

(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
2019 年 10 月 17 日 (17.10.2019)



(10) 国际公布号
WO 2019/196286 A1

(51) 国际专利分类号:

G16H 50/00 (2018.01)

(21) 国际申请号:

PCT/CN2018/099847

(22) 国际申请日:

2018 年 8 月 10 日 (10.08.2018)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201810321868.X 2018年4月11日 (11.04.2018) 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: 阮晓雯(RUAN, Xiaowen); 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)。 徐亮(XU,

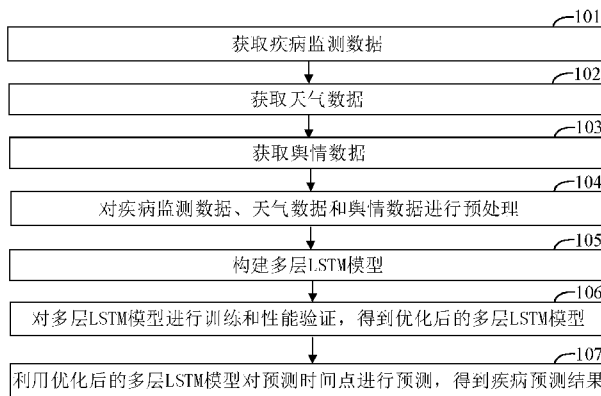
Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 肖京(XIAO, Jing); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市赛恩倍吉知识产权代理有限公司(SHENZHEN SCIENBIZIP INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市龙华新区龙观东路 83 号荣群大厦 9 楼, Guangdong 518109 (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,

(54) Title: ILLNESS PREDICTION METHOD AND DEVICE, COMPUTER DEVICE, AND READABLE STORAGE MEDIUM

(54) 发明名称: 疾病预测方法及装置、计算机装置及可读存储介质



101 Obtain illness monitoring data

102 Obtain weather data

103 Obtain public sentiment data

104 Preprocess illness monitoring data, weather data, and public sentiment data

105 Construct multilayer LSTM model

106 Train and verify performance of multilayer LSTM model to obtain optimized multilayer LSTM model

107 Use optimized multilayer LSTM model to predict predicted time point, obtaining illness prediction result at predicted time point

图 1

(57) Abstract: Provided is an illness prediction method, said method comprising: obtaining illness monitoring data, weather data, and public sentiment data; preprocessing said illness monitoring data, weather data, and public sentiment data; constructing a multilayer LSTM model; training and verifying the performance of the multilayer LSTM model to obtain an optimized multilayer LSTM model; using the optimized multilayer LSTM model to predict a predicted time point, obtaining an illness prediction result at the predicted time point. The present application also provides an illness prediction device, a computer device, and a readable storage medium. The present application achieves high-accuracy illness prediction.

(57) 摘要: 一种疾病预测方法, 所述方法包括: 获取疾病监测数据、天气数据和舆情数据; 对所述疾病监测数据、天气数据和舆情数据进行预处理; 构建多层LSTM模型; 对所述多层LSTM模型进行训练和性能验证, 得到优化后的多层LSTM模型; 利用所述优化后的多层LSTM模型对预测时间点进行预测, 得到所述预测时间点的疾病预测结果。本申请还提供一种疾病预测装置、计算机装置及可读存储介质。本申请可以实现高准确率的疾病预测。

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))。

疾病预测方法及装置、计算机装置及可读存储介质

本申请要求于2018年04月11日提交中国专利局,申请号为201810321868.X发明名称为“疾病预测方法及装置、计算机装置及可读存储介质”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

本申请涉及预测技术领域,具体涉及一种疾病预测方法及装置、计算机装置和非易失性可读存储介质。

背景技术

随着全球经济一体化进程的加快,经济与交流活动增加,人群流动日益频繁,为疾病的传播与爆发提供了有利环境,公共卫生健康问题越来越严峻。同时,社会与自然环境也发生着变化,环境污染、自然灾害等影响公众健康事件的增多也增加了突发公共卫生事件爆发的可能性。

如何能早期识别到疾病突发公共卫生事件,及时发出预警,尽早采取相应的控制措施,将突发公共卫生事件造成的损失降到最低,是公共卫生领域长期以来关注的焦点,也是卫生应急工作的重要内容。突发公共卫生事件预警,是通过对有关数据的收集,整理、分析和整合,运用计算机、网络、通讯等现代先进的技术,对事件的征兆进行监测、识别、诊断与评价,及时报警,告知有关部门和公众做好相关的应对和准备工作,及时采取有效的防控措施,尽可能阻止或减缓突发事件的发生或减少事件的危害。

突发公共卫生事件预警中的一个重要工作是疾病预测,即根据历史的疾病监测数据(即病患数据)预测未来的疾病监测数据。随着机器学习技术的发展,越来越多的机器学习方法应用在疾病预测上。然而,应用于疾病预测的传统的机器学习往往需要人为去定义特征集,然后从定义好的特征集中搜寻最好的特征组合,且效果往往都不够好,从而影响了疾病预测的准确率。

发明内容

鉴于以上内容,有必要提出一种疾病预测方法及装置、计算机装置和非易失性可读存储介质,其可以实现高准确率的疾病预测。

本申请的第一方面提供一种疾病预测方法,所述方法包括:

获取疾病监测数据,所述疾病监测数据是时间序列数据;获取所述疾病监测数据相关的天气数据,所述天气数据是与所述疾病监测数据对应的时间序列数据;获取所述疾病监测数据相关的舆情数据,所述舆情数据是与所述疾病监测数据对应的时间序列数据;对所述疾病监测数据、天气数据和舆情数据进行预处理;构建多层长短时记忆递归神经网络模型,即多层 LSTM 模型;从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据,利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证,得到优化后的多层 LSTM 模型;从预处理后的所述疾病监测数据、天气数据和舆情

数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

本申请的第二方面提供一种疾病预测装置，所述装置包括：

第一获取单元，用于获取疾病监测数据，所述疾病监测数据是时间序列数据；

第二获取单元，用于获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；

第三获取单元，用于获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据；

预处理单元，用于对所述疾病监测数据、天气数据和舆情数据进行预处理；

构建单元，用于构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；

优化单元，用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；

预测单元，用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

本申请的第三方面提供一种计算机装置，所述计算机装置包括存储器及处理器，所述存储器用于存储至少一个计算机可读指令，所述处理器用于执行所述至少一个计算机可读指令以实现以下步骤：

获取疾病监测数据，所述疾病监测数据是时间序列数据；获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据；对所述疾病监测数据、天气数据和舆情数据进行预处理；构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

本申请的第四方面提供一种非易失性可读存储介质，所述非易失性可读存储介质存储有至少一个计算机可读指令，所述至少一个计算机可读指令被处理器执行时实现以下步骤：

获取疾病监测数据，所述疾病监测数据是时间序列数据；获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监

测数据对应的时间序列数据；对所述疾病监测数据、天气数据和舆情数据进行预处理；构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

本申请获取疾病监测数据，所述疾病监测数据是时间序列数据；获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据；对所述疾病监测数据、天气数据和舆情数据进行预处理；构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

本申请通过多层 LSTM 模型对患病数据进行预测。LSTM 模型可以从数据中直接去提取知识，构造出有利于预测的特征向量，提高了预测精度。同时，与传统的 RNN (Recurrent Neural Networks, 循环神经网络) 模型相比，LSTM 模型解决了在时序数据量过大情况下引起的长期依赖时梯度消失的问题。并且，本申请将天气数据、舆情数据作为影响因素在加入到疾病预测中，提高了疾病预测的准确性。因此，本申请实现了高准确率的疾病预测。

附图说明

图 1 是本申请实施例一提供的疾病预测方法的流程图。

图 2 是本申请实施例二提供的疾病预测方法中获取疾病监测数据相关的天气数据的细化流程图。

图 3 是本申请实施例三提供的疾病预测装置的结构图。

图 4 是本申请实施例四提供的疾病预测装置第二获取单元的细化结构图。

图 5 是本申请实施例五提供的计算机装置的示意图。

具体实施方式

优选地，本申请的疾病预测方法应用在一个或者多个计算机装置中。所述计算机装置是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路 (Application Specific Integrated Circuit, ASIC)、可编程门阵列 (Field Programmable Gate Array, FPGA)、数字处理器 (Digital Signal Processor, DSP)、嵌入式设备等。

实施例一

图 1 是本申请实施例一提供的疾病预测方法的流程图。所述疾病预测方法应用于计算机装置。所述疾病预测方法利用长短时记忆递归神经网络模型对疾病监测数据进行预测，获得高准确率的疾病预测结果。

如图 1 所示，所述疾病预测方法具体包括以下步骤：

步骤 101，获取疾病监测数据，所述疾病监测数据是时间序列数据。

所述疾病监测数据可以包括流感、手足口病、麻疹、流行性腮腺炎等疾病的患病数据。

可以在预设区域（例如省市、地区）建立由多个监测点组成的疾病监测网络，从所述监测点获取疾病监测数据，由所述疾病监测数据构成疾病监测的时间序列数据。可以选择医疗机构、学校和幼托机构、药店等作为监测点，分别对相应的目标人群进行疾病监测及数据采集。可以选择满足预设条件的场所作为监测点。所述预设条件可以包括人数、规模等。例如，选择学生人数达到预设数量的学校和幼托机构作为监控点。又如，选择规模（例如以日营业额统计）达到预设规模的药店作为监控点。再如，选择规模（例如以日就医人数统计）达到预设规模的医院作为监控点。

不同时间的疾病监测数据构成疾病监测的时间序列数据。例如，可以将以日为单位采集到的疾病监测数据构成疾病监测的时间序列数据。或者，可以将以周为单位采集到的疾病监测数据构成疾病监测的时间序列数据。

医疗机构（主要包括医院）是最能捕捉疾病早期暴发预兆的场所，是开展疾病监测的首选。可以根据病人就诊情况，获取疾病监测数据。

一部分病人会自行去药店购药来缓解早期症状，因此，可以根据药店的药品销售情况，获取疾病监测数据。

儿童和青少年是疾病的高危人群以及疾病传播过程中的重要环节，也应该加强对该人群的监测。学校和幼托机构是监测儿童和青少年疾病发病情况的较佳场所。可以根据学校和幼托机构的儿童和青少年的请假情况，获得疾病监测数据。

因此，本申请中主要选择医疗机构、学校和幼托机构、药店这三类场所进行疾病监测数据的采集。当然，上述对数据源的选择，并不能限制在另外的实施方案中增加或替换其他重点关注人群或场所作为监测的数据源。例如，可以将宾馆纳入疾病监测范围，获取宾馆入住人员的疾病监测数据。

根据需要，可以取任意一类监控点（例如医疗机构）采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以取医院采集的疾病监测数据构成疾病监测的时间序列数据。或者，可以结合多类监控点采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以以医院采集的疾病监测数据为主，以药店参加的疾病监测数据作为补充，构成疾病监测的时间序列数据。

疾病监测数据可以包括疾病的就诊数、就诊率、发病数、发病率等患病数据。例如，可以从医疗机构（例如医院）获取疾病（例如流感）的每日就诊数，将疾病（例如流感）的每日就诊数作为疾病监测数据。又如，可以从学校获取学生的疾病（例如流感）的每日发病数，将疾病（例如流感）的每

日发病数作为疾病监测数据。

步骤 102, 获取所述疾病监测数据相关的天气数据, 所述天气数据是与所述疾病监测数据对应的时间序列数据。

疾病监测数据相关的天气数据是指对疾病监测数据(即疾病的患病数据)有影响的天气数据。可以预先分析不同天气数据对所述疾病监测数据的影响, 根据分析结果确定对所述疾病监测数据有影响或影响较大的天气数据。

所述天气数据可以包括湿度、气温、气压、降水量、水汽压、风速、风向、日照时数。在一具体实施例中, 所述天气数据可以包括每日的平均气温、平均气压、最高气温、最低气温、平均相对湿度、最小相对湿度、降水量、平均风速、日照时数、平均水汽压。

所述天气数据与所述疾病监测数据对应的时间段相同, 并且, 所述天气数据与所述疾病监测数据的统计周期(例如每日、每周)相同。例如, 所述疾病监测数据为 2018 年 1-2 月的每日就诊数, 所述天气数据为 2018 年 1-2 月的每日天气数据。又如, 所述疾病监测数据为 2017 年 1-12 月的每周就诊数, 所述天气数据为 2017 年 1-12 月的每周天气数据(例如周平均气温)。

可以从天气信息网站(例如中国天气网、新浪天气、搜狐天气等)抓取所述天气数据, 以提高天气数据的可靠性。可以理解, 可以从任意网页中抓取所述天气数据。

可以抓取预定区域的天气数据。所述预定区域可以包括省、市、地区等。例如, 抓取深圳市的天气数据。

可以抓取预定时间的天气数据。所述预定时间可以包括年、月、日等。例如, 抓取 2018 年 1-2 月的每日天气数据。

可以通过网络爬虫抓取所述天气数据。网络爬虫是一个可以自动提取网页数据信息内容的应用程序。网络爬虫通常是从一个或者是若干个初始网页的 URL(也称种子 URL)开始, 获取初始网页的 URL, 依照特定的算法和策略(例如深度优先搜索策略), 在对网页进行抓取的过程中, 不断地从当前的网页中抽取新的 URL 放入到相应的队列中, 直到满足停止条件为止。URL 为 Uniform Resource Locator 的缩写, 即统一资源定位符。

可以利用天气信息网站开放的 API 接口(例如中国天气网开放的 API 接口)抓取所述天气数据。API 是应用程序接口(application interface)的缩写, 通过 API 接口可以实现计算机软件之间的相互通信。天气信息网站开放的 API 接口可以返回 JSON 格式或者 XML 格式的数据。

在一具体实施例中, 可以利用天气信息网站开放的 API 接口, 通过网络爬虫抓取所述天气数据。利用天气信息网站开放的 API 接口, 通过网络爬虫抓取所述天气数据的具体过程参见图 2。

步骤 103, 获取所述疾病监测数据相关的舆情数据, 所述舆情数据是与所述疾病监测数据对应的时间序列数据。

疾病监测数据相关的舆情数据是指体现所述疾病监测数据的舆情数据。举例来说, 当疾病(例如流感)进入流行期时, 随着患病人数增多, 很多人会上网搜索疾病相关的词语(例如流感、达菲、高烧等特定词), 这些词语

的搜索量大大增加。又如，当疾病（例如流感）进入流行期时，随着患病人数增多，新闻、论坛、博客、贴吧等舆情网站上发布的疾病相关内容（例如患病信息、治疗信息等）增多。因此，可以利用疾病监测数据相关的舆情数据辅助进行疾病预测。

所述舆情数据可以包括特定词的搜索次数。例如，可以统计预设搜索引擎对特定词的搜索次数（例如特定地区预设定搜索引擎对特定词的每日搜索次数）。

所述舆情数据还可以包括特定舆情网站（例如新闻、论坛、博客、贴吧等）包含特定词的舆情信息的数量。

所述特定词是与预测的疾病相关的词语，例如，所述特定词是疾病症状相关的词语，当预测的疾病为流感时，所述特定词可以包括：发病突然、高烧、畏寒、头痛、无力、喉咙发炎、肌肉酸痛、干咳等。再如，当预测的疾病为手足口时，所述特定词可以包括：口痛、厌食、低热、手部小疱疹、口部小溃疡等。

所述舆情数据与所述疾病监测数据对应的时间段相同，并且，所述舆情数据与所述疾病监测的统计周期（例如每日、每周）相同。例如，所述疾病监测数据为 2018 年 1-2 月的每日就诊数，则所述舆情数据为 2018 年 1-2 月的每日舆情数据（例如特定词日搜索次数）。又如，所述疾病监测数据为 2017 年 1-12 月的每周就诊数，则所述舆情数据为 2017 年 1-12 月的每周舆情数据（例如特定词周搜索次数）。

可以理解，步骤 101-103 可以以任意顺序执行，也可以并行执行。

步骤 104，对所述疾病监测数据、天气数据和舆情数据进行预处理。

疾病监测数据、天气数据和舆情数据的预处理可以包括异常数据处理。对疾病监测数据、天气数据和舆情数据进行异常数据处理，是为了修正所述疾病监测数据、天气数据和舆情数据中的异常数据，提高疾病预测的可靠性和准确性。

所述异常数据处理可以包括填补所述疾病监测数据、天气数据和舆情数据中的缺失值。可以通过缺失值前后数据的平均值或中值来对缺失值进行填充，或者，可以通过回归拟合的方法对缺失值进行填充。

所述异常数据处理还可以包括修正所述疾病监测数据、天气数据和舆情数据中的异常值。所述异常值是明显偏离其他数据的数值。可以采用插值法修正所述异常值。

疾病监测数据、天气数据和舆情数据的预处理还可以包括对所述疾病监测数据、天气数据和舆情数据进行数据格式转换。例如，对疾病监测数据、天气数据和舆情数据进行标准化处理，使得疾病监测数据、天气数据和舆情数据具有一致性的标准格式，以适合作为 LSTM 模型的输入数据。

步骤 105，构建多层长短时记忆递归神经网络（Long Short-term Memory Recurrent Neural Network）模型，即多层 LSTM 模型。所述多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据（例如所述疾病监测数据、天气数据和舆情数据构成的输入数据）构造特

征, 得到第一隐藏层单元, 所述第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合, 得到第二隐藏层单元, 所述全连接层用于根据所述第二隐藏层单元得到预测结果 (例如疾病预测结果), 每个 LSTM 单元层包括遗忘门、输入门、输出门, 所述遗忘门、输入门、输出门控制所述 LSTM 单元层的记忆状态。

LSTM 模型是一种时间递归神经网络模型。相对于传统的循环神经网络 (Recurrent Neural Network, RNN) 模型, LSTM 模型通过在 LSTM 单元层构建一些门来存储信息, 因此其在模型训练的过程中, 梯度不会很快消失。

本方法使用的多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层, 第一层 LSTM 单元层用于对输入数据 (例如疾病监测数据、天气数据和舆情数据构成的输入数据) 构造特征, 得到第一隐藏层单元, 所述第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合, 得到第二隐藏层单元。所述全连接层根据所述第二隐藏层单元得到预测值。所述第一隐藏层单元为局部特征, 所述第二隐藏层单元为全局特征。也就是说, 第一层 LSTM 单元层用于提取局部信息, 第二层 LSTM 单元层用于结合局部特征得到全局特征, 所述全连接层用于根据全局特征得到预测结果 (例如疾病预测结果)。

LSTM 单元层包括遗忘门、输入门、输出门, 所述遗忘门、输入门、输出门控制 LSTM 单元层的记忆状态。输入门决定是否接收当前时刻的输入。输出门决定是否是否输出记忆状态。

在一实施例中, LSTM 单元层的遗忘门 f_t , 输入门 i_t 、输出门 o_t 、记忆状态 c_t 和隐藏层单元 h_t 可以计算如下:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f); \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i); \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o); \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c); \\ h_t &= o_t \odot \tanh(c_t). \end{aligned}$$

其中, W_f 、 U_f 、 b_f 为遗忘门的参数, W_i 、 U_i 、 b_i 为输入门的参数, W_o 、 U_o 、 b_o 为输出门的参数, W_c 、 U_c 、 b_c 为记忆单元的参数。

在另一实施例中, LSTM 单元层的遗忘门 f_t , 输入门 i_t 、输出门 o_t 、记忆状态 c_t 和隐藏层单元 h_t 可以计算如下:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f c_{t-1} + b_f); \\ i_t &= \sigma(W_i x_t + U_i c_{t-1} + b_i); \\ o_t &= \sigma(W_o x_t + U_o c_{t-1} + b_o); \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + b_c); \\ h_t &= o_t \odot \tanh(c_t). \end{aligned}$$

步骤 106, 从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据, 利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证, 得到优化后的多层 LSTM 模型。

可以从预处理后的所述疾病监测数据、天气数据和舆情数据中截取时间

序列数据，构成所述训练数据和所述验证数据。

所述多层 LSTM 模型的输入数据是一个预设维度（例如 1000 维）的向量。可以从截取的时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的向量，按照时间顺序，将各个时间点对应的向量依次输入所述多层 LSTM 模型，用来对所述多层 LSTM 模型进行训练或验证。

例如，从预处理后的所述疾病监测数据、天气数据和舆情数据中截取用于训练所述多层 LSTM 模型的第一时间序列数据；从截取的第一时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第一向量，按照时间顺序，将各个时间点对应的第一向量依次输入所述多层 LSTM 模型，用于对所述多层 LSTM 模型进行训练。从预处理后的所述疾病监测数据、天气数据和舆情数据中截取用于验证所述多层 LSTM 模型的第二时间序列数据；从截取的第二时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第二向量，按照时间顺序，将各个时间点对应的第二向量依次输入所述多层 LSTM 模型，用于对所述多层 LSTM 模型进行验证。

在对所述多层 LSTM 模型进行训练时，所述多层 LSTM 模型的损失函数可以定义为均方差，调整所述多层 LSTM 模型的参数，使所述均方差取得最小值。训练的过程可以采用 RMSprop 算法。RMSprop 是一种改进的随机梯度下降算法。均方差与 RMSprop 算法是现有技术，此处不再赘述。

步骤 107，从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

获取的预测时间点之前的疾病监测数据、天气数据和舆情数据为时间序列数据。可以从获取的预测时间点之前的疾病监测数据、天气数据和舆情数据中，将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第三向量，按照时间顺序，将各个时间点对应的第三向量依次输入所述多层 LSTM 模型，以对预测时间点进行疾病预测。

在进行疾病预测时，从初始时间点开始，优化后的多层 LSTM 模型通过当前时间点的输入数据及前一时间点的隐藏层单元逐层组合得到当前时间点的各隐藏层单元，根据当前时间点的隐藏层单元得到当前时间点的预测值，并根据时间顺序，不断递归获取下一时间点的隐藏层单元以及预测值，直到得到所述给定时间点的预测值。

实施例一通过多层 LSTM 模型对患病数据进行预测。LSTM 模型可以从数据中直接去提取知识，构造出有利于预测的特征向量，提高了预测精度。同时，与传统的 RNN 模型相比，LSTM 模型解决了在时序数据量过大情况下引起的长期依赖时梯度消失的问题。并且，实施例一将天气数据、舆情数据作为影响因素在加入到疾病预测中，提高了疾病预测的准确性。

实施例二

图 2 是本申请实施例二提供的疾病预测方法中获取疾病监测数据相关的天气数据（即图 1 中步骤 102）的细化流程图。

可以利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据。参阅图 2 所示，具体可以包括以下步骤：

步骤 201，生成面向所述天气信息网站的 API 接口的种子 URL 以及后续的 URL。

种子 URL 是网络爬虫进行一切工作的基础和前提。种子 URL 可以是一个也可以是多个。可以对天气信息网站的 URL 的结构特点进行分析，根据 URL 的结构特点得到后续的 URL。

步骤 202，向所述天气信息网站的 API 接口发送 HTTP 请求，请求访问所述 API 接口。

可以以 GET 方式向所述天气信息网站的 API 接口发送 HTTP 请求。当天气信息网站同意获取其提供的天气数据时，返回 HTTP 响应，以告知可以进行获取天气数据的操作。

步骤 203，对所述天气信息网站提供的数据进行分析和识别，以查看所述数据内容。

天气信息网站提供特定格式的数据内容，需要对天气信息网站提供的特定格式的数据内容进行分析 and 识别，来查看所述数据内容。例如，所述天气信息网站的 API 接口提供的数据格式为 JSON 格式。JSON 是一种数据交换格式，使用了类似于 C 语言的语法习惯。对该 JSON 格式的数据内容进行分析 and 识别，来查看所述数据内容。

步骤 204，判断所述数据内容是否为预定信息内容。

为了得到特定的天气数据，需要判断所述数据内容是否为预定信息内容。若所述数据内容是否不是预定信息内容，则舍弃该数据内容，否则执行下一步骤。

步骤 205，若所述数据内容为预定信息内容，则抓取所述数据内容。

数据抓取的最终目的是将网络数据内容抓取到本地。对于 JSON 格式的数据内容，在抓取所述数据内容时可以采用深度优先搜索策略进行状态空间搜索。

步骤 206，将抓取的数据内容作为所述天气数据保存到本地。

可以在计算设备上创建数据库，将所述天气数据保存到所述数据库中。

传统的网络爬虫都是首先设定一个或者多个入口 URL，在抓取网页的过程中，按照抓取的策略，从当前网页上提取出新的 URL 放入队列，以便获取 URL 对应的网页内容，将网页内容保存到本地，然后，再提取有效地址作为下一次的入口 URL，直到爬行完毕。随着网页数量的剧增，传统的网络爬虫会下载大量的无关网页。利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据，可以避免下载无关网页，高效地获取天气数据，从而提高疾病预测的效率。

实施例三

图 3 为本申请实施例三提供的疾病预测装置的结构图。如图 3 所示，所

述疾病预测装置 10 可以包括：第一获取单元 301、第二获取单元 302、第三获取单元 303、预处理单元 304、构建单元 305、优化单元 306、预测单元 307。

第一获取单元 301，用于获取疾病监测数据，所述疾病监测数据是时间序列数据。

所述疾病监测数据可以包括流感、手足口病、麻疹、流行性腮腺炎等疾病的患病数据。

可以在预设区域（例如省市、地区）建立由多个监测点组成的疾病监测网络，从所述监测点获取疾病监测数据，由所述疾病监测数据构成疾病监测的时间序列数据。可以选择医疗机构、学校和幼托机构、药店等作为监测点，分别对相应的目标人群进行疾病监测及数据采集。可以选择满足预设条件的场所作为监测点。所述预设条件可以包括人数、规模等。例如，选择学生人数达到预设数量的学校和幼托机构作为监控点。又如，选择规模（例如以日营业额统计）达到预设规模的药店作为监控点。再如，选择规模（例如以日就医人数统计）达到预设规模的医院作为监控点。

不同时间的疾病监测数据构成疾病监测的时间序列数据。例如，可以将以日为单位采集到的疾病监测数据构成疾病监测的时间序列数据。或者，可以将以周为单位采集到的疾病监测数据构成疾病监测的时间序列数据。

医疗机构（主要包括医院）是最能捕捉疾病早期暴发预兆的场所，是开展疾病监测的首选。可以根据病人就诊情况，获取疾病监测数据。

一部分病人会自行去药店购药来缓解早期症状，因此，可以根据药店的药品销售情况，获取疾病监测数据。

儿童和青少年是疾病的高危人群以及疾病传播过程中的重要环节，也应该加强对该人群的监测。学校和幼托机构是监测儿童和青少年疾病发病情况的较佳场所。可以根据学校和幼托机构的儿童和青少年的请假情况，获得疾病监测数据。

因此，本申请中主要选择医疗机构、学校和幼托机构、药店这三类场所进行疾病监测数据的采集。当然，上述对数据源的选择，并不能限制在另外的实施方案中增加或替换其他重点关注人群或场所作为监测的数据源。例如，可以将宾馆纳入疾病监测范围，获取宾馆入住人员的疾病监测数据。

根据需要，可以取任意一类监控点（例如医疗机构）采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以取医院采集的疾病监测数据构成疾病监测的时间序列数据。或者，可以结合多类监控点采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以以医院采集的疾病监测数据为主，以药店参加的疾病监测数据作为补充，构成疾病监测的时间序列数据。

疾病监测数据可以包括疾病的就诊数、就诊率、发病数、发病率等患病数据。例如，可以从医疗机构（例如医院）获取疾病（例如流感）的每日就诊数，将疾病（例如流感）的每日就诊数作为疾病监测数据。又如，可以从学校获取学生的疾病（例如流感）的每日发病数，将疾病（例如流感）的每日发病数作为疾病监测数据。

第二获取单元 302，用于获取所述疾病监测数据相关的天气数据，所述

天气数据是与所述疾病监测数据对应的时间序列数据。

疾病监测数据相关的天气数据是指对疾病监测数据（即疾病的患病数据）有影响的天气数据。可以预先分析不同天气数据对所述疾病监测数据的影响，根据分析结果确定对所述疾病监测数据有影响或影响较大的天气数据。

所述天气数据可以包括湿度、气温、气压、降水量、水汽压、风速、风向、日照时数。在一具体实施例中，所述天气数据可以包括每日的平均气温、平均气压、最高气温、最低气温、平均相对湿度、最小相对湿度、降水量、平均风速、日照时数、平均水汽压。

所述天气数据与所述疾病监测数据对应的时间段相同，并且，所述天气数据与所述疾病监测数据的统计周期（例如每日、每周）相同。例如，所述疾病监测数据为 2018 年 1-2 月的每日就诊数，所述天气数据为 2018 年 1-2 月的每日天气数据。又如，所述疾病监测数据为 2017 年 1-12 月的每周就诊数，所述天气数据为 2017 年 1-12 月的每周天气数据（例如周平均气温）。

可以从天气信息网站（例如中国天气网、新浪天气、搜狐天气等）抓取所述天气数据，以提高天气数据的可靠性。可以理解，可以从任意网页中抓取所述天气数据。

可以抓取预定区域的天气数据。所述预定区域可以包括省、市、地区等。例如，抓取深圳市的天气数据。

可以抓取预定时间的天气数据。所述预定时间可以包括年、月、日等。例如，抓取 2018 年 1-2 月的每日天气数据。

可以通过网络爬虫抓取所述天气数据。

可以利用天气信息网站开放的 API 接口（例如中国天气网开放的 API 接口）抓取所述天气数据。

在一具体实施例中，可以利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据。利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据的具体过程参见图 2。

第三获取单元 303，用于获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据。

疾病监测数据相关的舆情数据是指体现所述疾病监测数据的舆情数据。举例来说，当疾病（例如流感）进入流行期时，随着患病人数增多，很多人会上网搜索疾病相关的词语（例如流感、达菲、高烧等特定词），这些词语的搜索量大大增加。又如，当疾病（例如流感）进入流行期时，随着患病人数增多，新闻、论坛、博客、贴吧等舆情网站上发布的疾病相关内容（例如患病信息、治疗信息等）增多。因此，可以利用疾病监测数据相关的舆情数据辅助进行疾病预测。

所述舆情数据可以包括特定词的搜索次数。例如，可以统计预设搜索引擎对特定词的搜索次数（例如特定地区预设定搜索引擎对特定词的每日搜索次数）。

所述舆情数据还可以包括特定舆情网站（例如新闻、论坛、博客、贴吧等）包含特定词的舆情信息的数量。

所述特定词是与预测的疾病相关的词语，例如，所述特定词是疾病症状相关的词语，当预测的疾病为流感时，所述特定词可以包括：发病突然、高烧、畏寒、头痛、无力、喉咙发炎、肌肉酸痛、干咳等。再如，当预测的疾病为手足口时，所述特定词可以包括：口痛、厌食、低热、手部小疱疹、口部小溃疡等。

所述舆情数据与所述疾病监测数据对应的时间段相同，并且，所述舆情数据与所述疾病监测的统计周期（例如每日、每周）相同。例如，所述疾病监测数据为 2018 年 1-2 月的每日就诊数，则所述舆情数据为 2018 年 1-2 月的每日舆情数据（例如特定词日搜索次数）。又如，所述疾病监测数据为 2017 年 1-12 月的每周就诊数，则所述舆情数据为 2017 年 1-12 月的每周舆情数据（例如特定词周搜索次数）。

预处理单元 304，用于对所述疾病监测数据、天气数据和舆情数据进行预处理。

疾病监测数据、天气数据和舆情数据的预处理可以包括异常数据处理。对疾病监测数据、天气数据和舆情数据进行异常数据处理，是为了修正所述疾病监测数据、天气数据和舆情数据中的异常数据，提高疾病预测的可靠性和准确性。

所述异常数据处理可以包括填补所述疾病监测数据、天气数据和舆情数据中的缺失值。可以通过缺失值前后数据的平均值或中值来对缺失值进行填充，或者，可以通过回归拟合的方法对缺失值进行填充。

所述异常数据处理还可以包括修正所述疾病监测数据、天气数据和舆情数据中的异常值。所述异常值是明显偏离其他数据的数值。可以采用插值法修正所述异常值。

疾病监测数据、天气数据和舆情数据的预处理还可以包括对所述疾病监测数据、天气数据和舆情数据进行数据格式转换。例如，对疾病监测数据、天气数据和舆情数据进行标准化处理，使得疾病监测数据、天气数据和舆情数据具有一致性的标准格式，以适合作为 LSTM 模型的输入数据。

构建单元 305，用于构建多层长短时记忆递归神经网络（Long Short-term Memory Recurrent Neural Network）模型，即多层 LSTM 模型。所述多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据（例如所述疾病监测数据、天气数据和舆情数据构成的输入数据）构造特征，得到第一隐藏层单元，所述第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合，得到第二隐藏层单元，所述全连接层用于根据所述第二隐藏层单元得到预测结果（例如疾病预测结果），每个 LSTM 单元层包括遗忘门、输入门、输出门，所述遗忘门、输入门、输出门控制所述 LSTM 单元层的记忆状态。

本方法使用的多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据（例如疾病监测数据、天气数据和舆情数据构成的输入数据）构造特征，得到第一隐藏层单元，所述第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合，得到第二隐藏层单元。所述全

连接层根据所述第二隐藏层单元得到预测值。所述第一隐藏层单元为局部特征，所述第二隐藏层单元为全局特征。也就是说，第一层 LSTM 单元层用于提取局部信息，第二层 LSTM 单元层用于结合局部特征得到全局特征，所述全连接层用于根据全局特征得到预测结果（例如疾病预测结果）。

LSTM 单元层包括遗忘门、输入门、输出门，所述遗忘门、输入门、输出门控制 LSTM 单元层的记忆状态。输入门决定是否接收当前时刻的输入。输出门决定是否是否输出记忆状态。

在一实施例中，LSTM 单元层的遗忘门 f_t ，输入门 i_t 、输出门 o_t 、记忆状态 c_t 和隐藏层单元 h_t 可以计算如下：

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f); \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i); \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o); \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c); \\ h_t &= o_t \odot \tanh(c_t). \end{aligned}$$

其中， W_f 、 U_f 、 b_f 为遗忘门的参数， W_i 、 U_i 、 b_i 为输入门的参数， W_o 、 U_o 、 b_o 为输出门的参数， W_c 、 U_c 、 b_c 为记忆单元的参数。

在另一实施例中，LSTM 单元层的遗忘门 f_t ，输入门 i_t 、输出门 o_t 、记忆状态 c_t 和隐藏层单元 h_t 可以计算如下：

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f c_{t-1} + b_f); \\ i_t &= \sigma(W_i x_t + U_i c_{t-1} + b_i); \\ o_t &= \sigma(W_o x_t + U_o c_{t-1} + b_o); \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + b_c); \\ h_t &= o_t \odot \tanh(c_t). \end{aligned}$$

优化单元 306，用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型。

可以从预处理后的所述疾病监测数据、天气数据和舆情数据中截取时间序列数据，构成所述训练数据和所述验证数据。

所述多层 LSTM 模型的输入数据是一个预设维度（例如 1000 维）的向量。可以从截取的时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的向量，按照时间顺序，将各个时间点对应的向量依次输入所述多层 LSTM 模型，用来对所述多层 LSTM 模型进行训练或验证。

例如，从预处理后的所述疾病监测数据、天气数据和舆情数据中截取用于训练所述多层 LSTM 模型的第一时间序列数据；从截取的第一时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第一向量，按照时间顺序，将各个时间点对应的第一向量依次输入所述多层 LSTM 模型，用于对所述多层 LSTM 模型进行训练。从预处理后的所述疾病监测数据、天气数据和舆情数据中截取用于验证所述多层

LSTM 模型的第二时间序列数据；从截取的第二时间序列数据中将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第二向量，按照时间顺序，将各个时间点对应的第二向量依次输入所述多层 LSTM 模型，用于对所述多层 LSTM 模型进行验证。

在对所述多层 LSTM 模型进行训练时，所述多层 LSTM 模型的损失函数可以定义为均方差，调整所述多层 LSTM 模型的参数，使所述均方值取得最小值。训练的过程可以采用 RMSprop 算法。RMSprop 是一种改进的随机梯度下降算法。均方差与 RMSprop 算法是现有技术，此处不再赘述。

预测单元 307，用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

获取的预测时间点之前的疾病监测数据、天气数据和舆情数据为时间序列数据。可以从获取的预测时间点之前的疾病监测数据、天气数据和舆情数据中，将每个时间点对应的预处理后的疾病监测数据、天气数据和舆情数据构造一个预设维度的第三向量，按照时间顺序，将各个时间点对应的第三向量依次输入所述多层 LSTM 模型，以对预测时间点进行疾病预测。

在进行疾病预测时，从初始时间点开始，优化后的多层 LSTM 模型通过当前时间点的输入数据及前一时间点的隐藏层单元逐层组合得到当前时间点的各隐藏层单元，根据当前时间点的隐藏层单元得到当前时间点的预测值，并根据时间顺序，不断递归获取下一时间点的隐藏层单元以及预测值，直到得到所述给定时间点的预测值。

实施例三通过多层 LSTM 模型对患病数据进行预测。LSTM 模型可以从数据中直接去提取知识，构造出有利于预测的特征向量，提高了预测精度。同时，与传统的 RNN 模型相比，LSTM 模型解决了在时序数据量过大情况下引起的长期依赖时梯度消失的问题。并且，实施例三将天气数据、舆情数据作为影响因素在加入到疾病预测中，提高了疾病预测的准确性。

实施例四

图 4 是本申请实施例四提供的疾病预测装置中第二获取单元（即图 3 中 302）的细化结构图。

第二获取单元 302 可以利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据。参阅图 4 所示，第二获取单元 302 可以包括：生成子单元 3021、请求子单元 3022、分析子单元 3023、判断子单元 3024、抓取子单元 3025、存储子单元 3026。

生成子单元 3021，用于生成面向所述天气信息网站的 API 接口的种子 URL 以及后续的 URL。

种子 URL 是网络爬虫进行一切工作的基础和前提。种子 URL 可以是一个也可以是多个。可以对天气信息网站的 URL 的结构特点进行分析，根据 URL 的结构特点得到后续的 URL。

请求子单元 3022，用于向所述天气信息网站的 API 接口发送 HTTP 请

求，请求访问所述 API 接口。

可以以 GET 方式向所述天气信息网站的 API 接口发送 HTTP 请求。当天气信息网站同意获取其提供的天气数据时，返回 HTTP 响应，以告知可以进行获取天气数据的操作。

分析子单元 3023，用于对所述天气信息网站提供的的数据内容进行分析和识别，以查看所述数据内容。

天气信息网站提供特定格式的数据内容，需要对天气信息网站提供的特定格式的数据内容进行分析和识别，来查看所述数据内容。例如，所述天气信息网站的 API 接口提供的的数据格式为 JSON 格式。JSON 是一种数据交换格式，使用了类似于 C 语言的语法习惯。对该 JSON 格式的数据内容进行分析和识别，来查看所述数据内容。

判断子单元 3024，用于判断所述数据内容是否为预定信息内容。

为了得到特定的天气数据，需要判断所述数据内容是否为预定信息内容。若所述数据内容是否不是预定信息内容，则舍弃该数据内容，否则执行下一步骤。

抓取子单元 3025，用于若所述数据内容为预定信息内容，则抓取所述数据内容。

数据抓取的最终目的是将网络数据内容抓取到本地。对于 JSON 格式的数据内容，在抓取所述数据内容时可以采用深度优先搜索策略进行状态空间搜索。

存储子单元 3026，用于将抓取的数据内容作为所述天气数据保存到本地。

可以在计算设备上创建数据库，将所述天气数据保存到所述数据库中。

随着网页数量的剧增，传统的网络爬虫会下载大量的无关网页。第二获取单元 302 利用天气信息网站开放的 API 接口，通过网络爬虫抓取所述天气数据，可以避免下载无关网页，高效地获取天气数据，从而提高疾病预测的效率。

实施例五

图 5 为本申请实施例五提供的计算机装置的示意图。所述计算机装置 1 包括存储器 20、处理器 30 以及存储在所述存储器 20 中并可在所述处理器 30 上运行的计算机可读指令 40，例如疾病预测程序。所述处理器 30 执行所述计算机可读指令 40 时实现上述疾病预测方法实施例中的步骤，例如图 1 所示的步骤 101-107。或者，所述处理器 30 执行所述计算机可读指令 40 时实现上述装置实施例中各模块/单元的功能，例如图 3 中的单元 301-307。

示例性的，所述计算机可读指令 40 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 20 中，并由所述处理器 30 执行，以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令 40 在所述计算机装置 1 中的执行过程。如，所述计算机可读指令 40 可以被分割成图 3 中的单元 301、302、303、304、305、306、307，各单元功能参见实施例三。

所述计算机装置 1 可以是桌上型计算机、笔记本、掌上电脑及云端服务

器等计算设备。本领域技术人员可以理解，所述示意图 5 仅是计算机装置 1 的示例，并不构成对计算机装置 1 的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述计算机装置 1 还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 30 可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (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。通用处理器可以是微处理器或者该处理器 30 也可以是任何常规的处理器等，所述处理器 30 是所述计算机装置 1 的控制中心，利用各种接口和线路连接整个计算机装置 1 的各个部分。

所述存储器 20 可用于存储所述计算机可读指令 40 和/或模块/单元，所述处理器 30 通过运行或执行存储在所述存储器 20 内的计算机可读指令和/或模块/单元，以及调用存储在存储器 20 内的数据，实现所述计算机装置 1 的各种功能。所述存储器 20 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据计算机装置 1 的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 20 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如硬盘、内存、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）、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

所述计算机装置 1 集成的模块/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，可以存储在一个非易失性可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性可读取存储介质中，该计算机可读指令在被处理器执行时，可实现上述各个方法实施例的步骤。所述非易失性可读取存储介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（ROM, Read-Only Memory）、随机存取存储器（RAM, Random Access Memory）、电载波信号、电信信号以及软件分发介质等。

此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。计算机装置权利要求中陈述的多个单元或计算机装置也可以由同一个单元或计算机装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。

权 利 要 求 书

1. 一种疾病预测方法，其特征在于，所述方法包括：
获取疾病监测数据，所述疾病监测数据是时间序列数据；
获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；
获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据；
对所述疾病监测数据、天气数据和舆情数据进行预处理；
构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；
从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；
从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。
2. 如权利要求 1 所述的方法，其特征在于，所述从网页中抓取天气数据包括：
生成面向天气信息网站的 API 接口的种子 URL 以及后续的 URL；
向所述天气信息网站的 API 接口发送 HTTP 请求，请求访问所述 API 接口；
对所述天气信息网站提供的数据进行分析和识别，以查看所述数据内容；
判断所述数据内容是否为预定信息内容；
若所述数据内容为预定信息内容，则抓取所述数据内容；
将抓取的数据内容作为所述天气数据保存到本地。
3. 如权利要求 1 所述的方法，其特征在于，所述舆情数据包括：
特定词的搜索次数；或者
特定舆情网站包含特定词的舆情信息的数量。
4. 如权利要求 1 所述的方法，其特征在于，所述对所述疾病监测数据、天气数据和舆情数据进行预处理包括：
填补所述疾病监测数据、天气数据和舆情数据中的缺失值；
修正对所述疾病监测数据、天气数据和舆情数据中的异常值；
对所述疾病监测数据、天气数据和舆情数据进行数据格式转换。
5. 如权利要求 1-4 中任一项所述的方法，其特征在于，所述天气数据包括湿度、气温、气压、降水量、水汽压、风速、风向、日照时数。
6. 如权利要求 1-4 中任一项所述的方法，其特征在于，所述多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据构造特征，得到第一隐藏层单元，第二层 LSTM 单元层用于对所述第一隐藏

层单元进行组合,得到第二隐藏层单元,所述全连接层用于根据所述第二隐藏层单元得到预测结果,每个 LSTM 单元层包括遗忘门、输入门、输出门,所述遗忘门、输入门、输出门控制所述 LSTM 单元层的记忆状态。

7. 如权利要求 1-4 中任一项所述的方法,其特征在于,所述多层 LSTM 模型训练过程中使用的损失函数为均方差,使用的算法为 RMSprop 算法。

8. 一种疾病预测装置,其特征在于,所述装置包括:

第一获取单元,用于获取疾病监测数据,所述疾病监测数据是时间序列数据;

第二获取单元,用于获取所述疾病监测数据相关的天气数据,所述天气数据是与所述疾病监测数据对应的时间序列数据;

第三获取单元,用于获取所述疾病监测数据相关的舆情数据,所述舆情数据是与所述疾病监测数据对应的时间序列数据;

预处理单元,用于对所述疾病监测数据、天气数据和舆情数据进行预处理;

构建单元,用于构建多层长短时记忆递归神经网络模型,即多层 LSTM 模型;

优化单元,用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据,利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证,得到优化后的多层 LSTM 模型;

预测单元,用于从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据,将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型,得到所述预测时间点的疾病预测结果。

9. 一种计算机装置,其特征在于,所述计算机装置包括存储器及处理器,所述存储器用于存储至少一个计算机可读指令,所述处理器用于执行所述至少一个计算机可读指令以实现以下步骤:

获取疾病监测数据,所述疾病监测数据是时间序列数据;

获取所述疾病监测数据相关的天气数据,所述天气数据是与所述疾病监测数据对应的时间序列数据;

获取所述疾病监测数据相关的舆情数据,所述舆情数据是与所述疾病监测数据对应的时间序列数据;

对所述疾病监测数据、天气数据和舆情数据进行预处理;

构建多层长短时记忆递归神经网络模型,即多层 LSTM 模型;

从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据,利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证,得到优化后的多层 LSTM 模型;

从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据,将所述预测时间点之前的疾病监测数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型,得到所述预测时间点的疾病预测结果。

10.如权利要求 9 所述的计算机装置，其特征在于，所述从网页中抓取天气数据包括：

生成面向天气信息网站的 API 接口的种子 URL 以及后续的 URL；

向所述天气信息网站的 API 接口发送 HTTP 请求，请求访问所述 API 接口；

对所述天气信息网站提供的数据内容进行分析和识别，以查看所述数据内容；

判断所述数据内容是否为预定信息内容；

若所述数据内容为预定信息内容，则抓取所述数据内容；

将抓取的数据内容作为所述天气数据保存到本地。

11. 如权利要求 9 所述的计算机装置，其特征在于，所述舆情数据包括：

特定词的搜索次数；或者

特定舆情网站包含特定词的舆情信息的数量。

12. 如权利要求 9 所述的计算机装置，其特征在于，所述对所述疾病监测数据、天气数据和舆情数据进行预处理包括：

填补所述疾病监测数据、天气数据和舆情数据中的缺失值；

修正对所述疾病监测数据、天气数据和舆情数据中的异常值；

对所述疾病监测数据、天气数据和舆情数据进行数据格式转换。

13. 如权利要求 9-12 中任一项所述的计算机装置，其特征在于，所述天气数据包括湿度、气温、气压、降水量、水汽压、风速、风向、日照时数。

14. 如权利要求 9-12 中任一项所述的计算机装置，其特征在于，所述多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据构造特征，得到第一隐藏层单元，第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合，得到第二隐藏层单元，所述全连接层用于根据所述第二隐藏层单元得到预测结果，每个 LSTM 单元层包括遗忘门、输入门、输出门，所述遗忘门、输入门、输出门控制所述 LSTM 单元层的记忆状态。

15. 一种非易失性可读存储介质，其特征在于，所述非易失性可读存储介质存储有至少一个计算机可读指令，所述至少一个计算机可读指令被处理器执行时实现以下步骤：

获取疾病监测数据，所述疾病监测数据是时间序列数据；

获取所述疾病监测数据相关的天气数据，所述天气数据是与所述疾病监测数据对应的时间序列数据；

获取所述疾病监测数据相关的舆情数据，所述舆情数据是与所述疾病监测数据对应的时间序列数据；

对所述疾病监测数据、天气数据和舆情数据进行预处理；

构建多层长短时记忆递归神经网络模型，即多层 LSTM 模型；

从预处理后的所述疾病监测数据、天气数据和舆情数据中获取训练数据和验证数据，利用所述训练数据和所述验证数据对所述多层 LSTM 模型进行训练和性能验证，得到优化后的多层 LSTM 模型；

从预处理后的所述疾病监测数据、天气数据和舆情数据中获取预测时间点之前的疾病监测数据、天气数据和舆情数据，将所述预测时间点之前的疾病监测

数据、天气数据和舆情数据输入所述优化后的多层 LSTM 模型，得到所述预测时间点的疾病预测结果。

16. 如权利要求 15 所述的存储介质，其特征在于，所述从网页中抓取天气数据包括：

生成面向天气信息网站的 API 接口的种子 URL 以及后续的 URL；

向所述天气信息网站的 API 接口发送 HTTP 请求，请求访问所述 API 接口；

对所述天气信息网站提供的数据内容进行分析和识别，以查看所述数据内容；

判断所述数据内容是否为预定信息内容；

若所述数据内容为预定信息内容，则抓取所述数据内容；

将抓取的数据内容作为所述天气数据保存到本地。

17. 如权利要求 15 所述的存储介质，其特征在于，所述舆情数据包括：

特定词的搜索次数；或者

特定舆情网站包含特定词的舆情信息的数量。

18. 如权利要求 15 所述的存储介质，其特征在于，所述对所述疾病监测数据、天气数据和舆情数据进行预处理包括：

填补所述疾病监测数据、天气数据和舆情数据中的缺失值；

修正对所述疾病监测数据、天气数据和舆情数据中的异常值；

对所述疾病监测数据、天气数据和舆情数据进行数据格式转换。

19. 如权利要求 15-18 中任一项所述的存储介质，其特征在于，所述天气数据包括湿度、气温、气压、降水量、水汽压、风速、风向、日照时数。

20. 如权利要求 15-18 中任一项所述的存储介质，其特征在于，所述多层 LSTM 模型包括两层 LSTM 单元层和一层全连接层，第一层 LSTM 单元层用于对输入数据构造特征，得到第一隐藏层单元，第二层 LSTM 单元层用于对所述第一隐藏层单元进行组合，得到第二隐藏层单元，所述全连接层用于根据所述第二隐藏层单元得到预测结果，每个 LSTM 单元层包括遗忘门、输入门、输出门，所述遗忘门、输入门、输出门控制所述 LSTM 单元层的记忆状态。

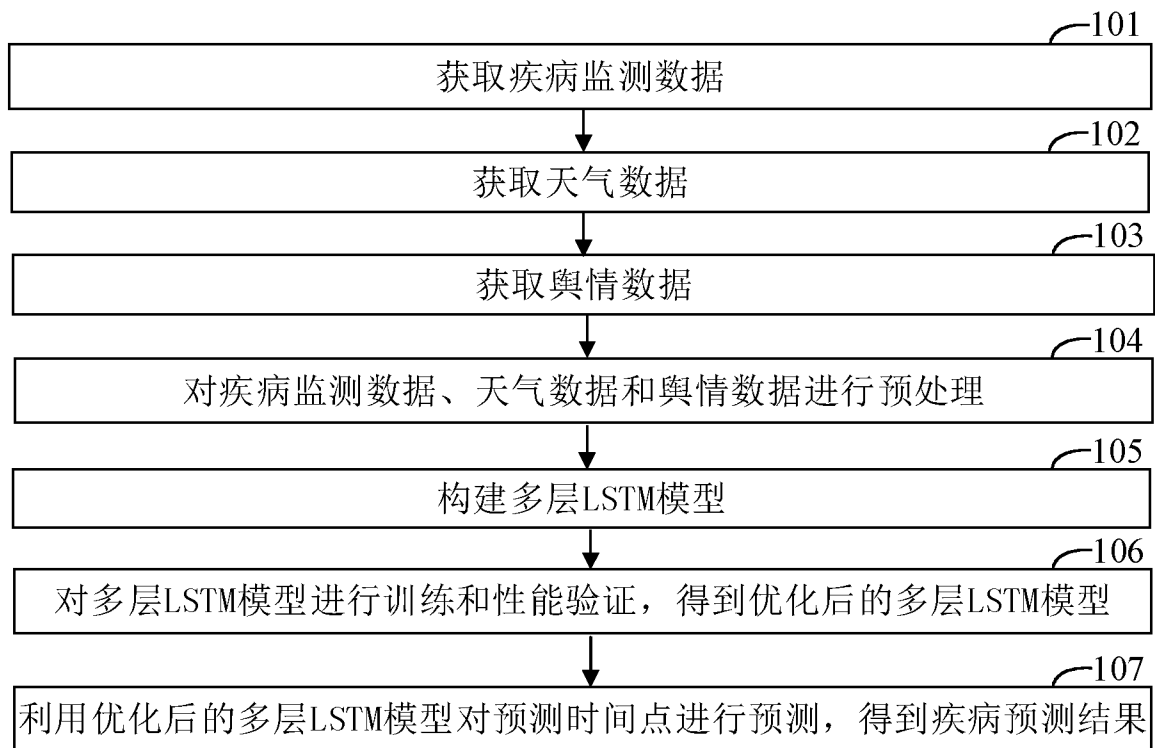


图 1

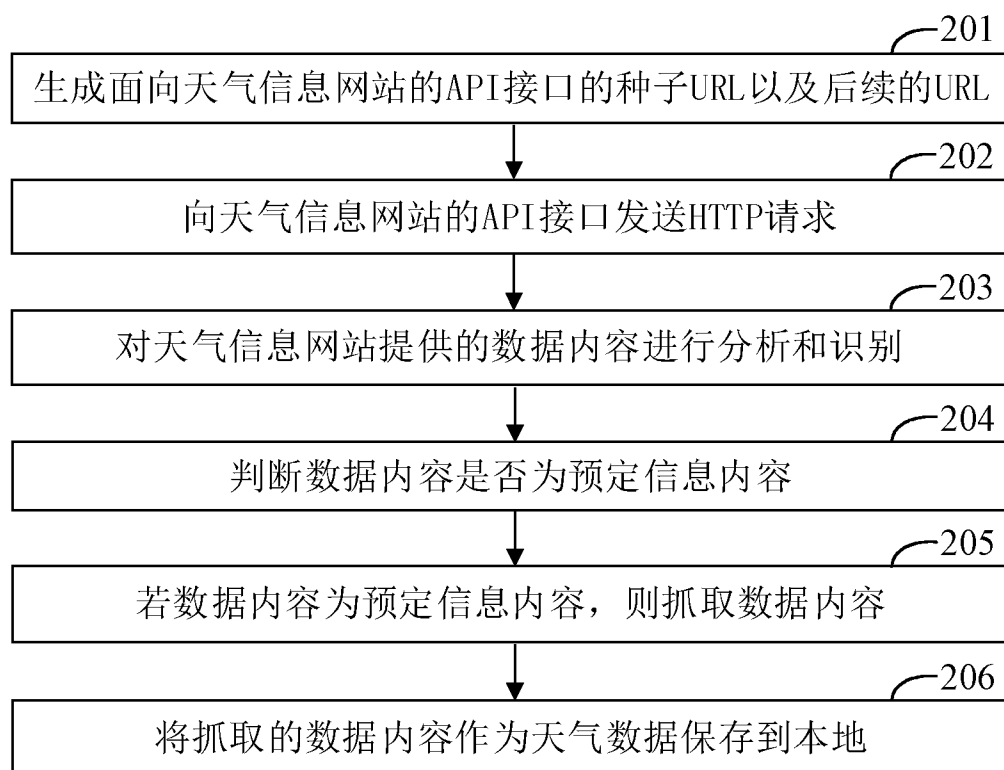


图 2

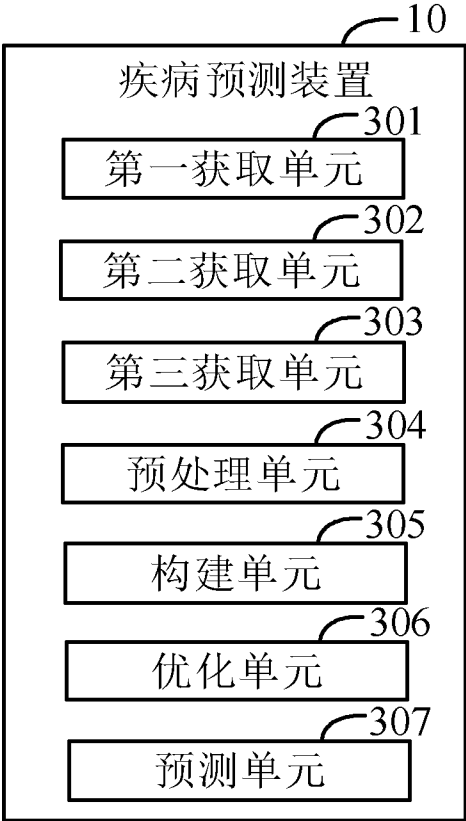


图 3

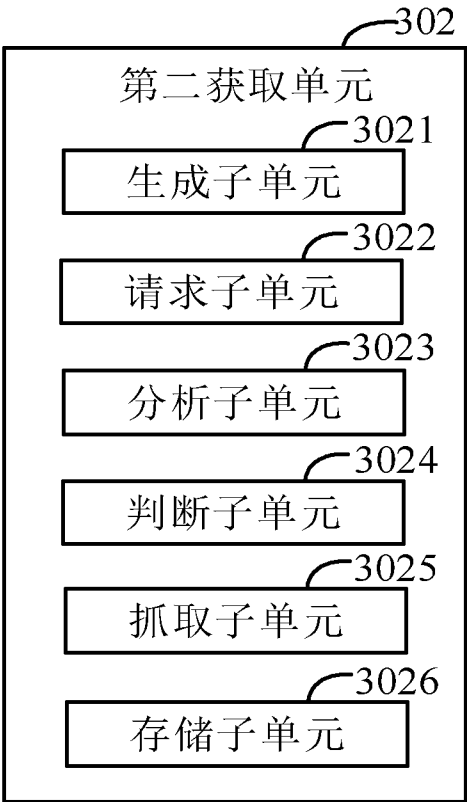


图 4

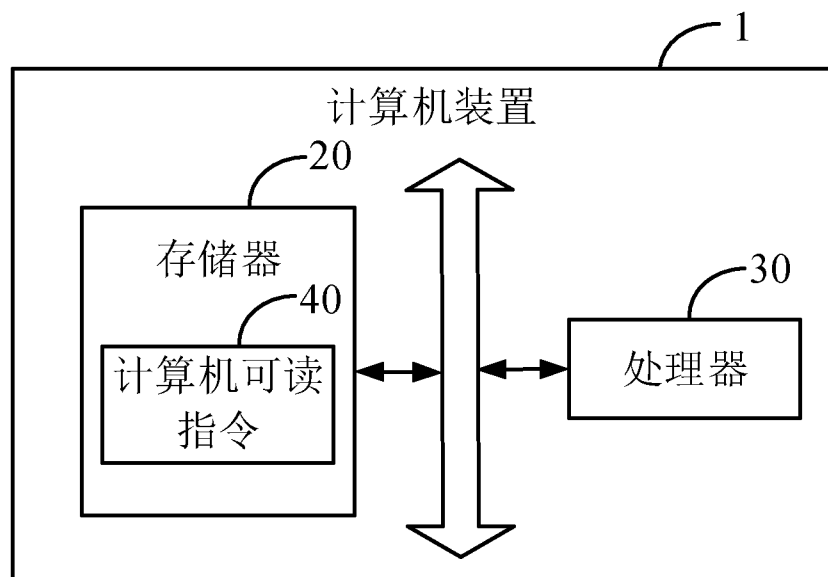


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/099847

A. CLASSIFICATION OF SUBJECT MATTER

G16H 50/00(2018.01)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/-; G06Q/-; G16H/-

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)

CNABS; SIPOABS, CNTXT, CNKI: 预测疾病, 天气, 疾病预测, 舆情, 神经网络, 监测, 数据, 模型, 训练, 优化, 多层, forecast, prognostication, disease predicts, weather, public sentiment, public option, neural networks, monitor, data, model, train, optimiz+, multilayer, layers

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 108288502 A (PING AN TECHNOLOGY SHENZHEN CO., LTD.) 17 July 2018 (2018-07-17) entire document	1-20
A	CN 107180152 A (NEC CORPORATION) 19 September 2017 (2017-09-19) entire document	1-20
A	CN 106030589 A (HRL LABORATORIES, LLC) 12 October 2016 (2016-10-12) entire document	1-20
A	CN 105812463 A (SHENZHEN QIANHAI ANYCHECK INFORMATION TECHNOLOGIES CO., LTD.) 27 July 2016 (2016-07-27) 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

09 January 2019

Date of mailing of the international search report

22 January 2019

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

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/099847

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108288502	A	17 July 2018	None			
CN	107180152	A	19 September 2017	None			
CN	106030589	A	12 October 2016	WO	2015127065	A1	27 August 2015
				EP	3108393	A1	28 December 2016
				US	2017308678	A1	26 October 2017
				EP	3108393	A4	01 November 2017
CN	105812463	A	27 July 2016	WO	2017152636	A1	14 September 2017

国际检索报告

国际申请号

PCT/CN2018/099847

A. 主题的分类

G16H 50/00(2018.01)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/-; G06Q/-; G16H/-

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))

CNABS;SIPOABS,CNXT,CNKI:预测疾病, 天气, 疾病预测, 舆情, 神经网络, 监测, 数据, 模型, 训练, 优化, 多层, forecast, prognostication, disease predicts, weather, public sentiment, public option, neural networks, monitor, data, model, train, optimiz+, multilayer, layers

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 108288502 A (平安科技深圳有限公司) 2018年 7月 17日 (2018 - 07 - 17) 全文	1-20
A	CN 107180152 A (日本电气株式会社) 2017年 9月 19日 (2017 - 09 - 19) 全文	1-20
A	CN 106030589 A (赫尔实验室有限公司) 2016年 10月 12日 (2016 - 10 - 12) 全文	1-20
A	CN 105812463 A (深圳市前海安测信息技术有限公司) 2016年 7月 27日 (2016 - 07 - 27) 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

2019年 1月 9日

国际检索报告邮寄日期

2019年 1月 22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

石志昕

电话号码 62411826

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/099847

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	108288502	A	2018年 7月 17日	无			
CN	107180152	A	2017年 9月 19日	无			
CN	106030589	A	2016年 10月 12日	WO	2015127065	A1	2015年 8月 27日
				EP	3108393	A1	2016年 12月 28日
				US	2017308678	A1	2017年 10月 26日
				EP	3108393	A4	2017年 11月 1日
CN	105812463	A	2016年 7月 27日	WO	2017152636	A1	2017年 9月 14日

表 PCT/ISA/210 (同族专利附件) (2015年1月)