(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104850624 B
(45)授权公告日 2018.06.22

(21)申请号 201510259365.0

(22)申请日 2015.05.20

(65)同一申请的已公布的文献号
申请公布号 CN 104850624 A

(43)申请公布日 2015.08.19

(73)专利权人 华东师范大学
地址 200241 上海市闵行区东川路500号

(72)发明人 兰曼 赵江

(74)专利代理机构 上海蓝迪专利商标事务所（普通合伙）31215
代理人 徐筱梅 张翔

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

(56)对比文件
CN 101937506 A,2011.01.05,
CN 102591978 A,2012.07.18,
US 2014156606 A1,2014.06.05,
王继奎 等.基于可信度模型的重复主数据检测算法.《计算机工程》.2014,第40卷(第5期),31-35,40.

审查员 邱川

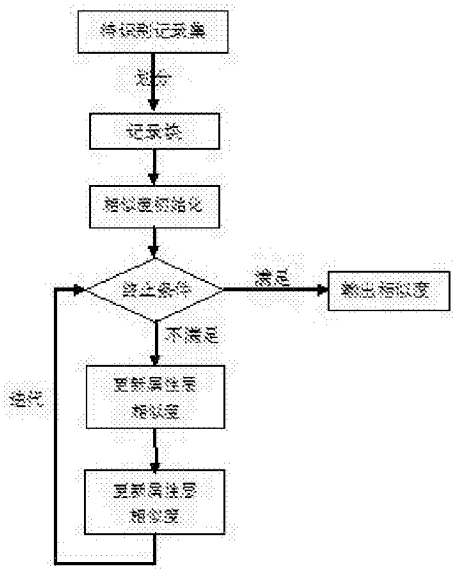
权利要求书2页 说明书5页 附图1页

(54)发明名称

近重复记录的相似度评估方法

(57)摘要

本发明公开了一种近重复记录相似度评估方法，该方法包括：步骤一：对待消重的大数据集进行分块操作，得到许多较小的数据块；步骤二：针对每个数据块，初始化属性层和记录层的相似度；步骤三：如果未满足迭代停止条件，则使用记录层相似度去更新属性层相似度和使用属性层相似度去更新记录层的相似度；步骤四：输出属性层和记录层的相似度。本发明采用了迭代地在属性和记录层传播相似度，从而克服了现实生产中记录存在缺失值和噪音值的问题，更准确地评估记录间的相似度。本发明还是一个无监督的方法，克服了需要标注数据带来的成本，并且其输出还可以灵活地集成到一些现存的基于聚类的或者基于距离的消重系统框架中。



CN 104850624 B

1.一种近重复记录相似度评估方法，其特征在于，包括如下步骤：

步骤一：对待消重的大数据集进行分块操作，得到许多较小的数据块；

步骤二：针对每个数据块，初始化属性层和记录层的相似度；

步骤三：如果未满足迭代停止条件，则使用记录层相似度去更新属性层的相似度和使用属性层相似度去更新记录层的相似度；

步骤四：输出属性层和记录层的相似度；其中：

所述步骤三中更新属性层相似度和更新记录层的相似度操作包括如下步骤：

步骤a1：检查迭代停止条件，如果满足条件，转到本方法的步骤四，否则继续以下步骤；

步骤a2：查找相似的属性簇并找到相关的记录，将记录间的相似度添加到计算属性层相似度的过程中；

步骤a3：查找相似的记录簇，使用更新的属性相似度和相似记录间的相似度去更新记录间相似度，转到步骤a1；其中：

更新属性层相似度使用如下表达式：

$$s(r_i^k, r_j^k) = \alpha T(r_i^k, r_j^k) + (1-\alpha)F(r_i^k, r_j^k)$$

式中，r_i^k和r_j^k为第i，j个记录的第k个属性，$T(r_i^k, r_j^k)$为传统属性相似度计算方式，$F(r_i^k, r_j^k)$为记录层反馈相似度，α∈[0,1]是一个权衡参数，用于决定传统相似度和反馈相似度的相对重要性，$F(r_i^k, r_j^k)$的计算如下：

$$F(r_i^k, r_j^k) = \frac{1}{1+|N(r_i^k)|+|N(r_j^k)|}\left(f(r_i^k, r_j^k) + \sum_{r_m^k \in N(r_i^k)} f(r_m^k, r_j^k) + \sum_{r_m^k \in N(r_j^k)} f(r_m^k, r_i^k)\right)$$

式中，$N(r_i^k)$为属性r_i^k的相似属性集合，$N(r_j^k)$为属性r_j^k的相似属性集合；$f(r_i^k, r_j^k)$为第i，j个记录的第k个属性的相似度；

更新记录层相似度使用如下表达式：

$$s(r_i, r_j) = \beta T(r_i, r_j) + (1-\beta)G(r_i, r_j)$$

式中，$T(r_i, r_j)$为传统记录相似度计算方式，$G(r_i, r_j)$为相似记录的反馈相似度，β∈[0,1]是一个权衡参数，用于决定传统相似度和反馈相似度的相对重要性，$G(r_i, r_j)$的计算方法如下：

$$G(r_i, r_j) = \frac{1}{|N(r_i)|+|N(r_j)|}\left(\sum_{r_{m \in N(r_j)}} s(r_i, r_m) + \sum_{r_{n \in N(r_i)}} s(r_j, r_n)\right)$$

式中，$N(r_i)$为记录r_i的相似记录集合，$N(r_j)$为记录r_j的相似记录集合。

2.如权利要求1所述的近重复记录相似度评估方法，其特征在于，所述步骤一中分块操作包括如下步骤：

步骤b1：评估每个记录中属性字段的重要性，人工设定每个属性的重要性或者使用自动化的方式设定，选取一个或者多个属性作为关键属性；

步骤b2：根据关键属性，使用合并聚类算法对记录进行快速聚类，每一簇的数据划分成为一个数据块。

3.如权利要求1所述的近重复记录相似度评估方法，其特征在于，所述步骤二中初始化操作包括如下步骤：

步骤c1：选择相似度度量函数来计算属性的相似度，如果属性值存在缺失，使用其他属性值的相似度来评估该属性的相似度；

步骤c2：根据上一步计算出来的属性相似度，计算记录间的相似度。

近重复记录的相似度评估方法

技术领域

[0001] 本发明涉及到大数据下近重复记录识别技术，涉及到记录间相似度的评估方法。

背景技术

[0002] 在大数据时代下，集成各种各样不同来源的数据是产生数据价值最基础的一环，而对近重复记录识别的消重工作是最核心的步骤。通常，一个记录通常由多个属性值构成，现有的识别方法主要可以归为以下几类：(1)基于概率匹配的方法，该方法使用条件独立假设或者广义的期望最大化(EM,Expectation Maximization)算法来推断单个记录对之间是否匹配的概率，每个观察值就是记录中属性的值；(2)基于距离的方法，它使用不同的相似度度量去计算属性层之间的相似性并通过为属性设置不同的权重来获得记录间的相似度，然后使用一个适当的匹配阈值去判断记录是否一样；(3)基于机器学习的方法，该方法从记录数据中抽取相似度特征，然后使用机器学习方法去学习如何匹配记录；(4)基于聚类的方法，它使用记录的相似矩阵把记录归入到不同的簇中，而在同一个簇中的记录则认为是近似重复记录或者潜在的近似重复记录；上述这些方法其实质上是计算记录的各属性的相似性度，为了克服一词多写，错写等情况，许多高容忍度的属性相似度度量方式被提出，例如，针对声音匹配的Soundex相似度度量。然而，每种方法都只针对特定的变量类型较为有效，对于缺失值或者噪声值的处理效果不好，尤其是互联网上的数据。

发明内容

[0003] 本发明的目的是针对现有技术的不足而提供的一种近重复记录相似度评估方法，该方法使用了属性间和记录间相似度相互传播的方法来评估记录间潜在的相似度，提升相似度评估的准确性，克服了缺失值，噪音值等无法修正错误所带来的影响。

[0004] 实现本发明目的的具体技术方案是：

[0005] 一种近重复记录相似度评估方法，包括如下步骤：

[0006] 步骤一：对待消重的大数据集进行分块操作，得到许多较小的数据块；

[0007] 步骤二：针对每个数据块，初始化属性层和记录层的相似度；

[0008] 步骤三：如果未满足迭代停止条件，则使用记录层相似度去更新属性层相似度和使用属性层相似度去更新记录层的相似度；

[0009] 步骤四：输出属性层和记录层的相似度。

[0010] 本发明提出所述的近重复记录相似度评估方法中，步骤一中的分块操作包括以下步骤：

[0011] 步骤a1：评估记录属性字段的重要性，可以人工设定每个属性的重要性或者使用自动化的方式设定，选取一个或者多个属性作为关键属性；

[0012] 步骤a2：根据关键属性，使用合并聚类(agglomerative clustering)来对记录进行快速聚类，每一簇的数据划分成为一个数据块。

[0013] 本发明提出所述的近重复记录相似度评估方法中，步骤二中的初始化包括以下步

骤：

[0014] 步骤b1：选择合适的相似度度量函数来计算属性的相似度，如果属性值存在缺失，则使用其他属性值对的相似度来评估该属性对的相似度；

[0015] 步骤b2：根据上一步计算出来的属性相似度，计算记录间的相似度。

[0016] 本发明提出所述的近重复记录相似度评估方法中，步骤三中更新属性层和记录层的相似度操作包括以下步骤：

[0017] 步骤c1：检查迭代停止条件，如果满足条件，转到本方法的步骤四，否则继续以下步骤；

[0018] 步骤c2：查找相似的属性簇并找到对应的记录，将记录间的相似度添加到计算属性相似度的过程中；

[0019] 步骤c3：查找相似的记录簇，使用更新的属性相似度和相似记录对的相似度去更新记录间相似度，转到步骤c1。

[0020] 本发明与现有技术不同之处有：一、本方法通过属性层的相似度估计和记录层的相似性估计之间的互相提升来达到更准确地估计记录间的相似性的目的，克服由缺失值和噪声值带来的相似度计算不准确的问题。在计算属性层相似度时，通过考虑相似属性簇的记录对的相似度，从而完成记录层和属性层相似度的传播。二、本方法是个无监督的算法，不像基于机器学习的方法需要训练数据，从而避免了人工标注数据所带来的成本，并且通过本方法得到的记录间的相似度可以灵活地集成到一些现存的基于聚类的或者基于距离的消重系统框架中。

[0021] 本发明的有益效果包括：使用了属性间和记录间相似度相互传播的方法来评估记录间潜在的相似度，提升相似度评估的准确性，克服了缺失值，噪音值等无法修正错误所带来的影响。并且该方法也可以得到属性间的相似度，可以被许多下游应用所使用，比如挖掘同义词。

附图说明

[0022] 图1是本发明方法的近重复记录的相似度评估流程图；

[0023] 图2是本发明方法中一个包含复杂文本类型的记录示例图。

具体实施方式

[0024] 结合以下具体实施例和附图，对本发明作进一步的详细说明。实施本发明的过程、条件、实验方法等，除以下专门提及的内容之外，均为本领域的普遍知识和公知常识，本发明没有特别限制内容。

[0025] 本发明中所涉及的专业术语的定义如下：

[0026] 记录(record)由一些属性构成，用来反映自然界中的一个实体(entity)，图2展示了一个包含复杂文本类型的记录的示例图。

[0027] 属性(attribute)是记录的一部分，用来刻画实体固有的性质，也可以称为字段(field)。

[0028] 消重(deduplication)是指在记录集合中，找到指向同一实体的记录的操作。

[0029] 属性层相似度是指属性间的相似度。

[0030] 记录层相似度是指记录间的相似度。

[0031] 由于在现实生产环境中，记录数据的量往往很大，在所有的记录两两之间进行完全的重复检查的计算成本巨大，所以本发明的第一步使用了合并聚类把大数据集分成许多较小的有交集的数据块，只有在同一数据块中的记录才进行两两比较。合并聚类算法如下：初始时每个记录都视为一块，如果两个块中存在任意两个记录的相似度大于阈值，则合并这两个块，最终直到不能再合并为止。为了加速分块的过程，在计算记录的相似度时，本发明并不考虑所有的属性而只考虑关键属性，通常关键属性只有1，2个。另外，本发明采用简单快速的相似度度量来计算相似度，例如考虑相同字数比率的戴斯(Dice)系数。算法描述如下：

[0032] 输入：记录集合$R=\{r_1,r_2,\cdots,r_n\}$，关键属性集合A，相似度函数Sim，阈值T

[0033] 输出：数据块$Bucket=\{b_1,b_2,\cdots,b_m\}$

[0034] 过程：

[0035] 步骤a1：初始化Bucket，将r_1视为一个数据块放入Bucket中。

[0036] 步骤a2：从第二个记录开始，依次遍历R，依据关键属性和相似度函数计算其与Bucket中数据块的相似度，如果相似度大于T，则加入到相应的数据块中，如果当前记录没有加入到任何数据块中，这它单独成为一个数据块加入到Bucket中。

[0037] 本发明的第二步骤针对每个数据块，进行属性层和记录层的初始化操作。考虑到不同属性有不同的重要性，因此本发明给不同的属性赋予不同的权重。记r_i^k为第i个记录的第k个属性，权重向量w，其中w_k表示第k个属性的相对重要性，并且$\Sigma_i w_i=1$，这一步的初始化如下：

[0038] (1)、属性层相似度初始化：当计算属性对的相似度的时候，往往会遇到缺失值的情况。直观的说，在一对记录中，含有缺失值属性对的相似度应该和那些不含缺失值属性对的相似对一致。所以本发明使用了插值的方法来评估含有缺失值属性对的相似度。给定一对记录(r_i,r_j)，令V为含有m(m为属性个数)个相似度值的相似度向量，这些相似度由普通的相似度函数度量；令I为指示向量，如果第k个属性值对含有缺失值，那么$I_k=0$，否则$I_k=1$。因此属性层的相似度初始化如下：

[0039] $$s(r_i^k,r_j^k)=\begin{cases} sim(r_i^k,r_j^k) & r_i^k,r_j^k \text{ 不为空} \\ \frac{r^T w}{I^T w}, & \text{为空} \end{cases} \quad (1)$$

[0040] (2)、记录层相似度初始化：本发明使用了传统的计算(即权重模式)方法来计算记录层的相似度，计算方法如下，

[0041] $$s(r_i,r_j)=\sum_{k=1}^{m} w_k s(r_i^k,r_j^k) \quad (2)$$

[0042] 本发明的第三步骤对属性层和记录层的相似度进行更新，分为以下步骤：

[0043] 步骤b1：属性层的相似度更新

[0044] 对于属性层相似度的计算由2部分构成：传统相似度和属性组层(field-group-level)的反馈相似度。首先定义属性对(r_i^k,r_j^k)的反馈信息如下：

[0045] $$f(r_i^k, r_j^k) = s(r_i^k, r_j^k) \quad (3)$$

[0046] 也就是等于他们记录层的相似度。接着我们定义在给定的属性对(r_i^k, r_j^k)时，属性r_i^k的属性组：

[0047] $$N(r_i^k) = \{r_m^k | s(r_i^k, r_m^k) > \theta \text{ 并且 } m \neq j\} \quad (4)$$

[0048] 其中，参数θ为近似重复的阈值。因此对于属性对(r_i^k, r_j^k)记录层的反馈相似度可以如下计算：

[0049] $$F(r_i^k, r_j^k) = \frac{1}{1 + |N(r_i^k)| + |N(r_j^k)|}\left(f(r_i^k, r_j^k) + \sum_{r_m^k \in N(r_i^k)} f(r_m^k, r_j^k) + \sum_{r_m^k \in N(r_j^k)} f(r_m^k, r_i^k)\right) \quad (5)$$

[0050] 也就是不同反馈相似度的平均。结合传统属性相似度计算方法(记为$T(r_i^k, r_j^k)$)，最终属性对的相似度可以使用如下方式计算：

[0051] $$s(r_i^k, r_j^k) = \alpha T(r_i^k, r_j^k) + (1-\alpha) F(r_i^k, r_j^k) \quad (6)$$

[0052] 其中，$T(r_i^k, r_j^k)$和$F(r_i^k, r_j^k)$由公式(1)，(5)计算而来，α∈[0,1]是一个权衡参数，用于决定传统相似度和反馈相似度的相对重要性。

[0053] 步骤b2：记录层相似度的更新

[0054] 类似的，记录层的相似度也由2部分组成：利用更新过后属性对的相似度重新计算的传统相似度和记录组层(record-group-level)的相似度。

[0055] 对于一对记录(r_i, r_j)，定义记录r_i的记录组为：

[0056] $N(r_i) = \{r_m | s(r_i, r_m) > \theta$并且$m \neq j\}$ (7)

[0057] 利用这个记录组来计算(r_i, r_j)的记录组层的相似度：

[0058] $$G(r_i, r_j) = \frac{1}{|N(r_i)| + |N(r_j)|}\left(\sum_{r_m \in N(r_j)} s(r_i, r_m) + \sum_{r_n \in N(r_i)} s(r_j, r_n)\right) \quad (8)$$

[0059] 最终，(r_i, r_j)的相似度可以使用如下的公式计算：

[0060] $s(r_i, r_j) = \beta T(r_i, r_j) + (1-\beta) G(r_i, r_j)$ (9)

[0061] 其中$T(r_i, r_j)$和$G(r_i, r_j)$可以由公式(2)，(8)计算，β∈[0,1]是一个权衡参数。

[0062] 步骤b3：迭代终止条件判断

[0063] 正如公式(6)(9)显示的，属性层和记录层的相似性可以相互传递。初始的，使用公式(1)(2)来初始化这两者相似度，接着使用公式(6)和(9)来依次地更新其相似度，最后当其值稳定的时候停止计算。通过计算变化率来衡量稳定度，第K次迭代后的稳定度定义如下：

[0064] $$\Delta s^{(k)} = \frac{\left|\sum_i^y \sum_j^y s^{(k)}(r_i, r_j) - \sum_i^y \sum_j^y s^{(k-1)}(r_i, r_j)\right|}{\sum_i^y \sum_j^y s^{(k-1)}(r_i, r_j)} \quad (10)$$

[0065] 其中$s^{(0)}(r_i, r_j)$为初始值。当变化率小于预定义阈值时就停止计算。

[0066] 最终，本发明的第四步骤输出属性层和记录层的相似度。

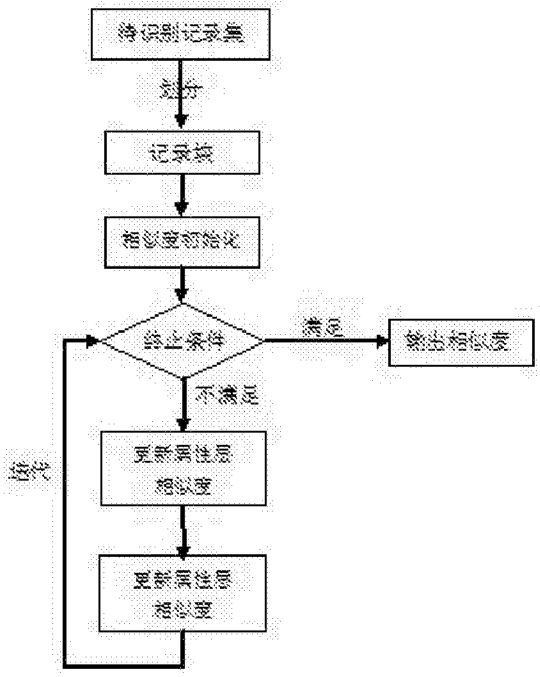


图1

属性 1	属性 2	属性 3		属性 n
简单类型	简单类型	复杂文本类型	其他类型	其他类型
是/否	100	大自然很神奇…,	…	三级甲等

图2