



(12) 发明专利申请

(10) 申请公布号 CN 112667617 A

(43) 申请公布日 2021. 04. 16

(21) 申请号 202011617367.X

(22) 申请日 2020.12.30

(71) 申请人 南京诚勤教育科技有限公司

地址 210000 江苏省南京市鼓楼区广东路
38号

(72) 发明人 尹源

(74) 专利代理机构 北京中建联合知识产权代理
事务所(普通合伙) 11004

代理人 王晓艳

(51) Int. Cl.

G06F 16/215 (2019.01)

G06F 16/242 (2019.01)

G06F 16/26 (2019.01)

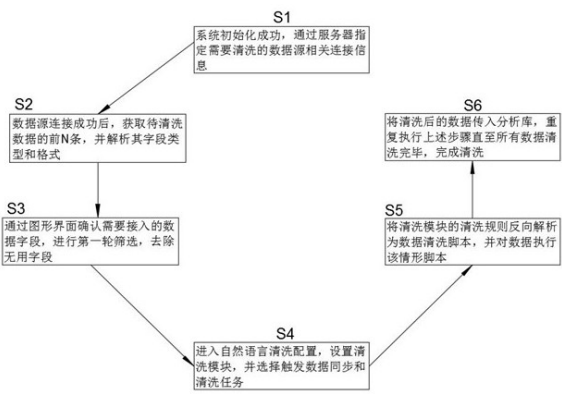
权利要求书2页 说明书4页 附图1页

(54) 发明名称

一种基于自然语言的可视化数据清洗系统
及方法

(57) 摘要

本发明涉及数据处理技术领域,具体涉及一种基于自然语言的可视化数据清洗系统及方法,本发明通过服务器指定需要清洗的数据源相关连接信息;获取待清洗数据的前N条,并解析其字段类型和格式去除无用字段;设置清洗模块,并选择触发数据同步和清洗任务;将清洗模块的清洗规则反向解析为数据清洗脚本,并对数据执行该情形脚本,将清洗后的数据传入分析库,重复执行直至所有数据清洗完毕,完成清洗。本发明实现对数据的清洗,无需掌握数据清洗工具的开发和使用方法,降低了大数据应用服务的技术门槛,提升了用户对大数据服务的体验,解决了传统的数据清洗系统的灵活性及可维护性问题,使的数据清洗工作人员的使用成本得到降低,效率得到提升。



1. 一种基于自然语言的可视化数据清洗方法,其特征在于,所述方法包括以下步骤:
S1系统初始化成功,通过服务器指定需要清洗的数据源相关连接信息;
S2数据源连接成功后,获取待清洗数据的前N条,并解析其字段类型和格式;
S3通过图形界面确认需要接入的数据字段,进行第一轮筛选,去除无用字段;
S4进入自然语言清洗配置,设置清洗模块,并选择触发数据同步和清洗任务;
S5将清洗模块的清洗规则反向解析为数据清洗脚本,并对数据执行该情形脚本;
S6将清洗后的数据传入分析库,重复执行上述步骤直至所有数据清洗完毕,完成清洗。
2. 根据权利要求1所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法中,指定需要清洗的数据源相关连接信息包括,对于远程数据源提供服务器主机信息、用户名密码和数据库;对于本地数据源提供对应的目录和文件路径。
3. 根据权利要求1所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法中,每添加一个清洗模块,系统根据解析出的源字段信息给出自然语言提示,辅助用户利用自然语言配置清洗规。
4. 根据权利要求1所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法中,对数据集中异常属性的进行识别,其先为每个属性赋予相应的权重,然后统计每个属性字段值的平均值以及标准差,依据此为每一个属性设置一个置信区间,根据属性值是否在置信区间内来判断属性是否异常。
5. 根据权利要求1所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法使用基于属性重要性的约简算法作为逻辑规则对数据属性进行清洗,对决策表 $S = \{U, Q, V, F\}$ 进行可辨识别阵计算,通过将可辨识别阵中的核属性赋给属性约减后得到的属性集,其中 U, Q, V, F 均为数据的属性。
6. 根据权利要求5所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法中,去掉可辨识别阵中核属性约减剩余所有属性组合项;计算各条件属性出现的频率,对所有的属性频率进行降序排列,选取属性频率最高的属性记为 a , $RED = RED \cup \{a\}$,在可变矩阵的所有组合项中删除包含条件属性 a 的组合项;判断可辨矩阵是否为空,若可辨矩阵不为空则继续删除包含条件属性 a 的组合项,若可辨矩阵为空则结束,其中 Red 为最后得到的约减结果。
7. 根据权利要求1所述的基于自然语言的可视化数据清洗方法,其特征在于,所述方法中,对基于属性重要性的约简算法制定的数据逻辑约束规则,利用数学方法获得规则闭集,并自动判断字段值是否违反规则约束,进而判断逻辑规则的正误。
8. 一种基于自然语言的可视化数据清洗系统,其特征在于,包括可视化清洗流程画布、自然语言转换模块、服务器以及存储有执行指令的存储器,当所述服务器执行所述存储器存储的所述执行指令时,所述服务器执行如权利要求1至7中任一项所述的自然语言的可视化数据清洗方法。
9. 根据权利要求8所述的基于自然语言的可视化数据清洗系统,其特征在于,所述可视化清洗流程画布支持拖曳方式绘制数据的清洗逻辑,添加不同的数据清洗组件块,并使用数据流转路径连接组件块。
10. 根据权利要求8所述的基于自然语言的可视化数据清洗系统,其特征在于,所述自然语言转换模块在添加数据清洗组件块时,使用自然语言描述清洗逻辑,对用户输入的自然语言

句进行校验和解析,如果解析成功,会将此语句转换为对应的底层数据过滤查询语句,并传输给底层的数据清洗执行模块;如果解析失败,则会返回异常状态代码给所述可视化清洗流程画布显示对应的异常信息提示用户。

一种基于自然语言的可视化数据清洗系统及方法

技术领域

[0001] 本发明涉及数据处理技术领域,具体涉及一种基于自然语言的可视化数据清洗系统及方法。

背景技术

[0002] 近年来随着大数据技术的发展,为原来海量的日志、上网记录、历史数据等提供了新的分析技术手段,通过分析这些海量数据能够发现很多平时发现不到的有价值信息,要做大数据分析,第一步就是要把分散在各处的数据采集上来,进行清洗,清洗好的数据进行入库。这个过程又叫ETL,涉及extract数据抽取、Transformation数据转换、Load数据装载三个步骤。

[0003] 过去数据清洗的手段针对不同的数据源需要采用不同的清洗工具,不同数据源的清洗均需要编写不同的程序和脚本,这些清洗手段要求用户掌握多种清洗工具的使用方法,具备较高的清洗工具的开发能力;导致数据清洗系统使用门槛高(需要学习针对数据源或清洗工具的相关专业知识),且数据清洗流程的维护成本高。

[0004] 在专利号为CN201710011044.8的发明文件中,公开了一种数据清洗方法及数据清洗装置,该数据清洗方法包括:获取待清洗的原始样本数据;确定对所述原始样本数据进行清洗的至少一种数据筛选机制,并获取用户根据所述原始样本数据对每种所述数据筛选机制设定的筛选值;根据所述至少一种数据筛选机制和用户设定的所述筛选值对所述原始样本数据进行筛选,以对所述原始样本数据进行清洗。本发明的技术方案能够实现对原始样本数据的全面清洗,并且能够降低数据清洗过程对操作人员的依赖,确保数据清洗结果的准确性与稳定性,同时也能够有效缩短数据清洗的时长。

[0005] 在专利号为CN201810143012的发明文件中,公开一种数据清洗方法和数据清洗系统。该数据清洗方法包括:步骤S10:通过图形化界面从异构数据源中选定待清洗的数据源;其中,异构数据源包括文本文件和数据库数据;步骤S11:通过图形化界面编辑数据清洗规则;步骤S12:通过图形化界面执行数据清洗。该数据清洗方法通过图形化界面从异构数据源中选定待清洗的数据源,能够实现对不同数据源的融合清洗,同时,用户通过在图形化界面上简单操作即可实现对数据的清洗,无需掌握数据清洗工具的开发和使用方法,降低了大数据应用服务的技术门槛,提升了用户对大数据服务的体验。

[0006] 综上,传统数据清洗系统多采用脚本编写,配置文件或控件拖曳式方式,实现简单,但学习和维护成本都较高,灵活度偏低。

发明内容

[0007] 针对现有技术的不足,本发明公开了一种基于自然语言的可视化数据清洗系统及方法,用于解决传统数据清洗系统多采用脚本编写,配置文件或控件拖曳式方式,实现简单,但学习和维护成本都较高,灵活度偏低的问题。

[0008] 本发明通过以下技术方案予以实现:

[0009] 第一方面,本发明公开了一种基于自然语言的可视化数据清洗方法,包括以下步骤:

[0010] S1系统初始化成功,通过服务器指定需要清洗的数据源相关连接信息;

[0011] S2数据源连接成功后,获取待清洗数据的前N条,并解析其字段类型和格式;

[0012] S3通过图形界面确认需要接入的数据字段,进行第一轮筛选,去除无用字段;

[0013] S4进入自然语言清洗配置,设置清洗模块,并选择触发数据同步和清洗任务;

[0014] S5将清洗模块的清洗规则反向解析为数据清洗脚本,并对数据执行该情形脚本;

[0015] S6将清洗后的数据传入分析库,重复执行上述步骤直至所有数据清洗完毕,完成清洗。

[0016] 更进一步的,所述方法中,指定需要清洗的数据源相关连接信息包括,对于远程数据源提供服务器主机信息、用户名密码和数据库;对于本地数据源提供对应的目录和文件路径。

[0017] 更进一步的,所述方法中,每添加一个清洗模块,系统根据解析出的源字段信息给出自然语言提示,辅助用户利用自然语言配置清洗规。

[0018] 更进一步的,所述方法中,对数据集中异常属性的进行识别,其先为每个属性赋予相应的权重,然后统计每个属性字段值的平均值以及标准差,依据此为每一个属性设置一个置信区间,根据属性值是否在置信区间内来判断属性是否异常。

[0019] 更进一步的,所述方法使用基于属性重要性的约简算法作为逻辑规则对数据属性进行清洗,对决策表 $S = \{U, Q, V, F\}$ 进行可辨识别阵计算,通过将可辨识别阵中的核属性赋给属性约减后得到的属性集,其中 U, Q, V, F 均为数据的属性。

[0020] 更进一步的,所述方法中,去掉可辨识别阵中核属性约减剩余所有属性组合项;计算各条件属性出现的频率,对所有的属性频率进行降序排列,选取属性频率最高的属性记为 a , $RED = RED \cup \{a\}$,在可变矩阵的所有组合项中删除包含条件属性 a 的组合项;判断可辨矩阵是否为空,若可辨矩阵不为空则继续删除包含条件属性 a 的组合项,若可辨矩阵为空则结束,其中 Red 为最后得到的约减结果。

[0021] 更进一步的,所述方法中,对基于属性重要性的约简算法制定的数据逻辑约束规则,利用数学方法获得规则闭集,并自动判断字段值是否违反规则约束,进而判断逻辑规则的正误。

[0022] 第二方面,本发明公开了一种基于自然语言的可视化数据清洗系统,包括可视化清洗流程画布、自然语言转换模块、服务器以及存储有执行指令的存储器,当所述服务器执行所述存储器存储的所述执行指令时,所述服务器执行第一方面所述的自然语言的可视化数据清洗方法。

[0023] 更进一步的,所述可视化清洗流程画布支持拖曳方式绘制数据的清洗逻辑,添加不同的数据清洗组件块,并使用数据流转路径连接组件块。

[0024] 更进一步的,所述自然语言转换模块在添加数据清洗组件块时,使用自然语言描述清洗逻辑,对用户输入的语句进行校验和解析,如果解析成功,会将此语句转换为对应的底层数据过滤查询语句,并传输给底层的数据清洗执行模块;如果解析失败,则会返回异常状态代码给所述可视化清洗流程画布显示对应的异常信息提示用户。

[0025] 本发明的有益效果为:

[0026] 本发明通过图形化界面以及自然语言引擎的结合,使用户通过在图形化界面上简单操作即可实现对数据的清洗,无需掌握数据清洗工具的开发和使用方法,降低了大数据应用服务的技术门槛,提升了用户对大数据服务的体验,进而解决了传统的数据清洗系统的灵活性及可维护性问题,利用可视化技术及自然语言交互方式使数据清洗工作人员的使用成本得到降低,效率得到提升,具有很强的市场应用前景。

附图说明

[0027] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅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的前提下,还可以根据这些附图获得其他的附图。

[0028] 图1是一种基于自然语言的可视化数据清洗方法的原理步骤图。

具体实施方式

[0029] 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。

[0030] 实施例1

[0031] 本实施例公开如图1所示的一种基于自然语言的可视化数据清洗方法,包括以下步骤:

[0032] S1系统初始化成功,通过服务器指定需要清洗的数据源相关连接信息;

[0033] S2数据源连接成功后,获取待清洗数据的前N条,并解析其字段类型和格式;

[0034] S3通过图形界面确认需要接入的数据字段,进行第一轮筛选,去除无用字段;

[0035] S4进入自然语言清洗配置,设置清洗模块,并选择触发数据同步和清洗任务;

[0036] S5将清洗模块的清洗规则反向解析为数据清洗脚本,并对数据执行该情形脚本;

[0037] S6将清洗后的数据传入分析库,重复执行上述步骤直至所有数据清洗完毕,完成清洗。

[0038] 本实施例中,指定需要清洗的数据源相关连接信息包括,对于远程数据源,提供服务器主机信息,用户名密码,数据库;对于本地数据源,如Excel,日志文件,则要提供对应的目录、文件路径。

[0039] 本实施例中,每添加一个清洗模块,系统根据解析出的源字段信息给出自然语言提示,辅助用户利用自然语言配置清洗规则,如数据过滤:仅导入课程类别不等于专业选修课的数据。

[0040] 本实施例对数据集中异常属性的进行识别,其先为每个属性赋予相应的权重,然后统计每个属性字段值的平均值以及标准差,依据此为每一个属性设置一个置信区间,根据属性值是否在置信区间内来判断属性是否异常。

[0041] 实施例2

[0042] 本实施例使用基于属性重要性的约简算法作为逻辑规则对数据属性进行清洗,对

决策表 $S = \{U, Q, V, F\}$ 进行可辨识别阵计算,通过将可辨识别阵中的核属性赋给属性约减后得到的属性集,其中 U, Q, V, F 均为数据的属性。

[0043] 本实施例中,去掉可辨识别阵中核属性约减剩余所有属性组合项;计算各条件属性出现的频率,对所有的属性频率进行降序排列,选取属性频率最高的属性记为 a , $RED = RED \cup \{a\}$,在可变矩阵的所有组合项中删除包含条件属性 a 的组合项;判断可辨矩阵是否为空,若可辨矩阵不为空则继续删除包含条件属性 a 的组合项,若可辨矩阵为空则结束,其中 Red 为最后得到的约减结果。

[0044] 本实施例对基于属性重要性的约简算法制定的数据逻辑约束规则,利用数学方法获得规则闭集,并自动判断字段值是否违反规则约束,进而判断逻辑规则的正误。

[0045] 本实例从条件属性的分类中,通过条件属性的累积去除,计算和比较相对的正区域,可以判断核心属性和所有重要属性是否被去除。然后,将符合要求的条件属性添加到属性约简集合中,输出最终属性约简集合。

[0046] 本实施例对基于属性约简算法的思想进行改进,对条件属性进行细分,然后通过比较去除条件属性后的相对正区域直接输出属性约简集合。算法改进后该算法主要是从条件属性的分类中,通过条件属性的累积去除,计算和比较相对的正区域,可以判断核心属性和所有重要属性是否被去除。最终将符合要求的条件属性添加到属性约简集合中,输出最终属性约简集合。

[0047] 实施例3

[0048] 本实施例公开一种基于自然语言的可视化数据清洗系统,包括可视化清洗流程画布、自然语言转换模块、服务器以及存储有执行指令的存储器,当所述服务器执行所述存储器存储的所述执行指令时,所述服务器执行自然语言的可视化数据清洗方法。

[0049] 本实施例区别于传统的数据清洗系统,在用户界面上,用户可以直接在可视化清洗流程画布上采用拖曳方式绘制数据的清洗逻辑,即:添加不同的数据清洗组件块,并使用数据流转路径连接组件块。

[0050] 在添加数据清洗组件块时,可使用自然语言描述清洗逻辑,如:输入“创建日期大于2019”,自然语言转换模块会对用户输入的语句进行校验和解析,如果解析成功,会将此语句转换为对应的底层数据过滤查询语句,并传输给底层的数据清洗执行模块;如果解析失败,则会返回异常状态代码给画布模块,画布模块会显示对应的异常信息提示用户。

[0051] 综上,本发明通过图形化界面以及自然语言引擎的结合,使用户通过在图形化界面上简单操作即可实现对数据的清洗,无需掌握数据清洗工具的开发和使用方法,降低了大数据应用服务的技术门槛,提升了用户对大数据服务的体验,进而解决了传统的数据清洗系统的灵活性及可维护性问题,利用可视化技术及自然语言交互方式使数据清洗工作人员的使用成本得到降低,效率得到提升,具有很强的市场应用前景。

[0052] 以上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精神和范围。

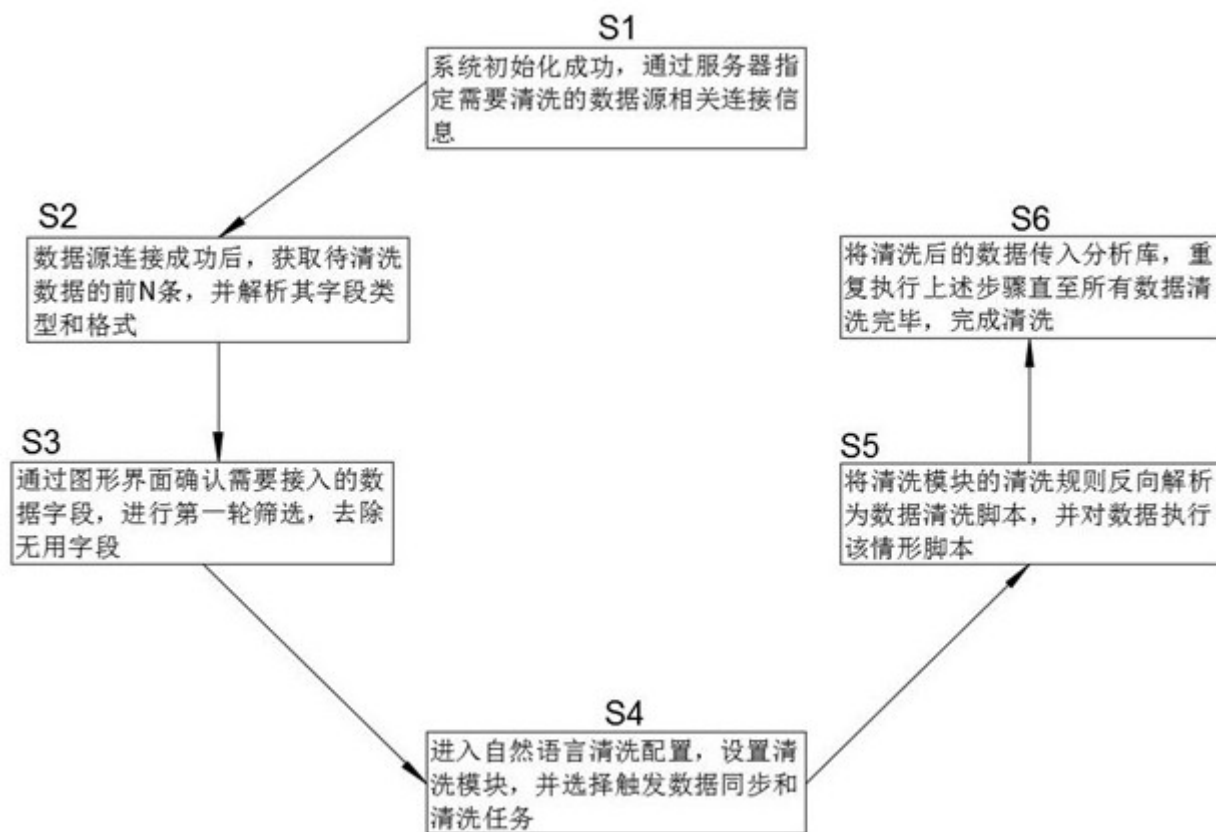


图1