(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110941663 B
(45)授权公告日 2022.12.23

(21)申请号 201911121294.2

(22)申请日 2019.11.15

(65)同一申请的已公布的文献号
申请公布号 CN 110941663 A

(43)申请公布日 2020.03.31

(73)专利权人 杭州数梦工场科技有限公司
地址 310024 浙江省杭州市转塘科技经济区块16号4幢326室

(72)发明人 徐鹏飞 单军

(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
专利代理师 王茹

(51)Int.Cl.
G06F 16/26(2019.01)

(56)对比文件
CN 110427739 A,2019.11.08
US 2011282856 A1,2011.11.17
CN 110347564 A,2019.10.18
蓝孙科.巧用规则深挖证件号码中的审计疑点.《中国审计》.2019,(第6期),第39页.
屈怀忠等.公民身份号码纠错浅谈.《警察技术》.2008,(第6期),第44-45页.
韩雪涛.身份证号码中的数学.《初中生学习·博闻》.2015,(第10期),第8页.

审查员 初星妍

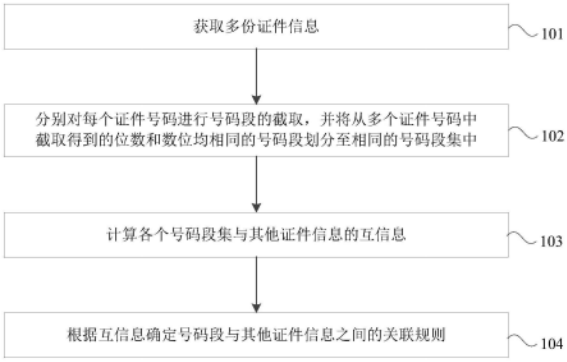
权利要求书2页 说明书8页 附图4页

(54)发明名称

证件信息的关联规则获取方法及装置

(57)摘要

本发明公开了证件信息的关联规则获取方法及装置、电子设备、存储介质。关联规则获取方法包括:获取多份证件信息,每份证件信息包括证件号码和其他证件信息;分别对每个证件号码进行号码段的截取,并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中;计算各个号码段集与所述其他证件信息的互信息;根据所述互信息确定号码段与其他证件信息之间的关联规则。本发明基于互信息确定证件信息的关联规则,能够挖掘出证件号码与其他证件信息之间的潜在规则,以此建立规则库,可为证件鉴伪提供参考。



CN 110941663 B

1.一种证件信息的关联规则获取方法,其特征在于,所述关联规则获取方法包括:

获取多份证件信息,每份证件信息包括证件号码和其他证件信息;

分别对每个证件号码进行号码段的截取,并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中;

分别计算每一号码段集与所述其他证件信息之间的互信息;

根据所述互信息确定号码段与所述其他证件信息之间的关联规则。

2.如权利要求1所述的证件信息的关联规则获取方法,其特征在于,分别计算每一号码段集与其他证件信息之间的互信息,包括:

从多个包含位数相同的号码段的号码段集中选取目标号码段集,所述目标号码段集为与所述其他证件信息的互信息为最大值的号码段集;

将所述目标号码段集中的号码段与其他号码段集中对应的号码段进行组合,并计算组合后的号码段集与所述其他证件信息的互信息。

3.如权利要求1所述的证件信息的关联规则获取方法,其特征在于,分别计算每一号码段集与其他证件信息之间的互信息,包括:

分别计算每个号码段集与所述其他证件信息的互信息;

按照所述互信息由大到小的顺序对所述号码段集进行排序;

选取排序靠前的若干号码段集,并将所述若干号码段集中对应的号码段进行组合;

计算组合后的号码段集与所述其他证件信息的互信息。

4.如权利要求1-3任一项所述的证件信息的关联规则获取方法,其特征在于,计算所述号码段集与其他证件信息的互信息,包括:

统计所述多份证件信息中,所述号码段集中的号码段与对应的证件信息的出现数量,根据所述出现数量确定号码段的权重;

将号码段集中的号码段赋予所述权重后,计算所述号码段集与所述其他证件信息的互信息。

5.如权利要求1所述的证件信息的关联规则获取方法,其特征在于,根据所述互信息确定号码段与其他证件信息之间的关联规则,包括:

计算所述互信息大于互信息阈值的号码段集的置信度;

根据置信度大于置信度阈值的号码段集与对应的其他证件信息确定所述关联规则。

6.如权利要求1所述的证件信息的关联规则获取方法,其特征在于,所述关联规则获取方法还包括:

使用正则表达式表示所述关联规则。

7.如权利要求1所述的证件信息的关联规则获取方法,其特征在于,所述其他证件信息包括以下信息中的至少一项:

证件所属用户的用户信息、证件签发地、证件签发时间、证件有效期、证件签发机关、证件类型。

8.一种证件信息的关联规则获取装置,其特征在于,所述关联规则获取装置包括:

获取模块,用于获取多份证件信息,每份证件信息包括证件号码和其他证件信息;

截取模块,用于分别对每个证件号码进行号码段的截取,并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中;

计算模块，用于分别计算每一号码段集与所述其他证件信息的互信息；

确定模块，用于根据所述互信息确定号码段与所述其他证件信息之间的关联规则。

9.如权利要求8所述的证件信息的关联规则获取装置，其特征在于，所述计算模块具体用于：

从多个包含位数相同的号码段的号码段集中选取目标号码段集，所述目标号码段集为与所述其他证件信息的互信息为最大值的号码段集；

将所述目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与所述其他证件信息的互信息。

10.如权利要求8所述的证件信息的关联规则获取装置，其特征在于，所述计算模块具体用于：

分别计算每个号码段集与所述证件信息的互信息；

按照所述互信息由大到小的顺序对所述号码段集进行排序；

选取排序靠前的若干号码段集，并将所述若干号码段集中对应的号码段进行组合；

计算组合后的号码段集与所述其他证件信息的互信息。

11.如权利要求8-10任一项所述的证件信息的关联规则获取装置，其特征在于，在计算所述号码段集与其他证件信息的互信息时，所述计算模块还用于：

统计所述多份证件信息中，所述号码段集中的号码段与对应的证件信息的出现数量，根据所述出现数量确定号码段的权重；

将号码段集中的号码段赋予所述权重后，计算所述号码段集与所述其他证件信息的互信息。

12.如权利要求8所述的证件信息的关联规则获取装置，其特征在于，所述确定模块具体用于：

计算所述互信息大于互信息阈值的号码段集的置信度；

根据置信度大于置信度阈值的号码段集与对应的其他证件信息确定所述关联规则。

13.如权利要求12所述的证件信息的关联规则获取装置，其特征在于，所述确定模块还用于：

使用正则表达式表示所述关联规则。

14.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述的证件信息的关联规则获取方法。

15.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的证件信息的关联规则获取方法的步骤。

证件信息的关联规则获取方法及装置

技术领域

[0001] 本发明涉及数据挖掘技术领域，特别涉及证件信息的关联规则获取方法及装置、电子设备、存储介质。

背景技术

[0002] 目前，证件鉴伪主要通过业务专家长期工作的实践，总结出证件信息的关联规则，进而基于结构化文本的关联规则逻辑匹配实现。然而，基于人工方式挖掘证件信息的关联规则，不仅需要大量的人力投入，且人工并不能挖掘出除经验之外的潜在规则，会导致因证件信息的关联规则挖掘不全面，影响证件鉴伪的准确性。

发明内容

[0003] 本发明提供一种证件信息的关联规则获取方法及装置、电子设备、存储介质，以挖掘出证件信息之间的潜在关联规则，提高证件鉴伪的准确性。

[0004] 具体地，本发明是通过如下技术方案实现的：

[0005] 第一方面，提供一种证件信息的关联规则获取方法，所述关联规则获取方法包括：

[0006] 获取多份证件信息，每份证件信息包括证件号码和其他证件信息；

[0007] 分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中；

[0008] 分别计算每一号码段集与其他证件信息之间的互信息；

[0009] 根据所述互信息确定号码段与其他证件信息之间的关联规则。

[0010] 可选地，分别计算每一号码段集与其他证件信息之间的互信息，包括：

[0011] 从多个包含位数相同的号码段的号码段集中选取目标号码段集，所述目标号码段集为与其他证件信息的互信息为最大值的号码段集；

[0012] 将所述目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与所述其他证件信息的互信息。

[0013] 可选地，分别计算每一号码段集与其他证件信息之间的互信息，包括：

[0014] 分别计算每个号码段集与所述其他证件信息的互信息；

[0015] 按照所述互信息由大到小的顺序对所述号码段集进行排序；

[0016] 选取排序靠前的若干号码段集，并将所述若干号码段集中对应的号码段进行组合；

[0017] 计算组合后的号码段集与所述其他证件信息的互信息。

[0018] 可选地，计算所述号码段集与其他证件信息的互信息，包括：

[0019] 统计所述多份证件信息中，所述号码段集中的号码段与对应的证件信息的出现数量，根据所述出现数量确定号码段的权重；

[0020] 将号码段集中的号码段赋予所述权重后，计算所述号码段集与所述其他证件信息的互信息；

[0021] 或，分别计算每个号码段集与所述其他证件信息的初始互信息，并将所述初始互信息大于互信息阈值的号码段集与对应的其他证件信息作为正样本，将所述初始互信息小于等于所述互信息阈值的号码段集与对应的其他证件信息作为负样本；

[0022] 根据所述正样本和所述负样本拟合所述号码段的权重；

[0023] 对所述号码段集中的号码段赋予所述权重后，再次计算所述号码段集与所述其他证件信息的互信息。

[0024] 可选地，根据所述互信息确定号码段与其他证件信息之间的关联规则，包括：

[0025] 计算所述互信息大于互信息阈值的号码段集的置信度；

[0026] 根据置信度大于置信度阈值的号码段集与对应的证件信息确定所述关联规则。

[0027] 可选地，所述关联规则获取方法还包括：

[0028] 使用正则表达式表示所述关联规则。

[0029] 可选地，所述其他证件信息包括以下信息中的至少一项：

[0030] 证件所属用户的用户信息、证件签发地、证件签发时间、证件有效期、证件签发机关、证件类型。

[0031] 第二方面，提供一种证件信息的关联规则获取装置，所述关联规则获取装置包括：

[0032] 获取模块，用于获取多份证件信息，每份证件信息包括证件号码和其他证件信息；

[0033] 截取模块，用于分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中；

[0034] 计算模块，用于计算各个号码段集与其他证件信息的互信息；

[0035] 确定模块，用于根据所述互信息确定号码段与其他证件信息之间的关联规则。

[0036] 可选地，所述计算模块具体用于：

[0037] 从多个包含位数相同的号码段的号码段集中选取目标号码段集，所述目标号码段集为与其他证件信息的互信息为最大值的号码段集；

[0038] 将所述目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与所述其他证件信息的互信息。

[0039] 可选地，所述计算模块具体用于：

[0040] 分别计算每个号码段集与所述其他证件信息的互信息；

[0041] 按照所述互信息由大到小的顺序对所述号码段集进行排序；

[0042] 选取排序靠前的若干号码段集，并将所述若干号码段集中对应的号码段进行组合；

[0043] 计算组合后的号码段集与所述其他证件信息的互信息。

[0044] 可选地，在计算所述号码段集与其他证件信息的互信息时，所述计算模块还用于：

[0045] 统计所述多份证件信息中，所述号码段集中的号码段与对应的证件信息的出现数量，根据所述出现数量确定号码段的权重；

[0046] 将号码段集中的号码段赋予所述权重后，计算所述号码段集与所述其他证件信息的互信息；

[0047] 或，分别计算每个号码段集与所述其他证件信息的初始互信息，并将所述初始互信息大于互信息阈值的号码段集与对应的其他证件信息作为正样本，将所述初始互信息小于等于所述互信息阈值的号码段集与对应的其他证件信息作为负样本；

[0048] 根据所述正样本和所述负样本拟合所述号码段的权重；

[0049] 对所述号码段集中的号码段赋予所述权重后，再次计算所述号码段集与所述其他证件信息的互信息。

[0050] 可选地，所述确定模块具体用于：

[0051] 计算所述互信息大于互信息阈值的号码段集的置信度；

[0052] 根据置信度大于置信度阈值的号码段集与对应的证件信息确定所述关联规则。

[0053] 可选地，所述确定模块还用于：

[0054] 使用正则表达式表示所述关联规则。

[0055] 第三方面，提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述任一项所述的证件信息的关联规则获取方法。

[0056] 第四方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述任一项所述的证件信息的关联规则获取方法的步骤。

[0057] 本发明的实施例提供的技术方案可以包括以下有益效果：

[0058] 本发明实施例中，基于互信息确定证件信息的关联规则，能够挖掘出证件号码与其他证件信息之间的潜在规则，以此建立规则库，可为证件鉴伪提供参考。

[0059] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本发明。

附图说明

[0060] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0061] 图1是本发明一示例性实施例示出的一种证件信息的关联规则获取方法的流程图；

[0062] 图2是本发明另一示例性实施例示出的一种证件信息的关联规则获取方法的流程图；

[0063] 图3是本发明另一示例性实施例示出的一种证件信息的关联规则获取方法的流程图；

[0064] 图4是本发明一示例性实施例示出的一种证件信息的关联规则获取的模块示意图；

[0065] 图5是本发明一示例性实施例示出的一种电子设备的结构示意图。

具体实施方式

[0066] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0067] 在本发明使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数

形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0068] 应当理解，尽管在本发明可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本发明范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0069] 图1是本发明一示例性实施例示出的一种证件信息的关联规则获取方法的流程图，该获取方法包括以下步骤：

[0070] 步骤101、获取多份证件信息。

[0071] 其中，每份证件信息包括证件号码和其他证件信息。其他证件信息可以是以下信息中的一项或多项组合：证件所属用户的用户信息(例如，用户性别、名族、出生日期、住址等)、证件签发地、证件签发时间、证件有效期、证件签发机关、证件类型。可以理解地，本实施例中针对每类证件建立对应的证件信息的关联规则，步骤101中获取的是同类证件的多份证件信息。

[0072] 步骤102、分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中。

[0073] 步骤102中，需采用相同截取规则对每个证件号码进行截取，截取规则可以但不限于是，先对证件号码的每个数位进行截取，再分别截取相邻的2位、3位号码段。

[0074] 以下表1示出的证件信息为例，对证件号码的每个数位进行截取，并将截取得到的数位相同的号码段划分至相同的号码段集中，结果为{4,3,3,3,5}、{4,1,3,3,1}、{0,0,0,1,0}、{5,3,7,6,4}、{8,9,9,5,2}、{6,5,7,6,3}；对证件号码的相邻2位进行截取，并将截取得到的数位相同的号码段划分至相同的号码段集中，结果为{44,31,33,33,51}、{40,10,30,31,10}、{05,03,07,16,04}、{58,39,79,65,42}、{86,95,97,56,23}；对证件号码的相邻3位进行截取，并将截取得到的数位相同的号码段划分至相同的号码段集中，结果为{440,310,330,331,510}、{405,103,307,316,104}、{058,039,079,165,042}、{586,395,797,656,423}。

[0075] 表1

[0076]

证件信息	证件号码	证件签发地	证件签发时间	用户性别
证件a	440586	广东省	2000年	男
证件b	310395	上海市	2001年	男
证件c	330797	浙江省	1995年	女
证件d	331656	浙江省	2018年	男
证件e	510423	重庆市	2018年	女

[0077] 步骤103、分别计算每一号码段集与其他证件信息的互信息。

[0078] 步骤103中计算互信息，也即计算步骤102获得的每个号码段集与其他证件信息之间的互信息，计算公式可以但不限于表示如下：

[0079] $$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)};$$

[0080] 其中，I(X;Y)表示互信息;X为号码段集的向量表示;Y为一项证件信息或多项证件信息组合的向量表示;p(x,y)为(x,y)同时出现的概率;p(x)为x在向量X中出现的概率;p(y)为y在向量Y中出现的概率。

[0081] 步骤104、根据互信息确定号码段与其他证件信息之间的关联规则。

[0082] 重复步骤103和步骤104，使证件号码中的每个号码段依次与证件信息中的所有其他证件信息均计算过互信息，通过对证件号码与其他证件信息的全面分析，即可得到号码段与某项证件信息和/或多项证件信息组合的相关性，若计算得到的互信息值比较大，说明该号码段集中的号码段与该项证件信息或该多项证件信息组合相关性较大，则可依据互信息较大的若干号码段与对应的其他证件信息确定证件信息的关联规则。

[0083] 本实施例的证件信息的关联规则获取方法适用各种类型的证件，例如身份证、驾驶证、护照等。本实施例中，基于互信息确定证件信息的关联规则，能够挖掘出各类证件的证件号码与其他证件信息之间的潜在关联规则，以此建立规则库，可为证件鉴伪提供参考。

[0084] 在图1示出的关联规则获取方法的流程图的基础上，图2示出了本发明一示例性实施例的另一种证件信息的关联规则获取方法的流程图，该获取方法包括以下步骤：

[0085] 步骤201、获取多份证件信息。

[0086] 步骤202、分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中。

[0087] 其中，步骤201和步骤202与步骤101和步骤102的具体实现方式类似，此处不再赘诉。

[0088] 步骤203-1、从多个包含位数相同的号码段的号码段集中选取目标号码段集。

[0089] 其中，目标号码段集为与其他证件信息的互信息为最大值的号码段集。

[0090] 步骤203-2、将目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与其他证件信息的互信息。

[0091] 在一个实现方式中，可重复执行步骤203-1和步骤203-2，直至证件号码中所有数位的数字均被截取并计算过互信息或者对组合后的号码段集计算互信息不再增大为止。以下还是以表1示出的证件信息为例，对重复计算互信息的具体实现过程进行说明：

[0092] 从包含位数最少的号码段的号码段集开始，也即分别将以下号码段集{4,3,3,3,5}、{4,1,3,3,1}、{0,0,0,1,0}、{5,3,7,6,4}、{8,9,9,5,2}、{6,5,7,6,3}与每项证件信息(或多项证件信息组合)计算互信息，针对每项证件信息(或证件信息组合)，将互信息为最大值的号码段集确定为本轮迭代的目标号码段集，假设包含位数为1的号码段的号码段集中，{4,3,3,3,5}与证件签发地的互信息为最大值，则将{4,3,3,3,5}确定为本次迭代过程中，针对证件签发地的目标号码段集，并进行下一轮迭代，选择证件号码中其他数位的数字与{4,3,3,3,5}进行组合，得到组合后的号码段集{44,31,33,33,51}、{40,30,30,30,50}、{45,33,37,36,54}、{48,39,39,35,52}、{46,35,37,36,53}，分别将组合后的号码段集与证件签发地该项证件信息计算互信息;重复执行上述步骤，直至证件号码中所有数位的数字均被截取并计算过互信息，或者组合后的号码段集计算互信息不再增大为止(本次迭代的互信息最大值大于下一次迭代的互信息最大值)。步骤204中则根据本次迭代中互信息为最大值的号码段集与对应的证件信息(或多项证件信息组合)确定关联规则。本实施例中，计算出最大值后只会在最大值的基础上进行号码段的组合，例如首次出现的最大互信息的号

码段集中号码段为a，那么组合后该号码段为ab，本实施例中无需计算关联性较小的号码段集与证件信息的互信息，可以提高计算的效率。

[0093] 在另一个实现方式中，互信息计算时，可加入权重。例如，统计多份证件信息中，号码段集中的号码段与对应的证件信息的出现数量，并根据出现数量确定号码段的权重。还是以表1为例，号码段集{44,31,33,33,51}中的各项元素44、31、33、51与证件签发地中的各元素广东省、上海市、浙江省、重庆市的出现数量分别为1、1、2和1，可将上述出现数量作为权重，并将号码段集中的号码段赋予对应的权重后，计算号码段集与其他证件信息的互信息，互信息计算公式可以被修改为：

[0094] $$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}n(x,y)p(x,y)\log\frac{p(x,y)}{p(x)p(y)};$$

[0095] 其中，n(x,y)表示(x,y)同时出现的出现数量。

[0096] 在另一个实现方式中，权重可以通过拟合得到，具体的：分别计算每个号码段集与所述其他证件信息的初始互信息，并将所述初始互信息大于互信息阈值的号码段集与对应的其他证件信息作为正样本，将所述初始互信息小于等于所述互信息阈值的号码段集与对应的其他证件信息作为负样本；根据所述正样本和所述负样本拟合所述号码段的权重。对所述号码段集中的号码段赋予所述权重后，再次计算所述号码段集与所述其他证件信息的互信息。

[0097] 步骤204、根据互信息确定号码段与其他证件信息之间的关联规则。

[0098] 在一个实现中，步骤204具体包括：计算互信息大于互信息阈值的号码段集的置信度，并根据置信度大于置信度阈值的号码段集与对应的证件信息确定关联规则。

[0099] 在另一个实现方式中，还可以使用正则表达式表示关联规则。

[0100] 本实施例中，通过互信息的迭代计算，可以进一步提高关联规则挖掘的准确性。进一步地，可使用本实施例的证件信息的关联规则获取方法，对不同国家、不同类型的证件信息进行关联规则挖掘，形成统一的规则库，有助于提升对不同类型的证照鉴伪工作的效率。

[0101] 图3示出了本发明一示例性实施例的另一种证件信息的关联规则获取方法的流程图，本实施例的关联规则获取方法与图2示出的关联规则获取方法基本相同，不同之处在于，本实施例中计算各个号码段集与其他证件信息的互信息的具体实现方式与图2示出的不同，参见图3，针对其他证件信息中的每项证件信息或多项证件信息组合，计算各个号码段集与其他证件信息的互信息的步骤具体包括：

[0102] 步骤303-1、分别计算每个号码段集与证件信息的互信息。

[0103] 步骤303-2、按照互信息由大到小的顺序对号码段集进行排序。

[0104] 步骤303-3、选取排序靠前的若干号码段集，并将若干号码段集中对应的号码段进行组合。

[0105] 其中，若干号码段集的数量可以根据实际需求自行选择，例如选择2和或者3个。

[0106] 步骤303-4、计算组合后的号码段集与其他证件信息的互信息。

[0107] 步骤304中，则根据步骤303-1和步骤303-4中计算的互信息大于互信息阈值的号码段集和对应的证件信息确定关联规则。

[0108] 本实施例中，将排序靠前的若干号码段集中对应的号码段进行组合，并计算互信息，对于本身互信息值较小的号码段集不再进行组合计算互信息，从而可以提高计算的效

率。

[0109] 在一种实现方式中，步骤303-1和/或步骤303-4中计算互信息时，也可加入权重，具体实现方式与步骤203-2的实现方式类似，此处不再赘诉。

[0110] 与前述证件信息的关联规则获取方法实施例相对应，本发明还提供了证件信息的关联规则获取装置的实施例。

[0111] 图4示出了本发明一示例性实施例的一种证件信息的关联规则获取装置的模块示意图，该关联规则获取装置包括：获取模块41、截取模块42、计算模块43和确定模块44。

[0112] 获取模块41用于获取多份证件信息，每份证件信息包括证件号码和其他证件信息；

[0113] 截取模块42用于分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中；

[0114] 计算模块43用于计算各个号码段集与其他证件信息的互信息；

[0115] 确定模块44用于根据所述互信息确定号码段与其他证件信息之间的关联规则。

[0116] 可选地，所述计算模块具体用于：

[0117] 从多个包含位数相同的号码段的号码段集中选取目标号码段集，所述目标号码段集为与其他证件信息中的某一项证件信息或多项证件信息组合的互信息为最大值的号码段集；

[0118] 将所述目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与所述某一项证件信息或多项证件信息组合的互信息。

[0119] 可选地，所述计算模块具体用于：

[0120] 分别计算每个号码段集与所述证件信息的互信息；

[0121] 按照所述互信息由大到小的顺序对所述号码段集进行排序；

[0122] 选取排序靠前的若干号码段集，并将所述若干号码段集中对应的号码段进行组合；

[0123] 计算组合后的号码段集与所述证件信息的互信息。

[0124] 可选地，在计算所述号码段集与其他证件信息的互信息时，所述计算模块还用于：

[0125] 统计所述多份证件信息中，所述号码段集中的号码段与对应的证件信息的出现数量，根据所述出现数量确定号码段的权重；

[0126] 将号码段集中的号码段赋予所述权重后，计算所述号码段集与所述其他证件信息的互信息；

[0127] 或，分别计算每个号码段集与所述其他证件信息的初始互信息，并将所述初始互信息大于互信息阈值的号码段集与对应的其他证件信息作为正样本，将所述初始互信息小于等于所述互信息阈值的号码段集与对应的其他证件信息作为负样本；

[0128] 根据所述正样本和所述负样本拟合所述号码段的权重；

[0129] 对所述号码段集中的号码段赋予所述权重后，再次计算所述号码段集与所述其他证件信息的互信息。

[0130] 可选地，所述确定模块具体用于：

[0131] 计算所述互信息大于互信息阈值的号码段集的置信度；

[0132] 根据置信度大于置信度阈值的号码段集与对应的证件信息确定所述关联规则。

[0133] 可选地，所述确定模块还用于：

[0134] 使用正则表达式表示所述关联规则。

[0135] 图5为本发明实施例提供的一种电子设备的结构示意图，示出了适于用来实现本发明实施方式的示例性电子设备50的框图。图5显示的电子设备50仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0136] 如图5所示，电子设备50可以以通用计算设备的形式表现，例如其可以为服务器设备。电子设备50的组件可以包括但不限于：上述至少一个处理器51、上述至少一个存储器52、连接不同系统组件(包括存储器52和处理器51)的总线53。

[0137] 总线53包括数据总线、地址总线和控制总线。

[0138] 存储器52可以包括易失性存储器，例如随机存取存储器(RAM)521和/或高速缓存存储器522，还可以进一步包括只读存储器(ROM)523。

[0139] 存储器52还可以包括具有一组(至少一个)程序模块524的程序工具525(或实用工具)，这样的程序模块524包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0140] 处理器51通过运行存储在存储器52中的计算机程序，从而执行各种功能应用以及数据处理，例如上述任一实施例提供的方法。

[0141] 电子设备50也可以与一个或多个外部设备54(例如键盘、指向设备等)通信。这种通信可以通过输入/输出(I/O)接口55进行。并且，模型生成的电子设备50还可以通过网络适配器56与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器56通过总线53与模型生成的电子设备50的其它模块通信。应当明白，尽管图中未示出，可以结合模型生成的电子设备50使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID(磁盘阵列)系统、磁带驱动器以及数据备份存储系统等。

[0142] 应当注意，尽管在上文详细描述中提及了电子设备的若干单元/模块或子单元/模块，但是这种划分仅仅是示例性的并非强制性的。实际上，根据本发明的实施方式，上文描述的两个或更多单元/模块的特征和功能可以在一个单元/模块中具体化。反之，上文描述的一个单元/模块的特征和功能可以进一步划分为由多个单元/模块来具体化。

[0143] 本发明实施例还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述任一项所述的证件信息的关联规则获取方法的步骤。

[0144] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

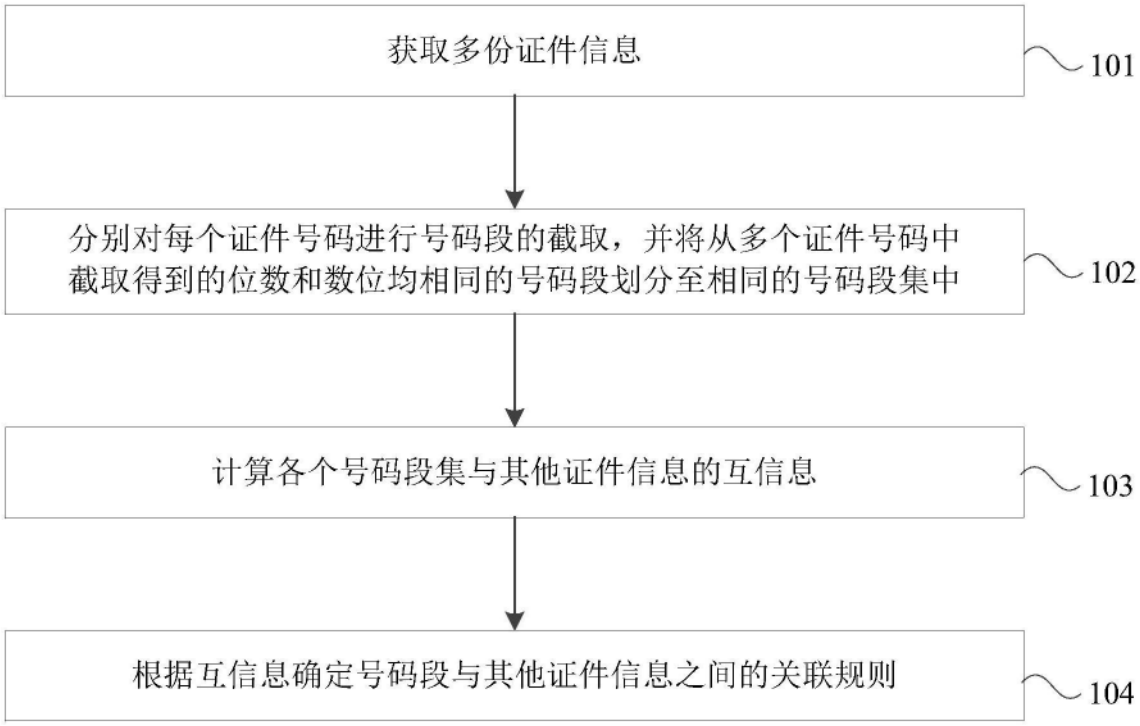

获取多份证件信息
101
分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中
102
计算各个号码段集与其他证件信息的互信息
103
根据互信息确定号码段与其他证件信息之间的关联规则
104

图1


获取多份证件信息
201
分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中
202
从多个包含位数相同的号码段的号码段集中选取目标号码段集
203-1
将目标号码段集中的号码段与其他号码段集中对应的号码段进行组合，并计算组合后的号码段集与其他证件信息的互信息
203-2
根据互信息确定号码段与其他证件信息之间的关联规则
204

图2

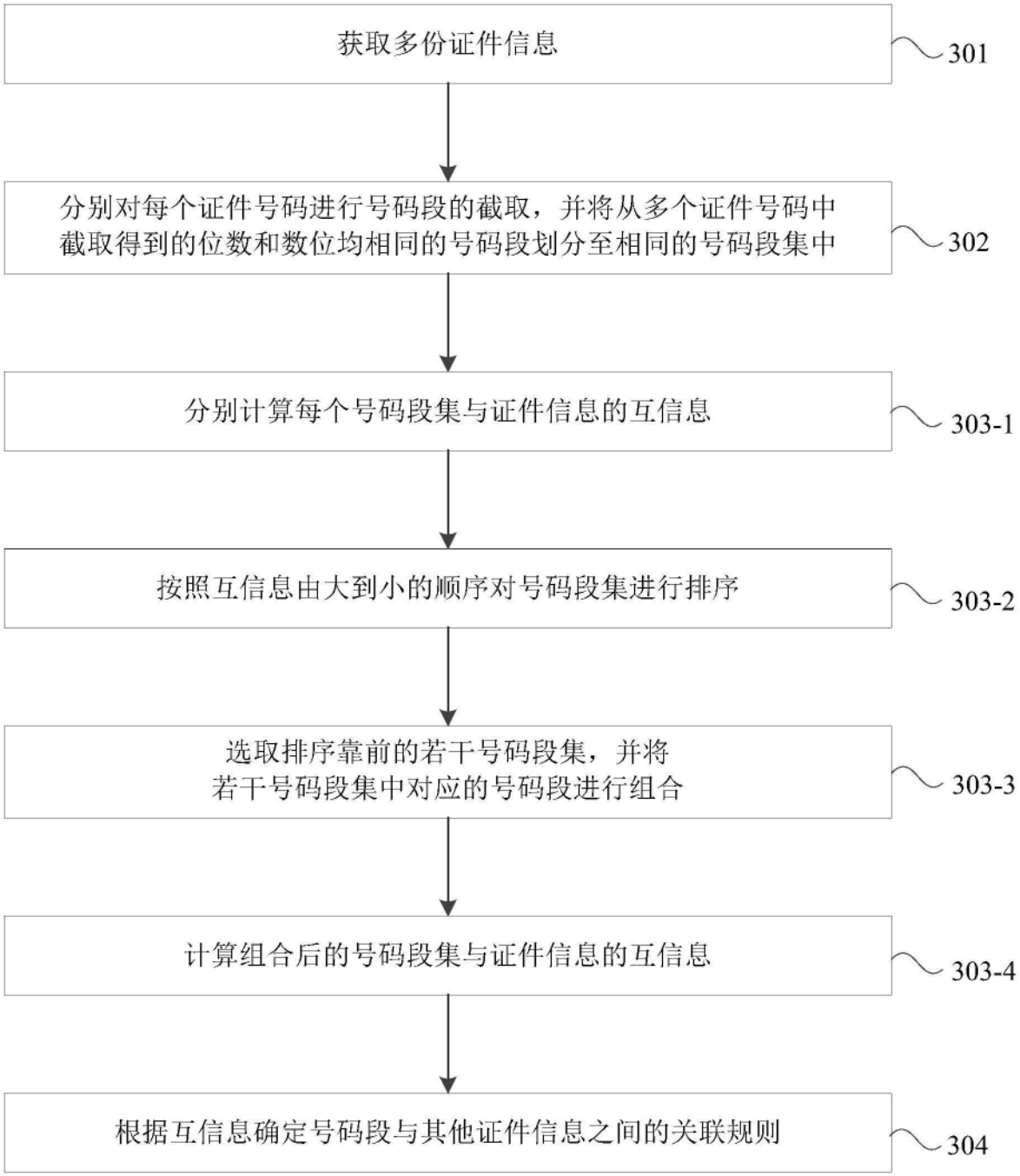

获取多份证件信息
301
分别对每个证件号码进行号码段的截取，并将从多个证件号码中截取得到的位数和数位均相同的号码段划分至相同的号码段集中
302
分别计算每个号码段集与证件信息的互信息
303-1
按照互信息由大到小的顺序对号码段集进行排序
303-2
选取排序靠前的若干号码段集，并将若干号码段集中对应的号码段进行组合
303-3
计算组合后的号码段集与证件信息的互信息
303-4
根据互信息确定号码段与其他证件信息之间的关联规则
304

图3

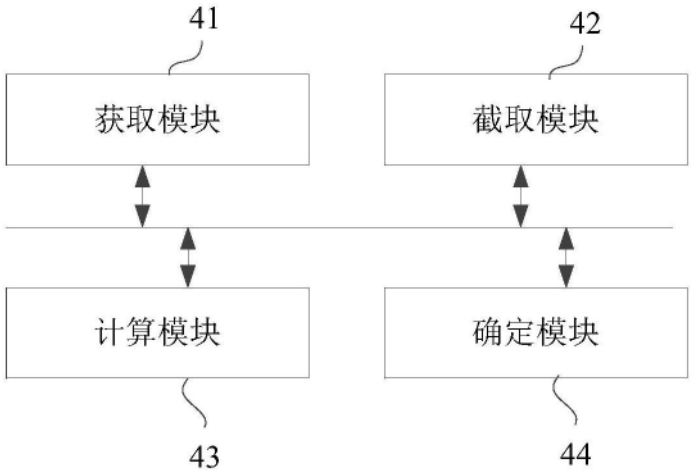

41
42
获取模块
截取模块
计算模块
确定模块
43
44

图4

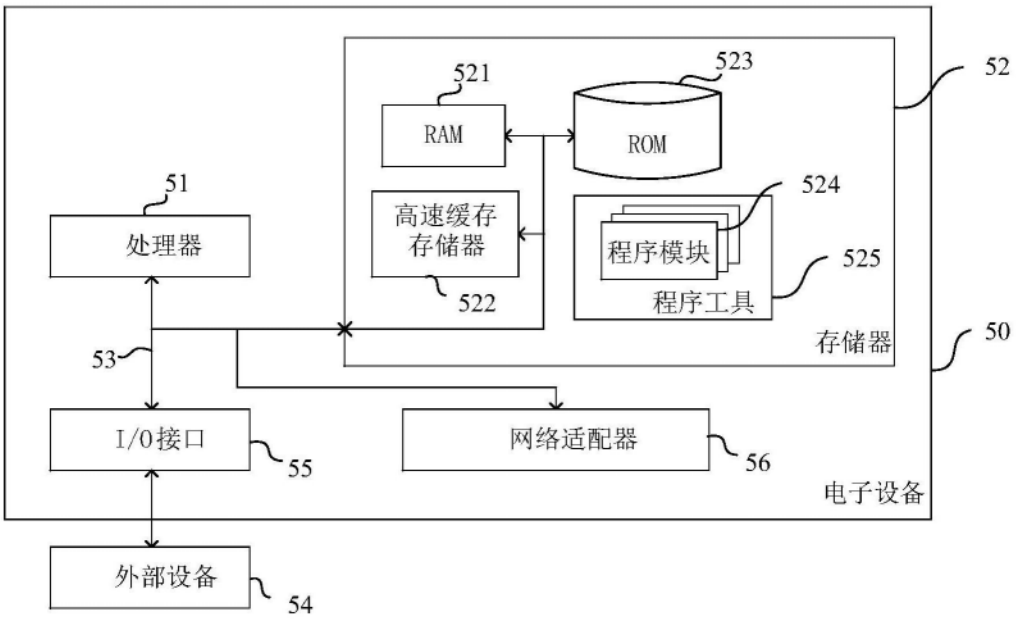

521
523
52
RAM
ROM
51
524
高速缓存
存储器
处理器
程序模块
525
522
程序工具
存储器
50
53
I/O接口
网络适配器
55
56
电子设备
外部设备
54

图5