

(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
2021 年 1 月 28 日 (28.01.2021)



(10) 国际公布号
WO 2021/012913 A1

(51) 国际专利分类号:
G06Q 40/08 (2012.01) **G06F 16/906** (2019.01)

(21) 国际申请号: PCT/CN2020/099572

(22) 国际申请日: 2020 年 6 月 30 日 (30.06.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910664820.3 2019 年 7 月 23 日 (23.07.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 程旺(CHENG, Wang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司(SCIHEAD IP LAW FIRM); 中国广东省广州市

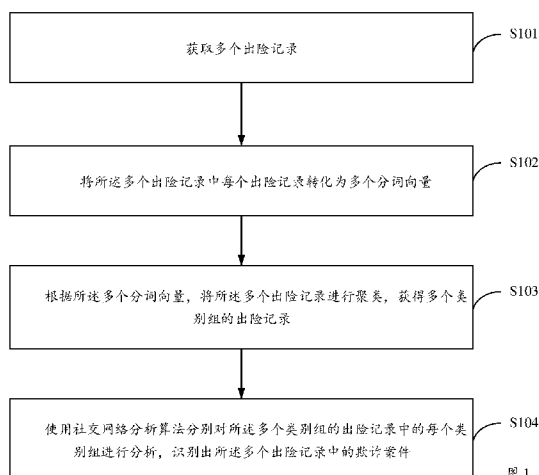
越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(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, JO, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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: DATA RECOGNITION METHOD AND SYSTEM, ELECTRONIC DEVICE AND COMPUTER STORAGE MEDIUM

(54) 发明名称: 数据识别的方法、系统、电子设备及计算机存储介质



S101 Acquire a plurality of accident records
S102 Convert each of the plurality of accident records into a plurality of segmented word vectors.
S103 Cluster the plurality of accident records according to the plurality of segmented word vectors to obtain a plurality of category groups of accident records
S104 Respectively analyze each category group of the plurality of category groups of accident records by using a social network analysis algorithm, to recognize fraudulent cases from the plurality of accident records

图 1

(57) Abstract: Disclosed are a data recognition method and system, an electronic device and a computer storage medium. The method comprises: acquiring a plurality of accident records (S101); converting each of the plurality of accident records into a plurality of segmented word vectors (S102); clustering the plurality of accident records according to the plurality of segmented word vectors to obtain a plurality of category groups of accident records (S103); and respectively analyzing each category group of the plurality of category groups of accident records by using a social network analysis algorithm, to recognize fraudulent cases from the plurality of accident records (S104).

(57) 摘要: 一种数据识别的方法、系统、电子设备及计算机存储介质。所述方法包括: 获取多个出险记录 (S101); 将所述多个出险记录中每个出险记录转化为多个分词向量 (S102); 根据所述多个分词向量, 将所述多个出险记录进行聚类, 获得多个类别组的出险记录 (S103); 使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析, 识别出所述多个出险记录中的欺诈案件 (S104)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

数据识别的方法、系统、电子设备及计算机存储介质

[0001] 本申请要求于2019年07月23日提交中国专利局、申请号为2019106648203，发明名称为“数据识别的方法、系统、电子设备及计算机存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及计算机领域，尤其涉及数据识别的方法、系统、电子设备及计算机存储介质，具体涉及了金融科技的欺诈检测技术。

背景技术

[0003] 目前，货运车辆保险行业反欺诈的判断，主要是通过人工分析报案描述信息和现场勘查等方式实现，无法实现群体欺诈的识别，即使通过社交网络分析（Social Network Analysis, SNA）算法构建网络来识别群体欺诈，在现实的复杂场景中，仅依赖离散型特征和少量的连续型特征，有时模型的表现会比较差。发明人意识到目前的反欺诈识别缺乏技术手段，欺诈行为识别率低。

发明概述

技术问题

[0004] 本申请提供了数据识别的方法、系统、电子设备及计算机存储介质，用于解决群体欺诈行为识别率低的问题。

问题的解决方案

技术解决方案

[0005] 第一方面，本申请提供了数据识别的方法，所述方法包括以下步骤。获取多个出险记录。将所述多个出险记录中每个出险记录转化为多个分词向量。根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录。使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。由于对出险记录进行了高精度的聚类，每个类别的出险记录是相关性非常高的数据，使用聚类后的出险记录进行SNA分析时，能够更好的建立出险案件关系网络，从而提高了群体欺诈识

别的正确率。

[0006] 第二方面，提供了一种数据识别系统，所述系统包括获取单元、转化单元、聚类单元以及识别单元，其中，所述获取单元用于获取多个出险记录。所述转化单元用于将所述多个出险记录中每个出险记录转化为多个分词向量。所述聚类单元用于根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录。所述识别单元用于使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。由于该系统对出险记录进行了高精度的聚类，每个类别的出险记录是相关性非常高的数据，使用聚类后的出险记录进行SNA分析时，所以基于该系统能够更好的建立出险案件关系网络，从而提高了群体欺诈识别的正确率。

[0007] 第三方面，提供了一种电子设备，包括处理器、输入设备、输出设备和存储器，所述处理器、输入设备、输出设备和存储器相互连接，其中，所述存储器用于存储计算机程序，所述计算机程序包括程序指令，所述处理器被配置用于调用所述程序指令，执行上述第一方面所述的方法。

[0008] 第四方面，提供了一种计算机可读存储介质，所述计算机存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行上述第一方面的方法。

发明的有益效果

有益效果

[0009] 可以看到，基于本申请提供的数据识别的方法、系统、电子设备及计算机存储介质，通过获取多个出险记录，将所述多个出险记录中每个出险记录转化为多个分词向量，根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录，使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。由于对出险记录进行了高精度的聚类，每个类别的出险记录是相关性非常高的数据，使用聚类后的出险记录进行SNA分析时，能够更好的建立出险案件关系网络，从而提高了群体欺诈识别的正确率。

对附图的简要说明

附图说明

[0010] 为了更清楚地说明本申请实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0011] 图1是本申请提供的一种数据识别方法的流程示意图。

[0012] 图2是本申请提供的一场景下使用SNA算法建立的案件关系网络示意图。

[0013] 图3是本申请提供的另一场景下使用SNA算法建立的案件关系网络示意图。

[0014] 图4是本申请提供的一种数据识别系统的结构示意图。

[0015] 图5是本申请提供的一种电子设备结构示意图。

发明实施例

本发明的实施方式

[0016] 下面通过具体实施方式结合附图对本申请作进一步详细说明。在以下的实施方式中，很多细节描述是为了使得本申请能被更好的理解。然而，本领域技术人员可以毫不费力的认识到，其中部分特征在不同情况下是可以省略的，或者可以由其他方法所替代。在某些情况下，本申请相关的一些操作并没有在说明书中显示或描述，这是为了避免本申请的核心部分被过多的描述所淹没。对于本领域技术人员而言，详细描述这些相关操作并不是必要的，他们根据说明书中的描述以及本领域的一般技术知识即可完整了解相关操作。

[0017] 应当理解，当在本说明书和所附权利要求书中使用术语时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0018] 需要说明的是，在本申请实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请实施例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。

[0019] 图1是本申请提供的一种数据识别方法的流程示意图。由图1可知，本申请提供

的数据识别方法包括以下步骤。

[0020] S101: 获取多个出险记录。

[0021] 在本申请实施例中, 出险指的是保险合同上规定或约定的赔偿或给付条件出现的情况, 例如, 车辆在保险期间, 发生意外事故后, 通知或者报告保险公司的过程就是汽车出险。其中, 所述出险记录可以是数据库中的出险记录数据, 例如, 对于汽车保险来说, 所述出险信息可以是事故记录, 包括事故车牌, 事故地点, 车险保单号, 保单代理人, 理赔记录、保险购买记录、车辆涉案人员、包括司机、报案人、受益人和伤者, 以及修理厂、报案电话、检修地点、GPS信息等数据、疾病诊断记录等涉及保险方面的记录。可以理解的是, 将出险记录作为原始数据进行反欺诈识别, 相比于普通的投保数据来说, 可以大大的提高反欺诈识别的准确度。应理解, 上述举例仅用于说明, 并不能构成具体限定。

[0022] S102: 将所述多个出险记录中每个出险记录转化为多个分词向量。

[0023] 在本申请实施例中, 将所述多个出险记录中的每个出险记录转化为多个分词向量包括。对所述多个出险记录中的每个出险记录进行切词处理, 将所述多个出险记录中的每个出险记录分为多个分词词语。将所述多个分词词语中, 与停用词集中的词语相同的分词词语删除, 与保留词集中的词语相同的分词词语保留, 获得筛选后的分词词语, 其中, 所述停用词集是多个与出险记录信息无关的分词词语的集合, 所述保留词集是预先设定的不能筛选掉的词语的集合。将所述筛选后的分词词语映射为多个分词向量。应理解, 出险记录包括了多个信息, 比如, 包括了具体的事故发生经过, 具体的人员伤亡情况, 疾病诊断记录, 具体的交警事故认定记录等, 各个出险记录的数据长短也不同, 如果直接进行数据预处理, 将会是一个很大的工作量, 因此首先对获取的出险记录进行切词处理, 从而提高数据处理效率。

[0024] 具体地, 切词处理可以是将每条出险记录的汉字序列切分成一个一个单独的词, 从形式上看, 词是稳定的字的组合, 因此在上下文中, 相邻的字同时出现的次数越多, 就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计, 计算它们的互现信息。定义两个字的互现信息, 计算两个汉字X、Y的相邻

共现概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词，进而可以进行切词的操作，但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，浪费存储空间，也使得搜索效率低下。因此可以结合停用词集和保留词集，将出现频度高、不是出险相关信息的词语过滤出去，例如语气助词、副词、介词、连接词等自身并无明确意义，只有将其放入一个完整句子中才有一定作用的词语。同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。并且，出险记录在进行分词操作后，还需要进行向量化的操作将每个出险记录转化为词向量。可以理解的是，“开心”和“幸福”对人来说，是两个非常接近的词语，而计算机是无法知道这两个词是相近的，因此，需要将每个词语用计算机能够理解的语言去表示，也就是说，将词语向量化，将单词表征为多维的浮点数，每一维的浮点数的数值大小表示了它与另一个单词之间的距离，表征的结果就是语义相近的词被映射到相近的集合空间上，从而使得计算机可以计算每个单词之间相似度，换句话说，使得计算机能够理解语言想要表达的含义后，再进行进一步反欺诈识别处理。

[0025] S103：根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录。

[0026] 在本申请实施例中，所述根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录包括。从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率。通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。也就是说，将出险记录转化为分词向量后，会出现很多个词语向量，虽然经过了停用词级和保留词级的筛选，但是很多词语向量对于接下来的反欺诈识别来说，数据量还是很大，应理解，因此，如果再将数据进行进一步的筛选，将易于

识别的数据进行聚类，可以大大提高聚类的准确度，将准确的聚类结果输入SNA网络，可以得到更准确的识别结果。其中，更易于识别的数据可以是，例如，如果一个词汇在某一案件出险记录中出现次数较多，而在其它案件出险信息中出现较少，则认为该词汇具有很好的区分能力。再例如，如果出险信息字段中包含某词汇的案件数较少，则说明该词汇具有很好的区分能力。应理解，上述举例仅用于说明，并不能构成具体限定。

[0027] 在本申请实施例中，目标分词向量可以使用词频-逆文本频率指数（Term Frequency - Inverse Document Frequency, TF-IDF）方法进行筛选，其中，TF-IDF方法是一种统计方法，用于评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。可以理解的是，对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取TF词频作为测度，就可以体现同类文本的特点，另外考虑到单词区别不同类别的能力，TF-IDF方法认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度IDF的概念，以TF和IDF的乘积作为特征空间坐标系的取值测度，并用它完成对权值TF的调整，调整权值的目的在于突出重要单词，抑制次要单词。也就是说，如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。因此，与直接将出险记录进行聚类相比，可以获得更好地聚类结果，筛选出适合分类的分词向量后，根据出险记录地重要程度矩阵进行聚类，可以更加准确地将出险信息相似的案件聚集在一起，将聚类结果作为SNA网络的输入数据，可以大大提高反欺诈案件识别地准确率。

[0028] 在本申请实施例中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种，本申请不作具体限定。可以理解的是，聚类的结果可以是同省市或者是购买同类型报销的出险记录聚集在一起，例如，将微损案件的出险记录聚集在一起，或者，将同省的案件聚集在一起等，此处不作具体限定。

[0029] S104: 对所述多个类别组的出险记录中的每个类别组, 使用社交网络分析SNA算法, 识别出所述多个出险记录中的欺诈案件。

[0030] 在本申请实施例中, 所述使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析, 识别出所述多个出险记录中的欺诈案件包括。获取所述多个类别组的出险记录。基于所述多个类别组的出险记录, 使用社交网络分析算法建立出险案件关系网络, 其中, 一个类别组的出险记录对应一个或者多个出险案件关系网络, 所述出险案件关系网络包括多个节点, 所述节点代表所述出险记录中的个体、组织或虚拟个体, 所述多个节点之间的连线表示所述多个节点之间存在社交关系。通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析, 提取出各个节点对应的群体出险特征。将所述各个节点对应的群体出险特征输入分类模型, 获得各个节点的欺诈率, 其中, 所述分类模型是使用样本集对神经网络进行训练得到的模型, 所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据。根据所述各个节点的欺诈率, 将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件, 将欺诈率高于第一阈值的多个节点识别为欺诈团伙。其中, SNA网络是由多个点(社会行动者)和各个点之间的连线(行动者之间的关系)组成的集合。每个节点可以是组织、个人、网络ID等不同含义的实体或虚拟个体, 而个体之间的关系可以是亲友、动作行为、收发消息等多种关系。通过SNA网络分析, 可以从杂乱的数据和连接关系中, 找到我们需要的关键信息, 也就是各个节点的群体性出险行为。以医疗保险的出险数据为例, 各个节点的出险特征可以是。病患所在的区域, 病患看病的医院、病患采购药品项目的数量和具体时间, 病患患得的疾病, 病患看诊的医生等行为。对病患的群体性出险行为进行分析, 就相当于对病患所在的区域、病患采购药品项目的数量和具体时间、病患患得的疾病等进行综合分析。若查到病患多次在不同的医院购买大量的药品, 且药品的种类各不相同, 可确定群体性出险特征为。用户的药品购买量大、药品类型多等等。再例如, 以车辆保险出险数据为例, 各个节点的出险特征可以是。车辆所处的城市、车牌号、车辆购买的保险类别、处理事故的交警、肇事者身份信息以及受害人身份信息等等。若查到车辆多次在

不同地点出险并且均为微损案件，由于微损案件金额低，可以快速报案处理，因此可确定群体性出险特征为用户多次参与微损案件出险。若查到车辆多次在不同地点出险并且受害人身份均为同一人，可确定群体性出险特征为用户多次与他人协同骗保。同理，可以获得商业保险、意外保险等其他出险数据的群体性出险特征。应理解，上述举例仅用于说明，并不能构成具体限定。

[0031] 在本申请实施例，除了使用神经网络根据所述各个节点的欺诈率，可以识别出所述多个出险记录中的欺诈案件，还可以根据出险率识别高风险案件。例如，图2是本申请提供的一场景下使用SNA算法建立的案件关系网络示意图，其中，灰色的点代表出险记录较多的用户，黑色的点代表有出险记录但是次数较少的穿越用户，白色的点代表没有出险记录的用户。通过数据计算分析可以得出，该团伙的高出险率达到66.8%，说明该团伙的平均出险率较低。出险用户占有所有用户的91.4%，进一步验证了该团伙的欺诈性。可以理解的，确认出欺诈团伙后，该团伙参与的出险记录即可确认为欺诈案件。

[0032] 在本申请实施例，使用SNA构建案件关系网络后，还可以根据网络结构，与预设的网络模型进行匹配，从而识别出高风险案件。例如，图3是本申请提供的另一场景下使用SNA算法建立的案件关系网络示意图，其中，黑色的点代表有出现记录但是次数较少的穿越用户，虽然该案件关系网络中各个节点的欺诈率不高，并且出险率也不高，但是由于案件关系网络中，有关系的两个节点会进行连接，因此该场景下的案件关系网络具有更高的风险性，该网络的背后通常是多人协作的团伙作案，其两两互通表示两两认识，背后的目的多为相互勾结，伪造信息以达到出险要求，需要重点关注。应理解，上述举例仅用于说明，SNA算法建立的案件关系网络还可以是除图2、图3外的其他网络结构，分析方法也与图2、图3使用的分析方法不同，本申请不作具体限定。

[0033] 上述方法中，通过获取多个出险记录，将所述多个出险记录中每个出险记录转化为多个分词向量，根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录，使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。由于对出险记录进行了高精度的聚类，每个类别的出险记录是相关性非常高的

数据，使用聚类后的出险记录进行SNA分析时，能够更好的建立出险案件关系网络，从而提高了群体欺诈识别的正确率。

[0034] 图4是本申请提供的一种数据识别系统的结构示意图。由图4可知，本申请提供的数据识别系统包括获取单元410、转化单元420、聚类单元430以及识别单元440，其中，

[0035] 所述获取单元410用于获取多个出险记录。

[0036] 所述转化单元420用于将所述多个出险记录中每个出险记录转化为多个分词向量。

[0037] 所述聚类单元430用于根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录。

[0038] 所述识别单元440用于使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。

[0039] 在本申请实施例中，出险指的是保险合同上规定或约定的赔偿或给付条件出现的情况，例如，车辆在保险期间，发生意外事故后，通知或者报告保险公司的过程就是汽车出险。其中，所述出险记录可以是数据库中的出险记录数据，例如，对于汽车保险来说，所述出险信息可以是事故记录，包括事故车牌，事故地点，车险保单号，保单代理人，理赔记录、保险购买记录、车辆涉案人员、包括司机、报案人、受益人和伤者，以及修理厂、报案电话、检修地点、GPS信息等数据、疾病诊断记录等涉及保险方面的记录。可以理解的是，将出险记录作为原始数据进行反欺诈识别，相比于普通的投保数据来说，可以大大的提高反欺诈识别的准确度。应理解，上述举例仅用于说明，并不能构成具体限定。

[0040] 在本申请实施例中，所述转化单元420具体用于：对所述多个出险记录中的每个出险记录进行切词处理，将所述多个出险记录中的每个出险记录分为多个分词词语。结合停用词集和保留词集，对所述多个分词词语进行筛选，获得筛选后的分词词语，其中，所述停用词集是多个与出险记录信息无关的分词词语的集合，所述保留词集是预先设定的不能筛选掉的词语的集合。对所述筛选后的分词词语进行向量化处理，获得所述多个出险记录中的每个出险记录的多个分词向量。应理解，出险记录包括了多个信息，比如，包括了具体的事故发生经

过，具体的人员伤亡情况，疾病诊断记录，具体的交警事故认定记录等，各个出险记录的数据长短也不同，如果直接进行数据预处理，将会是一个很大的工作量，因此首先对获取的出险记录进行切词处理，从而提高数据处理效率。

[0041] 具体地，切词处理可以是将每条出险记录的汉字序列切分成一个一个单独的词，从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息，计算两个汉字X、Y的相邻共现概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词，进而可以进行切词的操作，但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，浪费存储空间，也使得搜索效率低下。因此可以结合停用词集和保留词集，将出现频度高、不是出险相关信息的词语过滤出去，例如语气助词、副词、介词、连接词等自身并无明确意义，只有将其放入一个完整句子中才有一定作用的词语。同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。并且，出险记录在进行分词操作后，还需要进行向量化的操作将每个出险记录转化为词向量。可以理解的是，“开心”和“幸福”对人来说，是两个非常接近的词语，而计算机是无法知道这两个词是相近的，因此，需要将每个词语用计算机能够理解的语言去表示，也就是说，将词语向量化，将单词表征为多维的浮点数，每一维的浮点数的数值大小表示了它与另一个单词之间的距离，表征的结果就是语义相近的词被映射到相近的集合空间上，从而使得计算机可以计算每个单词之间相似度，换句话说，使得计算机能够理解语言想要表达的含义后，再进行进一步反欺诈识别处理。

[0042] 在本申请实施例，所述聚类单元430具体用于从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在

所属的出险记录中出现的频率低于在其他出险记录中出现的频率。所述聚类单元430具体用于通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。也就是说，将出险记录转化为分词向量后，会出现很多个词语向量，虽然经过了停用词级和保留词级的筛选，但是很多词语向量对于接下来的反欺诈识别来说，数据量还是很大，应理解，因此，如果再将数据进行进一步的筛选，将易于识别的数据进行聚类，可以大大提高聚类的准确度，将准确的聚类结果输入SNA网络，可以得到更准确的识别结果。其中，更易于识别的数据可以是，例如，如果一个词汇在某一案件出险记录中出现次数较多，而在其它案件出险信息中出现较少，则认为该词汇具有很好的区分能力。再例如，如果出险信息字段中包含某词汇的案件数较少，则说明该词汇具有很好的区分能力。应理解，上述举例仅用于说明，并不能构成具体限定。

[0043] 在本申请实施例中，目标分词向量可以使用TF-IDF方法进行筛选，其中，TF-IDF方法是一种统计方法，用于评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。可以理解的是，对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取TF词频作为测度，就可以体现同类文本的特点，另外考虑到单词区别不同类别的能力，TF-IDF方法认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度IDF的概念，以TF和IDF的乘积作为特征空间坐标系的取值测度，并用它完成对权值TF的调整，调整权值的目的在于突出重要单词，抑制次要单词。也就是说，如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。因此，与直接将出险记录进行聚类相比，可以获得更好地聚类结果，筛选出适合分类的分词向量后，根据出险记录地重要程度矩阵进行聚类，可以更加准确地将出险信息相似的案件聚集在一起，将聚类结果作为SNA网络的输入数据，可以大大提高反欺诈案件识别地准确率。

[0044] 在本申请实施例中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚

层次聚类算法中的一种或者多种，本申请不作具体限定。可以理解的是，聚类的结果可以是同省市或者是购买同类型报销的出险记录聚集在一起，例如，将微损案件的出险记录聚集在一起，或者，将同省的案件聚集在一起等，此处不作具体限定。

[0045] 在本申请实施例中，所述识别单元440具体用于获取所述多个类别组的出险记录。所述识别单元440具体用于基于所述多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，所述出险案件关系网络包括多个节点，所述节点代表所述出险记录中的个体、组织或虚拟个体，所述多个节点之间的连线表示所述多个节点之间存在社交关系。所述识别单元440具体用于通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征。所述识别单元440具体用于将所述各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据。所述识别单元440具体用于根据所述各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。其中，SNA网络是由多个点（社会行动者）和各个点之间的连线（行动者之间的关系）组成的集合。每个节点可以是组织、个人、网络ID等不同含义的实体或虚拟个体，而个体之间的关系可以是亲友、动作行为、收发消息等多种关系。通过SNA网络分析，可以从杂乱的数据和连接关系中，找到我们需要的关键信息，也就是各个节点的群体性出险行为。以医疗保险的出险数据为例，各个节点的出险特征可以是：病患所在的区域，病患看病的医院、病患采购药品项目的数量和具体时间，病患患得的疾病，病患看诊的医生等行为。对病患的群体性出险行为进行分析，就相当于对病患所在的区域、病患采购药品项目的数量和具体时间、病患患得的疾病等进行综合分析。若查到病患多次在不同的医院购买大量的药品，且药品的种类各不相同，可确定群体性出险特征为。用户的药品购买量大、药品类型多等等。再例如，以车辆保险出险数据为例，各个节点的

出险特征可以是。车辆所处的城市、车牌号、车辆购买的保险类别、处理事故的交警、肇事者身份信息以及受害人身份信息等等。若查到车辆多次在不同地点出险并且均为微损案件，由于微损案件金额低，可以快速报案处理，因此可确定群体性出险特征为用户多次参与微损案件出险。若查到车辆多次在不同地点出险并且受害人身份均为同一人，可确定群体性出险特征为用户多次与他人协同骗保。同理，可以获得商业保险、意外保险等其他出险数据的群体性出险特征。应理解，上述举例仅用于说明，并不能构成具体限定。

[0046] 在本申请实施例，除了使用神经网络根据所述各个节点的欺诈率，可以识别出所述多个出险记录中的欺诈案件，还可以根据出险率识别高风险案件。例如，图2是一个SNA算法建立的案件关系网络，其中，灰色的点代表出险记录较多的用户，黑色的点代表有出险记录但是次数较少的穿越用户，白色的点代表没有出险记录的用户。通过数据计算分析可以得出，该团伙的高出险率达到66.8%，说明该团伙的平均出险率较低。出险用户占有所有用户的91.4%，进一步验证了该团伙的欺诈性。可以理解的，确认出欺诈团伙后，该团伙参与的出险记录即可确认为欺诈案件。

[0047] 在本申请实施例，使用SNA构建案件关系网络后，还可以根据网络结构，与预设的网络模型进行匹配，从而识别出高风险案件。例如，图3是本申请提供的另一场景下使用SNA算法建立的案件关系网络示意图，其中，黑色的点代表有出现记录但是次数较少的穿越用户，虽然该案件关系网络中各个节点的欺诈率不高，并且出险率也不高，但是由于案件关系网络中，有关系的两个节点会进行连接，因此该网络具有更高的风险性，这种网络的背后通常是多人协作的团伙作案，其两两互通表示两两认识，背后的目的多为相互勾结，伪造信息以达到出险要求，需要重点关注。应理解，上述举例仅用于说明，SNA算法建立的案件关系网络还可以是除图2、图3外的其他网络结构，分析方法也与图2、图3使用的分析方法不同，本申请不作具体限定。

[0048] 上述系统中，通过获取多个出险记录，将所述多个出险记录中每个出险记录转化为多个分词向量，根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录，使用社交网络分析算法分别对所述多个类别组的

出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。由于对出险记录进行了高精度的聚类，每个类别的出险记录是相关性非常高的数据，使用聚类后的出险记录进行SNA分析时，能够更好的建立出险案件关系网络，从而提高了群体欺诈识别的正确率。

[0049] 参见图5，图5是本申请提供的一种电子设备的结构示意图。如图所示的本实施例中的电子设备可以包括。一个或者多个处理器511、存储器512和通信接口513。其中，处理器511、存储器512和通信接口513之间可以通过总线514连接。

[0050] 处理器511包括一个或者多个通用处理器，其中，通用处理器可以是能够处理电子指令的任何类型的设备，包括中央处理器(Central Processing Unit, CPU)、图像处理器(Graphics Processing Unit, GPU)、微处理器、微控制器、主处理器、控制器以及专用集成电路(Application Specific Integrated Circuit, ASIC)、数字信号处理器(Digital Signal Processor, DSP)、可编程门阵列(Field -Programmable Gate Array, FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。处理器511用于执行存储器512存储的程序指令。

[0051] 存储器512可以包括易失性存储器，例如随机存取存储器(Random Access Memory, RAM)。存储器也可以包括非易失性存储器，例如只读存储器(Read-Only Memory, ROM)、快闪存储器(Flash Memory)、硬盘(Hard Disk Drive, HDD)或固态硬盘(Solid-State Drive, SSD)，存储器还可以包括上述种类的存储器的组合。存储器512可以采用集中式存储，也可以采用分布式存储，此处不作具体限定。可以理解的是，存储器512用于存储计算机程序，例如：计算机程序指令等。在本申请实施例中，存储器512可以向处理器511提供指令和数据。

[0052] 通信接口513可以为有线接口（例如以太网接口）或无线接口（例如蜂窝网络接口或使用无线局域网接口），用于与其他计算机设备或用户进行通信。当通信接口513为有线接口时，通信接口513可以采用网络通讯协议(Transmission Control Protocol/Internet Protocol, TCP/IP)之上的协议族，例如，远程函数调用(Remote Function Call, RFC)协议、简单对象访问协议(Simple Object

Access Protocol, SOAP) 协议、简单网络管理协议(Simple Network Management Protocol, SNMP)、公共对象请求代理体系结构协议(Common Object Request Broker Architecture, CORBA) 以及分布式协议等等。当通信接口513为无线接口时, 可以根据全球移动通信系统(Global System for Mobile Communication, GSM) 或者码分多址(Code Division Multiple Access, CDMA) 标准利用蜂窝通信, 因此包括用于数据传输的无线调制解调器、电子处理设备、一个或多个数字存储器设备以及双天线。

[0053] 在本申请实施例中, 处理器511、存储器512、通信接口513和总线514可执行本申请实施例提供的数据识别方法的任一实施例中所描述的实现方式, 具体的, 处理器511可用于调用存储器中的指令, 执行如下方法: 获取多个出险记录。将多个出险记录中每个出险记录转化为多个分词向量。根据多个分词向量, 将所述多个出险记录进行聚类, 获得多个类别组的出险记录。使用社交网络分析算法分别对多个类别组的出险记录中的每个类别组进行分析, 识别出多个出险记录中的欺诈案件。

[0054] 在一具体实施例中, 处理器511具体用于执行: 对多个出险记录中的每个出险记录进行切词处理, 将多个出险记录中的每个出险记录分为多个分词词语。将多个分词词语中, 与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留, 获得筛选后的分词词语, 其中, 停用词集是多个与出险记录信息无关的分词词语的集合, 保留词集是预先设定的不能筛选掉的词语的集合。将筛选后的分词词语映射为多个分词向量。

[0055] 在一具体实施例中, 处理器511具体用于执行: 从每个出险记录的多个分词向量中, 筛选出目标分词向量, 其中, 目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率, 或者, 目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率。通过聚类算法, 将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。

[0056] 在一具体实施例中, 处理器511具体用于执行: 获取多个类别组的出险记录。基于多个类别组的出险记录, 使用社交网络分析算法建立出险案件关系网络, 其中, 一个类别组的出险记录对应一个或者多个出险案件关系网络, 出险案件

关系网络包括多个节点，节点代表所述出险记录中的个体、组织或虚拟个体，多个节点之间的连线表示所述多个节点之间存在社交关系。通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征。将各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据。根据各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

[0057] 在一具体实施例中，聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

[0058] 需要说明的是，处理器511执行方法的具体内容可参考前文方法实施例的相关描述，为了说明书简洁，这里不再赘述。

[0059] 在本申请的另一实施例中提供一种计算机可读存储介质，所述计算机可读存储介质可以是非易失性，也可以是易失性。所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令被处理器执行时实现本申请提供的数据识别方法的任一实施例中所描述的实现方式，在此不再赘述。

[0060] 所述计算机可读存储介质可以是前述任一实施例所述的终端的内部存储单元，例如终端的硬盘或内存。所述计算机可读存储介质也可以是所述终端的外部存储设备，例如所述终端上配备的插接式硬盘，智能存储卡(Smart Media Card, SMC)，安全数字(Secure Digital, SD)卡，闪存卡(Flash Card)等。进一步地，所述计算机可读存储介质还可以既包括所述终端的内部存储单元也包括外部存储设备。所述计算机可读存储介质用于存储所述计算机程序以及所述终端所需的其他程序和数据。所述计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

[0061] 本申请实施例中，计算机可读存储介质存储有计算机程序，计算机程序被处理器执行以实现如下方法。获取多个出险记录。将所述多个出险记录中每个出险

记录转化为多个分词向量。根据多个分词向量，将多个出险记录进行聚类，获得多个类别组的出险记录。使用社交网络分析算法分别对多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。

[0062] 在一具体实施例中，计算机程序被处理器执行时具体实现如下方法：对多个出险记录中的每个出险记录进行切词处理，将多个出险记录中的每个出险记录分为多个分词词语。将多个分词词语中，与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留，获得筛选后的分词词语，其中，停用词集是多个与出险记录信息无关的分词词语的集合，保留词集是预先设定的不能筛选掉的词语的集合。将所述筛选后的分词词语映射为多个分词向量。

[0063] 在一具体实施例中，计算机程序被处理器执行时具体实现如下方法。从每个出险记录的多个分词向量中，筛选出目标分词向量，其中，目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率。通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。

[0064] 在一具体实施例中，计算机程序被处理器执行时具体实现如下方法。获取多个类别组的出险记录。基于多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，出险案件关系网络包括多个节点，节点代表所述出险记录中的个体、组织或虚拟个体，多个节点之间的连线表示所述多个节点之间存在社交关系。通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征。将各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，分类模型是使用样本集对神经网络进行训练得到的模型，样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据。根据各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

[0065] 在一具体实施例中，聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

[0066] 需要说明的是，计算机程序被处理器执行方法的具体内容可参考前文方法实施例的相关描述，为了说明书简洁，这里不再赘述。在本申请所提供的几个实施例中，应该理解到，所揭露的方法及装置，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另外，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接，也可以是电的，机械的或其它的形式连接。

[0067] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本申请实施例方案的目的。

[0068] 另外，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单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0069] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0070] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

- [权利要求 1] 一种数据识别的方法，其中，所述方法包括：
- 获取多个出险记录；
- 将所述多个出险记录中每个出险记录转化为多个分词向量；
- 根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录；
- 使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。
- [权利要求 2] 根据权利要求1所述的方法，其中，将所述多个出险记录中的每个出险记录转化为多个分词向量包括：
- 对所述多个出险记录中的每个出险记录进行切词处理，将所述多个出险记录中的每个出险记录分为多个分词词语；
- 将所述多个分词词语中，与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留，获得筛选后的分词词语，其中，所述停用词集是多个与出险记录信息无关的分词词语的集合，所述保留词集是预先设定的不能筛选掉的词语的集合；
- 将所述筛选后的分词词语映射为多个分词向量。
- [权利要求 3] 根据权利要求2所述的方法，其中，所述根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录包括：
- 从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率；
- 通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。
- [权利要求 4] 根据权利要求1所述的方法，其中，所述使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件包括：

获取所述多个类别组的出险记录；

基于所述多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，所述出险案件关系网络包括多个节点，所述节点代表所述出险记录中的个体、组织或虚拟个体，所述多个节点之间的连线表示所述多个节点之间存在社交关系；

通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征；

将所述各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据；

根据所述各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

[权利要求 5] 根据权利要求3所述的方法，其中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

[权利要求 6] 一种数据识别的系统，其中，所述系统包括获取单元、转化单元、聚类单元以及识别单元，其中，

所述获取单元用于获取多个出险记录；

所述转化单元用于将所述多个出险记录中每个出险记录转化为多个分词向量；

所述聚类单元用于根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录；

所述识别单元用于使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。

- [权利要求 7] 根据权利要求6所述的方法，其中，所述转化单元具体用于：
对所述多个出险记录中的每个出险记录进行切词处理，将所述多个出险记录中的每个出险记录分为多个分词词语；
将所述多个分词词语中，与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留，获得筛选后的分词词语，其中，所述停用词集是多个与出险记录信息无关的分词词语的集合，所述保留词集是预先设定的不能筛选掉的词语的集合；
将所述筛选后的分词词语映射为多个分词向量。
- [权利要求 8] 根据权利要求7所述的系统，其中，
所述聚类单元具体用于从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率；
所述聚类单元具体用于通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。
- [权利要求 9] 根据权利要求6所述的系统，其中，
所述识别单元具体用于获取所述多个类别组的出险记录；
所述识别单元具体用于基于所述多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，所述出险案件关系网络包括多个节点，所述节点代表所述出险记录中的个体、组织或虚拟个体，所述多个节点之间的连线表示所述多个节点之间存在社交关系；
所述识别单元具体用于通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，以提取出各个节点对应的群体出险特征；
所述识别单元具体用于将所述各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体

出险特征数据以及对应的已知的欺诈率数据；

所述识别单元具体用于将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

[权利要求 10] 根据权利要求8所述的系统，其中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

[权利要求 11] 一种电子设备，其中，所述电子设备包括处理器和存储器；所述存储器用于存储指令；所述处理器用于调用存储器中的指令，执行如下方法：

获取多个出险记录；

将所述多个出险记录中每个出险记录转化为多个分词向量；

根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类别组的出险记录；

使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。

[权利要求 12] 根据权利要求11所述的电子设备，其中，所述处理器具体用于执行：对所述多个出险记录中的每个出险记录进行切词处理，将所述多个出险记录中的每个出险记录分为多个分词词语；

将所述多个分词词语中，与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留，获得筛选后的分词词语，其中，所述停用词集是多个与出险记录信息无关的分词词语的集合，所述保留词集是预先设定的不能筛选掉的词语的集合；

将所述筛选后的分词词语映射为多个分词向量。

[权利要求 13] 根据权利要求12所述的电子设备，其中，所述处理器具体用于执行：从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在所属的出险记录中出

现的频率低于在其他出险记录中出现的频率；

通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。

- [权利要求 14] 根据权利要求11所述的电子设备，其中，所述处理器具体用于执行：
- 获取所述多个类别组的出险记录；
- 基于所述多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，所述出险案件关系网络包括多个节点，所述节点代表所述出险记录中的个体、组织或虚拟个体，所述多个节点之间的连线表示所述多个节点之间存在社交关系；
- 通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征；
- 将所述各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据；
- 根据所述各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

- [权利要求 15] 根据权利要求13所述的电子设备，其中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

- [权利要求 16] 一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行以实现如下方法：
- 获取多个出险记录；
- 将所述多个出险记录中每个出险记录转化为多个分词向量；
- 根据所述多个分词向量，将所述多个出险记录进行聚类，获得多个类

别组的出险记录；

使用社交网络分析算法分别对所述多个类别组的出险记录中的每个类别组进行分析，识别出所述多个出险记录中的欺诈案件。

[权利要求 17] 根据权利要求16所述的计算机可读存储介质，其中，所述计算机程序被处理器执行时具体实现如下方法：

对所述多个出险记录中的每个出险记录进行切词处理，将所述多个出险记录中的每个出险记录分为多个分词词语；

将所述多个分词词语中，与停用词集中的词语相同的分词词语删除、与保留词集中的词语相同的分词词语保留，获得筛选后的分词词语，其中，所述停用词集是多个与出险记录信息无关的分词词语的集合，所述保留词集是预先设定的不能筛选掉的词语的集合；

将所述筛选后的分词词语映射为多个分词向量。

[权利要求 18] 根据权利要求17所述的计算机可读存储介质，其中，所述计算机程序被处理器执行时具体实现如下方法：

从所述每个出险记录的多个分词向量中，筛选出目标分词向量，其中，所述目标分词向量在所属的出险记录中出现的频率高于在其他出险记录中出现的频率，或者，所述目标分词向量在所属的出险记录中出现的频率低于在其他出险记录中出现的频率；

通过聚类算法，将包含相同或相近目标分词向量的多个出险记录聚类为多个类别组的出险记录。

[权利要求 19] 根据权利要求16所述的计算机可读存储介质，其中，所述计算机程序被处理器执行时具体实现如下方法：

获取所述多个类别组的出险记录；

基于所述多个类别组的出险记录，使用社交网络分析算法建立出险案件关系网络，其中，一个类别组的出险记录对应一个或者多个出险案件关系网络，所述出险案件关系网络包括多个节点，所述节点代表所述出险记录中的个体、组织或虚拟个体，所述多个节点之间的连线表示所述多个节点之间存在社交关系；

通过社交网络分析算法对所述出险案件关系网络中的各个节点与其他节点之间的关系进行分析，提取出各个节点对应的群体出险特征；将所述各个节点对应的群体出险特征输入分类模型，获得各个节点的欺诈率，其中，所述分类模型是使用样本集对神经网络进行训练得到的模型，所述样本集包括已知的多个维度群体出险特征数据以及对应的已知的欺诈率数据；根据所述各个节点的欺诈率，将欺诈率高于第一阈值的多个节点所属的出险记录识别为欺诈案件，将欺诈率高于第一阈值的多个节点识别为欺诈团伙。

[权利要求 20] 根据权利要求18所述的计算机可读存储介质，其中，所述聚类算法是K均值聚类算法、均值漂移聚类算法、具有噪声的基于密度的聚类算法、用高斯混合模型的最大期望聚类算法以及凝聚层次聚类算法中的一种或者多种。

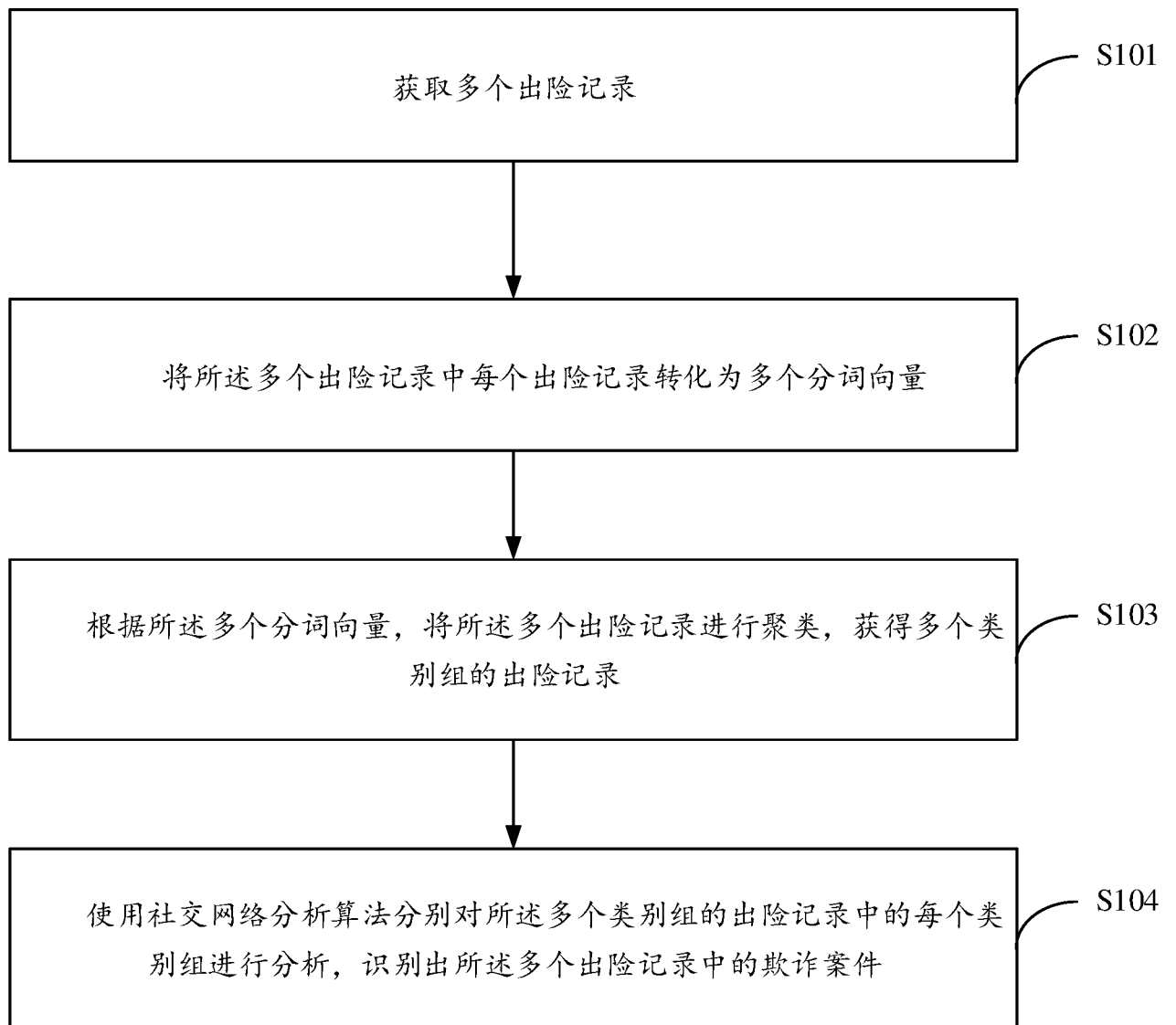


图 1

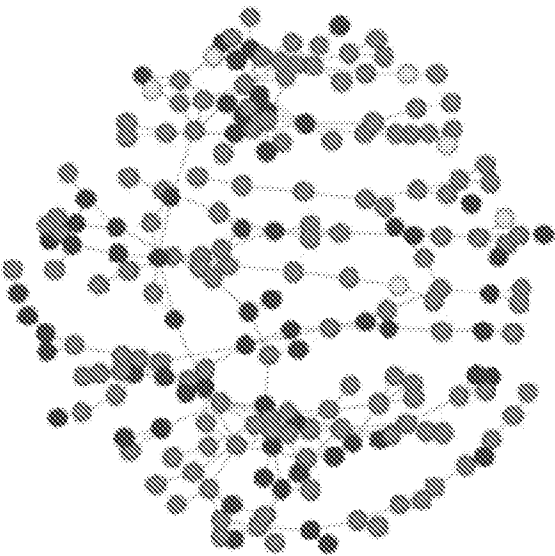


图 2

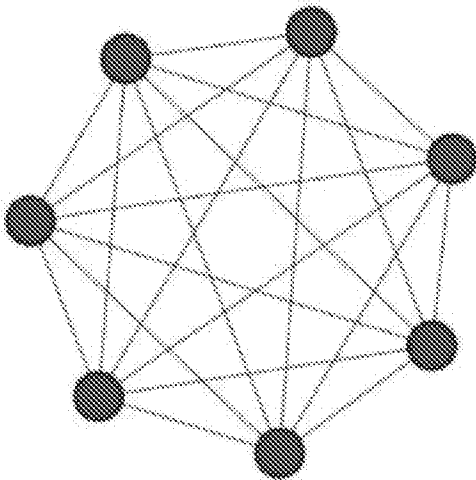


图 3

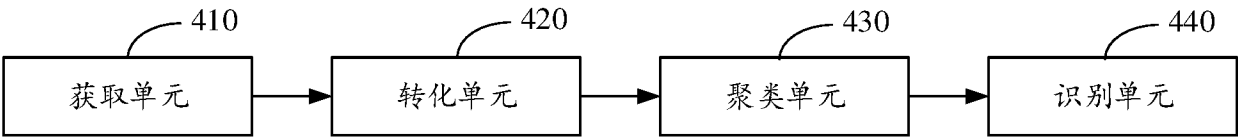


图 4

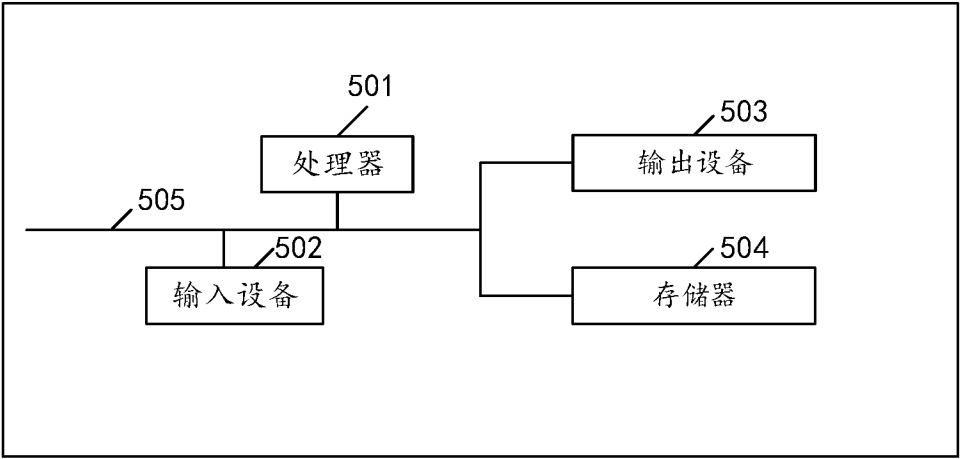


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/099572

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 40/08(2012.01)i; G06F 16/906(2019.01)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)

G06Q; 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: 保险, 出险, 记录, 分词, 向量, 聚类, 类别, 分组, 欺诈, 诈骗, 欺骗, 案件, 风险, insurance, record, word, segmentation, vector, cluster, category, group, fraud, case, risk

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110490750 A (PING'AN TECHNOLOGY (SHENZHEN) CO., LTD.) 22 November 2019 (2019-11-22) claims 1-10	1-20
X	CN 109919781 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 June 2019 (2019-06-21) description, paragraphs [0036]-[0067]	1-20
A	CN 109446528 A (NANJING ZHONGFU INFORMATION TECHNOLOGY CO., LTD.) 08 March 2019 (2019-03-08) entire document	1-20
A	CN 109242499 A (BANK OF CHINA LIMITED) 18 January 2019 (2019-01-18) entire document	1-20
A	CN 109685647 A (SUNSHINE PROPERTY & CASUALTY INSURANCE CO., LTD.) 26 April 2019 (2019-04-26) entire document	1-20



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

26 August 2020

Date of mailing of the international search report

09 October 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/099572

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110490750	A	22 November 2019	None	
CN	109919781	A	21 June 2019	None	
CN	109446528	A	08 March 2019	None	
CN	109242499	A	18 January 2019	None	
CN	109685647	A	26 April 2019	None	

国际检索报告

国际申请号

PCT/CN2020/099572

A. 主题的分类

G06Q 40/08(2012.01)i; G06F 16/906(2019.01)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)

G06Q; 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: 保险, 出险, 记录, 分词, 向量, 聚类, 类别, 分组, 欺诈, 诈骗, 欺骗, 案件, 风险, insurance, record, word, segmentation, vector, cluster, category, group, fraud, case, risk

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
PX	CN 110490750 A (平安科技深圳有限公司) 2019年 11月 22日 (2019 - 11 - 22) 权利要求1-10	1-20
X	CN 109919781 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 说明书第[0036]-[0067]段	1-20
A	CN 109446528 A (南京中孚信息技术有限公司) 2019年 3月 8日 (2019 - 03 - 08) 全文	1-20
A	CN 109242499 A (中国银行股份有限公司) 2019年 1月 18日 (2019 - 01 - 18) 全文	1-20
A	CN 109685647 A (阳光财产保险股份有限公司) 2019年 4月 26日 (2019 - 04 - 26) 全文	1-20

☐ 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

2020年 8月 26日

国际检索报告邮寄日期

2020年 10月 9日

ISA/CN的名称和邮寄地址

中国国家知识产权局(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

受权官员

王晓敏

电话号码 86-(10)-53961359

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/099572

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	110490750	A	2019年 11月 22日	无	
CN	109919781	A	2019年 6月 21日	无	
CN	109446528	A	2019年 3月 8日	无	
CN	109242499	A	2019年 1月 18日	无	
CN	109685647	A	2019年 4月 26日	无	