^{m}(P_{\mathrm{ji}}-\overline{P})^2\sum_{\mathrm{i=1}}^{\mathrm{m}}(D_{\mathrm{j}}-\overline{D})^2}}\times e^{\sum_{\mathrm{i=1}}^{m}(1-\frac{\mathrm{d_{ji}}}{d_{ji}+\eta_{ji}})\frac{2\mathrm{d}_{ji}}{\mathrm{d_{ji}}+\delta_{ji}}} \quad (2);$$

[0145] 其中，$\overline{P}$表示第j个目标数据列中的m个元素的平均相似值；D_j表示第N个目标数据列的平均占比值，e表示自然常数；d_{ji}表示第j个目标数据列中得第i个元素占用第j个目标数据列总存储空间的存储值；η_{ji}表示第j个目标数据列中的第一修正参数；δ_{ji}表示第j个目标数据列中中第二修正参数；

[0146] 步骤S4：根据公式(3)对修正处理值F_j进行从高到低的排序，确定最高修正处理值A对应的目标数据列为最优目标数据列，并获得最优计算方式；

[0147] $A=\max(F_j, j=1、2、3...N)$ (3)；

[0148] 其中，max()表示获取F_j中的最高修正处理值A。

[0149] 上述技术方案的有益效果是：当映射关系为一对一或多对一的情况时，直接获取目标数据列的计算方式，其操作简单，且当映射关系为多对一或多对多的情况时，获取目标数据列中的最优目标数据列，进而得到最优计算方式，可以有效的节省选择目标序列在采用计算方式的便捷性，降低选择目标序列计算过程中的复杂性，同时还节省选择目标序列的计算时间，使得计算过程得到优化。

[0150] 本发明实施例提供一种数据预处理通用流程系统，如图2所示，包括：

[0151] 定义模块，用于定义数据结构到目标数据集合；

[0152] 输入模块，用于接收输入的选择数据集合；

[0153] 映射模块，用于映射所述输入模块所输入的选择数据集合中的选择数据列，到所述定义模块所定义的所述目标数据集合中的目标数据列，获得相关的映射关系；

[0154] 确定模块，用于根据所述映射模块所获得的映射关系，确定所述目标数据集合中的每个所述目标数据列的计算方式；

[0155] 获得模块，用于根据所述确定模块所确定的计算方式，对所述选择数据列的数据进行粒度处理，并获得粒度处理后的选择数据。

[0156] 上述技术方案的有益效果是：通过定义数据结构和相关映射，来提高处理流程的扩张性和灵活性。

[0157] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

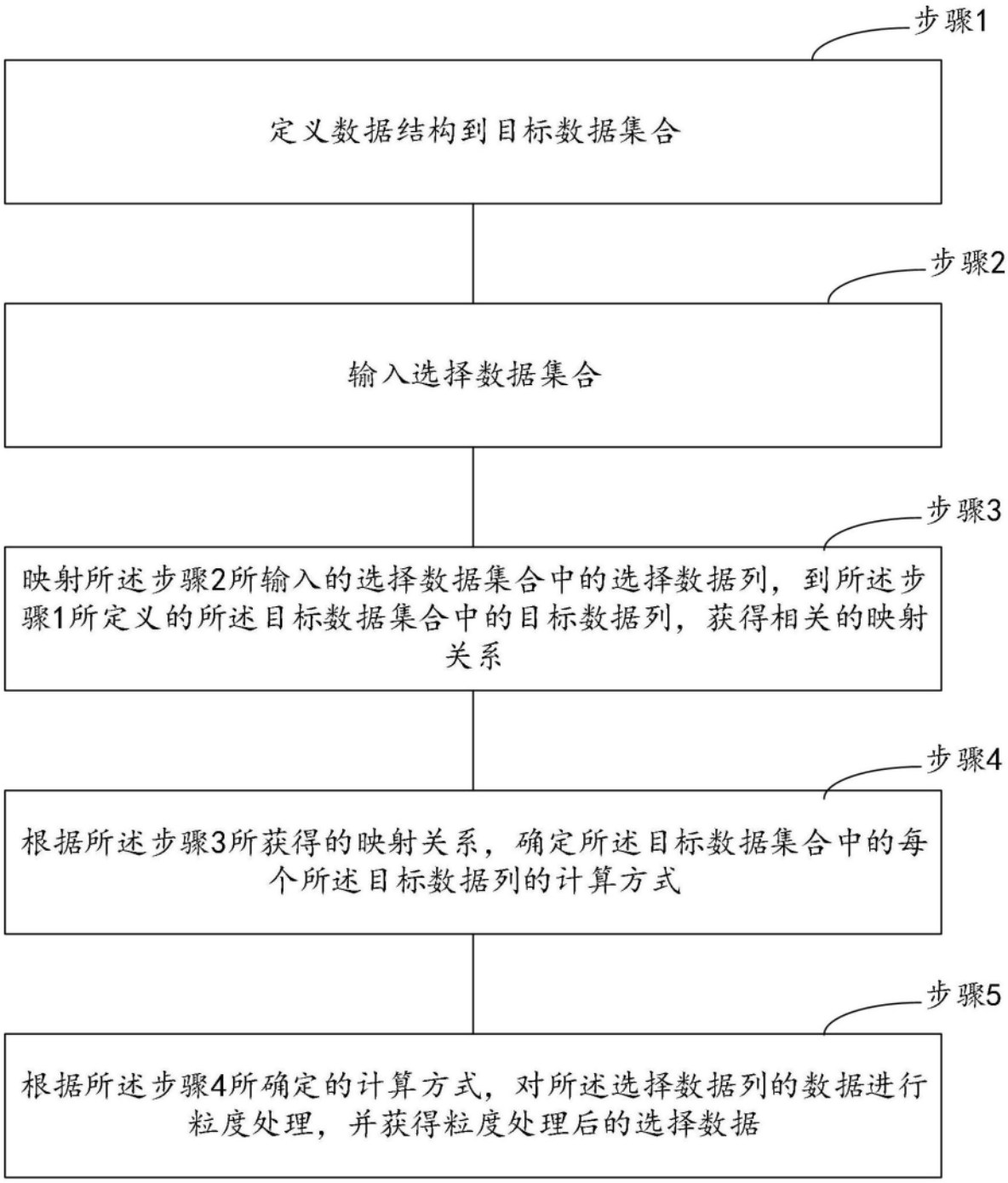

步骤1
定义数据结构到目标数据集合
步骤2
输入选择数据集合
步骤3
映射所述步骤2所输入的选择数据集合中的选择数据列，到所述步骤1所定义的所述目标数据集合中的目标数据列，获得相关的映射关系
步骤4
根据所述步骤3所获得的映射关系，确定所述目标数据集合中的每个所述目标数据列的计算方式
步骤5
根据所述步骤4所确定的计算方式，对所述选择数据列的数据进行粒度处理，并获得粒度处理后的选择数据

图1

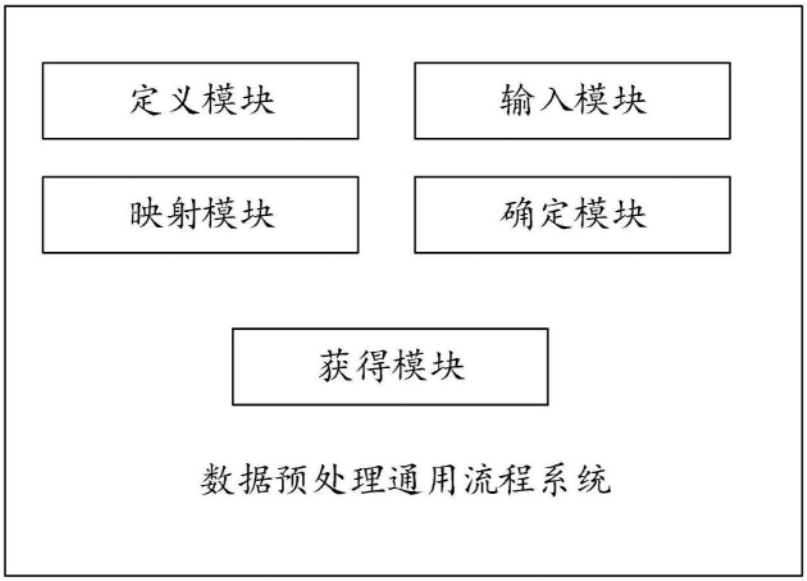


图2