(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113590599 A
(43)申请公布日 2021.11.02

(21)申请号 202110831880.7

(22)申请日 2021.07.22

(71)申请人 创意信息技术股份有限公司
地址 610000 四川省成都市青羊区万和路99号丽阳天下7－9室

(72)发明人 段红兵 王震 王勇 黄磊 王苹 周礼 钟明坤

(74)专利代理机构 成都金英专利代理事务所(普通合伙) 51218
代理人 郭肖凌

(51)Int.Cl.
G06F 16/215(2019.01)
G06Q 50/26(2012.01)

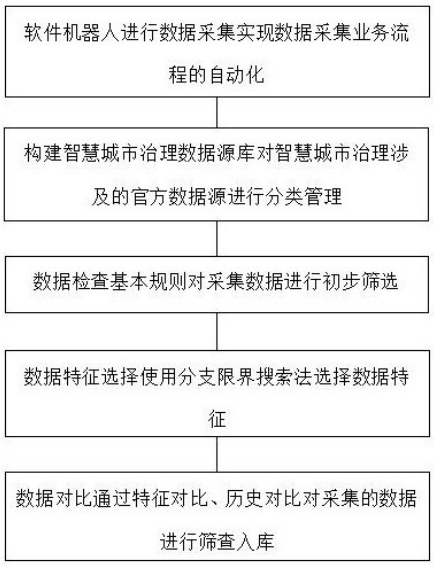
权利要求书1页 说明书6页 附图1页

(54)发明名称

基于多元复杂数据环境的数据检查方法

(57)摘要

本发明公开了基于多元复杂数据环境的数据检查方法，其特征在于，包括以下步骤：步骤1：使用自动化数据采集系统进行数据采集，实现数据采集业务流程的自动化；步骤2：构建智慧城市治理数据源库，对智慧城市治理的数据源进行分类管理；步骤3：数据检查基本规则，对采集数据进行初步筛选；步骤4：数据特征选择，使用分支限界搜索法选择数据特征；步骤5：通过数据对比对采集的数据进行筛查入库。本发明适用于城市治理，结合城市治理调度的业务需求，检查从多元复杂数据中获取到的各类业务数据，利用数据检查机制，确认数据的及时性、准确性，并能契合城市治理调度的核心业务，且确保城市治理有迹可循、有据可依。



CN 113590599 A

1.基于多元复杂数据环境的数据检查方法,其特征在于,包括以下步骤:

步骤1:使用自动化数据采集系统进行数据采集,实现数据采集业务流程的自动化;

步骤2:构建智慧城市治理数据源库,对智慧城市治理的数据源进行分类管理;

步骤3:数据检查基本规则,对采集数据进行初步筛选;

步骤4:数据特征选择,使用分支限界搜索法选择数据特征;

步骤5:通过数据对比对采集的数据进行筛查入库。

2.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述软件机器人包括模拟系统登录模块、连接系统接口API模块、读写数据库模块、读取excel文件模块和规则定制操作模块;以完成从复杂的数据环境采集数据,兼容更多的数据采集场景。

3.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述数据源包括生态环境、经济运行、城市安全、社会稳定和营商环境这5个方面的官方数据,软件机器人在进行数据采集时,通过扫描数据源,获取相关领域的数据,在数据源有新增、更改和废气时,及时更新数据库。

4.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述数据检查基本规则包括空值检查、枚举值检查、模式检查、唯一性检查、正确性检查和普及性检查,对采集数据进行初步筛选,缩小后续进行数据对比的数据量。

5.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述步骤4包括以下子步骤:

步骤401:对根据前期收集的各方面样本数据设置分支限界,使用分支限界搜索法对样本数据进行全面扫描,剪除不符合限界的特征;

步骤402:使用保留下的数据特征对样本数据进行特征匹配度,确认匹配度平均值是否达到阈值;若达到阈值,则保留数据特征做为后续数据对比的依据;否则设置新的分支限界,继续对剩余数据特征进行分支限界搜索,直到匹配度平均值达到阈值。

6.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述数据对比包括特征对比或历史对比,以及同时进行特征对比和历史对比;

特征对比:依据前面筛选出的数据特征,对采集的数据进行数据特征比对,计算特征匹配度,保留特征匹配度在阈值之上的数据;

历史对比:提取最近3个统计周期的数据,计算各数据项的变化幅度,对变化幅度超过阈值的数据进行标记,由人工确认数据的合理性。

7.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,筛查入库通过率低于阈值时,以现有数据为样本,使用分支限界搜索重新选择数据特征。

8.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述自动化数据采集系统为软件机器人。

9.根据权利要求1所述的基于多元复杂数据环境的数据检查方法,其特征在于,所述自动化数据采集系统为可配置数据采集模板。

基于多元复杂数据环境的数据检查方法

技术领域

[0001] 本发明涉及大数据治理领域，尤其涉及一种基于多元复杂数据环境的数据检查方法。

背景技术

[0002] 在城市建设中，国家、省、市、县存在统建、自建多种系统建设形式，系统与系统之间相互独立，形成大量的系统孤岛和数据孤岛。而城市治理需要整合各类系统的核心业务数据，构建城市治理大数据资源池，并基于大数据资源池，形成城市治理的业务逻辑，从中发现城市运行过程中存在的问题，包括但不限于生态、经济、安全和稳定。

[0003] 同时，由于统建、自建系统的特性，无法做到形成统一的数据标准和标准的数据接入方式，存在着大量的多元、复杂数据，必须采用不同的数据接入方式，包括但不限于excel表格导入、接口、库表等形式。

[0004] 在专利申请号CN201410447568.8中公开了一种以数据更新周期进行预操作提高数据查询效率的装置及方法，是通过根据数据更新周期以及业务数据查询需求，对各类查询进行分解、优化组合为最小次数查询操作，并提前进行查询操作，从而有效地降低被查询系统的压力、提高了业务数据查询的效率。本发明通过依据数据更新周期进行预先操作，从而减轻被查询系统的压力，并提高查询结果的递送速度。

[0005] 在专利申请号CN202011621928.3中公开了一种基于点位标记的高炉实时数据更新方法和装置，所述方法包括：在服务器端执行：定时获取最近一个周期的数据集合，并根据最近一个周期的数据集合预测下一周期的模拟数据集合；在一个周期后获取下一周期的实际数据集合；比较下一周期的实际数据集合和下一周期的模拟数据集合的差异，获得更新点数据集合，并将所述更新点数据集合发送给客户端；在客户端执行：将最近一个周期的数据集合复制为下一周期的临时数据集合，并根据接收的更新点数据对下一周期的临时数据集合进行数据修正，以得到下一周期的实际数据集合，再根据下一周期的实际数据集合执行图形显示或图表动态显示。适用于高炉数据等缓慢变化的实时数据的显示刷新。

[0006] 以上专利虽然对数据的查询和更新提出了高效的方法手段，但是它们并没有对比多元复杂数据的接入、查询和更新提出一种有效的实现方法，在面对数据源必须采用不同的接入方式，无法做到形成统一的数据标准和标准的数据接入方式。

[0007] 为了确保数据的及时性、准确性，且确保城市治理有迹可循、有据可依，就必须建立基于多元复杂数据构建数据检查机制。

发明内容

[0008] 本发明的目的在于克服现有技术的不足，依据智慧城市治理调度业务需求，检查从多元复杂数据环境采集到的数据的数量、质量和准确性，为解决不同数据接入形式下存在无法检查多元复杂数据的及时性、准确性问题，提供一种基于多元复杂数据环境的数据检查方法。

[0009] 本发明的目的是通过以下技术方案来实现的：

基于多元复杂数据环境的数据检查方法，包括以下步骤：

步骤1：使用自动化数据采集系统进行数据采集，实现数据采集业务流程的自动化；

步骤2：构建智慧城市治理数据源库，对智慧城市治理的数据源进行分类管理；

步骤3：数据检查基本规则，对采集数据进行初步筛选；

步骤4：数据特征选择，使用分支限界搜索法选择数据特征；

步骤5：通过数据对比对采集的数据进行筛查入库。

[0010] 进一步的，所述软件机器人包括模拟系统登录模块、连接系统接口API模块、读写数据库模块、读取excel文件模块和规则定制操作模块；以完成从复杂的数据环境采集数据，兼容更多的数据采集场景。

[0011] 进一步的，所述数据源包括生态环境、经济运行、城市安全、社会稳定和营商环境这5个方面的官方数据，软件机器人在进行数据采集时，通过扫描数据源，获取相关领域的数据，在数据源有新增、更改和废气时，及时更新数据库。

[0012] 进一步的，所述数据检查基本规则包括空值检查、枚举值检查、模式检查、唯一性检查、正确性检查和普及性检查，对采集数据进行初步筛选，缩小后续进行数据对比的数据量。

[0013] 进一步的，所述步骤4包括以下子步骤：

步骤401：对根据前期收集的各方面样本数据设置分支限界，使用分支限界搜索法对样本数据进行全面扫描，剪除不符合限界的特征；

步骤402：使用保留下的数据特征对样本数据进行特征匹配度，确认匹配度平均值是否达到阈值；若达到阈值，则保留数据特征做为后续数据对比的依据；否则设置新的分支限界，继续对剩余数据特征进行分支限界搜索，直到匹配度平均值达到阈值。

[0014] 进一步的，所述数据对比包括特征对比或历史对比，以及同时进行特征对比和历史对比；

特征对比：依据前面筛选出的数据特征，对采集的数据进行数据特征比对，计算特征匹配度，保留特征匹配度在阈值之上的数据；

历史对比：提取最近3个统计周期的数据，计算各数据项的变化幅度，对变化幅度超过阈值的数据进行标记，由人工确认数据的合理性。

[0015] 进一步的，筛查入库通过率低于阈值时，以现有数据为样本，使用分支限界搜索重新选择数据特征。

[0016] 进一步的，所述自动化数据采集系统为软件机器人。

[0017] 进一步的，所述自动化数据采集系统为可配置数据采集模板。

[0018] 本发明的有益效果：本发明适用于城市治理，结合城市治理调度的业务需求，检查从多元复杂数据中获取到的各类业务数据，利用数据检查机制，确认数据的及时性、准确性，并能契合城市治理调度的核心业务，且确保城市治理有迹可循、有据可依。

附图说明

[0019] 图1是本发明的设备原理框图。

具体实施方式

[0020] 为了对本发明的技术特征、目的和效果有更加清楚的理解，现对照附图说明本发明的具体实施方式。

[0021] 实施例1，如图1所示，基于多元复杂数据环境的数据检查方法，包括以下步骤：

步骤1：使用软件机器人进行数据采集实现数据采集业务流程的自动化；

步骤2：构建智慧城市治理数据源库对智慧城市治理涉及的官方数据源进行分类管理；

步骤3：数据检查基本规则对采集数据进行初步筛选；

步骤4：数据特征选择使用分支限界搜索法选择数据特征；

步骤5：数据对比通过特征对比、历史对比对采集的数据进行筛查入库。

[0022] 其中，所述软件机器人包括模拟系统登录模块、连接系统接口API模块、读写数据库模块、读取excel文件模块和规则定制操作模块；以完成从复杂的数据环境采集数据，兼容更多的数据采集场景。

[0023] 其中，所述数据源包括生态环境、经济运行、城市安全、社会稳定和营商环境这5个方面的官方数据，软件机器人在进行数据采集时，通过扫描数据源，获取相关领域的数据，在数据源有新增、更改和废气时，及时更新数据库。

[0024] 在本实施例中，使用软件机器人进行数据采集具体为：

通过可视化平台设计大批量、可重复性的数据采集任务，实现数据采集业务流程的自动化，按照一定的规则持续不断的重复操作；支持从复杂的数据环境(包括业务系统数据接口、excel文件、数据回流平台库表)采集数据，兼容更多的数据采集场景。

[0025] 在本实施例中，构建智慧城市治理数据源库具体为：

对智慧城市治理涉及的生态环境、经济运行、城市安全、社会稳定、营商环境这5个方面的官方数据源进行分类管理，确保数据来源的可靠性、安全性、有效性。

[0026] 其中生态环境细分为空气质量、地表水质、土壤环境、生态状况、城市环境、绿色节能；

经济运行细分为经济总量、投资与消费、进出口、企业活力、财政收支、存款贷款、居民收入、重点行业；

城市安全细分为社会诉求、社会管理、公共服务、社会救助与服务；

民生稳定细分为市场监管、源头管控、安全和应急防控、执法监督、安全共治。

[0027] 营商环境细分为企业服务质量、创新创业、生产要素获取便利度、产业环境、人才发展、法制环境。

[0028] 在本实施例中，数据检查基本规则具体为：

使用数据检查基本规则，包括空值检查、枚举值检查、模式检查、唯一性检查、正确性检查、普及性检查，对采集数据进行初步筛选，缩小后续进行数据对比的数据量。

[0029] 在本实施例中，数据特征选择具体为：

数据采集完成后，需要对数据进行筛查，为此需要分别从生态环境、经济运行、城市安全、社会稳定、营商环境5个方面提取相关的数据特征。

[0030] 对根据前期收集的各方面样本数据设置分支限界，使用分支限界搜索法对样本数据进行全面扫描，剪除不符合限界的特征。

[0031] 使用保留下的数据特征对样本数据进行特征匹配度，确认匹配度平均值是否达到阈值。如达到阈值则保留数据特征做为后续数据对比的依据。否则设置新的分支限界，继续对剩余数据特征进行分支限界搜索，直到匹配度平均值达到阈值。

[0032] 在本实施例中，数据对比具体为：

数据对比分为特征对比和历史对比两步：特征对比：依据前面筛选出的数据特征，对采集的数据进行数据特征比对，计算特征匹配度，保留特征匹配度在阈值之上的数据；历史对比：提取最近3个统计周期的数据，计算各数据项的变化幅度，对变化幅度超过阈值的数据进行标记，由人工确认数据的合理性。当人工确认为合理的数据比例达到阈值时，即数据通过率过低时，需要以现有数据为样本，使用分支限界搜索重新选择数据特征。

[0033] 在进行了特征对比的检查后，需要再进行历史对比的检查，由于数据已经经过一轮特征对比筛选，即筛选出的数据是符合数据的。与历史对比有2个方面的作用，一是确认提取的特征是否对历史数据有效；二是对于偏差较大的历史数据，是后续做数据分析的素材。

[0034] 实施例2，基于多元复杂数据环境的数据检查方法，包括以下步骤：

步骤1：通过可配置数据采集软件进行数据采集实现数据采集业务流程的自动化；

步骤2：构建智慧城市治理数据源库对智慧城市治理涉及的官方数据源进行分类管理；

步骤3：数据检查基本规则对采集数据进行初步筛选；

步骤4：数据特征选择使用分支限界搜索法选择数据特征；

步骤5：数据对比通过特征对比、历史对比对采集的数据进行筛查入库。

[0035] 其中，所述软件机器人包括模拟系统登录模块、连接系统接口API模块、读写数据库模块、读取excel文件模块和规则定制操作模块；以完成从复杂的数据环境采集数据，兼容更多的数据采集场景。

[0036] 其中，所述数据源包括生态环境、经济运行、城市安全、社会稳定和营商环境这5个方面的官方数据，软件机器人在进行数据采集时，通过扫描数据源，获取相关领域的数据，在数据源有新增、更改和废气时，及时更新数据库。

[0037] 在本实施例中，使用可配置数据采集模板进行数据采集具体为：

通过可视化平台设计大批量、可重复性的数据采集任务，可赋予管理用户一定权限，根据用户的选择完成数据采集业务，或者针对各种情况设置判断值，实现数据采集业务流程的自动化。

[0038] 在本实施例中，构建智慧城市治理数据源库具体为：

对智慧城市治理涉及的生态环境、经济运行、城市安全、社会稳定、营商环境这5个方面的官方数据源进行分类管理，确保数据来源的可靠性、安全性、有效性。

[0039] 其中生态环境细分为空气质量、地表水质、土壤环境、生态状况、城市环境、绿色节能；

经济运行细分为经济总量、投资与消费、进出口、企业活力、财政收支、存款贷款、居民收入、重点行业；

城市安全细分为社会诉求、社会管理、公共服务、社会救助与服务；

民生稳定细分为市场监管、源头管控、安全和应急防控、执法监督、安全共治。

[0040] 营商环境细分为企业服务质量、创新创业、生产要素获取便利度、产业环境、人才发展、法制环境。

[0041] 在本实施例中，数据检查基本规则具体为：

使用数据检查基本规则，包括空值检查、枚举值检查、模式检查、唯一性检查、正确性检查、普及性检查，对采集数据进行初步筛选，缩小后续进行数据对比的数据量。

[0042] 在本实施例中，数据特征选择具体为：

数据采集完成后，需要对数据进行筛查，为此需要分别从生态环境、经济运行、城市安全、社会稳定、营商环境5个方面提取相关的数据特征。

[0043] 对根据前期收集的各方面样本数据设置分支限界，使用分支限界搜索法对样本数据进行全面扫描，剪除不符合限界的特征。

[0044] 使用保留下的数据特征对样本数据进行特征匹配度，确认匹配度平均值是否达到阈值。如达到阈值则保留数据特征做为后续数据对比的依据。否则设置新的分支限界，继续对剩余数据特征进行分支限界搜索，直到匹配度平均值达到阈值。

[0045] 在本实施例中，数据对比具体为：

数据对比分为特征对比和历史对比两步：特征对比：依据前面筛选出的数据特征，对采集的数据进行数据特征比对，计算特征匹配度，保留特征匹配度在阈值之上的数据；历史对比：提取最近3个统计周期的数据，计算各数据项的变化幅度，对变化幅度超过阈值的数据进行标记，由人工确认数据的合理性。当人工确认为合理的数据比例达到阈值时，即数据通过率过低时，需要以现有数据为样本，使用分支限界搜索重新选择数据特征。

[0046] 以上述流程依次设置软件模块，实现的系统只需用户配置一定的参数，可以在无人值守的情况下，实现数据库的自动更新。同时，同人工进行确认，可以进一步地保证数据的合理性。

[0047] 实施例3，基于多元复杂数据环境的数据检查方法，包括以下步骤：

步骤1：使用软件机器人进行数据采集实现数据采集业务流程的自动化；

步骤2：构建智慧城市治理数据源库对智慧城市治理涉及的官方数据源进行分类管理；

步骤3：数据检查基本规则对采集数据进行初步筛选；

步骤4：数据特征选择使用分支限界搜索法选择数据特征；

步骤5：数据对比通过特征对比对采集的数据进行筛查入库，当数据通过率过低时，启动数据特征重新选择。

[0048] 其中，所述软件机器人包括模拟系统登录模块、连接系统接口API模块、读写数据库模块、读取excel文件模块和规则定制操作模块；以完成从复杂的数据环境采集数据，兼容更多的数据采集场景。

[0049] 其中，所述数据源包括生态环境、经济运行、城市安全、社会稳定和营商环境这5个方面的官方数据，软件机器人在进行数据采集时，通过扫描数据源，获取相关领域的数据，在数据源有新增、更改和废气时，及时更新数据库。

[0050] 在本实施例中，使用软件机器人进行数据采集具体为：

通过可视化平台设计大批量、可重复性的数据采集任务，实现数据采集业务流程的自动化，按照一定的规则持续不断的重复操作；支持从复杂的数据环境(包括业务系统数

据接口、excel文件、数据回流平台库表)采集数据，兼容更多的数据采集场景。

[0051] 在本实施例中，数据检查基本规则具体为：

使用数据检查基本规则，包括空值检查、枚举值检查、模式检查、唯一性检查、正确性检查、普及性检查，对采集数据进行初步筛选，缩小后续进行数据对比的数据量。

[0052] 在本实施例中，数据特征选择具体为：

数据采集完成后，需要对数据进行筛查，为此需要分别从生态环境、经济运行、城市安全、社会稳定、营商环境5个方面提取相关的数据特征。

[0053] 对根据前期收集的各方面样本数据设置分支限界，使用分支限界搜索法对样本数据进行全面扫描，剪除不符合限界的特征。

[0054] 使用保留下的数据特征对样本数据进行特征匹配度，确认匹配度平均值是否达到阈值。如达到阈值则保留数据特征做为后续数据对比的依据。否则设置新的分支限界，继续对剩余数据特征进行分支限界搜索，直到匹配度平均值达到阈值。

[0055] 在本实施例中，数据对比仅通过特征对比进行筛查入库；其中，特征对比：依据前面筛选出的数据特征，对采集的数据进行数据特征比对，计算特征匹配度，保留特征匹配度在阈值之上的数据。由于本实施例中仅通过特征对比，而缺少历史对比进行检查，因此此时入库的数据只是满足了符合特征，但是，其中的数据可能存在与以往数据变化较大、数据聚合度以及产生误差数据等。因此，在进行了特征对比的检查后，再进行历史对比的检查是十分必要的。

[0056] 本发明适用于城市治理，结合城市治理调度的业务需求，检查从多元复杂数据中获取到的各类业务数据，利用数据检查机制，确认数据的及时性、准确性，并能契合城市治理调度的核心业务，且确保了城市治理有迹可循、有据可依。

[0057] 以上显示和描述了本发明的基本原理和主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护的范围由所附的权利要求书及其等效物界定。

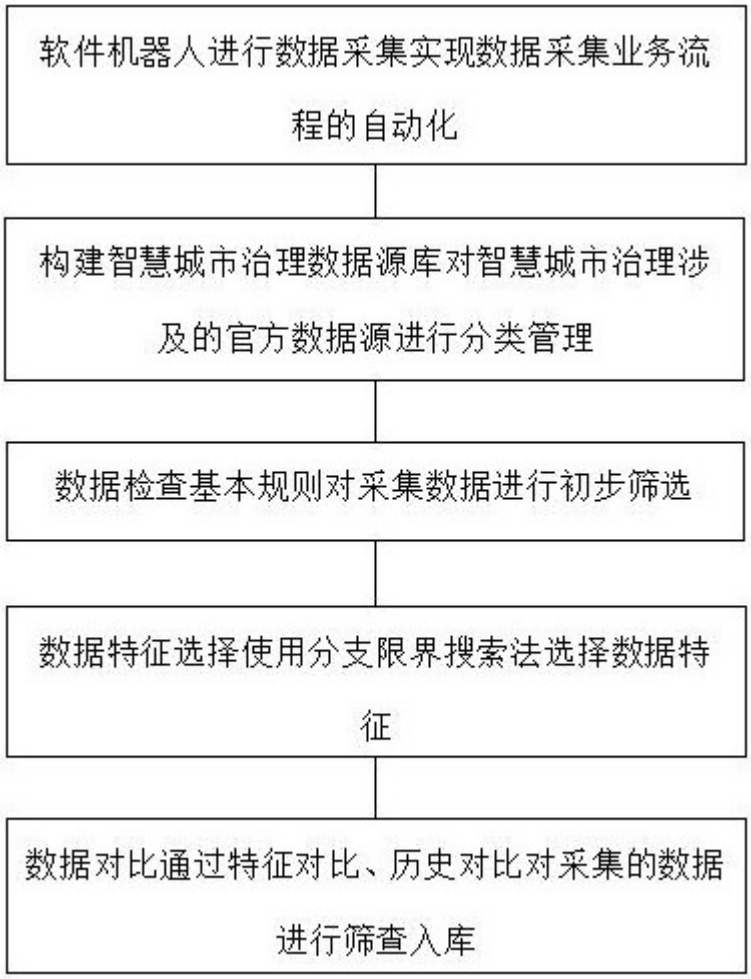

软件机器人进行数据采集实现数据采集业务流程的自动化
构建智慧城市治理数据源库对智慧城市治理涉及的官方数据源进行分类管理
数据检查基本规则对采集数据进行初步筛选
数据特征选择使用分支限界搜索法选择数据特征
数据对比通过特征对比、历史对比对采集的数据进行筛查入库

图1