(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115409433 B
(45)授权公告日 2023.04.07

(21)申请号 202211359353.1

(22)申请日 2022.11.02

(65)同一申请的已公布的文献号
申请公布号 CN 115409433 A

(43)申请公布日 2022.11.29

(73)专利权人 成都宏恒信息科技有限公司
地址 610000 四川省成都市天府新区正兴街道湖畔路北段269号1栋1单元5楼

(72)发明人 王刚 彭保

(74)专利代理机构 四川域策汇智知识产权代理有限公司 51351
专利代理师 郭禾苗

(51)Int.Cl.
G06Q 10/0639(2023.01)
G06Q 50/26(2012.01)
G06N 3/0464(2023.01)
G06N 3/0442(2023.01)
G06N 3/045(2023.01)
G06N 3/048(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 111967494 A,2020.11.20
CN 112527861 A,2021.03.19
WO 2019191810 A1,2019.10.10

审查员 毛冉

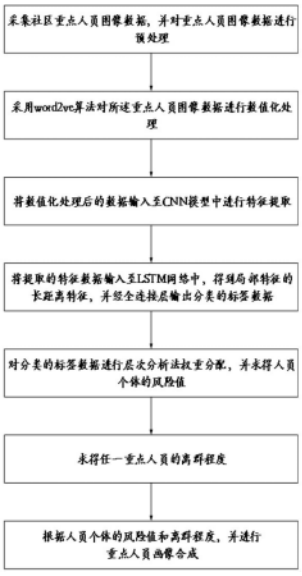
权利要求书3页 说明书7页 附图1页

(54)发明名称

基于深度NLP的社区重点人员画像分析方法及装置

(57)摘要

本发明公开了基于深度NLP的社区重点人员画像分析方法及装置，包括：采集社区重点人员数据，并对重点人员数据进行预处理；采用word2ve算法对所述重点人员数据进行数值化处理；将数值化处理后的数据输入至CNN模型中进行特征提取；将提取的特征数据输入至LSTM网络中，得到局部特征的长距离特征，并经全连接层输出分类的标签数据；对分类的标签数据进行层次分析法权重分配，并求得人员个体的风险值；求得任一重点人员的离群程度；根据人员个体的风险值和离群程度，采用多标签聚类算法进行重点人员画像合成。通过上述方案，本发明具有逻辑简单、准确可靠等优点。



CN 115409433 B

1.基于深度NLP的社区重点人员画像分析方法,其特征在于,包括以下步骤:

采集社区重点人员数据,并对重点人员数据进行预处理;

采用word2ve算法对所述重点人员数据进行数值化处理;

将数值化处理后的数据输入至CNN模型中进行特征提取;

将提取的特征数据输入至LSTM网络中,得到局部特征的长距离特征,并经全连接层输出分类的标签数据;

对分类的标签数据进行层次分析法权重分配,并求得人员个体的风险值;

求得任一重点人员的离群程度;

根据人员个体的风险值和离群程度,采用多标签聚类算法进行重点人员画像合成,包括:

抽取人员个体的风险值和离群程度的标签,并计算任一标签对应的用户人数;

采用余弦相似度函数计算标签之间的相关性,其表达式为:

$$\cos\theta = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{T}(A_i)^2} \times \sqrt{\sum_{i=1}^{T}(B_i)^2}}$$

其中,A和B表示标签;A_i表示A标签对应个体风险值与离群程度对应的T维向量;B_i表示B标签对应个体风险值与离群程度对应的T维向量;

筛选出与每个A类标签相关性最大的B类标签,并将该A类标签归类到该B类标签,得到最终的用户画像。

2.根据权利要求1所述的基于深度NLP的社区重点人员画像分析方法,其特征在于,对重点人员的数据的进行预处理,采用基于词典规则的中文分词方法对重点人员的数据进行处理。

3.根据权利要求1所述的基于深度NLP的社区重点人员画像分析方法,其特征在于,所述人员个体的风险值的表达式为:

$$R = \sum_{i=1}^{n} Score_i \times Weight_i$$

其中,R表示个体的风险值,n表示标签的数量;$Score_i$表示标签风险度;$Weight_i$表示标签对应的风险权重。

4.根据权利要求3所述的基于深度NLP的社区重点人员画像分析方法,其特征在于,所述风险权重采用以下步骤获取:

将数个标签作为输入,并搭建层次结构模型;

构造判断矩阵;

对判断矩阵的任一列向量进行归一化后将任一行向量求和,并进行归一化处理得到列向量w_i,其表达式为:

$$W_i = \bar{W}_i / \sum_{n=1}^{n} \bar{W}_i$$

$$\bar{W}_i = \sum_{j=1}^{n} \bar{A}_{ij}$$

$$\bar{A}_{ij} = A_{ij} / \sum_{k=1}^{n} A_{kj}$$

其中，A_{ij}表示构造的判断矩阵，$\overline{A_{ij}}$表示归一化的判断矩阵，A_{kj}表示第k行第j列的元素，k表示行数；i＝1,2,3.....n；j＝1,2,3.....n；

对列向量w_i的层次总排序，求得第K层元素相对于总目标的排序为：

$$\varphi_j^K = \sum_{i=1}^{n} p_{ij}^K W_j^K$$

其中，p_{ij}^K表示第K层元素相对于第K-1层元素的排序；W_j^K表示k-1层元素对总目标的权重。

5.根据权利要求3所述的基于深度NLP的社区重点人员画像分析方法，其特征在于，所述重点人员的离群程度的局部离群因子的表达式为：

$$LOF_k(p) = \frac{\sum_{0 \in N_k(p)} \frac{lrd_k(\mathrm{o})}{lrd_k(\mathrm{p})}}{|N_k(p)|} = \frac{\sum_{0 \in N_k(p)} lrd_k(\mathrm{o})}{|N_k(p)|} / lrd_k(\mathrm{p})$$

$$d(p,o) = \sum_{n=1}^{n} AHP_i \times d_i(p,o)$$

其中，$d_k(p)$表示对于点p的第k距离，$d_k(p)$＝d(p,o)；$N_k(p)$表示距离领域点p的第k距离领域；$lrd_k(p)$表示点p的局部可达密度；$lrd_k(o)$表示点o的局部可达密度；AHP_i表示第i个标签的权重值；$d_i(p,o)$表示第i个标签欧氏距离；d(p,o)表示p和o两点之间的距离；$LOF_k(p)$表示局部离群因子。

6.一种基于深度NLP的社区重点人员画像分析装置，其特征在于，包括：

数据采集预处理模块，采集社区重点人员数据，并对重点人员数据进行预处理；

数值化处理模块，与数据采集预处理模块连接，采用word2ve算法对所述重点人员数据进行数值化处理；

CNN模型，与数值化处理模块连接，将数值化处理后的数据输入至CNN模型中进行特征提取；

LSTM网络，与CNN模型连接，将提取的特征数据输入至LSTM网络中，得到局部特征的长距离特征，并经全连接层输出分类的标签数据；

层次分析模块，与LSTM网络连接，对分类的标签数据进行层次分析法权重分配，并求得人员个体的风险值；

离群程度分析模块，与LSTM网络连接，求得任一重点人员的离群程度；

以及，画像合成模块，与层次分析模块和离群程度分析模块连接，根据人员个体的风险值和离群程度，采用多标签聚类算法进行重点人员画像合成，包括：

抽取人员个体的风险值和离群程度的标签，并计算任一标签对应的用户人数；

采用余弦相似度函数计算标签之间的相关性，其表达式为：

$$\cos\theta = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} (A_i \times B_i)}{\sqrt{\sum_{i=1}^{T} (A_i)^2} \times \sqrt{\sum_{i=1}^{T} (B_i)^2}}$$

其中，A和B表示标签；A_i表示A标签对应个体风险值与离群程度对应的T维向量；B_i表示B标签对应个体风险值与离群程度对应的T维向量；

筛选出与每个A类标签相关性最大的B类标签，并将该A类标签归类到该B类标签，得到最终的用户画像。

7.一种电子设备，包括存储器、处理器及存储在所述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要求1至5任一项所述的基于深度NLP的社区重点人员画像分析方法。

8.一种计算机可读存储介质，其存储有计算机程序，其特征在于，所述计算机程序被处理器执行时，实现权要求1至5任一项所述的基于深度NLP的社区重点人员画像分析方法的步骤。

基于深度NLP的社区重点人员画像分析方法及装置

技术领域

[0001] 本发明涉及大数据分析技术领域，尤其是基于深度NLP的社区重点人员画像分析方法及装置。

背景技术

[0002] 重点人员是指存在危害国家和社会安全嫌疑的人员，而重点人员管控工作是有关部门按照有关法律法规对重点人员进行一项带有秘密性的基础工作。在大数据时代的背景下，以往传统重点人员管控的模式受到了挑战，传统的重点人员管控的模式已经无法满足相关部门的正常需要，再加上当今的信息高速流通使得漏控、脱控现象时有发生，有关部门之间信息壁垒的存在，使得已经掌握的数据无法最大程度发挥作用，因而可以满足当前社会需求且可以全面、精准的对重点人员进行管控的方法已势在必行。

[0003] 例如专利公开号为“CN110727688A”、名称为“重点人员网格化服务管理系统”的中国发明专利，其包括对管辖区域进行网格化划分和规划的区域划分单元、对重点人员进行实时状态信息采集的信息采集单元、对整个区域内进行摄像监控的摄像监控单元、保存重点人员信息并在监控视频中标注出重点人员，对重点人员实时状态信息进行分析得出是否存在危险或犯罪的结果的后台处理中心，以及将后台处理中心得到的信息反馈给对应负责人的信息反馈单元，所述区域划分单元、信息采集单元、摄像监控单元、信息反馈单元均与后台处理中心连接。其采用摄像监控单元进行监控，并进行人工标注的方式，其工作量较大，且存在漏控、脱控现象。

[0004] 再如专利公开号为“CN107133646A”、名称为“一种通过人员行为轨迹识别重点人员的方法及系统”的中国发明专利，其收集重点人员和非重点人员的行为轨迹数据，建立重点人员轨迹信息集和非重点人员轨迹信息集；对重点人员轨迹信息集和非重点人员轨迹信息集的信息进行特征提取；从重点人员轨迹信息集和非重点人员轨迹信息集中提取每个人行为轨迹信息中的位置特征和时间特征，将位置特征和时间特征作为TOKEN串，并统计提取出的TOKEN串出现的次数和字频；建立动态模型；计算每个哈希表TOKEN串ti出现的概率；利用样本人员的轨迹特征进行学习训练，计算样本人员为重点人员的概率；建立新表；根据建立哈希表估计新人员为重点人员的可能性。该技术只从轨迹信息中获得数据，未从其他维度(社交维度等)进行综合考量，有可能出现误判或者漏判情况。

[0005] 再如专利公开号为“CN112330742A”、名称为“公共区域重点人员活动路线记录方法及装置”的中国发明专利，其包括：获取视频监控信息中行人的生物特征信息，生物特征信息包括：人脸特征、体型特征、步态特征和行为特征中的至少一项；将生物特征信息与数据库中预存的重点人员特征信息进行匹配，重点人员特征信息包括生物特征信息和身份信息；在生物特征信息与数据库中预存的重点人员特征信息匹配时，实时记录重点人员的活动位置信息；根据活动位置信息生成对应的重点人员活动轨迹。该技术只从个体角度进行分析来预估重点人员的风险状况，未进行群体分析来得出更可靠、准确的结论。

[0006] 因此，急需要提出一种逻辑简单、准确可靠的基于深度NLP的社区重点人员画像分

析方法及装置。

发明内容

[0007] 针对上述问题,本发明的目的在于提供基于深度NLP的社区重点人员画像分析方法,本发明采用的技术方案如下:

[0008] 第一部分,本技术提供了一种基于深度NLP的社区重点人员画像分析方法,其包括以下步骤:

[0009] 采集社区重点人员数据,并对重点人员数据进行预处理;

[0010] 采用word2ve算法对所述重点人员数据进行数值化处理;

[0011] 将数值化处理后的数据输入至CNN模型中进行特征提取;

[0012] 将提取的特征数据输入至LSTM网络中,得到局部特征的长距离特征,并经全连接层输出分类的标签数据;

[0013] 对分类的标签数据进行层次分析法权重分配,并求得人员个体的风险值;

[0014] 求得任一重点人员的离群程度;

[0015] 根据人员个体的风险值和离群程度,采用多标签聚类算法进行重点人员画像合成。

[0016] 第二部分,本技术提供了一种基于深度NLP的社区重点人员画像分析装置,其包括:

[0017] 数据采集预处理模块,采集社区重点人员数据,并对重点人员数据进行预处理;

[0018] 数值化处理模块,与数据采集预处理模块连接,采用word2ve算法对所述重点人员图像数据进行数值化处理;

[0019] CNN模型,与数值化处理模块连接,将数值化处理后的数据输入至CNN模型中进行特征提取;

[0020] LSTM网络,与CNN模型连接,将提取的特征数据输入至LSTM网络中,得到局部特征的长距离特征,并经全连接层输出分类的标签数据;

[0021] 层次分析模块,与LSTM网络连接,对分类的标签数据进行层次分析法权重分配,并求得人员个体的风险值;

[0022] 离群程度分析模块,与LSTM网络连接,求得任一重点人员的离群程度;

[0023] 以及,画像合成模块,与层次分析模块和离群程度分析模块连接,根据人员个体的风险值和离群程度,采用多标签聚类算法进行重点人员画像合成。

[0024] 与现有技术相比,本发明具有以下有益效果:

[0025] (1)本发明巧妙地采用word2ve算法对所述重点人员的数据进行数值化处理。其中,word2ve是一种无监督的学习模型,其主要思想是:在相似邻近词分布的中心词之间存在一定的语义相似度,它可以在一个语料集上实现词汇信息到语义空间的映射,最终获得一个词向量模型。本发明采用word2ve算法进行数值化处理,其从多个维度来刻画重点人员画像,以保证获取充足的数据集,便于对其进行个体分析。

[0026] (2)本发明巧妙地采用CNN模型进行特征提取,其使用CNN模型可以提取到数据的局部特征;并且,CNN模型中权值是共享的,从而减少参数数量,降低训练难度。

[0027] (3)本发明巧妙地采用LSTM网络对CNN模型取的特征向量进行处理,得到局部特征

的长距离特征，并经全连接层输出分类的标签数据；其好处在于，LSTM网络具有长时记忆功能，解决了长序列训练过程中存在的梯度消失和梯度爆炸问题。本发明巧妙地对分类的标签数据进行层次分析法权重分配，并求得人员个体的风险值；在本发明中，由于群体中存在不平衡现象，即在不同维度中不同风险程度重点人员的人数及所占比例存在较大差异；因此，本发明采用层次分析法，用于个体目标分析，其经过一个不同维度权重的风险权重分配，以得到一个综合权重，最终得到风险值。

[0028] (4)本发明通过求得任一重点人员的离群程度，该过程不同于个体目标分析，群体目标分析的目标是基于多个标签的数据，并计算每一个重点人员的离群程度。所谓离群程度指的是在全局管控工作中，某一重点人员与其他人员存在一定的偏离，其产生的原因是全局管控工作的动态变化。本发明通过计算每一个重点人员的离群程度，有利于本领域技术人员把握当前全局管控工作，实时地调整当前管控工作。

[0029] (5)本发明巧妙地结合人员个体的风险值和离群程度，并采用多标签聚类算法进行重点人员画像合成，其通过多标签聚类算法，可以对已经得到的个体风险值和离群程度进行整合，从而得到完整的重点人员画像。

[0030] 综上所述，本发明具有逻辑简单、准确可靠等优点，在大数据分析技术领域具有很高的实用价值和推广价值。

附图说明

[0031] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例中所需使用的附图作简单介绍，应当理解，以下附图仅示出了本发明的某些实施例，因此不应被看作是对保护范围的限定，对于本领域技术人员来说，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相关的附图。

[0032] 图1为本发明的逻辑流程图。

具体实施方式

[0033] 为使本申请的目的、技术方案和优点更为清楚，下面结合附图和实施例对本发明作进一步说明，本发明的实施方式包括但不限于下列实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0034] 如图1所示，本实施例提供了一种基于深度NLP的社区重点人员画像分析方法，本技术所提出的网络模型由三部分组成，第一部分是社区重点人员初步画像建立，通过获取海量的社区人员静态数据与动态数据样本，按照自然属性、社交属性、主题属性、经济状况等标签维度将数据分类，建立社区人员画像标签模型。第二部分是社区重点人员画像分析，将社区重点人员画像分析分为个体目标分析和群体目标分析。采用机器学习算法结合不平衡学习技术构建个体目标分析模型，用来判断个体人员的风险程度。构建基于局部异常因子算法用于群体目标分析，最终得到每一个重点人员的离群程度。第三部分将前两个部分的结果进行融合，得到最终的人员画像。

[0035] 具体来说，本技术的分析方法包括以下步骤：

[0036] 第一步，采集社区重点人员数据，并对重点人员数据进行预处理；该步骤是主要针

对剔除无意义的符号信息或其它的冗余信息。

[0037] 本步骤中使用的分词方法为基于词典规则的中文分词。主要是使用已经建立好的词库,通过词典匹配的方式使用正向最大匹配法对句子进行划分,其具体思想如下:

[0038] (1)按从左往右的顺序,从句子中取出m(m指词典中最长词的长度)个字作为匹配字段。

[0039] (2)查找字典,和取出的字段进行匹配;

[0040] (3)进入判断:

[0041] 匹配成功:将该字段作为一个词分出去;

[0042] 匹配不成功:将该字段最后一个字去掉,剩下的字作为新的匹配字段,再次进行匹配;

[0043] (4)循环以上过程直到分完为止。

[0044] 第二步,采用word2ve算法对所述重点人员数据进行数值化处理,为了让计算机能够理解词汇,需要将词汇信息映射到一个数值化的语义空间中一一词向量空间,便于后续操作。

[0045] 第三步,将数值化处理后的数据输入至CNN模型中进行特征提取;具体来说,该步骤使用了textCNN模型,其结构如下:

[0046] 第一层:将第二步中得到的词向量作为输入;

[0047] 第二层(卷积层):使用多个过滤器对词向量进行卷积,使用的激活函数为Relu;

[0048] 第三层(池化层):将卷积层的结果进行池化,并添加dropout正则来防止过拟合;

[0049] 第四层:输出提取的特征向量。

[0050] 第四步,将提取的特征数据输入至LSTM网络中,得到局部特征的长距离特征,并经全连接层输出分类的标签数据。在LSTM的每个单元中有三种类型的门:遗忘门、输入门和输出门。遗忘门决定上一时刻的单元状态 c_{t-1} 有多少保留到当前时刻的单元状态 c_t ;输入门决定当前时刻的隐藏状态输入 x_t 和上一时刻的隐藏状态输出 h_{t-1} 有多少保存到当前时刻的单元状态 c_t ; 输出门决定当前时刻的单元状态 c_t 有多少作为当前时刻的隐藏状态输出 h_t ; 遗忘门和输入门控制当前时刻t的 LSTM 单元状态 c_t 。

[0051] 遗忘门的公式如下:

[0052] $$f_t = \sigma(W_f[h_t, x_t] + b_f)$$

[0053] 其中, W_f 表示遗忘门的权重参数矩阵, b_f 表示遗忘门的偏置向量, σ 表示Sigmoid函数,其值域为(0,1)。

[0054] 输入门的公式如下:

[0055] $$i_t = \sigma(W_i[h_{t-1}, x_t] + b_f)$$
$$a_t = tanh(W_C[h_{t-1}, x_t] + b_C)$$

[0056] 其中，W_i 表示输入门的权重参数矩阵，W_C 表示输出门的权重参数矩阵，b_i 表示输入门的偏置向量，b_c 表示输出门的偏置向量，$tanh$ 表示激活函数，其值域为(-1，1)。

[0057] 单元状态公式如下：

[0058] $$c_t = f_t \odot c_{t-1} + i_t \odot \overline{c}_t)$$

[0059] 其中，$\overline{C_t}$ 表示候选单元状态，$\odot$表示按元素相乘。

[0060] 输出门公式如下：

[0061] $$o_t = \sigma(W_o[h_{t-1}, x_t] + b_0)$$

[0062] 其中，W_o 表示输出门的权重参数矩阵，b_o 表示输出门的偏置向量。

[0063] 隐藏状态 / 最终输出公式为：

[0064] $h_t = o_t \odot tanh(c_t)$ 。

[0065] 第五步，对分类的标签数据进行层次分析法权重分配，并求得人员个体的风险值。该步骤通过输入个体的不同标签的数据，最终得到该个体的风险状况。个体目标分析的目标是基于某一维度，判定重点人员的风险程度。所谓的风险程度是指存在实施领域专家定义的危险行为的可能。

[0066] 由于群体中存在不平衡现象，即在不同维度中不同风险程度重点人员的人数及所占比例存在较大差异，因而本发明使用基于AHP方法，用于个体目标分析，经过一个不同维度权重的风险权重分配，来得到一个综合权重，最终得到风险值。

[0067] 在此，人员个体的风险值的表达式为：

[0068] $$R = \sum_{i=1}^{n} Score_i \times Weight_i$$

[0069] 其中，R表示个体的风险值；n 表示标签的数量；$Score_i$ 表示标签风险度；$Weight_i$ 表示标签对应的风险权重。

[0070] AHP方法(层次分析法)可有效地将定性问题定量化，利用判断矩阵的最大特征值及特征向量，计算某层指标或因子相对于上层各指标或因子的权重值。其步骤如下：

[0071] 将数个标签作为输入，并搭建层次结构模型；

[0072] 构造判断矩阵，其是为了通过相对尺度的值来判定要比较的两个元素之间的重要程度，该值越大，则表明对比的两个元素中前者相对于后者更重要，如表1所示。

[0073] 表1判断矩阵表

[0074]

标度	含义
1	表示两个因素相比，具有相同重要性
3	表示两个因素相比，前者比后者稍重要
5	明显重要
7	强烈重要
9	极端重要
2,4,6,8	表示上述相邻判断的中间值

倒数	若因素 i 与因素 j 的重要性比为 α_{ij}，那么因素 j 与因素 i 重要性之比为 $\alpha_{ji}=\frac{1}{\alpha_{ij}}$

[0075] 对判断矩阵的任一列向量进行归一化后将任一行向量求和，并进行归一化处理得到列向量 w_i，其表达式为：

[0076] $$\overline{A_{ij}}=A_{ij}\Big/\sum_{k=1}^{n}A_{kj}(i=1,2,3...n;j=1,2,3...n)$$

[0077] $$\overline{W_i}=\sum_{j=1}^{n}\overline{A_{ij}}$$

[0078] $$W_i=\overline{W_i}\Big/\sum_{n=1}^{n}\overline{W_i}$$

[0079] 其中，A_{ij} 表示构造的判断矩阵，$\overline{A_{ij}}$ 表示归一化的判断矩阵，A_{kj} 表示第 k 行第j列的元素，k 表示行数。

[0080] 对列向量 w_i 的层次总排序，求得第K层元素相对于总目标的排序为：

[0081] $$\varphi_j^K=\sum_{i=1}^{n}p_{ij}^K W_j^K$$

[0082] 其中，p_{ij}^K 表示第K层元素相对于第K-1层元素的排序；W_j^K 表示K-1层元素对总目标的权重。

[0083] 第六步，求得任一重点人员的离群程度；本步骤不同于个体目标分析，群体目标分析的目标是基于多个标签的数据，计算每一个重点人员的离群程度。

[0084] LOF 主要是通过比较每个点p和其邻域点的密度来判断该点是否为异常点，如果点p的密度越低，越可能被认定是异常点。其密度通过点之间的欧氏距离来计算的，点之间欧氏距离越远，密度越低，距离越近，密度越高。

[0085] 在此，本发明中对该距离进行了优化，使用多个标签的欧氏距离来进行测算，重点人员的离群程度的局部离群因子的表达式为：

[0086] $$LOF_k(p)=\frac{\sum_{0\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}=\frac{\sum_{0\in N_k(p)}lrd_k(o)}{|N_k(p)|}\Big/lrd_k(p)$$

[0087] $$d(p,o)=\sum_{n=1}^{n}AHP_i\times d_i(p,o)$$

[0088] 其中，$d_k(p)$ 表示对于点 p 的第 k 距离，$d_k(p)=d_{(p,o)}$；$N_k(p)$ 表示距离领域点 p 的第 k 距离领域；$lrd_k(p)$ 表示点 p 的局部可达密度；$lrd_k(o)$ 表示点 o 的局部可达密度；AHP_i 表示第 i 个标签的权重值；$d_i(p,o)$ 表示第 i 个标签欧

氏距离；$d(p,o)$表示 p 和 o两点之间的距离；$LOF_k(p)$ 表示局部离群因子。

[0089] 第七步，根据人员个体的风险值和离群程度，采用多标签聚类算法进行重点人员画像合成。具体如下：

[0090] (1)从用户数据中抽取两种类型的标签(个体风险和离群程度)；

[0091] (2)计算每类标签对应的用户人数，即人员甲身上既有高风险类标签，又有高低群程度标签则记为数字1，同时存在两个标签的人越多，则表明相关部门需要对该类人进行及时管控。

[0092] (3)用余弦相似度函数计算两两标签之间的相关性，余弦值越高，其相似度越大，其表达式为：

[0093] $$cos\theta=\frac{A\cdot B}{|A|\times|B|}=\frac{\sum_{i=1}^{T}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{T}(A_i)^2}\times\sqrt{\sum_{i=1}^{T}(B_i)^2}}$$

[0094] A和B对应个体风险值与离群程度对应的T维向量。

[0095] (4)筛选出与每个A类标签相关性最大的B类标签，即将该A类标签归类到该B类标签，得到最终的用户画像。

[0096] 上述实施例仅为本发明的优选实施例，并非对本发明保护范围的限制，但凡采用本发明的设计原理，以及在此基础上进行非创造性劳动而作出的变化，均应属于本发明的保护范围之内。

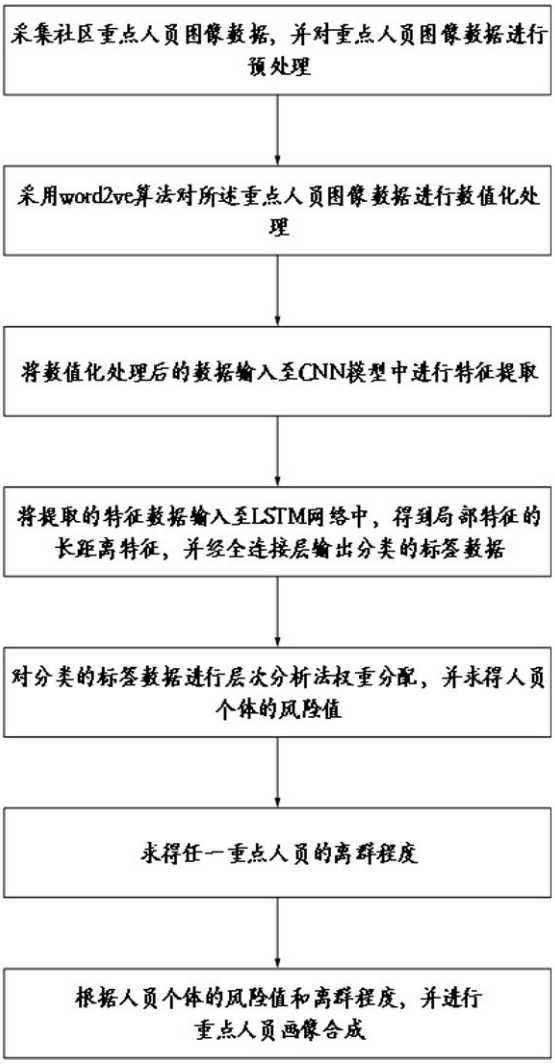

采集社区重点人员图像数据，并对重点人员图像数据进行预处理
采用word2vc算法对所述重点人员图像数据进行数值化处理
将数值化处理后的数据输入至CNN模型中进行特征提取
将提取的特征数据输入至LSTM网络中，得到局部特征的长距离特征，并经全连接层输出分类的标签数据
对分类的标签数据进行层次分析法权重分配，并求得人员个体的风险值
求得任一重点人员的离群程度
根据人员个体的风险值和离群程度，并进行重点人员画像合成

图1