(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2016 年 12 月 1 日 (01.12.2016)

WIPO | PCT

(10) 国际公布号
WO 2016/188283 A1

(51) 国际专利分类号:
*G06F 17/30* (2006.01)

(21) 国际申请号: PCT/CN2016/080019

(22) 国际申请日: 2016 年 4 月 22 日 (22.04.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510276123.2 2015 年 5 月 26 日 (26.05.2015) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 英属开曼群岛大开曼乔治城资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人; 及
(71) 申请人 (仅对美国): 王丰金 (WANG, Fengjin) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

(54) Title: REPEATED DATA IDENTIFICATION METHOD AND DEVICE

(54) 发明名称 : 重复数据识别方法和装置

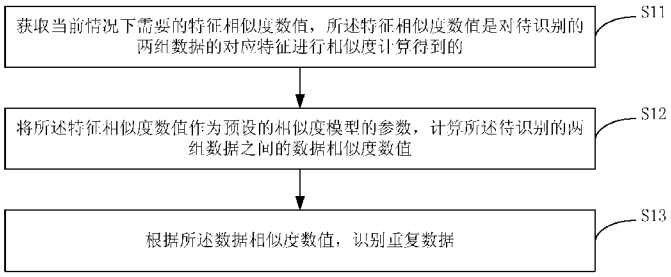


图 1

S11 ACQUIRE A SIMILARITY CHARACTER NUMERICAL VALUE WHICH IS REQUIRED UNDER A CURRENT SITUATION, WHEREIN THE SIMILARITY CHARACTER NUMERICAL VALUE IS OBTAINED BY PERFORMING SIMILARITY COMPUTATION ON CORRESPONDING CHARACTERS OF TWO GROUPS OF TO-BE-IDENTIFIED DATA

S12 USE THE SIMILARITY CHARACTER NUMERICAL VALUE AS A PARAMETER OF A PRESET SIMILARITY MODEL, AND COMPUTE A DATA SIMILARITY NUMERICAL VALUE BETWEEN THE TWO GROUPS OF TO-BE-IDENTIFIED DATA

S13 IDENTIFY REPEATED DATA ACCORDING TO THE DATA SIMILARITY NUMERICAL VALUE

(57) Abstract: A repeated data identification method and device. The repeated data identification method comprises: acquiring a similarity character numerical value which is required under a current situation, wherein the similarity character numerical value is obtained by performing similarity computation on corresponding characters of two groups of to-be-identified data (S11); using the similarity character numerical value as a parameter of a preset similarity model, and computing a data similarity numerical value between the two groups of to-be-identified data (S12); and identifying repeated data according to the data similarity numerical value (S13). The method can realize automatic identification of repeated data.

(57) 摘要: 一种重复数据识别方法和装置，该重复数据识别方法包括获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的（S11）；将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值（S12）；根据所述数据相似度数值，识别重复数据（S13）。该方法能够实现重复数据的自动识别。

WO 2016/188283 A1

**本国际公布**:

— 包括国际检索报告(条约第 21 条(3))。

# 重复数据识别方法和装置

## 技术领域

本申请涉及数据处理技术领域，尤其涉及一种重复数据识别方法和装置。

## 背景技术

在大数据时代，企业内部越来越多的业务需要使用大数据技术来分析业务、支撑业务，但是不同的业务团队在分析业务的过程中有很多相似的业务逻辑，加上各个业务团队之间沟通不及时，导致大规模离线数据处理平台上有很多相似数据，并且随着业务的发展，这种相似数据会越来越多，这不但浪费了大规模离线数据处理平台的存储资源，而且也浪费了大规模离线数据处理平台的计算资源。

现有技术中，一般都是开发人员在看到别的业务团队的相似数据后，才发现有重复数据。或者正好有开发人员对两边的业务都比较熟悉，所以了解业务两边的重复数据，平台层面并没有一个很好的方法来解决这个问题。

但是，这种方式存在如下问题：需要人工去熟悉所有的数据，才能完全识别出大规模数据处理平台上的重复数据；当大规模数据处理平台上的数据增长到一定级别以后，人工识别已经不可能。

## 发明内容

本申请旨在至少在一定程度上解决相关技术中的技术问题之一。

为此，本申请的一个目的在于提出一种重复数据识别方法，该方法可以实现重复数据的自动识别。

本申请的另一个目的在于提出一种重复数据识别装置。

为达到上述目的，本申请第一方面实施例提出的重复数据识别方法，包括：获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的；将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值；根据所述数据相似度数值，识别重复数据。

本申请第一方面实施例提出的重复数据识别方法，通过采用相似度模型对待识别的两组数据进行重复数据识别，可以在需要识别重复数据时，有一个统

一的标准，不需要人为识别，实现重复数据的自动识别。

为达到上述目的，本申请第二方面实施例提出的重复数据识别装置，包括：获取模块，用于获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的；计算模块，用于将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值；识别模块，用于根据所述数据相似度数值，识别重复数据。

本申请第二方面实施例提出的重复数据识别装置，通过采用相似度模型对待识别的两组数据进行重复数据识别，可以在需要识别重复数据时，有一个统一的标准，不需要人为识别，实现重复数据的自动识别。

本申请附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本申请的实践了解到。

## 附图说明

本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

图 1 是本申请一实施例提出的重复数据识别方法的流程示意图；

图 2 是本申请另一实施例提出的重复数据识别装置的结构示意图。

## 具体实施方式

下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的模块或具有相同或类似功能的模块。下面通过参考附图描述的实施例是示例性的，仅用于解释本申请，而不能理解为对本申请的限制。相反，本申请的实施例包括落入所附加权利要求书的精神和内涵范围内的所有变化、修改和等同物。

图 1 是本申请一实施例提出的重复数据识别方法的流程示意图，该方法包括：

S11：获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的。

其中，待识别的两组数据可以分别记录在两张表内，相应的，特性相似度数值是对两个表的特征进行相似度计算得到的。

可选的，所述待识别的两组数据分别记录在两张表内，所述相似度特征数

值包括如下项中的至少一项：

表血缘方面的相似度数值，表语义方面的相似度数值，表内容方面的相似度数值。

可选的，所述表血缘方面的相似度数值包括如下项中的至少一项：表血缘相似度数值，字段血缘相似度数值；或者，

所述表语义方面的相似度数值包括如下项中的至少一项：表结构（schema）相似度数值，表名相似度数值；或者，

所述表内容方面的相似度数值包括如下项中的至少一项：表记录数相似度数值，表分区大小相似度数值。

其中，不同情况下需要的相似度特征数值可以是不同的。可以根据当前情况获取相应的上述的六种相似度特征数值中的至少一项。

在当前情况下，可以确定当前需要的相似度特征数值，之后可以在线计算需要的相似度特征数值，或者，从已经计算得到的上述六种相似度特征数值中获取当前需要的相似度特征数值。

可选的，所述获取当前情况下需要的相似度特征数值，包括：

如果当前情况是进行上下游表比较，则获取如下的相似度特征数值：表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值；或者，

如果当前情况是进行相似表比较，则获取如下的相似度特征数值：表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值；或者，

如果当前情况是进行表来源相似比较，则获取如下的相似度特征数值：

表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，以及，表名相似度数值。

上述的六种相似度特征数值的计算公式可以分别表示为：

（1）表血缘的相似度数值 S1：使用余弦相似性来计算两张表的相似性。

具体如下：表 A 的父血缘是(a, b)、表 B 的父血缘是(b, c)，取两个并集并排序得出表 A 和表 B 的余弦相似向量 C=(a, b, c)，对比相似性向量可以得出表 A 的相似性向量 A1=(1,1,0)，表 B 的相似性向量 B1=(0,1,1)，表 A 表 B 的表血缘相似度数值的计算公式为：S1= A1*B1。

（2）字段血缘相似度数值 S2：同样使用余弦相似性来计算。不过首先要获得表 A 表 B 的字段血缘，然后使用表血缘类似的方法计算表 A 和表 B 的相

似性向量 A1、B1，进而计算字段血缘相似度数值，计算公式为：S2=A1*B1。

（3）表 schema 相似度数值 S3：schema 的相似性同样使用余弦相似性来计算，不过这里注意分区列不参与计算，同样的方法得到两张表的相似性向量 A1、B1，进而计算表 schema 相似度数值，计算公式为：S3=A1*B1。

（4）表名相似度数值 S4：需要先将表名按照下划线拆开，然后去除无用词，主要是纯数字无特殊意义的词，两张表根据剩下的词计算相似性向量 A1、B1，进而计算表名相似度数值，计算公式为：S4=A1*B1。

（5）表记录数相似度数值 S5：通过计算两张表分区记录数的波动性来衡量两张表记录数的相似性，计算公式为：

$$S5= -\frac{\log 10(\frac{\sum(x-\overline{x})(y-\overline{y})}{n})}{6}$$，其中，

$x$表示表 A 的一个分区的记录数，$\overline{x}$表示 A 参与计算的分区记录数的平均数，$y$表示表 B 的一个分区的记录数，$\overline{y}$表示表 B 参与计算的分区记录数的平均数，n 是统计的分区的个数，n 取值范围是（7,60），越大越精确。

（6）表分区大小相似度数值 S6：使用分区大小的波动相似性衡量两张表分区大小的相似性，计算公式同（5），不过这里计算时注意统一单位。

S12：将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值。

可选的，所述将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值，包括：

如果当前情况是进行上下游表比较，则采用如下计算公式计算所述数据相似度数值：

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4；或者，

如果当前情况是进行相似表比较，则采用如下计算公式计算所述数据相似度数值：

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5；或者，

如果当前情况是进行表来源相似比较，则采用如下计算公式计算所述数据相似度数值：

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35;

其中，S 是数据相似度数值，S1 是表血缘相似度数值，S2 是字段血缘相似度数值，S3 是表结构相似度数值，S4 是表名相似度数值，S5 是表记录数相似度数值，S6 是表分区大小相似度数值。

具体的，上下游表相似时，这里的上下游指的是离线数据加工层级，首先根据表血缘梳理出上下游表，然后使用表 schema 的相似度数值 S3、表名相似度数值 S4、表记录数相似度数值 S5、表分区大小相似度数值 S6 构建此类表的相似度模型，计算公式如下：

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4。

计算相似的表时，使用上述六个特征构建此类表的相似度模型，计算公式如下：

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5。

来源相似的表指的是流入离线数据处理平台的源头表，使用上面前四种特征构建相似度模型，计算公式如下：

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35。

S13：根据所述数据相似度数值，识别重复数据。

其中，这里的识别重复数据不限于识别出两组完全一致的数据，而是指识别出两组数据的相似程度。

可选的，所述根据所述数据相似度数值，识别重复数据，包括：

根据所述数据相似度数值，确定所述数据相似度数值属于的预设的数值阈值；

根据预设的数值阈值与相似程度的对应关系，确定所述数据相似度数值属于的数值阈值对应的相似程度，得到所述待识别的两组数据的相似程度。

具体的，上述三种场景的数据相似度数值 S 的取值范围均为[0,1]，取值越大相似性越大。例如，0.9 以上表示两张表数据重复，0.7~0.9 之间表示两张表数据重复性比较大，小于 0.7 表示两张表重复性比较低。

本实施例的方法可以应用到 Hadoop 集群，或者 odps 集群等。

本实施例中，通过采用相似度模型对待识别的两组数据进行重复数据识别，可以在需要识别重复数据时，有一个统一的标准，不需要人为识别，实现重复数据的自动识别。本实施例通过选择上述各种具体的相似度特征数值，以及根据不同情况采用不同的相似度特征数值进行识别，可以非常适用于大规模的重复数据识别。

图 2 是本申请另一实施例提出的重复数据识别装置的结构示意图，该装置

20 包括：获取模块 21，计算模块 22 和识别模块 23。

获取模块 21，用于获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的；

其中，待识别的两组数据可以分别记录在两张表内，相应的，特性相似度数值是对两个表的特征进行相似度计算得到的。

可选的，所述待识别的两组数据分别记录在两张表内，所述相似度特征数值包括如下项中的至少一项：

表血缘方面的相似度数值，表语义方面的相似度数值，表内容方面的相似度数值。

可选的，所述表血缘方面的相似度数值包括如下项中的至少一项：表血缘相似度数值，字段血缘相似度数值；或者，

所述表语义方面的相似度数值包括如下项中的至少一项：表结构（schema）相似度数值，表名相似度数值；或者，

所述表内容方面的相似度数值包括如下项中的至少一项：表记录数相似度数值，表分区大小相似度数值。

其中，不同情况下需要的相似度特征数值可以是不同的。可以根据当前情况获取相应的上述的六种相似度特征数值中的至少一项。

在当前情况下，可以确定当前需要的相似度特征数值，之后可以在线计算需要的相似度特征数值，或者，从已经计算得到的上述六种相似度特征数值中获取当前需要的相似度特征数值。

可选的，所述获取模块 21 具体用于：

如果当前情况是进行上下游表比较，则获取如下的相似度特征数值：表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值；或者，

如果当前情况是进行相似表比较，则获取如下的相似度特征数值：表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值；或者，

如果当前情况是进行表来源相似比较，则获取如下的相似度特征数值：

表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，以及，表名相似度数值。

上述的六种相似度特征数值的计算公式可以分别表示为：

（2）表血缘的相似度数值 S1：使用余弦相似性来计算两张表的相似性。

具体如下：表 A 的父血缘是(a, b)、表 B 的父血缘是(b, c)，取两个并集并排序得出表 A 和表 B 的余弦相似向量 C=(a, b, c)，对比相似性向量可以得出表 A 的相似性向量 A1=(1,1,0)，表 B 的相似性向量 B1=(0,1,1)，表 A 表 B 的表血缘相似度数值的计算公式为：S1= A1*B1。

（2）字段血缘相似度数值 S2：同样使用余弦相似性来计算。不过首先要获得表 A 表 B 的字段血缘，然后使用表血缘类似的方法计算表 A 和表 B 的相似性向量 A1、B1，进而计算字段血缘相似度数值，计算公式为：S2=A1*B1。

（3）表 schema 相似度数值 S3：schema 的相似性同样使用余弦相似性来计算，不过这里注意分区列不参与计算，同样的方法得到两张表的相似性向量 A1、B1，进而计算表 schema 相似度数值，计算公式为：S3=A1*B1。

（4）表名相似度数值 S4：需要先将表名按照下划线拆开，然后去除无用词，主要是纯数字无特殊意义的词，两张表根据剩下的词计算相似性向量 A1、B1，进而计算表名相似度数值，计算公式为：S4=A1*B1。

（5）表记录数相似度数值 S5：通过计算两张表分区记录数的波动性来衡量两张表记录数的相似性，计算公式为：

$$S5= -\frac{\log 10(\frac{\sum(x-\overline{x})(y-\overline{y})}{n})}{6}$$，其中，

$x$表示表 A 的一个分区的记录数，$\overline{x}$表示 A 参与计算的分区记录数的平均数，$y$表示表 B 的一个分区的记录数，$\overline{y}$表示表 B 参与计算的分区记录数的平均数，n 是统计的分区的个数，n 取值范围是（7,60），越大越精确。

（6）表分区大小相似度数值 S6：使用分区大小的波动相似性衡量两张表分区大小的相似性，计算公式同（5），不过这里计算时注意统一单位。

计算模块 22，用于将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值；

可选的，所述计算模块 22 具体用于：

如果当前情况是进行上下游表比较，则采用如下计算公式计算所述数据相似度数值：

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4；或者，

如果当前情况是进行相似表比较，则采用如下计算公式计算所述数据相似

度数值:

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5; 或者,

如果当前情况是进行表来源相似比较,则采用如下计算公式计算所述数据相似度数值:

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35;

其中,S 是数据相似度数值,S1 是表血缘相似度数值,S2 是字段血缘相似度数值,S3 是表结构相似度数值,S4 是表名相似度数值,S5 是表记录数相似度数值,S6 是表分区大小相似度数值。

具体的,上下游表相似时,这里的上下游指的是离线数据加工层级,首先根据表血缘梳理出上下游表,然后使用表 schema 的相似度数值 S3、表名相似度数值 S4、表记录数相似度数值 S5、表分区大小相似度数值 S6 构建此类表的相似度模型,计算公式如下:

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4。

计算相似的表时,使用上述六个特征构建此类表的相似度模型,计算公式如下:

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5。

来源相似的表指的是流入离线数据处理平台的源头表,使用上面前四种特征构建相似度模型,计算公式如下:

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35。

识别模块 23,用于根据所述数据相似度数值,识别重复数据。

可选的,所述识别模块 23 具体用于:

根据所述数据相似度数值,确定所述数据相似度数值属于的预设的数值阈值;

根据预设的数值阈值与相似程度的对应关系,确定所述数据相似度数值属于的数值阈值对应的相似程度,得到所述待识别的两组数据的相似程度。

具体的,上述三种场景的数据相似度数值 S 的取值范围均为[0,1],取值越大相似性越大。例如,0.9 以上表示两张表数据重复,0.7~0.9 之间表示两张表数据重复性比较大,小于 0.7 表示两张表重复性比较低。

本实施例的方法可以应用到 Hadoop 集群,或者 odps 集群等。

本实施例中,通过采用相似度模型对待识别的两组数据进行重复数据识别,可以在需要识别重复数据时,有一个统一的标准,不需要人为识别,实现重复数据的自动识别。本实施例通过选择上述各种具体的相似度特征数值,以

及根据不同情况采用不同的相似度特征数值进行识别，可以非常适用于大规模的重复数据识别。

需要说明的是，在本申请的描述中，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。此外，在本申请的描述中，除非另有说明，“多个”的含义是指至少两个。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本申请的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描

述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

# 权 利 要 求 书

1、一种重复数据识别方法，其特征在于，包括:

获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的;

将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值;

根据所述数据相似度数值，识别重复数据。

2、根据权利要求 1 所述的方法，其特征在于，所述待识别的两组数据分别记录在两张表内，所述相似度特征数值包括如下项中的至少一项:

表血缘方面的相似度数值，表语义方面的相似度数值，表内容方面的相似度数值。

3、根据权利要求 2 所述的方法，其特征在于，

所述表血缘方面的相似度数值包括如下项中的至少一项: 表血缘相似度数值，字段血缘相似度数值; 或者，

所述表语义方面的相似度数值包括如下项中的至少一项: 表结构相似度数值，表名相似度数值; 或者，

所述表内容方面的相似度数值包括如下项中的至少一项: 表记录数相似度数值，表分区大小相似度数值。

4、根据权利要求 1-3 任一项所述的方法，其特征在于，所述获取当前情况下需要的相似度特征数值，包括:

如果当前情况是进行上下游表比较，则获取如下的相似度特征数值: 表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值; 或者，

如果当前情况是进行相似表比较，则获取如下的相似度特征数值: 表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值; 或者，

如果当前情况是进行表来源相似比较，则获取如下的相似度特征数值:

表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，以及，表名相似度数值。

5、根据权利要求 4 所述的方法，其特征在于，所述将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似

度数值，包括:

如果当前情况是进行上下游表比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4; 或者，

如果当前情况是进行相似表比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5; 或者，

如果当前情况是进行表来源相似比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35;

其中，S 是数据相似度数值，S1 是表血缘相似度数值，S2 是字段血缘相似度数值，S3 是表结构相似度数值，S4 是表名相似度数值，S5 是表记录数相似度数值，S6 是表分区大小相似度数值。

6、根据权利要求 1 或 5 所述的方法，其特征在于，所述根据所述数据相似度数值，识别重复数据，包括:

根据所述数据相似度数值，确定所述数据相似度数值属于的预设的数值阈值;

根据预设的数值阈值与相似程度的对应关系，确定所述数据相似度数值属于的数值阈值对应的相似程度，得到所述待识别的两组数据的相似程度。

7、一种重复数据识别装置，其特征在于，包括:

获取模块，用于获取当前情况下需要的相似度特征数值，所述相似度特征数值是对待识别的两组数据的对应特征进行相似度计算得到的;

计算模块，用于将所述相似度特征数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值;

识别模块，用于根据所述数据相似度数值，识别重复数据。

8、根据权利要求 7 所述的装置，其特征在于，所述获取模块具体用于:

如果当前情况是进行上下游表比较，则获取如下的相似度特征数值: 表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值; 或者，

如果当前情况是进行相似表比较，则获取如下的相似度特征数值: 表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，表名相似度数值，表记录数相似度数值，以及，表分区大小相似度数值; 或者，

如果当前情况是进行表来源相似比较，则获取如下的相似度特征数值:

表血缘相似度数值，字段血缘相似度数值，表结构相似度数值，以及，表名相似度数值。

9、根据权利要求 8 所述的装置，其特征在于，所述计算模块具体用于:

如果当前情况是进行上下游表比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.8*S3+0.2*S4)*0.4+(0.7*S5+0.3*S6)*0.4; 或者，

如果当前情况是进行相似表比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.4*S1+0.6*S2)*0.4+(0.8*S3+0.2*S4)*0.1+(0.7*S5+0.3*S6)*0.5; 或者，

如果当前情况是进行表来源相似比较，则采用如下计算公式计算所述数据相似度数值:

S=(0.4*S1+0.6*S2)*0.65+(0.8*S3+0.2*S4)*0.35;

其中，S 是数据相似度数值，S1 是表血缘相似度数值，S2 是字段血缘相似度数值，S3 是表结构相似度数值，S4 是表名相似度数值，S5 是表记录数相似度数值，S6 是表分区大小相似度数值。

10、根据权利要求 7-9 任一项所述的装置，其特征在于，所述识别模块具体用于:

根据所述数据相似度数值，确定所述数据相似度数值属于的预设的数值阈值;

根据预设的数值阈值与相似程度的对应关系，确定所述数据相似度数值属于的数值阈值对应的相似程度，得到所述待识别的两组数据的相似程度。

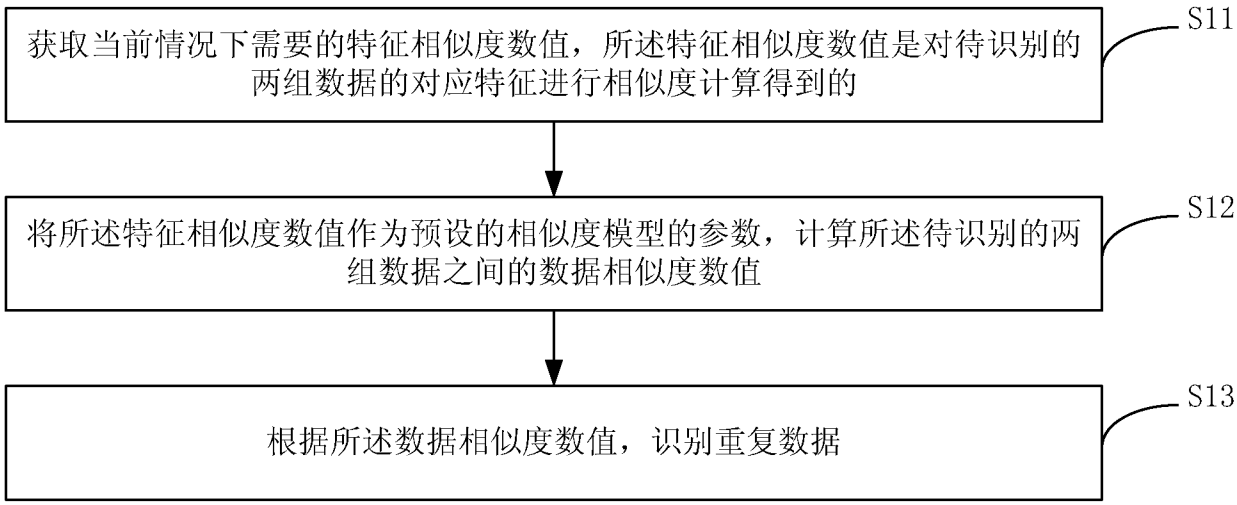
获取当前情况下需要的特征相似度数值，所述特征相似度数值是对待识别的两组数据的对应特征进行相似度计算得到的
S11
将所述特征相似度数值作为预设的相似度模型的参数，计算所述待识别的两组数据之间的数据相似度数值
S12
根据所述数据相似度数值，识别重复数据
S13

图 1

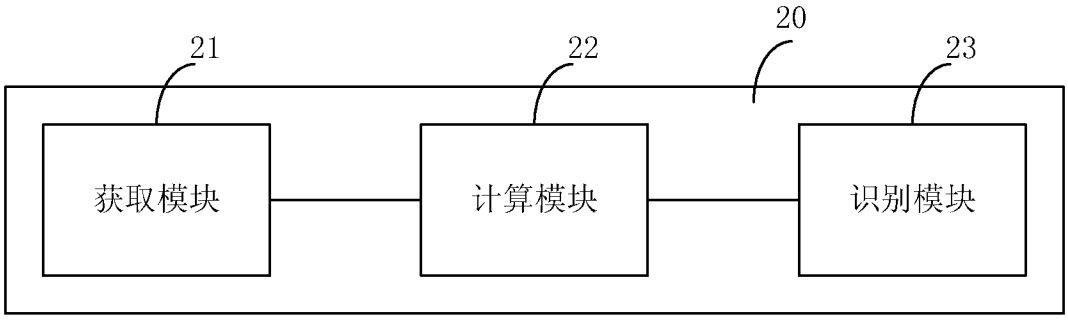
20
21
22
23
获取模块
计算模块
识别模块

图 2

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/080019

**A. CLASSIFICATION OF SUBJECT MATTER**

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, GOOGLE: paramrter?, model, similar+, same, dupli+, repeat+, identif+, detect+, struct+, content?, field?, table?, list?, semantic, source, information, feature, characteristic, date

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | CN 102768659 A (ALIBABA GROUP HOLDING LTD.) 07 November 2012 (07.11.2012) description, paragraphs [0034], [0052] and [0054] | 1-10 |
| A | CN 102033867 A (UNIV. NORTHWESTERN POLYTECHNIC) 27 April 2011 (27.04.2011) the whole document | 1-10 |
| A | US 2007112898 A1 (CLAIRVOYANCE CORPORATION) 17 May 2007 (17.05.2007) the whole document | 1-10 |

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 07 June 2016 | 01 July 2016 |
| Name and mailing address of the ISA<br>State Intellectual Property Office of the P. R. China<br>No. 6, Xitucheng Road, Jimenqiao<br>Haidian District, Beijing 100088, China<br>Facsimile No. (86-10) 62019451 | Authorized officer<br>CHENG, Xiaomei<br>Telephone No. (86-10) 62413966 |

**INTERNATIONAL SEARCH REPORT**
Information on patent family members

International application No.
PCT/CN2016/080019

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| CN 102768659 A | 07 November 2012 | HK 1172706 A1 | 18 December 2015 |
| CN 102033867 A | 27 April 2011 | None | |
| US 2007112898 A1 | 17 May 2007 | JP 2009521738 A | 04 June 2009 |
| | | WO 2007059232 A2 | 24 May 2007 |

# 国际检索报告

国际申请号
PCT/CN2016/080019

A. 主题的分类

G06F 17/30(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,CNKI,WPI,EPODOC,google:相似，相似度，模型，检测，识别，重复，去重，父，源，结构，语义，内容，表名，表，信息，参数，特征，表内容，数据，字段，parameter?，model，similar+，same，dupli+，repeat+，identif+，detect+，struct+，content?，field?，table?，list?，semantic

C. 相关文件

| 类 型* | 引用文件，必要时，指明相关段落 | 相关的权利要求 |
|---|---|---|
| X | CN 102768659 A (阿里巴巴集团控股有限公司) 2012年 11月 7日 (2012-11-07)<br>说明书第[0034]、[0052]、[0054]段 | 1-10 |
| A | CN 102033867 A (西北工业大学) 2011年 4月 27日 (2011-04-27)<br>全文 | 1-10 |
| A | US 2007112898 A1 (CLAIRVOYANCE CORPORATION) 2007年 5月 17日 (2007-05-17)<br>全文 | 1-10 |

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

| 国际检索实际完成的日期 | 国际检索报告邮寄日期 |
|---|---|
| 2016年 6月 7日 | 2016年 7月 1日 |

| ISA/CN的名称和邮寄地址 | 受权官员 |
|---|---|
| 中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路6号 100088 | 程小梅 |
| 传真号 (86-10)62019451 | 电话号码 (86-10)01062413966 |

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/080019

| 检索报告引用的专利文件 | | | 公布日<br>(年/月/日) | 同族专利 | | | 公布日<br>(年/月/日) |
|---|---|---|---|---|---|---|---|
| CN | 102768659 | A | 2012年 11月 7日 | HK | 1172706 | A1 | 2015年 12月 18日 |
| CN | 102033867 | A | 2011年 4月 27日 | | 无 | | |
| US | 2007112898 | A1 | 2007年 5月 17日 | JP | 2009521738 | A | 2009年 6月 4日 |
| | | | | WO | 2007059232 | A2 | 2007年 5月 24日 |