

(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
2017 年 12 月 21 日 (21.12.2017)



(10) 国际公布号
WO 2017/215346 A1

- (51) 国际专利分类号:
G06K 9/62 (2006.01)
- (21) 国际申请号: PCT/CN2017/081387
- (22) 国际申请日: 2017 年 4 月 21 日 (21.04.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610423480.1 2016年6月15日 (15.06.2016) CN
- (71) 申请人: 北京京东尚科信息技术有限公司 (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。北京京东世纪贸易有限公司 (BEIJING JINGDONG CENTURY TRADING CO., LTD.) [CN/CN]; 中国北京市大兴区北京经济技术开发区科创十一街18号C座2层201室, Beijing 100176 (CN)。

- (72) 发明人: 闫强(YAN, Qiang); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。王 晓(WANG, Xiao); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。葛胜利(GE, Shengli); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。李爱华(LI, Aihua); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。
- (74) 代理人: 中国国际贸易促进委员会专利商标事务所 (CCPIT PATENT AND TRADEMARK LAW OFFICE); 中国北京市西城区阜成门外大街2号万通新世界广场8层, Beijing 100037 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,

(54) Title: SERVICE DATA CLASSIFICATION METHOD AND APPARATUS

(54) 发明名称: 业务数据分类方法和装置

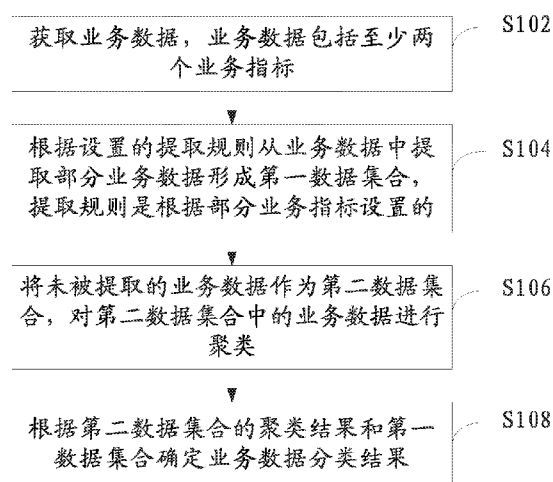


图 1

- S102 Acquiring service data, wherein the service data comprises at least two service indexes
- S104 According to a set extraction rule, extracting some service data from the service data so as to form a first data set, wherein the extraction rule is set according to some of the service indexes
- S106 Regarding service data which is not extracted as a second data set, and clustering the service data in the second data set
- S108 According to a cluster result of the second data set and the first data set, determining a classification result of the service data

(57) Abstract: Disclosed are a service data classification method and apparatus, which relate to the field of data processing. The service data classification method comprises: acquiring service data, wherein the service data comprises a plurality of service indexes (S102); according to a set extraction rule, extracting some service data, with a category attribute complying with a pre-set condition, from the service data so as to form a first data set, wherein the extraction rule is set according to some of the service indexes (S104); regarding service data which is not extracted as a second data set, and clustering the service data in the second data set (S106); and according to a cluster result of the second data set and the first data set, determining a classification result of the service data (S108). According to some of the service indexes of service data, an extraction rule is set, and in accordance with the extraction rule, service data with a clear category attribute is extracted, and according to a cluster result of service data, which is not extracted, together with the service data with a clear category attribute, a classification result of the service data is determined, thereby improving the accuracy of service data classification.

JP, KE, KG, KH, KN, KP, KR, KW, 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)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种业务数据分类方法和装置, 涉及数据处理领域。其中的业务数据分类方法包括: 获取业务数据, 业务数据包括多个业务指标 (S102); 根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合, 提取规则是根据部分业务指标设置的 (S104); 将未被提取的业务数据作为第二数据集合, 对第二数据集合中的业务数据进行聚类 (S106); 根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果 (S108)。根据业务数据的部分业务指标设置提取规则, 并按照提取规则提取类别属性明确的业务数据, 再根据未被提取的业务数据的聚类结果以及类别属性明确的业务数据共同确定业务数据分类的结果, 提高了业务数据分类的准确性。

业务数据分类方法和装置

技术领域

本发明涉及数据处理领域，特别涉及一种业务数据分类方法和装置。

背景技术

数据聚类是数据挖掘和数据分类中采用的一种常规的技术。但是在对业务数据进行分类的过程中，业务数据通常是具有一定业务属性的，而常规的聚类方法单纯从数据层面去解决问题，忽略了数据的业务含义。例如，某些业务指标特别高或者特别低的数据会被作为离群数据剔除，不再参与聚类，而这些被剔除的离群数据可能是业务价值较高的数据。

因此，按照常规的聚类方法进行分类，业务价值较高的数据无法体现其特有的价值，影响业务数据分类的准确性。

发明内容

本发明实施例所要解决的一个技术问题是：如何使业务数据分类的结果更准确。

根据本发明实施例的第一个方面，提供一种业务数据分类方法，包括：获取业务数据，业务数据包括多个业务指标；根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，提取规则是根据部分业务指标设置的；将未被提取的业务数据作为第二数据集合，对第二数据集合中的业务数据进行聚类；根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果。

在一个实施例中，根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合包括：根据设定的阈值提取某一业务指标数据离群的业务数据，形成第一数据集合；或者，根据业务数据的多个业务指标的逻辑运算结果提取出逻辑运算结果离群的业务数据，形成第一数据集合；或者，根据某一业务指标的数据分布情况，提取出某一业务指标数据离群的业务数据，形成第一数据集合。

在一个实施例中，根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果包括：将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。

在一个实施例中，将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并包括：确定第一数据集合和第二数据集合中提取规则中涉及的业务指标的平均特征，将第一数据集合和第二数据集合中平均特征最接近的类进行合并获得业务数据分类结果，其中，平均特征是每个类的提取规则涉及的业务指标的平均值或中心点；或者，将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。

在一个实施例中，根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合包括：分别根据设置的不同的提取规则从业务数据中提取部分业务数据以形成不同的类，不同的类形成第一数据集合。

在一个实施例中，在对第二数据集合中的业务数据进行聚类之前，方法还包括：从第二数据集合中筛选出离群的业务指标数据，将用于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据；或者，计算某一业务指标中所有非空数据的均值，将均值赋予第二数据集合中该业务指标的数据空值。

在一个实施例中，对第二数据集合中的业务数据进行聚类包括：按照各个预定聚类个数分别对第二数据集合中的业务数据进行预聚类，计算每个预定聚类个数对应的预聚类结果的轮廓系数；按照预定聚类个数递增的顺序排列相应的轮廓系数，获取轮廓系数中的若干极大值，并确定其中的最大值；将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数，预设条件为极大值与最大值的差值小于预设值；采用实际聚类个数对第二数据集合中的业务数据进行聚类。

根据本发明实施例的第二个方面，提供一种业务数据分类装置，包括：业务数据获取模块，用于获取业务数据，业务数据包括多个业务指标；业务数据提取模块，用于根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，提取规则是根据部分业务指标设置的；业务数据聚类模块，用于将未被提取的业务数据作为第二数据集合，对第二数据集合中的业务数据进行聚类；业务数据分类模块，用于根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果。

在一个实施例中，业务数据提取模块包括以下至少一个单元：第一提取单元，用于根据设定的阈值提取某一业务指标数据离群的业务数据，形成第一数据集合；第二

提取单元，用于根据业务数据的多个业务指标的逻辑运算结果提取出逻辑运算结果离群的业务数据，形成第一数据集合；第三提取单元，用于根据某一业务指标的数据分布情况，提取出某一业务指标数据离群的业务数据，形成第一数据集合。

在一个实施例中，业务数据分类模块用于将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。

在一个实施例中，业务数据分类模块包括平均特征获取单元和合并单元；平均特征获取单元用于确定第一数据集合和第二数据集合中提取规则中涉及的业务指标的平均特征，其中，平均特征是每个类的提取规则涉及的业务指标的平均值或中心点；合并单元用于将第一数据集合和第二数据集合中平均特征最接近的类进行合并获得业务数据分类结果；或者，业务数据分类模块用于将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。

在一个实施例中，业务数据提取模块用于分别根据设置的不同的提取规则从业务数据中提取部分业务数据以形成不同的类，不同的类形成第一数据集合。

在一个实施例中，装置还包括预处理模块，预处理模块包括：离群点处理单元，用于从第二数据集合中筛选出离群的业务指标数据，将用于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据；和/或，空值处理单元，用于计算第二数据集合中某一业务指标中所有非空数据的均值，将均值赋予第二数据集合中该业务指标的数据空值。

在一个实施例中，业务数据聚类模块包括：预聚类单元，用于按照各个预定聚类个数分别对第二数据集合中的业务数据进行预聚类；轮廓系数计算单元，用于根据于聚类结果计算每个预定聚类个数对应的预聚类结果的轮廓系数；实际聚类个数确定单元，用于按照预定聚类个数递增的顺序排列相应的轮廓系数，获取轮廓系数中的若干极大值，并确定其中的最大值，将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数，预设条件为极大值与最大值的差值小于预设值；实际聚类单元，用于采用实际聚类个数对第二数据集合中的业务数据进行聚类。

根据本发明实施例的第三个方面，提供一种业务数据分类装置，包括：存储器；以及耦接至存储器的处理器，处理器被配置为基于存储在存储器中的指令，执行前述任意一种业务数据分类方法。

根据本发明实施例的第四个方面，提供一种计算机非瞬时性可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述任意一种业务数据分类方法。

本发明根据业务数据的部分业务指标设置提取规则，并按照提取规则提取类别属性明确的业务数据，再根据未被提取的业务数据的聚类结果以及类别属性明确的业务数据共同确定业务数据分类的结果，提高了业务数据分类的准确性。

通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

附图说明

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性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本发明业务数据分类方法的一个实施例的流程图。

图 2 为本发明业务数据聚类方法的一个实施例的流程图。

图 3 为本发明业务数据分类装置的一个实施例的结构图。

图 4 为本发明业务数据分类装置的另一个实施例的结构图

图 5 为本发明业务数据分类装置的又一个实施例的结构图

图 6 为本发明业务数据分类装置的再一个实施例的结构图。

具体实施方式

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，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。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。

除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实

际的比例关系绘制的。

对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为授权说明书的一部分。

在这里示出和讨论的所有示例中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它示例可以具有不同的值。

应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

针对常规的聚类方法单纯从数据层面去解决问题，忽略了数据的业务含义，使得业务价值较高的数据无法体现其特有的价值，从而影响业务数据分类的准确性的问题，提出本发明。

图 1 为本发明业务数据分类方法的一个实施例的流程图。如图 1 所示，该实施例的方法包括：

步骤 S102，获取业务数据，业务数据包括多个业务指标。

在一个实施例中，可以根据业务分类目的设置相关的业务指标，并获取相关的业务数据，业务数据具有多个维度，各个维度是与业务分类目的相关的各个业务指标。

例如，某业务分类目的为根据用户的活跃度划分用户等级，也即对用户进行活跃度高低的分类，则业务数据例如可以包括用户近一段时间的 PV（Page View，页面浏览量）、订单总量、订单总额、持续 PV 时间、商品收藏数量、注册时间等能够反映用户活跃度的业务指标。

其中，可以通过 ETL（Extract-Transform-Load，抽取、转换、加载）技术进行数据提取和指标计算。例如，可以通过 SQL 语句中的 WHERE 条件限定所要获取的业务指标的条件，再将计算结果插入到目标表结构中。

其中，还可以为业务指标选择性地添加标识信息，标识该指标值的大小和业务目的是否为正相关的。例如，对于根据活跃度为用户进行分类的场景，PV 越大表示用户活跃度越高，相邻登录时间的间隔越大表示用户的活跃度越低。添加标识信息有助于后续步骤中提取规则的设置和使用。

步骤 S104，根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，提取规则是根据部分业务指标设置的。

在业务数据中，存在类别属性确定的业务数据，也即通过这些数据的部分业务指

标的数值可以直接确定其所属的业务类别。这类数据往往在某个或某些指标上的数值过大或过小，从而被判定为离群点。例如，某用户 **PV** 很小，但是订单总量很高，同时订单总额相对很低，即该用户经常进行购买，但是购买的商品为价格较低的日用快消品；另一用户的 **PV** 较高，同时持续 **PV** 时间很长，但相较于其他用户的订单总量为均值或者偏小。从数据来看，上述两位用户都存在某一指标极大或极小的情况。按照常规聚类方法，如果直接对包含上述两个用户的业务数据进行聚类，由于极端值的存在会使聚类效果较差；如果去除离群点后再聚类，虽然聚类效果较好，但是被去除的数据无法体现其特有的价值，影响业务数据分类的准确性。因此，本发明打破常规，将类别属性确定的部分业务数据提取出来，并用于参与后续的业务数据分类。

步骤 **S106**，将未被提取的业务数据作为第二数据集合，对第二数据集合中的业务数据进行聚类。

形成第二数据集合的一种实现方式可以如下所示。定义原始的所有业务数据组成数据集 **O**，第一数据集合为 **D**，则第二数据集合可以通过以下 **SQL** 语句获得：

```
SELECT  *
FROM    O
WHERE NOT EXISTS( SELECT NULL FROM D
                  WHERE O.ID = D.ID )
```

其中，针对第二数据集合的聚类方法例如可以采用 **Kmeans**、**Brich**、**Optics** 等聚类算法。以 **Kmeans** 算法为例，聚类过程如下：

- 1、随机选择 **k** 个类的初始中心，其中，**k** 的值为确定的实际聚类个数。
- 2、对所有数据点，计算其到 **k** 个中心的距离，将该数据点归属到距离最短的中心所在类。
- 3、更新各个类的中心点。
- 4、判断是否达到设定的收敛条件（或称停止条件），如果不满足，返回到 2-3 步骤进行继续迭代；如果满足，则停止迭代，该聚类中心为最优聚类中心，聚类结果为最终的聚类结果。

步骤 **S108**，根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果。

第一数据集合可以包括一个或多个类。

当第一数据集合包括一个类时，第一数据集合中的所有业务数据作为一个整体，

不再有更进一步的划分。此时，第一数据集中的业务数据具有相同的类别属性，例如其中均为活跃度高的业务数据或者均为活跃度低业务数据。

当第一数据集合包括多个类时，同一个类中的业务数据具有相同的类别属性。第一数据集合的类有不同的划分和形成方法。例如，第一数据集合中的类可以根据业务需要进行划分的，或者，分别根据设置的不同的提取规则从业务数据中提取部分业务数据以形成不同的类，各个不同的类形成第一数据集合。

例如，第一数据集合中可以包括类 A、类 B 和类 C，三个类是分别根据不同的提取规则提取的。类 A 是订单数量按照由大到小排序时位于前 5% 的业务数据，类 B 是订单数量按照由大到小排序时位于后 5% 的业务数据，类 C 是收藏商品数量大于 200 并且收藏店铺数量大于 150 的数据。显然，类 A 和类 C 中的类别属性是活跃度较高的，类 B 的类别属性是活跃度较低的。从而，可以保留根据各个提取规则所获取的类的特点，在后续步骤中能够进行有针对性的合并。本领域技术人员应当清楚，提取规则并不限于上述几种规则。根据需要，可以采用其他的提取规则，这里不再赘述。

在一个实施例中，可以根据第二数据集合和第一数据集合中类的类别属性近似程度，和/或，业务数据分类之间的业务数据数量的差异程度，对第二数据集合和第一数据集合中的类进行合并，确定出业务数据分类结果。例如，将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。或者，将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并，并使得分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。或者，如果两个集合的各类中业务数据数量差异较小，也可以不合并，将两个集合的各类直接作为最终的业务数据分类结果。

上述实施例通过根据业务数据的部分业务指标设置提取规则，并按照提取规则提取类别属性明确的业务数据，再根据未被提取的业务数据的聚类结果以及类别属性明确的业务数据共同确定业务数据分类的结果，提高了业务数据分类的准确性。

在步骤 S104 中，可以采用多种方法提取类别属性符合预设条件的部分业务数据。下面介绍三种示例性的提取方法。

第一种方法为根据设定的阈值提取某一业务指标数据离群的业务数据。该方法的一个应用例可以为，将业务指标超出预设的上限阈值和/或者低于预设的下限阈值的业务数据提取出来。

例如，连续登录天数是体现用户活跃度的一个十分直观的业务指标，因此，可以根据业务需求设置上限阈值和下限阈值。当用户的连续登录天数大多集中在一个月以内时，可以结合这一现状和业务需求，将连续登录天数超过 90 天的用户提取出来，提取的这些用户显然为活跃度很高的用户。

从而，可以根据与业务目最相关的业务指标，将最直观的高价值业务数据提取出来。

第二种方法为根据业务数据的多个业务指标的逻辑运算结果提取出逻辑运算结果离群的业务数据。该方法的一个应用例可以为，将逻辑运算结果超出预设的上限阈值和/或者低于预设的下限阈值的业务数据提取出来。

以订单总额和订单数量这两个业务指标为例，虽然它们可以分别体现用户的活跃度，然而，订单总额和订单数量之间的关系也可以反映用户的活跃度，例如可以计算各个业务数据中订单总额和订单数量的比值，即计算每个用户的平均订单单价，如果平均订单单价非常高，例如超过了 5 万元，即可以将用户划分到活跃度高的类别。即，根据所有业务数据中订单总额和订单数量的比值是否离群，决定是否提取离群的比值所对应的业务数据。

这种方法考虑了指标间的运算关系，可以更灵活地提取数据，扩展了提取规则的设置范围。

第三种方法为根据某一业务指标的数据分布情况，提取出某一业务指标数据离群的业务数据。

例如，可以计算同一业务指标上所有数据的均值和方差，将以均值为中心的预设浮动范围之外的业务指标数据确定为离群的业务指标数据，并提取出业务指标数据离群的业务数据，其中的预设浮动范围可以根据方差的预设倍数确定。例如，计算连续登录天数这一指标中所有数据的均值和方差，将连续登录天数大于均值+2*方差以及小于均值-2*方差的数据提取出来。

例如，还可以将同一业务指标上所有数据按照大小顺序依次排列，提取出业务指标数据在预设上分位数上侧和/或在预设下分位数下侧的业务数据。例如，将大于 95 分位数或小于 5 分位数的数据提取出来，即将同一业务指标的所有数据中最小的 5% 和最大的 5% 的数据提取出来。

这种方法通过业务数据的分布特性筛选出数值极大或极小的业务指标，适用于难

以根据业务情况设置具体数值阈值的应用场景。

在步骤 S108 中例如可以采用以下方法确定业务数据分类结果。

以下为如何根据业务指标选取第二数据集中的类与第一数据集中的类进行合并的一个应用例：首先，确定第一数据集合和第二数据集合中提取规则中涉及的业务指标的平均特征，平均特征是每个类的提取规则涉及的业务指标的平均值或中心点；然后，将第一数据集合和第二数据集合中平均特征最接近的类进行合并获得业务数据分类结果。

例如，第一数据集合中的类 D 是根据其中的订单总额大于 30 万元提取的。因此，在合并时将第二数据集合中的各个类按照订单总额的平均值或者中心点进行由大到小的排序，排序结果中最大的类即为可以与类 D 进行合并的类。该方法同样适用于提取规则涉及多个指标时所提取的类的合并，例如第一数据集合中的类 E 是根据订单总额指标除以订单数量指标大于 5 万确定的，则第二数据集合中的类的平均特征为类中各个业务数据的订单总额指标除以订单数量的计算结果的平均值或者中心点。

通过计算各个类的平均特征，能够客观地确定与待合并的类在类别属性上最相近的类，从而提高了合并的准确性。

以下方法为决定是否进行类之间的合并的一个应用例：将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。即，如果第一数据集合中各个类的业务数据的数量以及第二数据集合中各个类的业务数据的数量之间的差异没有超出预设范围，则无需进行合并；如果超过了预设范围，则合并类别属性最接近的类。

上述合并操作的条件不仅适用于第一数据集合中的类与第二数据集合中的类之间，还可以用于第一数据集合本身的类之间，即，如果第一数据集合中的各个类数量均远小于第二数据集合中的类，可以将第一数据集合中具有同样类别属性的类进行合并，以使结果的数据均匀。

通过类之间数据量的差异决定是否进行合并，能够使业务分类的结果更均匀，应用性更好。

本发明还提供了对第二数据集合中的业务数据进行聚类的方法。

图 2 为业务数据聚类方法的一个实施例的流程图。如图 2 所示，该实施例的方法包括：

步骤 S2062, 对第二数据集中的业务数据进行预处理。

其中, 预处理可以包括离群点处理、空值处理和标准化处理中的一个或多个。

离群点处理过程的一个应用例如下所示: 从第二数据集中筛选出离群的业务指标数据, 将用于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据。例如, 可以将大于同一业务指标中所有数据的均值+方差的业务指标赋予均值+方差, 小于同一业务指标中所有数据的均值-方差的业务指标赋予均值-方差; 还可以将大于同一业务指标中所有数据的上四分位数的业务指标赋予上四分位数, 小于同一业务指标中所有数据的下四分位数的业务指标赋予下四分位数。此外, 还可以将大于同一业务指标中所有数据的均值+方差的业务指标赋予上四分位数, 小于同一业务指标中所有数据的均值-方差的业务指标赋予下四分位数。

由于在进行业务数据聚类之前已经将业务类别属性确定的数据提取出来, 因此, 此处进行预处理的数据为数值较均匀的数据, 对其进行离群点处理会产生更好的聚类效果, 而不会影响业务数据分类的准确性。

空值处理过程的一个应用例如下所示: 计算第二数据集中某一业务指标中所有非空数据的均值, 将均值赋予第二数据集中该业务指标的数据空值。从而, 使具有空值的数据在同一指标的数据中处于平均水平, 以提高聚类的准确性。

对于完成离群点处理、空值处理等操作的数据, 还可以对其进行标准化或者规范化处理。标准化处理的一个应用例为, 对于同一业务指标, 可以首先计算该业务指标中所有数据的均值和方差, 再将其中的每个业务指标的原始数值替换为(原始数值-均值)/方差, 从而统一参与聚类的业务数据的各个业务指标的权重。

根据需要, 本领域技术人员还可以采用其他数据预处理的方法, 这里不再赘述。

步骤 S2064, 确定聚类的实际聚类个数。

在进行聚类之前, 可以根据业务需求人工指定实际聚类的个数, 也可以采用以下预聚类的方法确定实际聚类个数。通过预聚类方法确定聚类个数的一个应用例如下所示:

1. 按照各个预定聚类个数分别对第二数据集中的业务数据进行预聚类, 计算每个预定聚类个数对应的预聚类结果的轮廓系数;

设参与聚类的业务数据总数为 N 。当采用 n 作为预定聚类个数时, 聚类结果的轮廓系数为 $f(n)$, 聚类结果中第 i 个数据点的轮廓系数为 S_i , 聚类结果和数据点的轮廓系

数的计算方法分别如公式(1)和公式(2)所示:

$$f(n) = \sum S_i / N \quad (1)$$

$$S_i = (b_i - a_i) / \max(a_i, b_i) \quad (2)$$

其中, a_i 为第 i 个业务数据到类内每个业务数据的距离的平均值; 对于 b_i , 首先求第 i 个业务数据到不包含该业务数据的各个类中每个业务数据的距离的平均值, 将各个平均值的最小值作为 b_i 。

2. 按照预定聚类个数递增的顺序排列相应的轮廓系数, 获取轮廓系数中的若干极大值, 并确定其中的最大值;

上述步骤可以通过坐标系较为直观地表述出来。当横坐标为预定聚类个数, 纵坐标为轮廓系数时, 将各个聚类结果对应的数据点依次连接, 纵坐标大于相邻两点的数据点对应的轮廓系数即为极大值, 各个极大值中的最大值也是所有数据点中的极大值。

3. 将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数, 预设条件为极大值与最大值的差值小于预设值。

轮廓系数用于衡量聚类结果中各个类本身的凝聚度以及不同类之间的分离度, 因此通常将轮廓系数的最大值对应的聚类个数作为实际聚类个数。然而, 对于业务上的分类, 在满足了轮廓系数较大的情况下, 还需要使聚类个数较小, 以免聚类个数过多不利于业务分类结果的呈现。因此, 本发明选取了在特定阈值以上的若干极大值, 例如大于最大值减去 0.1 后所得值的极大值, 再将符合上述条件的若干极大值所对应的最小的聚类个数作为实际聚类个数。

确定实际聚类个数的一个应用例为: 首先判断第一个极大值是否为最大值, 如果是, 则将最大值对应的聚类个数作为实际聚类个数; 如果第一个极大值不是最大值, 则将首次出现的与所述最大值的差距小于预设值的极大值对应的聚类个数作为实际聚类个数。

步骤 S2066, 采用实际聚类个数对第二数据集合进行聚类。

通过采用上述方法, 优化了参与聚类的业务数据以及聚类方法, 使得聚类的结果更准确, 从而提高了业务数据分类的准确度。

图 3 为本发明业务数据分类装置的一个实施例的结构图。如图 3 所示, 该实施例的装

置包括：业务数据获取模块 32，用于获取业务数据，业务数据包括多个业务指标；业务数据提取模块 34，用于根据设置的提取规则从业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，提取规则是根据部分业务指标设置的；业务数据聚类模块 36，用于将未被提取的业务数据作为第二数据集合，对第二数据集合中的业务数据进行聚类；业务数据分类模块 38，用于根据第二数据集合的聚类结果和第一数据集合确定业务数据分类结果。

通过根据业务数据的部分业务指标设置提取规则，并按照提取规则提取类别属性明确的业务数据，再根据未被提取的业务数据的聚类结果以及类别属性明确的业务数据共同确定业务数据分类的结果，提高了业务数据分类的准确性。

其中，第一数据集合可以包括一个或多个类，业务数据分类模块 38 用于将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。从而，能够提高业务分类结果的准确性。

其中，业务数据提取模块 34 可以用于分别根据设置的不同的提取规则从业务数据中提取部分业务数据以形成不同的类，不同的类形成第一数据集合。从而，可以保留根据各个提取规则所获取的类的特点，在后续步骤中能够进行有针对性的合并。

图 4 为本发明业务数据分类装置的另一个实施例的结构图。如图 4 所示，该实施例的业务数据提取模块 34 可以包括第一提取单元 442、第二提取单元 444 和第三提取单元 446 中的至少一个。

其中，第一提取单元 442 用于根据设定的阈值提取某一业务指标数据离群的业务数据，形成第一数据集合，从而可以根据与业务目最相关的业务指标，将最直观的高价值业务数据提取出来。

第二提取单元 444 用于根据业务数据的多个业务指标的逻辑运算结果提取出逻辑运算结果离群的业务数据，形成第一数据集合，第二提取单元 444 考虑了指标间的运算关系，可以更灵活地提取数据，扩展了提取规则的设置范围。

第三提取单元 446 用于根据某一业务指标的数据分布情况，提取出某一业务指标数据离群的业务数据，形成第一数据集合，适用于难以根据业务情况设置具体数值阈值的应用场景。。

业务数据分类模块 38 可以包括平均特征获取单元 482 和合并单元 484。平均特征获取单元 482 用于确定第一数据集合和第二数据集合中提取规则中涉及的业务指标的

平均特征，其中，平均特征是每个类的提取规则涉及的业务指标的平均值或中心点；合并单元 484 用于将第一数据集合和第二数据集合中平均特征最接近的类进行合并获得业务数据分类结果。通过计算各个类的平均特征，能够客观地确定与待合并的类在类别属性上最相近的类，从而提高了合并的准确性。

或者，业务数据分类模块 38 也可以用于将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。通过类之间数据的差异决定是否进行合并，能够使业务分类的结果更均匀，应用性更好。

该装置还可以包括预处理模块 45，预处理模块 45 包括离群点处理单元 452 和/或空值处理单元 454。

离群点处理单元 452 用于从第二数据集合中筛选出离群的业务指标数据，将用于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据。由于在进行业务数据聚类之前已经将业务类别属性确定的数据提取出来，因此，此处进行预处理的数据为数值较均匀的数据，对其进行离群点处理会产生更好的聚类效果，而不会影响业务数据分类的准确性。

空值处理单元 454 用于计算第二数据集合中某一业务指标中所有非空数据的均值，将均值赋予第二数据集合中该业务指标的数据空值。从而，使具有空值的数据在同一指标的数据中处于平均水平，以提高聚类的准确性。

其中，业务数据聚类模块 36 可以包括预聚类单元 462、轮廓系数计算单元 464、实际聚类个数确定单元 466、实际聚类单元 468。预聚类单元 462 用于按照各个预定聚类个数分别对第二数据集合中的业务数据进行预聚类；轮廓系数计算单元 464 用于根据于聚类结果计算每个预定聚类个数对应的预聚类结果的轮廓系数；实际聚类个数确定单元 466 用于按照预定聚类个数递增的顺序排列相应的轮廓系数，获取轮廓系数中的若干极大值，并确定其中的最大值，将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数，预设条件为极大值与最大值的差值小于预设值；实际聚类单元 468 用于采用实际聚类个数对第二数据集合中的业务数据进行聚类。

通过采用上述方法，可以使聚类结果既具有较好的数学特性，也有较好的可用性。

图 5 为本发明业务数据分类装置的又一个实施例的结构图。如图 5 所示，该实施例的装置 500 包括：存储器 510 以及耦接至该存储器 510 的处理器 520，处理器 520 被配

置为基于存储在存储器 510 中的指令，执行前述任意一个实施例中的业务数据分类方法。

其中，存储器 510 例如可以包括系统存储器、固定非易失性存储介质等。系统存储器例如存储有操作系统、应用程序、引导装载程序 (Boot Loader) 以及其他程序等。

图 6 为本发明业务数据分类装置的再一个实施例的结构图。如图 6 所示，该实施例的装置 500 包括：存储器 510 以及处理器 520，还可以包括输入输出接口 630、网络接口 640、存储接口 650 等。这些接口 630，640，650 以及存储器 510 和处理器 520 之间例如可以通过总线 660 连接。其中，输入输出接口 630 为显示器、鼠标、键盘、触摸屏等输入输出设备提供连接接口。网络接口 640 为各种联网设备提供连接接口。存储接口 650 为 SD 卡、U 盘等外置存储设备提供连接接口。

本发明的实施例还提供一种计算机非瞬时性可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现前述任意一种业务数据分类方法。

本领域内的技术人员应当明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用非瞬时性存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解为可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计

算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

权 利 要 求

1. 一种业务数据分类方法，其特征在于，包括：

获取业务数据，所述业务数据包括多个业务指标；

根据设置的提取规则从所述业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，所述提取规则是根据部分业务指标设置的；

将未被提取的业务数据作为第二数据集合，对所述第二数据集合中的业务数据进行聚类；

根据所述第二数据集合的聚类结果和所述第一数据集合确定业务数据分类结果。

2. 根据权利要求 1 所述的方法，其特征在于，所述根据设置的提取规则从所述业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合包括：

根据设定的阈值提取某一业务指标数据离群的业务数据，形成第一数据集合；

或者，

根据业务数据的多个业务指标的逻辑运算结果提取出所述逻辑运算结果离群的业务数据，形成第一数据集合；

或者，

根据某一业务指标的数据分布情况，提取出所述某一业务指标数据离群的业务数据，形成第一数据集合。

3. 根据权利要求 1 所述的方法，其特征在于，所述根据所述第二数据集合的聚类结果和所述第一数据集合确定业务数据分类结果包括：

将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。

4. 根据权利要求 3 所述的方法，其特征在于，所述将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并包括：

确定第一数据集合和第二数据集合中所述提取规则中涉及的业务指标的平均特征，将第一数据集合和第二数据集合中所述平均特征最接近的类进行合并获得业务数据分类结果，其中，所述平均特征是每个类的所述提取规则涉及的业务指标的平均值或中心点；

或者，

将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得所述分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。

5. 根据权利要求 1 所述的方法，其特征在于，所述根据设置的提取规则从所述业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合包括：

分别根据设置的不同的提取规则从所述业务数据中提取部分业务数据以形成不同的类，所述不同的类形成第一数据集合。

6. 根据权利要求 1 所述的方法，其特征在于，在对所述第二数据集合中的业务数据进行聚类之前，所述方法还包括：

从所述第二数据集合中筛选出离群的业务指标数据，将用于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据；

或者，

计算第二数据集合中某一业务指标中所有非空数据的均值，将均值赋予第二数据集合中该业务指标的空值。

7. 根据权利要求 1 所述的方法，其特征在于，所述对所述第二数据集合中的业务数据进行聚类包括：

按照各个预定聚类个数分别对所述第二数据集合中的业务数据进行预聚类，计算每个预定聚类个数对应的预聚类结果的轮廓系数；

按照预定聚类个数递增的顺序排列相应的轮廓系数，获取所述轮廓系数中的若干极大值，并确定其中的最大值；

将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数，所述预设条件为极大值与所述最大值的差值小于预设值；

采用所述实际聚类个数对第二数据集合中的业务数据进行聚类。

8. 一种业务数据分类装置，其特征在于，包括：

业务数据获取模块，用于获取业务数据，所述业务数据包括多个业务指标；

业务数据提取模块，用于根据设置的提取规则从所述业务数据中提取类别属性符合预设条件的部分业务数据形成第一数据集合，所述提取规则是根据部分业务指标设置的；

业务数据聚类模块，用于将未被提取的业务数据作为第二数据集合，对所述第二

数据集合中的业务数据进行聚类；

业务数据分类模块，用于根据所述第二数据集合的聚类结果和所述第一数据集合确定业务数据分类结果。

9. 根据权利要求 8 所述的装置，其特征在于，所述业务数据提取模块包括以下至少一个单元：

第一提取单元，用于根据设定的阈值提取某一业务指标数据离群的业务数据，形成第一数据集合；

第二提取单元，用于根据业务数据的多个业务指标的逻辑运算结果提取出所述逻辑运算结果离群的业务数据，形成第一数据集合；

第三提取单元，用于根据某一业务指标的数据分布情况，提取出所述某一业务指标数据离群的业务数据，形成第一数据集合。

10. 根据权利要求 8 所述的装置，其特征在于，所述业务数据分类模块用于将第一数据集合包括的一个或多个类和第二数据集合中的各个类中类别属性最接近的类进行合并，获得业务数据分类结果。

11. 根据权利要求 10 所述的装置，其特征在于，

所述业务数据分类模块包括平均特征获取单元和合并单元；所述平均特征获取单元用于确定第一数据集合和第二数据集合中所述提取规则中涉及的业务指标的平均特征，其中，所述平均特征是每个类的所述提取规则涉及的业务指标的平均值或中心点；所述合并单元用于将第一数据集合和第二数据集合中所述平均特征最接近的类进行合并并获得业务数据分类结果；

或者，所述业务数据分类模块用于将第一数据集合中的各个类和第二数据集合中的各个类中类别属性最接近的类进行合并获得业务数据分类结果，并使得所述分类结果中的各个业务数据分类之间的业务数据的数量差异符合预设范围。

12. 根据权利要求 8 所述的装置，其特征在于，所述业务数据提取模块用于分别根据设置的不同的提取规则从所述业务数据中提取部分业务数据以形成不同的类，所述不同的类形成第一数据集合。

13. 根据权利要求 8 所述的装置，其特征在于，还包括预处理模块，所述预处理模块包括：

离群点处理单元，用于从所述第二数据集合中筛选出离群的业务指标数据，将用

于确定业务指标数据是否离群的业务指标分界值赋予离群的业务指标数据；和/或，

空值处理单元，用于计算第二数据集合中某一业务指标中所有非空数据的均值，将均值赋予第二数据集合中该业务指标的数据空值。

14. 根据权利要求 8 所述的装置，其特征在于，所述业务数据聚类模块包括：

预聚类单元，用于按照各个预定聚类个数分别对所述第二数据集合中的业务数据进行预聚类；

轮廓系数计算单元，用于根据于聚类结果计算每个预定聚类个数对应的预聚类结果的轮廓系数；

实际聚类个数确定单元，用于按照预定聚类个数递增的顺序排列相应的轮廓系数，获取所述轮廓系数中的若干极大值，并确定其中的最大值，将首次出现的符合预设条件的极大值对应的预聚类个数作为实际聚类个数，所述预设条件为极大值与所述最大值的差值小于预设值；

实际聚类单元，用于采用所述实际聚类个数对第二数据集合中的业务数据进行聚类。

15. 一种业务数据分类装置，其特征在于，包括：

存储器；以及

耦接至所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行如权利要求 1-7 中任一项所述的业务数据分类方法。

16. 一种计算机非瞬时性可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求 1-7 中任一项所述的业务数据分类方法。

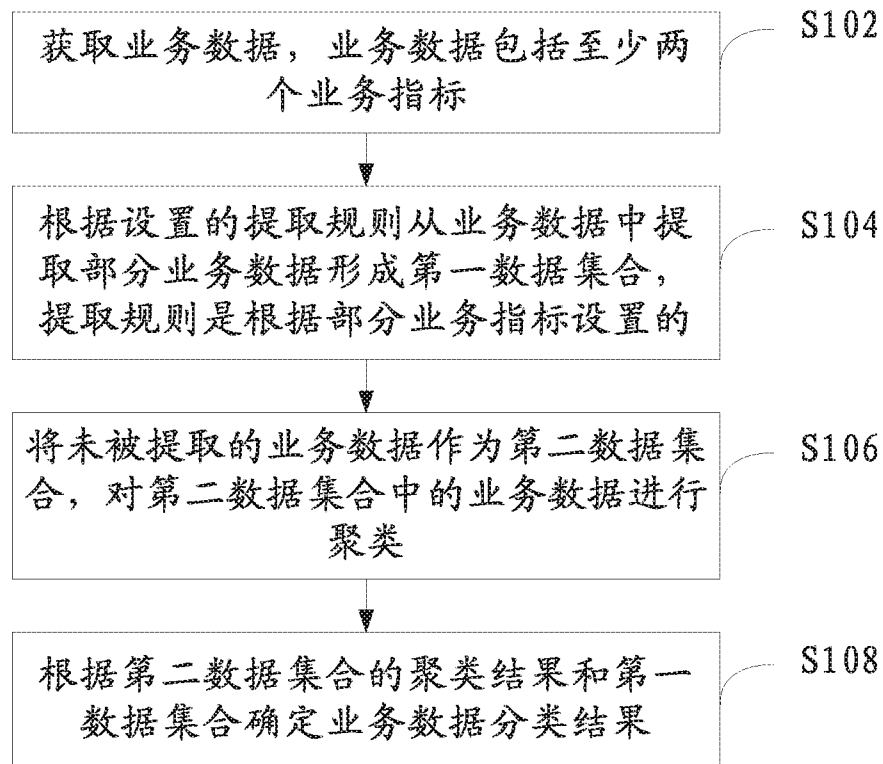


图 1

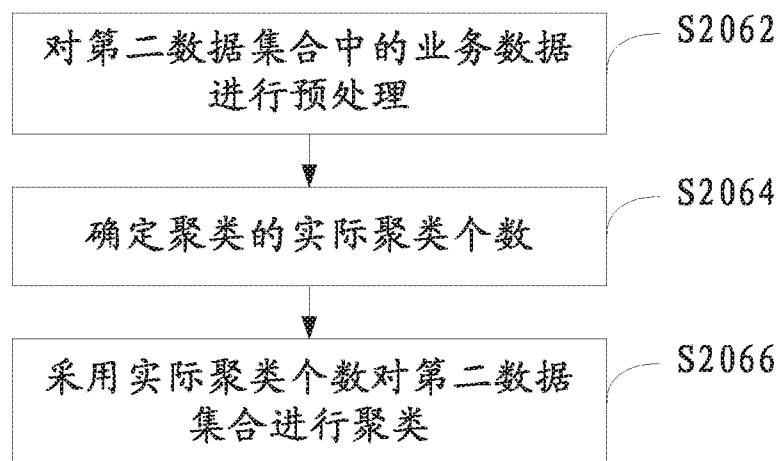


图 2



图 3

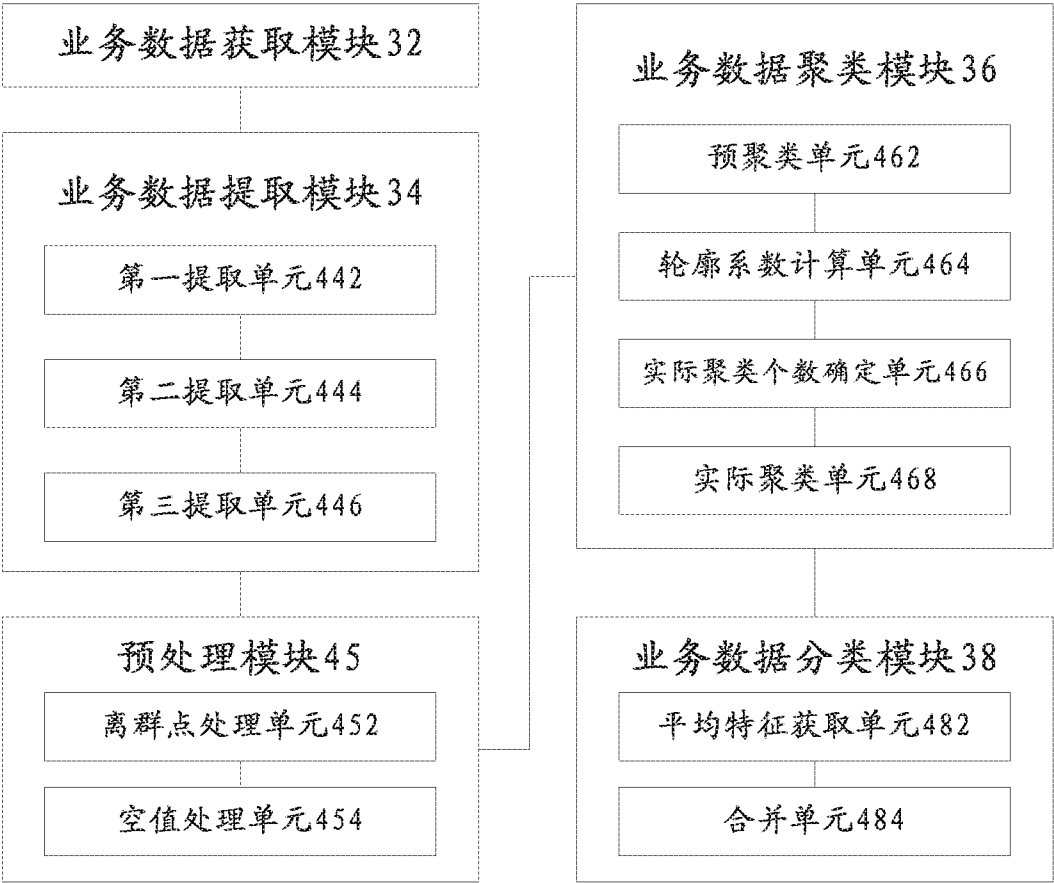


图 4

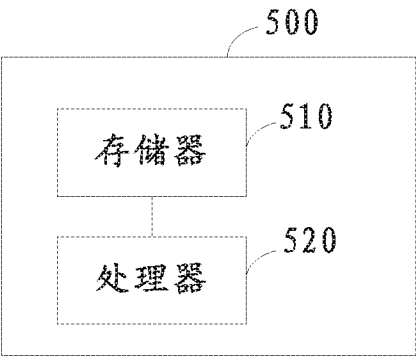


图 5

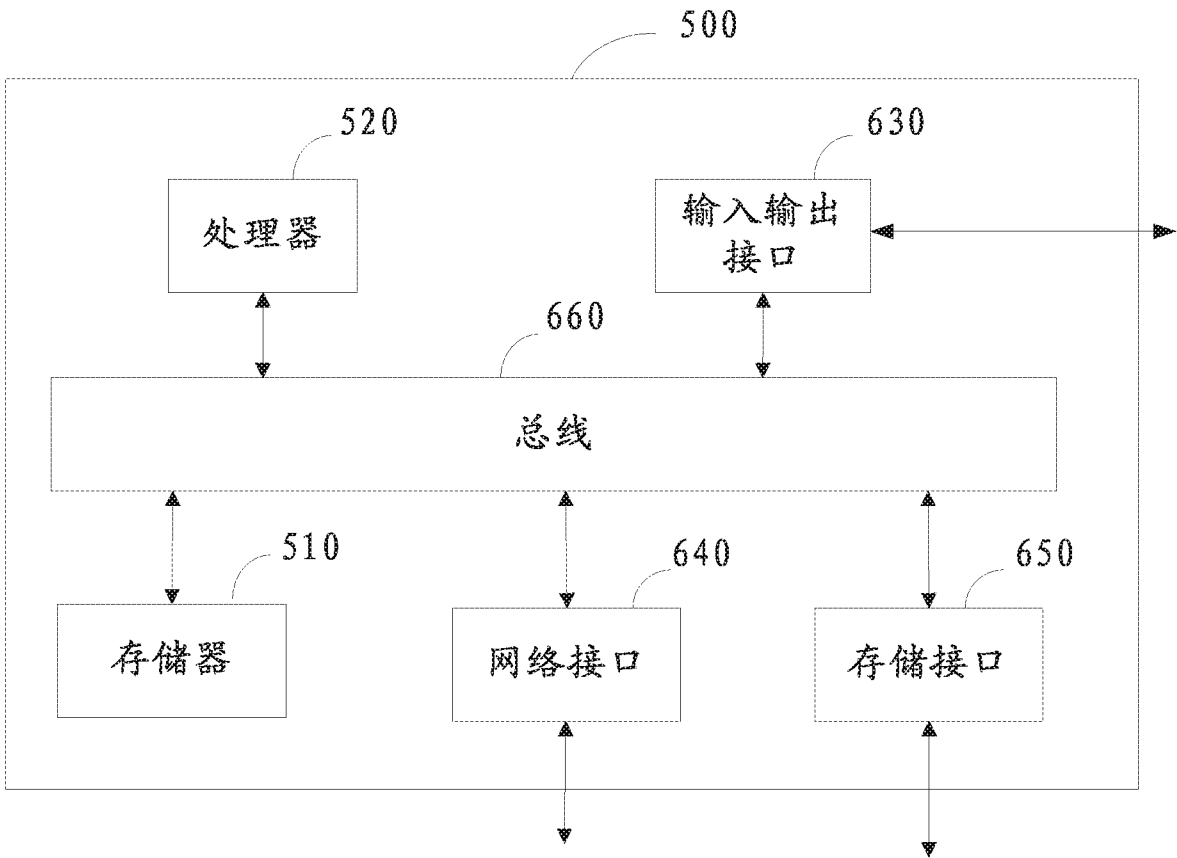


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/081387

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62 (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)

G06K; 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)

WPI, EPODOC, CNPAT, CNKI, IEEE: business data, delete, collect, correct, merge; class, cluster, outlier, classify, categorize, parameter, add

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106156791 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. et al.), 23 November 2016 (23.11.2016), claims 1-15, and description, paragraph [0118]	1-16
Y	CN 104181597 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA), 03 December 2014 (03.12.2014), description, paragraphs [0009]-[0030]	1-16
Y	EP 1455300 A2 (NEC CORPORATION et al.), 08 September 2004 (08.09.2004), description, paragraphs [0048]-[0051]	1-16
A	US 6735550 B1 (UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH), 11 May 2004 (11.05.2004), the whole document	1-16
A	CN 103810261 A (XI'AN UNIVERSITY OF TECHNOLOGY), 21 May 2014 (21.05.2014), the whole document	1-16
A	CN 104699702 A (CHINA UNIONPAY CO., LTD.), 10 June 2015 (10.06.2015), the whole document	1-16

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search
06 July 2017 (06.07.2017)

Date of mailing of the international search report
31 July 2017 (31.07.2017)

Name and mailing address of the ISA/CN:
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No.: (86-10) 62019451

Authorized officer

ZENG, Xuan

Telephone No.: (86-10) **62413976**

International application No.
PCT/CN2017/081387

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2017/081387

A. 主题的分类 G06K 9/62 (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) G06K; 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)) WPI, EPDOC, CNPAT, CNKI, IEEE: 业务数据, 分类, 聚类, 删除, 加入, 提取, 离群, 校正, 合并, 并入; class, cluster, outlier, classify, categorize, parameter, add																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 106156791 A (北京京东尚科信息技术有限公司 等) 2016年 11月 23日 (2016 - 11 - 23) 权利要求1-15, 说明书第[0118]段</td> <td>1-16</td> </tr> <tr> <td>Y</td> <td>CN 104181597 A (电子科技大学) 2014年 12月 3日 (2014 - 12 - 03) 说明书第[0009]-[0030]段</td> <td>1-16</td> </tr> <tr> <td>Y</td> <td>EP 1455300 A2 (NEC CORPORATION 等) 2004年 9月 8日 (2004 - 09 - 08) 说明书第[0048]-[0051]段</td> <td>1-16</td> </tr> <tr> <td>A</td> <td>US 6735550 B1 (UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH) 2004年 5月 11日 (2004 - 05 - 11) 全文</td> <td>1-16</td> </tr> <tr> <td>A</td> <td>CN 103810261 A (西安理工大学) 2014年 5月 21日 (2014 - 05 - 21) 全文</td> <td>1-16</td> </tr> <tr> <td>A</td> <td>CN 104699702 A (中国银联股份有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文</td> <td>1-16</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 106156791 A (北京京东尚科信息技术有限公司 等) 2016年 11月 23日 (2016 - 11 - 23) 权利要求1-15, 说明书第[0118]段	1-16	Y	CN 104181597 A (电子科技大学) 2014年 12月 3日 (2014 - 12 - 03) 说明书第[0009]-[0030]段	1-16	Y	EP 1455300 A2 (NEC CORPORATION 等) 2004年 9月 8日 (2004 - 09 - 08) 说明书第[0048]-[0051]段	1-16	A	US 6735550 B1 (UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH) 2004年 5月 11日 (2004 - 05 - 11) 全文	1-16	A	CN 103810261 A (西安理工大学) 2014年 5月 21日 (2014 - 05 - 21) 全文	1-16	A	CN 104699702 A (中国银联股份有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文	1-16
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 106156791 A (北京京东尚科信息技术有限公司 等) 2016年 11月 23日 (2016 - 11 - 23) 权利要求1-15, 说明书第[0118]段	1-16																					
Y	CN 104181597 A (电子科技大学) 2014年 12月 3日 (2014 - 12 - 03) 说明书第[0009]-[0030]段	1-16																					
Y	EP 1455300 A2 (NEC CORPORATION 等) 2004年 9月 8日 (2004 - 09 - 08) 说明书第[0048]-[0051]段	1-16																					
A	US 6735550 B1 (UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH) 2004年 5月 11日 (2004 - 05 - 11) 全文	1-16																					
A	CN 103810261 A (西安理工大学) 2014年 5月 21日 (2014 - 05 - 21) 全文	1-16																					
A	CN 104699702 A (中国银联股份有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文	1-16																					
☐ 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 2017年 7月 6日		国际检索报告邮寄日期 2017年 7月 31日																					
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 曾璇 电话号码 (86-10)62413976																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/081387

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106156791	A	2016年 11月 23日	无	
CN	104181597	A	2014年 12月 3日	无	
EP	1455300	A2	2004年 9月 8日	US 2004236742 A1	2004年 11月 25日
				JP 2004272350 A	2004年 9月 30日
US	6735550	B1	2004年 5月 11日	无	
CN	103810261	A	2014年 5月 21日	无	
CN	104699702	A	2015年 6月 10日	无	