(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 105045807 A
(43)申请公布日 2015.11.11

(21)申请号 201510305440.2

(22)申请日 2015.06.04

(71)申请人 浙江力石科技股份有限公司
地址 311121 浙江省杭州市余杭区文一西路998号海创园科研孵化区18号楼506、507室

(72)发明人 陈海江 吕浩 邵奇可 颜世航

(74)专利代理机构 上海汉声知识产权代理有限公司 31236
代理人 胡晶

(51)Int.Cl.
G06F 17/30(2006.01)

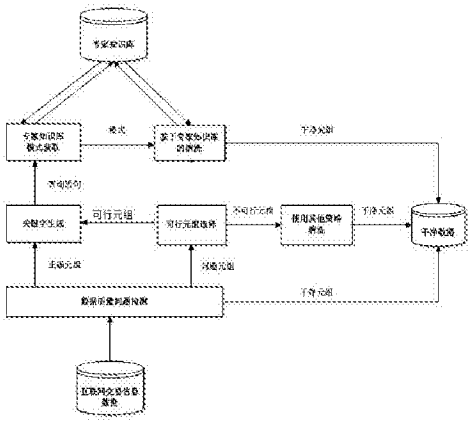
权利要求书2页 说明书9页 附图3页

(54)发明名称

互联网交易信息的数据清洗算法

(57)摘要

本发明提供了一种针对不同互联网交易平台来源的数据进行清洗的方法，首先将数据库中的元组进行分类，将其中确定正确的元组数据进行与专家知识库进行模式交互，以基于知识库检索内容的模糊匹配为工具，获得其相应的模式知识。然后利用找到的模式知识，对数据中存在质量问题且适用的数据进行清洗。同时，针对不同类型海量数据的质量错误也提出了适宜的高效检测方案。而采用BP神经网络方法实现自学习的专家知识库为互联网交易信息数据清洗提供了更加高效、安全的清洗方式。



CN 105045807 A

1.一种互联网交易信息的数据清洗算法,其特征在于,包括:

将待清洗的互联网交易信息数据进行数据质量问题检测获得干净元组、正确元组和问题元组;

对所述干净元组:直接送入干净数据库;

对所述正确元组:生成需要向专家知识库检索关键语句,根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式,所述专家知识库模式包括文本依赖关系语句,对所述专家知识库模式进行数据清洗后送入所述干净数据库;

对所述问题元组:进行可行元组的判断获得适合基于专家知识库模式清洗的可行元组和不适合基于专家知识库模式清洗的不可行元组,

对所述可行元组生成向所述专家知识库检索关键语句后从该专家知识库中查询获得专家知识库模式,再经过数据清洗后送入所述干净数据库,

对所述不可行元组进行其他策略数据清洗后送入所述干净数据库。

2.根据权利要求1所述的一种互联网交易信息的数据清洗算法,其特征在于,所述专家知识库采用BP神经网络算法实现自学习,所述BP神经网络算法具体为:

一个m层的神经网络,对于给定的互联网交易信息样本集X_i(i = 1,2, …,n),设第k层的i个神经元的输入总和表示为U_i^k,输出总和为X_i^k;从第k-1层的第j个神经元到第k层的第i个神经元的权系数为W_{ij},各个神经元的激发函数为f(•),则各个变量的关系可表示为:

$$X_i^k = f(U_i^k)$$

$$U_i^k = \sum_j W_{ij} X_j^{k-1}$$

式中,输入层节点数为n,隐藏层节点数为h,输出层节点数为o,分别确定输入层与隐藏层、隐藏层与输出层间的链接权值矩阵为W_h、W_o以及阀值b_h、b_o。

3.根据权利要求2所述的一种互联网交易信息的数据清洗算法,其特征在于,期望输出和实际输出之差的平方和为所述专家知识库的误差函数,所述专家知识库的误差函数为:

$$e = \frac{1}{2}\sum_i (X_i^m - Y_i)^2$$

Y_i是输出单元的期望值,第m层是输出层,X_i^m是实际输出;BP算法采用非线性规划中的最速下降方法,按误差函数e的负梯度方向修改权系数。

4.根据权利要求2所述的一种互联网交易信息的数据清洗算法,其特征在于,一个互联网交易信息样本中所有向量之间的差异采用机器学习中的马氏距离衡量;对于1个向量X_1～X_1,确立最合理向量X_k作为BP神经网络标准输出展开样本训练;一个所述样本中包含的向量的协方差矩阵记为S,向量X_i与X_j之间的马氏距离为:

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

$$\min\{\sum_{i=0}^{l} D(X_k, X_i)\}$$

所述协方差矩阵S中每个元素是各个矢量元素之间的协方差Cov(X,Y)，$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$，其中E为一个所述样本中包含的向量的数学期望。

5.根据权利要求1所述的一种互联网交易信息的数据清洗算法，其特征在于，所述问题元组包括丢失值，和/或错误值，和/或冲突值；

所述丢失值为数据属性存在空缺的值；对于丢失值的检测方法是：对于待清洗的互联网交易信息数据$D(T_1,T_2,\ldots,T_n)$中的每个元组$T(A_1,A_2,\ldots,A_m)$的属性A进行检测，若存在缺失的属性值则为包含丢失至的问题元组；

所述错误值为数据存在的属性被认定为是错误的值；对于错误值的检测方法是：对于待清洗的互联网交易信息数据$D(T_1,T_2,\ldots,T_n)$中的每个元组$T(A_1,A_2,\ldots,A_m)$进行基于条件依赖函数的条件依赖检测，若该数据的属性不满足所述条件依赖函数则该元组为包含错误值的问题元组；

所述冲突值为一个数据的属性值出现多个对应值；对于冲突值的检测方法是：首先对于待清洗的互联网交易信息数据进行元组匹配找出可能冲突的元组，然后对所述可能冲突的元组进行聚类得到包含冲突值的问题元组。

6.根据权利要求5所述的一种互联网交易信息的数据清洗算法，其特征在于，所述元组匹配具体为：

S1：对于待清洗的互联网交易信息数据中的元组进行相似性两两匹配，若元组对的相似程度达到预设的相似阈值则该元组对指向同一实体，将指向同一实体的元组作为一个组群；

S2：为所述组群创建与元组属性对应的Bloom Filter数组，检查所述组群中的各元组的逐项属性是否在对应的Bloom Filter数组中，在同一Bloom Filter数组的元组则累加该元组的权值，

S3，所述元组权值超过预设的上限则提取出来作为所述可能冲突的元组。

7.根据权利要求1所述的一种互联网交易信息的数据清洗算法，其特征在于，根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式的过程具体为：

将所述关键语句发送给专家知识库的搜索引擎，获取并解析专家知识库反馈的查询结果，采用最优模糊匹配方法进行模式挖掘获得所述专家知识库模式。

8.根据权利要求1所述的一种互联网交易信息的数据清洗算法，其特征在于，所述不可行元组包括：

数据属性数量少于预设属性数量下限值或属性之间的关联度弱于预设关联度下限值的元组；

属性存在质量问题并且无法通过专家知识库模式对应修复的元组；

不同属性的数据同时出现错误并且无法修复的元组。

9.根据权利要求8所述的一种互联网交易信息的数据清洗算法，其特征在于，所述其他策略数据清洗包括：

若数据集中提供一些约束条件的就使用约束函数方式直接清洗数据；

若对同一个实体的描述出现多种情况时，则采用真值发现算法选择进行清洗，即通过对数据源准确性的学习，给予不同的权值来使得各个描述不等价来实现真值发现。

互联网交易信息的数据清洗算法

技术领域

[0001] 本发明涉及计算机应用领域，具体地，涉及一种互联网交易信息的数据清洗算法。

背景技术

[0002] 近年来我国互联网交易持续保持高速发展，近5年来平均增速达到80％。2013年电子商务总交易额超过10万亿元人民币，网络零售市场规模已经超过美国成为世界最大的网络零售市场。随着电子商务的发展，也出现了一些市场自身难以解决的问题，包括产品虚假宣传、假货泛滥、网络诈骗及钓鱼网站很行、物流配送服务不规范、退货难及逆向物流不畅通以及网民个人信息泄露等问题。主要是由于不同的电商平台的信用评价体系的规范各不相同；同时电商信息系统中的数据越来越多，甚至达到了TB、PB以上的海量数据级别导致海量数据聚集之后由于内容过时、输入错误、重复输入、属性值冲突等严重影响着数据质量，进而导致无法保证系统中数据的质量能满足监管系统的需求。

[0003] 为了克服由于数据质量而引发的问题，采取数据处理的技术是非常必要的。目前很多通过处理数据来获得更高质量的数据的方法已经被提出，在这些技术当中，数据清洗技术至关重要。

[0004] 针对数据清洗的处理方法主要包括以下几种：

[0005] 1. 通过关系数据中键与键之间的函数依赖进行数据清洗是比较直接的方法，但是对于和互联网那个交易信息这种海量数据的规则挖掘并不充分。

[0006] 2. 基于条件函数依赖的数据方法采用函数依赖作为基础并且增加了语义上的约束条件，这样可以有效的清洗存在函数依赖的关系的数据元组，但是互联网交易信息来自不同的电商平台，很多数据的函数依赖并不明确，同时一些数据在进行清洗之前是无法获得函数关系的。

[0007] 3. 采用人为参与的数据清洗，即在数据清洗的过程中，若系统遭遇无法处理的情况时，需通过人的反馈操作进行下一步的清洗步骤。这种方法的优点是由于人的参与准确性会大大提高，但处理的时间消耗比较大；同时不同的人对于依赖关系的规则判断标准并不能保证完全一致，主观依赖性过强。

[0008] 4. 采用机器学习的反馈方式，即用机器学习的方法替代人的反馈过程，在清洗过程之前先让机器学习正确的清洗操作，然后在清洗过程中不断积累学习，这样可以提成算法的时间效率，但是精确度有所下降，并且学习过程会增加系统的额外开销，同时清洗过程中对数据之间的依赖关系要求依然比较高。

[0009] 综上所述，当前的数据清洗方法对于互联网交易信息的处理的需求存在着一定的局限性。

发明内容

[0010] 针对现有技术中的缺陷，本发明的目的是提供一种互联网交易信息的数据清洗算法。

[0011] 根据本发明提供的一种互联网交易信息的数据清洗算法，包括：

[0012] 将待清洗的互联网交易信息数据进行数据质量问题检测获得干净元组、正确元组和问题元组；

[0013] 对所述干净元组：直接送入干净数据库；

[0014] 对所述正确元组：生成需要向专家知识库检索关键语句，根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式，所述专家知识库模式包括文本依赖关系语句，对所述专家知识库模式进行数据清洗后送入所述干净数据库；

[0015] 对所述问题元组：进行可行元组的判断获得适合基于专家知识库模式清洗的可行元组和不适合基于专家知识库模式清洗的不可行元组，

[0016] 对所述可行元组生成向所述专家知识库检索关键语句后从该专家知识库中查询获得专家知识库模式，再经过数据清洗后送入所述干净数据库，

[0017] 对所述不可行元组进行其他策略数据清洗后送入所述干净数据库。

[0018] 作为一种优化方案，所述专家知识库采用BP神经网络算法实现自学习，所述BP神经网络算法具体为：

[0019] 一个m层的神经网络，对于给定的互联网交易信息样本集X_i(i = 1,2, …,n)，设第k层的i个神经元的输入总和表示为U_i^k，输出总和为X_i^k；从第k-1层的第j个神经元到第k层的第i个神经元的权系数为W_{ij}，各个神经元的激发函数为f(•)，则各个变量的关系可表示为：

[0020] $$X_i^k = f(U_j^k)$$

[0021] $$U_i^k = \sum_j W_{ij} X_j^{k-1}$$

[0022] 式中，输入层节点数为n，隐藏层节点数为h，输出层节点数为o，分别确定输入层与隐藏层、隐藏层与输出层间的链接权值矩阵为W_h、W_o以及阀值b_h、b_o；

[0023] 作为一种优化方案，期望输出和实际输出之差的平方和为所述专家知识库的误差函数，所述专家知识库的误差函数为：

[0024] $$e = \frac{1}{2}\sum_i (X_i^m - Y_i)^2$$

[0025] Y_i是输出单元的期望值，第m层是输出层，X_i^m是实际输出；BP算法采用非线性规划中的最速下降方法，按误差函数e的负梯度方向修改权系数。

[0026] 作为一种优化方案，一个互联网交易信息样本中所有向量之间的差异采用机器学习中的马氏距离衡量；对于l个向量X_1～X_l，确立最合理向量X_k作为BP神经网络标准输出展开样本训练；一个所述样本中包含的向量的协方差矩阵记为S，向量X_i与X_j之间的马氏距离为：

[0027] $$D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

[0028] $$\min\{\sum_{i=0}^{l} D(X_k, X_i)\}$$

[0029] 所述协方差矩阵S中每个元素是各个矢量元素之间的协方差Cov(X,Y)，

$Cov(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$，其中 E 为一个所述样本中包含的向量的数学期望。

[0030] 作为一种优化方案，所述问题元组包括丢失值，和 / 或错误值，和 / 或冲突值 ；

[0031] 所述丢失值为数据属性存在空缺的值 ；对于丢失值的检测方法是 ：对于待清洗的互联网交易信息数据 $D(T_1, T_2, \ldots, T_n)$ 中的每个元组 $T(A_1, A_2, \ldots, A_m)$ 的属性 A 进行检测，若存在缺失的属性值则为包含丢失至的问题元组 ；

[0032] 所述错误值为数据存在的属性被认定为是错误的值 ；对于错误值的检测方法是 ：对于待清洗的互联网交易信息数据 $D(T_1, T_2, \ldots, T_n)$ 中的每个元组 $T(A_1, A_2, \ldots, A_m)$ 进行基于条件依赖函数的条件依赖检测，若该数据的属性不满足所述条件依赖函数则该元组为包含错误值的问题元组 ；

[0033] 所述冲突值为一个数据的属性值出现多个对应值 ；对于冲突值的检测方法是 ：首先对于待清洗的互联网交易信息数据进行元组匹配找出可能冲突的元组对，然后对所述可能冲突的元组对进行聚类得到包含冲突值的问题元组。

[0034] 作为一种优化方案，所述元组匹配具体为 ：

[0035] S1 ：对于待清洗的互联网交易信息数据中的元组进行相似性两两匹配，若元组对的相似程度达到预设的相似阈值则该元组对指向同一实体，将指向同一实体的元组作为一个组群 ；

[0036] S2 ：为所述组群创建与元组属性对应的 Bloom Filter 数组，检查所述组群中的各元组的逐项属性是否在对应的 Bloom Filter 数组中，在同一 Bloom Filter 数组的元组则累加该元组的权值，

[0037] S3，所述元组权值超过预设的上限则提取出来作为所述可能冲突的元组。

[0038] 作为一种优化方案，根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式的过程具体为 ：

[0039] 将所述关键语句发送给专家知识库的搜索引擎，获取并解析专家知识库反馈的查询结果，采用最优模糊匹配方法进行模式挖掘获得所述专家知识库模式。

[0040] 作为一种优化方案，所述不可行元组包括 ：

[0041] 数据属性数量少于预设属性数量下限值或属性之间的关联度弱于预设关联度下限值的元组 ；

[0042] 属性存在质量问题并且无法通过专家知识库模式对应修复的元组 ；

[0043] 不同属性的数据同时出现错误并且无法修复的元组。

[0044] 作为一种优化方案，所述其他策略数据清洗包括 ：

[0045] 若数据集中提供一些约束条件的就使用约束函数方式直接清洗数据 ；

[0046] 若对同一个实体的描述出现多种情况时，选出这些描述中最准确的一个，则采用使用真值发现算法选择进行清洗，即通过对数据源准确性的学习，给予不同的权值来使得各个描述不等价来实现真值发现的。

[0047] 与现有技术相比，本发明具有如下的有益效果 ：

[0048] 本发明提出了一种针对不同互联网交易平台来源的数据进行清洗的方法，首先将数据库中的元组进行分类，将其中确定正确的元组数据进行与专家知识库进行模式交互，以基于知识库检索内容的模糊匹配为工具，获得其相应的模式知识。然后利用找到的模式

知识，对数据中存在质量问题且适用的数据进行清洗。同时，针对不同类型海量数据的质量错误也提出了适宜的高效检测方案。而采用 BP 神经网络方法实现自学习的专家知识库为互联网交易信息数据清洗提供了更加高效、安全的清洗方式。

附图说明

[0049] 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简单的介绍，显而易见，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。附图中：

[0050] 图 1 是可选实施例中的一种互联网交易信息的数据清洗算法流程；

[0051] 图 2 是采用 Bloom Filter 的元组匹配流程；

[0052] 图 3 是获取专家知识库模式的流程；

[0053] 图 4 是采用模式挖掘的基于专家知识库模式的清洗流程；

[0054] 图 5 是 BP 神经网络方法流程。

具体实施方式

[0055] 下文结合附图以具体实施例的方式对本发明进行详细说明。以下实施例将有助于本领域的技术人员进一步理解本发明，但不以任何形式限制本发明。应当指出的是，还可以使用其他的实施例，或者对本文列举的实施例进行结构和功能上的修改，而不会脱离本发明的范围和实质。

[0056] 在本发明提供的一种互联网交易信息的数据清洗算法的实施例中，如图 1 所示，将待清洗的互联网交易信息数据进行数据质量问题检测获得干净元组、正确元组和问题元组；

[0057] 对所述干净元组：直接送入干净数据库；

[0058] 对所述正确元组：生成需要向专家知识库检索关键语句，根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式，所述专家知识库模式包括文本依赖关系语句，对所述专家知识库模式进行数据清洗后送入所述干净数据库；

[0059] 对所述问题元组：进行可行元组的判断获得适合基于专家知识库模式清洗的可行元组和不适合基于专家知识库模式清洗的不可行元组，

[0060] 对所述可行元组生成向所述专家知识库检索关键语句后从该专家知识库中查询获得专家知识库模式，再经过数据清洗后送入所述干净数据库，

[0061] 对所述不可行元组进行其他策略数据清洗后送入所述干净数据库。

[0062] 本实施例中的互联网交易信息包括交易主体信息和交易行为信息。数据的质量是指在传输过程中出现数据丢失或乱码数据，或在数据抓取过程中抓取顺序错误或头文件出错等导致的数据质量下降。由于数据质量会严重影响监管系统的分析与决策，为了克服因数据质量引发的问题，本发明提出一种对互联网交易信息数据的清洗算法，算法内容主要包括数据质量问题的检测、专家知识库模式的交互以及清洗部分。

[0063] 为了实现上述算法模块采用的具体内容如图 1 所示的一种实施例，所述问题元组包括丢失值，和 / 或错误值，和 / 或冲突值。数据质量问题监测：监测互联网交易信息的数

据元组是否存在问题，数据中存在质量问题的情况主要包括：丢失值、错误值和冲突值。

[0064] 所述丢失值为数据属性存在空缺的值；对于丢失值的检测方法是：对于待清洗的互联网交易信息数据 $D(T_1, T_2, \ldots, T_n)$ 中的每个元组 $T(A_1, A_2, \ldots, A_m)$ 的属性 A 进行检测，若存在缺失的属性值则为包含丢失至的问题元组。丢失值就是从互联网获取的交易信息的数据属性存在空缺的值，这种错误的发生经常是由数据集成导致的，例如在将两个数据源进行集成时，两个数据源中的属性个数不同，就会导致部分元组的属性值空缺。

[0065] 所述错误值为数据存在的属性被认定为是错误的值；对于错误值的检测方法是：对于待清洗的互联网交易信息数据 $D(T_1, T_2, \ldots, T_n)$ 中的每个元组 $T(A_1, A_2, \ldots, A_m)$ 进行基于条件依赖函数的条件依赖检测，若该数据的属性不满足所述条件依赖函数则该元组为包含错误值的问题元组。错误值是指数据的属性是存在的，但是被认定为是错误的，这种错误经常是由于数据抓取或输入错误引起的。针对海量的互联网交易数据，采用条件依赖函数来进行数据错误值的判断。在数据错误值检测开始之前，条件依赖函数的知识 m，即标准集合 m，是已知的，然后对一个数据集 $D(T_1, T_2, \ldots, T_n)$ 中的每个元组 $T(A_1, A_2, \ldots, A_m)$ 进行条件依赖检测。例如该数据的属性对应的是交易行为，但是该数据却是交易主体的内容，如卖家店铺名称等，则无法满足该标准集合对应的条件依赖函数，该数据为错误值，该数据所在元组为问题元组。

[0066] 所述冲突值为一个数据的属性值出现多个对应值；对于冲突值的检测方法是：首先对于待清洗的互联网交易信息数据进行元组匹配找出可能冲突的元组对，然后对所述可能冲突的元组对进行聚类得到包含冲突值的问题元组。

[0067] 冲突值是数据的属性值存在多个可能的值，但是只有一个值是正确的，这种错误是通过实体识别的方法进行判断，需要选择或者制造一个真实的属性值来消除冲突值。为了针对海量数据来实现这个过程，首先进行元组匹配将可能冲突的元组对找出，然后将找出的元组进行聚类最终得到有冲突的元组集合。

[0068] (1) 所述元组匹配具体为：

[0069] S1：对于待清洗的互联网交易信息数据中的元组进行相似性两两匹配，若元组对的相似程度达到预设的相似阈值则该元组对指向同一实体，将指向同一实体的元组作为一个组群；

[0070] S2：为所述组群创建与元组属性对应的 Bloom Filter 数组，检查所述组群中的各元组的逐项属性是否在对应的 Bloom Filter 数组中，在同一 Bloom Filter 数组的元组则累加该元组的权值，

[0071] S3，所述元组权值超过预设的上限则提取出来作为所述可能冲突的元组。

[0072] 作为一种匹配过程的实施例，首先设定判断的上限和下限，如果两个元组的相似性达到上限就认为这两个元组指向同一实体；如果两个元组的相似性在第一次比较就低于下限值，就认为是指向不同的实体。然后为每个组群创建 m 个 Bloom Filter 数组，m 是第一类属性中属性的个数，对每个数组中的 N 个元组把属性插入到相应的 Bloom Filter 数组中。然后检查各个元组的第一类属性是否在相应的 Bloom Filter 重，如果返回结果是 Yes 那就说明这些属性在相应的组群中是冗余的；同时如果两个元组具有相似属性那么就累加权值，当权值超过上限就把他们从数据源中删除并且输出，如果没有达到上限就继续比较下一个属性。如果第一类的所有属性都经过比较之后，权值还没有达到下限则认为两组数

据是不同的，就不进行余下属性的比较。当权值在上限和下限之间，就按照时序的方式继续比较余下的属性来判断是否匹配。匹配的过程采用 Bloom Filter 的结构作为一个链表数组的形式，当一个属性被哈希成一个数字的时候，只需要增加一个节点到相应的链表中，通过这种方式比较过程将仅限于在 Bloom Filter 比较重返回 Yes 的元组，这样将大幅度提高比较的效率。采用 Bloom Filter 数组的比较过程如图 2 所示。

[0073] (2) 元组聚类的操作是基于匹配的结果进行的，对于两两相似的元组可以将所有重复的元组放在一起。把每个元组看成一个点，而之前已经匹配好的元组之间用一条线来连接，这样对元组的聚类就等价于对所有的点进行聚类划分，找到图上内部联系紧密的点，放入一个社区之内，进而完成元组的聚类操作。

[0074] 所述专家知识库采用 BP 神经网络算法实现自学习。专家知识库采用 BP 神经网络方法实现自学习的过程，然后根据训练好的知识库对互联网电子商务交易数据进行问题质量检测和清洗。

[0075] 在专家知识库学习的过程中采用 BP 神经网络方法，先获取互联网电子商务交易的初始数据，再输入检测指标数和输出检测指标数，计算各层输出，计算各层输出，计算实际输出和目标输出的误差 e，计算局部梯度，修正输出层权值，修正隐含层权值后判断训练集中是否还有未训练的样本，若还有未训练样本则返回至输入检测指标数和输出检测指标数的步骤继续执行，若样本全部训练结束则再判断误差是否满足条件，或迭代是否满足条件，任一满足则结束学习，都不满足则将 e 重置为 0 后返回至输入检测指标数和输出检测指标数的步骤重新执行。

[0076] BP 神经网络算法如图 5 所示，确立输入层节点数为 n，隐藏层节点数为 h，输出层节点数为 o，分别确定输入层与隐藏层、隐藏层与输出层间的链接权值矩阵为 W_h、W_o 以及阀值 b_h、b_o。一个 m 层的神经网络，对于给定的互联网交易信息样本集 X_i(i = 1,2, …,n)，设第 k 层的 i 个神经元的输入总和表示为 U_i^k，输出总和为 X_i^k；从第 k-1 层的第 j 个神经元到第 k 层的第 i 个神经元的权系数为 W_{ij}，各个神经元的激发函数为 f(•)，则各个变量的关系可表示为式 (1)、(2)：

[0077] $$X_i^k = f(U_i^k) \quad (1)$$

[0078] $$U_i^k = \sum_j W_{ij} X_j^{k-1} \quad (2)$$

[0079] 定义期望输出和实际输出之差的平方和作为所述专家知识库的误差函数如式 (3) 所示：

[0080] $$e = \frac{1}{2}\sum_i (X_i^m - Y_i)^2 \quad (3)$$

[0081] Y_i 是输出单元的期望值，第 m 层是输出层，X_i^m 是实际输出。BP 算法采用非线性规划中的最速下降方法，按误差函数 e 的负梯度方向修改权系数。

[0082] 本文采用机器学习中的马氏距离衡量某一样本中所有向量之间的差异，对于 1 个向量 X_1～X_1，确立最合理向量 X_k 作为 BP 神经网络标准输出展开样本训练。所述最合理向量 X_k 是指到其他各个向量距离和最短的向量。协方差矩阵记为 S，向量 X_i 与 X_j 之间的马氏距离定义如式 (4) 所示：

[0083] $$D(X_i,X_j)=\sqrt{(X_i-X_j)^T S^{-1}(X_i-X_j)} \quad (4)$$

[0084] $$\min\{\sum_{i=0}^{L} D(X_k,X_l)\} \quad (5)$$

[0085] 协方差矩阵中每个元素是各个矢量元素之间的协方差Cov(X,Y)，$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$，其中E为互联网交易信息的数学期望。

[0086] 针对数据清洗专家知识库的BP神经网络训练过程，专家知识库中数据元组之间的马氏距离的初始设定值，根据监测的互联网交易数据的内容分成交易主体和交易行为两大类的应用场景，该初始设定值在0.7效果最佳。

[0087] 专家知识库的模式交互是对经过质量检测的数据分成正确元组和问题元组两大类，均在专家知识库的指导下完成数据的清洗然后放入干净数据库中，为后续分析处理提供高质量的数据信息。

[0088] 对于正确元组，首先经过关键字生成模块生成需要向专家知识库检索的Query语句，然后获得文本依赖的关系语句，进而完成基于专家知识库的数据清洗；对于不正确的元组首先要通过可行元组选择的判断，可行元组进行基于专家知识库的数据清洗，对于不可行元组则采用其他策略进行清洗。

[0089] 作为一种实施例，根据所述关键语句在所述专家知识库中进行查询获得专家知识库模式的过程具体为：

[0090] 将所述关键语句发送给专家知识库的搜索引擎，获取并解析专家知识库反馈的查询结果，采用最优模糊匹配方法进行模式挖掘获得所述专家知识库模式。

[0091] 获取专家知识库模式的过程如图3所示，经过数据质量检测模块处理的正确元组将组装好的Query语句发送给专家知识库的搜索引擎，然后对专家知识库反馈的Results进行解析并把结果存储起来，然后采用最优模糊匹配方法进行模式挖掘的处理，最后进行基于专家知识库的数据清洗。对于给定属性值Attr1和模式知识Pattern，将两者组合内容进行搜索，记录搜索结果中包含当前索要清洗的属性值的信息并记录出现的次数，通过排序、模糊匹配和包含比较，找到最优模糊匹配断句Sentence，接着利用给定的属性值和模式知识进行拆分就可以获得推荐值，清洗过程见图4。

[0092] 对于问题元组首先要进行可行元组的判断，选择出适用于基于专库知识库模式清洗的数据元组，然后进行数据清洗。所述不可行元组包括：

[0093] 1、数据属性数量少于预设属性数量下限值或属性之间的关联度弱于预设关联度下限值的元组；

[0094] 2、属性存在质量问题并且无法通过专家知识库模式对应修复的元组；

[0095] 3、不同属性的数据同时出现错误并且无法修复的元组。以下三种情况判断为不可行元组：数据属性数量较少或者属性之间的关系过弱，无法通过专家知识库检索获取属性之间的依赖关系；存在质量问题的属性并且无法通过专家知识库模式对应修复的数据；以及当数据元组中不同属性的数据同时出现错误并且无法修复的。针对非可行元组需要采用其他清洗手段包括：如果数据集中提供一些约束条件的就使用约束函数方式直接清洗数据；当对同一个实体的描述出现多种情况是，需要选出这些描述中最准确的一个，则采用使用真值发现算法选择进行清洗，也就是通过对数据源准确性的学习，给予不同的权值来使

得各个描述不等价来实现真值发现的。

[0096] 真值发现是找到冲突值中对真实实体描述最为准确的那个。使用基于数据源的方法，其中包含两个重要的度量分别为：数据源的准确性和数据源的支配关系。

[0097] 数据源的准确性即此数据源中对所有数据的描述中，准确的比率高低。基于数据源的准确性的数据清洗技术需要不断的学习出各个数据源的准确性，若数据源 D 的准确性 A(D) 是最高的，那么将使用 D 对数据进行清洗。

[0098] 数据源的支配关系即数据源的传递关系，若数据源 D1 中元组 T 的描述与 D2 相同，那么就说 D2 支配 D1。本实施例中数据源为待清洗的互联网交易信息数据。

[0099] 所述其他策略数据清洗包括：

[0100] 若数据集中提供一些约束条件的就使用约束函数方式直接清洗数据；

[0101] 若对同一个实体的描述出现多种情况时，选出这些描述中最准确的一个，则采用真值发现算法选择进行清洗，即通过对数据源准确性的学习，给予不同的权值来使得各个描述不等价来实现真值发现。

[0102] 所述同一个实体的描述出现多种情况是指对同一个交易信息的表达出现多个数据元组的时候，需要通过对数据源准确性的学习，给予不同的权值来使得各个描述不等价来实现真值发现。

[0103] 作为一种实施例，京东商城的一个交易内容，“订单编号：8971959437”+“交易时间：2105 年 03 月 10 日 15：00”+“联想电脑官方旗舰店”+“联想 (Lenovo)G40-70MA 14.0 英寸笔记本电脑”+“i5-4258U 4G 500G 2G 独显 GT820M 显卡 DVD 刻录 Win8”+“金属黑”+“孙祥”+“浙江省杭州市西湖区”+“浙江工业大学”+“310023”+“151XXXXXXXX”+“货到付款”。

[0104] 根据获取的互联网电子商务交易信息的来源平台不同，首先根据已经训练好的知识库里的模式知识部分对数据信息进行关键字抽取。

[0105] 对于商品店铺和商品属性信息进行拼加，生成的查询语句就是类似 Query = {“联想电脑官方旗舰店 + 联想 (Lenovo)G40-70MA 14.0 英寸笔记本电脑”，“联想 (Lenovo)G40-70MA 14.0 英寸笔记本电脑 + 金属黑”}，如果有数据缺失产生的空缺值或由于传输过程中产生的错误值就根据 BP 专家知识库中的相应信息进行修复和清洗。

[0106] 对于订单编号丢失或错误，根据对应的电商平台订单命名规则按照订单交易时间可以进行修复，同时订单编号和交易时间也可以相互判断是否是错误值或冲突值。对应的电商平台订单命名规则是条件依赖函数的一种。

[0107] 本发明提出了一种针对不同互联网交易平台来源的数据进行清洗的框架，首先将数据库中的元组进行分类，将其中确定正确的元组数据进行与专家知识库进行模式交互，以基于知识库检索内容的模糊匹配为工具，获得其相应的模式知识。然后利用找到的模式知识，对数据中存在质量问题且适用的数据进行清洗。同时，针对不同类型海量数据的质量错误也提出了适宜的高效检测方案。

[0108] 目前针对互联网交易信息的清洗还停留在针对单一平台的处理，对于能处理不同平台来源的数据还处于摸索阶段，本发明通过专家知识库的学习可以很好的解决数据来源不一致带来的数据异构问题。

[0109] 以上所述仅为本发明的较佳实施例，本领域技术人员知悉，在不脱离本发明的精

神和范围的情况下，可以对这些特征和实施例进行各种改变或等同替换。另外，在本发明的教导下，可以对这些特征和实施例进行修改以适应具体的情况及材料而不会脱离本发明的精神和范围。因此，本发明不受此处所公开的具体实施例的限制，所有落入本申请的权利要求范围内的实施例都属于本发明的保护范围。

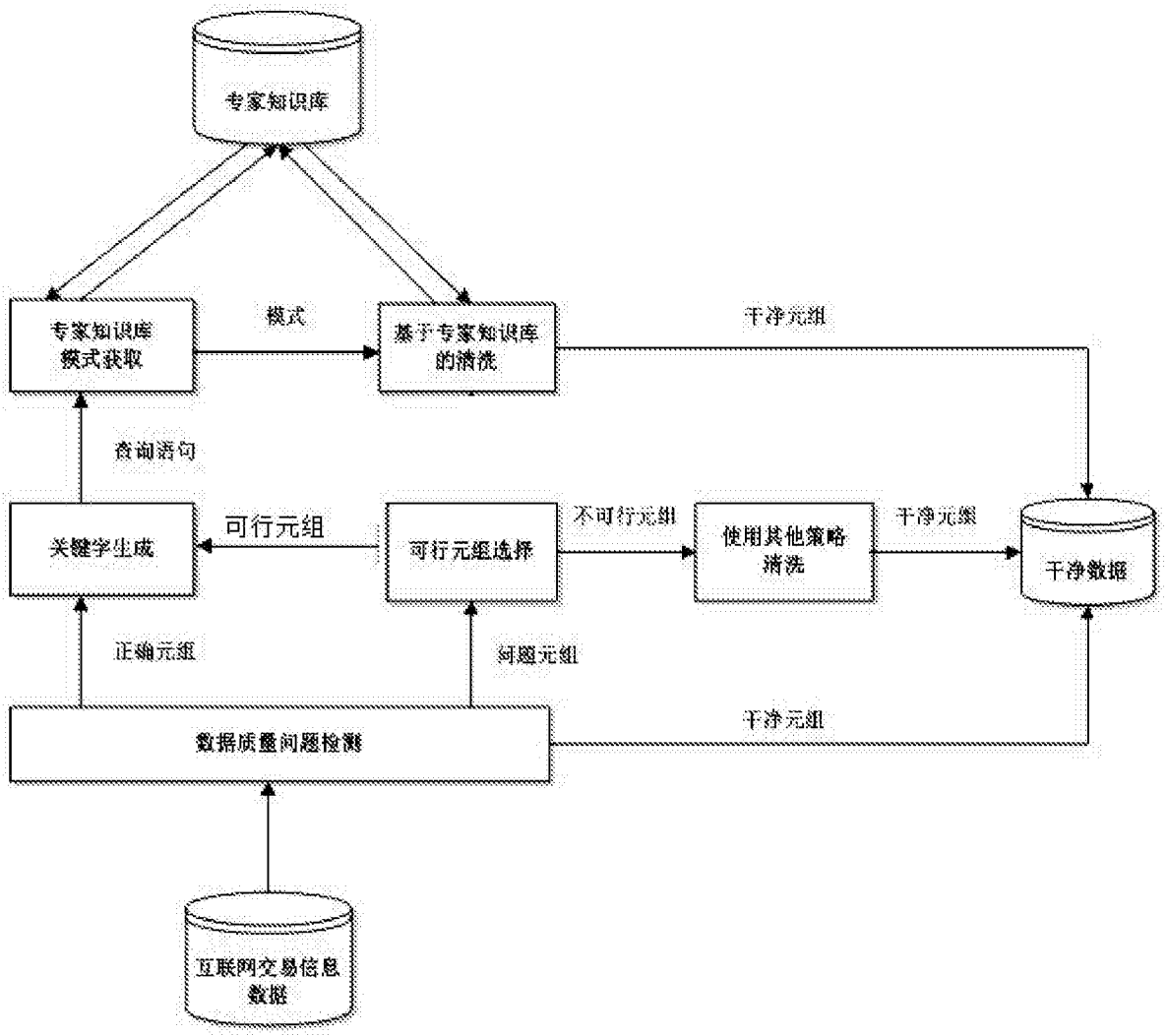

专家知识库
专家知识库模式获取
模式
基于专家知识库的清洗
干净元组
查询语句
关键字生成
可行元组
可行元组选择
不可行元组
使用其他策略清洗
干净元组
干净数据
正确元组
问题元组
数据质量问题检测
干净元组
互联网交易信息数据

图 1

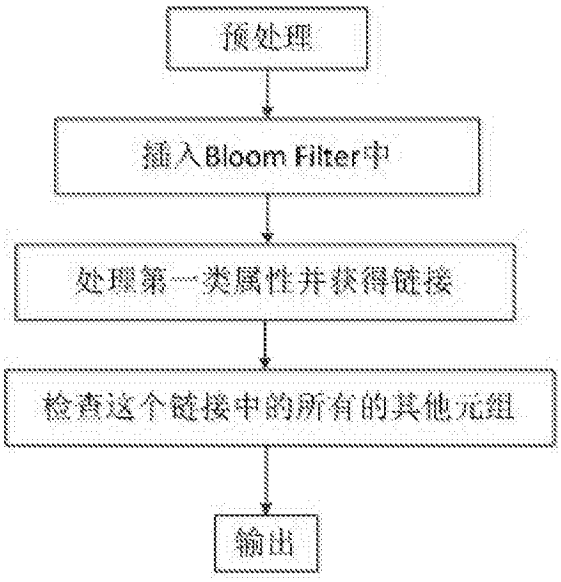

预处理
插入Bloom Filter中
处理第一类属性并获得链接
检查这个链接中的所有的其他元组
输出

图 2

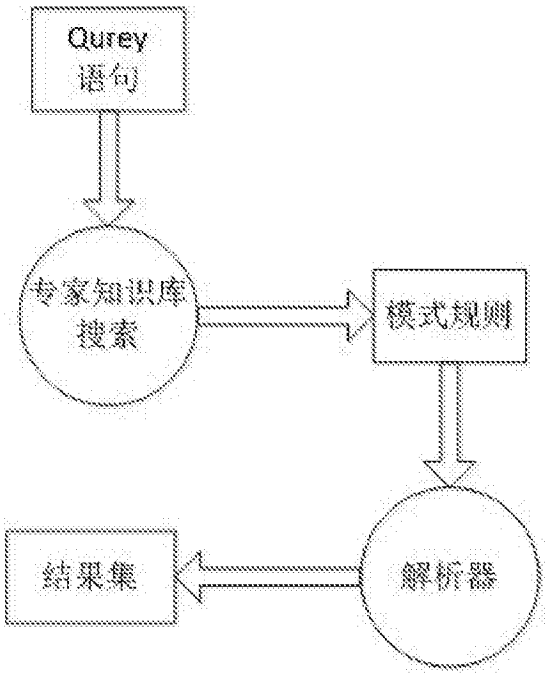

Qurey
语句
专家知识库
搜索
模式规则
结果集
解析器

图 3

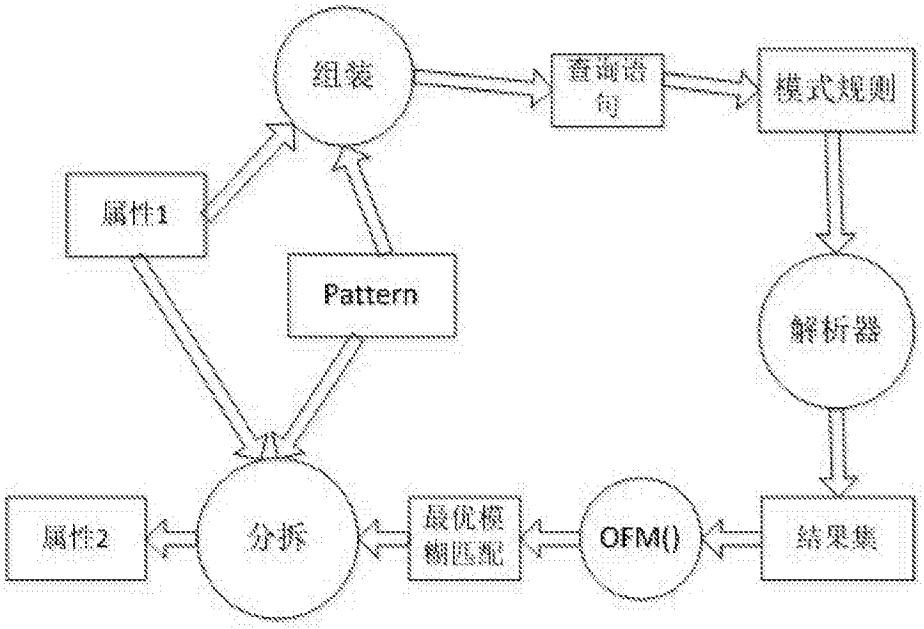

组装
查询语句
模式规则
属性1
Pattern
解析器
属性2
分拆
最优模
糊匹配
OFM()
结果集

图 4

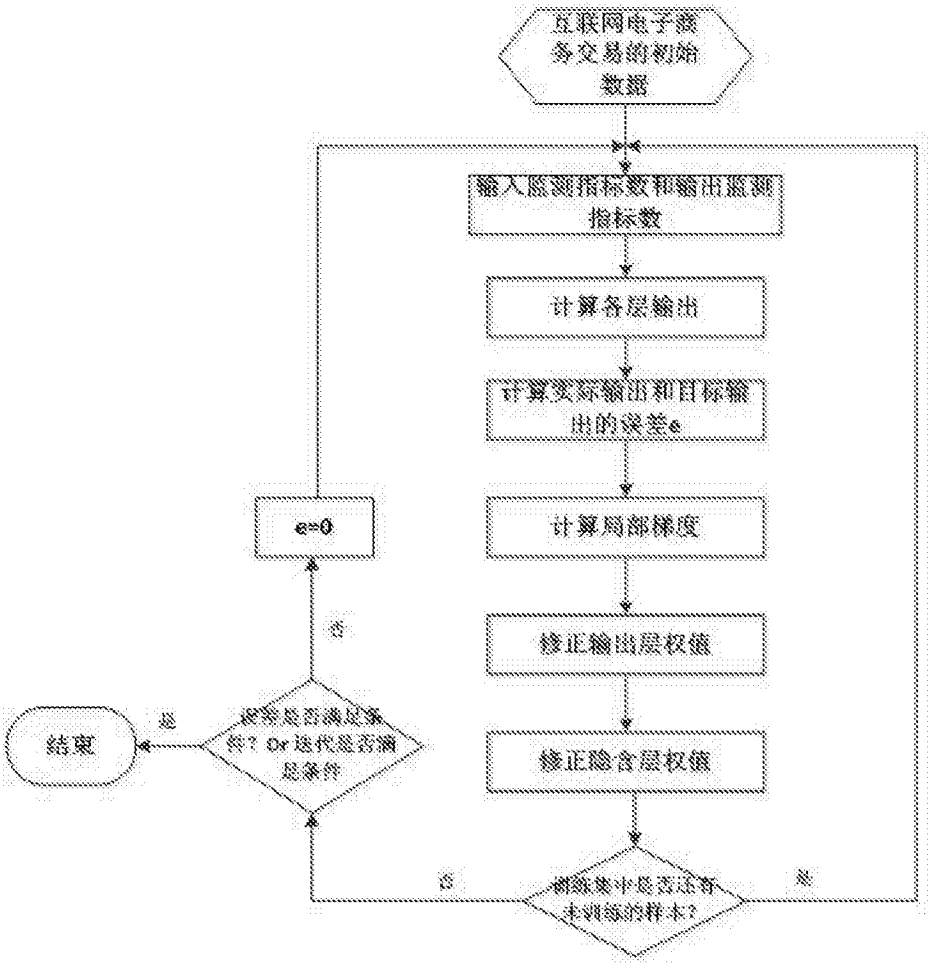

互联网电子商务交易的初始数据
输入监测指标数和输出监测指标数
计算各层输出
计算实际输出和目标输出的误差e
计算局部梯度
修正输出层权值
修正隐含层权值
样本集中是否还有未训练的样本?
否
是
e=0
否
网络是否满足条件? Or 迭代是否满足条件
是
结束

图 5