(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112699277 A
(43)申请公布日 2021.04.23

(21)申请号 201911013243.8

(22)申请日 2019.10.23

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 贺国秀 康杨杨 蒋卓人 孙常龙 张琼 司罗

(74)专利代理机构 北京博浩百睿知识产权代理有限责任公司 11134
代理人 谢湘宁 张文华

(51)Int.Cl.
G06F 16/901(2019.01)
G06F 16/9035(2019.01)
G06F 16/951(2019.01)
G06Q 30/02(2012.01)

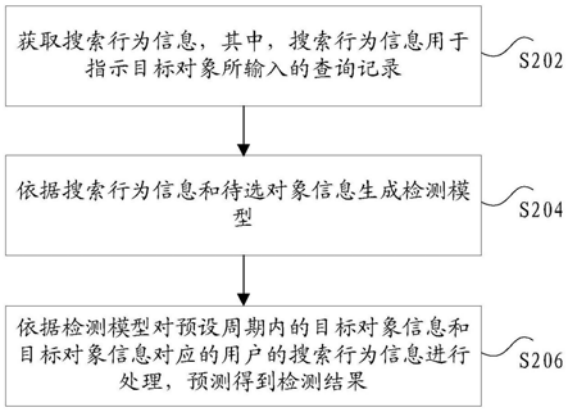
权利要求书2页 说明书18页 附图3页

(54)发明名称

数据检测的方法和装置

(57)摘要

本发明公开了一种数据检测的方法和装置。其中,该方法包括:获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。本发明解决了现有技术在对隐藏商品进行数据检测的过程中,由于技术本身的缺陷导致的检测效率低的技术问题。



CN 112699277 A

1.一种数据检测的方法,包括:

获取搜索行为信息,其中,所述搜索行为信息用于指示目标对象所输入的查询记录;

依据所述搜索行为信息和待选对象信息生成检测模型;

依据所述检测模型对预设周期内的目标对象信息和所述目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。

2.根据权利要求1所述的方法,其中,获取搜索行为信息包括:

获取每条所述用户确定所述目标对象的记录信息;

提取所述记录信息中所述用户的查询记录;

获取所述查询记录中的查询序列以及所述查询序列中的待选对象序列;

依据所述查询序列和所述待选对象序列得到所述搜索行为信息。

3.根据权利要求1或2所述的方法,其中,依据所述搜索行为信息和待选对象信息生成检测模型包括:

依据所述搜索行为信息中的查询数据进行向量计算,得到查询向量矩阵;

依据所述待选对象信息中的待选对象数据进行向量计算,得到待选对象向量矩阵;

依据所述查询向量矩阵和待选对象向量矩阵生成所述检测模型。

4.根据权利要求3所述的方法,其中,所述方法还包括:

依据所述查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息;

依据所述语义信息和所述意图信息获取向量;

依据所述语义信息和所述意图信息的向量进行拼接,得到标签。

5.根据权利要求4所述的方法,其中,所述方法还包括:

获取所述意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

6.根据权利要求4所述的方法,其中,所述方法还包括:

获取所述语义信息中的最后一个潜在语义状态;

依据所述最后一个潜在语义状态计算与剩余潜在语义状态的相似度;

依据所述相似度对所有语义状态进行池化。

7.根据权利要求3所述的方法,其中,所述依据所述检测模型对预设周期内的目标对象信息和所述目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果包括:

获取所述预设周期内的目标对象信息;

提取所述目标对象信息对应的用户的搜索行为信息;

依据所述检测模型对所述目标信息和所述搜索行为信息进行检测,获取所述目标对象信息中不满足预设检测条件的目标对象数量;

将所述目标对象数量作为所述检测结果。

8.根据权利要求1所述的方法,其中,所述方法还包括:根据特定时间段内的流行语进行商品推荐。

9.一种数据检测的方法,包括:

获取在线交易平台上的所有搜索行为信息,其中,所述搜索行为信息用于指示目标对象所输入的查询记录;

依据所述搜索行为信息和待选对象信息生成检测模型;

依据所述检测模型对目标对象信息和所述目标对象信息对应的搜索行为信息进行处

理，预测得到检测结果。

10.一种数据检测的装置，包括：

获取模块，用于获取搜索行为信息，其中，所述搜索行为信息用于指示目标对象所输入的查询记录；

模型生成模块，用于依据所述搜索行为信息和待选对象信息生成检测模型；

检测模块，用于依据所述检测模型对预设周期内的目标对象信息和所述目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果。

11.一种数据检测的装置，包括：

爬取模块，用于获取在线交易平台上的所有搜索行为信息，其中，所述搜索行为信息用于指示目标对象所输入的查询记录；

模型生成模块，用于依据所述搜索行为信息和待选对象信息生成检测模型；

检测模块，用于依据所述检测模型对目标对象信息和所述目标对象信息对应的搜索行为信息进行处理，预测得到检测结果。

12.一种存储介质，所述存储介质包括存储的程序，其中，在所述程序运行时控制所述存储介质所在设备执行权利要求1或9所述的数据检测的方法。

13.一种数据检测的装置，包括存储介质和处理器，所述处理器用于运行存储于所述存储介质中的程序，其中，所述程序运行时执行权利要求1或9所述的数据检测的方法。

数据检测的方法和装置

技术领域

[0001] 本发明涉及互联网技术领域，具体而言，涉及一种数据检测的方法和装置。

背景技术

[0002] 零售平台由于其平台属性，商家会发布一些隐性商品，这类隐性商品存在违规的特性。隐性商品的特点是常用一些同义词甚至完全和商品不相关的词来作为该商品的文本介绍内容，以此规避平台的防控机制。而这些商品可能会被一些经验丰富的买家通过自己的搜寻技术将这些商品从海量的相似商品中找出，造成了这类商品在一定程度上的传播。由于这些传播会对平台、用户的购物体验造成较大的影响，因此平台急需更加智能和高效的方法尽快发现这些隐蔽商品。

[0003] 现有的检测方法包括：基于关键词拦截结合人工判断的方法，该方法包括：由业务专家根据隐蔽商品的特性和已经确认的隐蔽商品信息总结、归纳、收集商品敏感字典。在敏感关键词拦截的基础上，进行人工校验，根据业务变化，调整增添敏感字典中的关键词。尽管这种方法可以快速拦截绝大多数隐蔽商品，但是这种方式不够灵活，相对比较滞后，很难应付卖家创造的新词或者新的描述，而且容易造成误判。

[0004] 此外现有的检测方法还包括：基于手工特征的机器学习的方法，该方法包括：利用商品的文本信息作为特征，训练机器学习分类模型以理解商品文本中蕴含的语义。然而，由于隐蔽商品的文本内容相对来说是比较隐晦或者与正常商品相似，所以已有的基于机器学习的模型很难捕捉到有效的语义。同样，由于这些模型需要大量的训练语料，而这些由商品本身的信息如商品标题或者描述的数据很难很快的捕捉到其中的变化，训练好的模型相对于最新的场景总是一个“老”模型，所以在线上的使用并不能非常好的满足业务需求。

[0005] 基于上述，现有的检测方法还包括：基于深度学习的方法，该方法还包括：与手工特征的机器学习方法的流程类似，区别在于用深度学习自动学习输入的文本特征。

[0006] 针对上述现有技术在对隐藏商品进行数据检测的过程中，由于技术本身的缺陷导致的检测效率低的问题，目前尚未提出有效的解决方案。

发明内容

[0007] 本发明实施例提供了一种数据检测的方法和装置，以至少解决现有技术在对隐藏商品进行数据检测的过程中，由于技术本身的缺陷导致的检测效率低的技术问题。

[0008] 根据本发明实施例的一个方面，提供了一种数据检测的方法，包括：获取搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录；依据搜索行为信息和待选对象信息生成检测模型；依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果。

[0009] 可选的，获取搜索行为信息包括：获取每条用户确定目标对象的记录信息；提取记录信息中用户的查询记录；获取查询记录中的查询序列以及查询序列中的待选对象序列；依据查询序列和所待选对象序列得到搜索行为信息。

[0010] 可选的，依据搜索行为信息和待选对象信息生成检测模型包括：依据搜索行为信息中的查询数据进行向量计算，得到查询向量矩阵；依据待选对象信息中的待选对象数据进行向量计算，得到待选对象向量矩阵；依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0011] 进一步地，可选的，该方法还包括：依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息；依据语义信息和意图信息获取向量；依据语义信息和意图信息的向量进行拼接，得到标签。

[0012] 可选的，该方法还包括：获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0013] 可选的，该方法还包括：获取语义信息中的最后一个潜在语义状态；依据最后一个潜在语义状态计算与剩余潜在语义状态的相似度；依据相似度对所有语义状态进行池化。

[0014] 可选的，依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果包括：获取预设周期内的目标对象信息；提取目标对象信息对应的用户的搜索行为信息；依据检测模型对目标信息和搜索行为信息进行检测，获取目标对象信息中不满足预设检测条件的目标对象数量；将目标对象数量作为检测结果。

[0015] 可选的，该方法还包括：根据特定时间段内的流行语进行商品推荐。

[0016] 根据本发明实施例的另一方面，还提供了一种数据检测的方法，包括：获取在线交易平台上的所有搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录；依据搜索行为信息和待选对象信息生成检测模型；依据检测模型对目标对象信息和目标对象信息对应的搜索行为信息进行处理，预测得到检测结果。

[0017] 根据本发明另一实施例的一方面，还提供了一种数据检测的装置，包括：获取模块，用于获取搜索行为信息，其中，搜索行为信息，用于指示目标对象所输入的查询记录；模型生成模块，用于依据搜索行为信息和待选对象信息生成检测模型；检测模块，用于依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果。

[0018] 根据本发明另一实施例的另一方面，还提供了一种数据检测的装置，包括：爬取模块，用于获取在线交易平台上的所有搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录；模型生成模块，用于依据搜索行为信息和待选对象信息生成检测模型；检测模块，用于依据检测模型对目标对象信息和目标对象信息对应的搜索行为信息进行处理，预测得到检测结果。

[0019] 根据本发明又一实施例的一方面，还提供了一种存储介质，存储介质包括存储的程序，其中，在程序运行时控制存储介质所在设备执行上述数据检测的方法。

[0020] 根据本发明又一实施例的一方面，还提供了一种数据检测的装置，包括存储介质和处理器，处理器用于运行存储于存储介质中的程序，其中，程序运行时执行上述数据检测的方法。

[0021] 在本发明实施例中，采用引入用户在搜索隐蔽商品时的搜索行为信息的方式，通过获取搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录；依据搜索行为信息和待选对象信息生成检测模型；依据检测模型对预设周期内的目标对象信息和目

标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果，达到了避免卖家刻意规避交易平台数据检测规则导致检测疏漏问题产生的目的，从而实现了提高对隐蔽商品检测的技术效果，进而解决了现有技术在对隐藏商品进行数据检测的过程中，由于技术本身的缺陷导致的检测效率低的技术问题。

附图说明

[0022] 此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

[0023] 图1是根据本发明实施例1的一种数据检测的方法的计算机终端(或移动设备)的硬件结构框图；

[0024] 图2是根据本发明实施例1的一种数据检测的方法的流程图；

[0025] 图3是根据本发明实施例1的一种基于原创递归神经网络和树剪枝机制检测隐蔽商品的原理图；

[0026] 图4是根据本发明实施例2的一种数据检测的方法的流程图；

[0027] 图5是根据本发明实施例3的一种数据检测的装置的示意图；

[0028] 图6是根据本发明实施例4的一种数据检测的装置的示意图；

[0029] 图7是根据本申请实施例5的一种计算机终端的结构框图。

具体实施方式

[0030] 为了使本技术领域的人员更好地理解本发明方案，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分的实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本发明保护的范围。

[0031] 需要说明的是，本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0032] 本申请实施例涉及的技术名词如下：

[0033] 用户搜寻行为：用户在产生购买商品的行为之前所做的努力，如提交查询、点击浏览商品等。

[0034] 采莓(Berrypicking)模型：一种信息搜寻行为模型，用来表示用户为了找到符合自己意图的结果，会根据搜索引擎返回的结果不断的迭代自己的查询。

[0035] 递归神经网络(Recursive Neural Network，简称RNN)：一种具有树状阶层结构且网络节点按其连接顺序对输入信息进行递归的人工神经网络。

[0036] 树剪枝机制：为了简化决策树模型，避免过拟合，减去决策树模型中的一些子树或

者叶结点,并将其根结点作为新的叶结点,从而实现模型的简化。

[0037] 实施例1

[0038] 根据本发明实施例,提供了一种数据检测的方法实施例,需要说明的是,在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行,并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0039] 本申请实施例一所提供的方法实施例可以在移动终端、计算机终端或者类似的运算装置中执行。以运行在计算机终端上为例,图1是本发明实施例的一种数据检测方法的计算机终端的硬件结构框图。如图1所示,计算机终端10可以包括一个或多个(图中仅示出一个)处理器102(处理器102可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器104、以及用于通信功能的传输模块106。本领域普通技术人员可以理解,图1所示的结构仅为示意,其并不对上述电子装置的结构造成限定。例如,计算机终端10还可包括比图1中所示更多或者更少的组件,或者具有与图1所示不同的配置。

[0040] 存储器104可用于存储应用软件的软件程序以及模块,如本发明实施例中的数据检测方法对应的程序指令/模块,处理器102通过运行存储在存储器104内的软件程序以及模块,从而执行各种功能应用以及数据处理,即实现上述的应用程序的数据检测方法。存储器104可包括高速随机存储器,还可包括非易失性存储器,如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中,存储器104可进一步包括相对于处理器102远程设置的存储器,这些远程存储器可以通过网络连接至计算机终端10。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0041] 传输模块106用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算机终端10的通信供应商提供的无线网络。在一个实例中,传输模块106包括一个网络适配器(Network Interface Controller,NIC),其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中,传输模块106可以为射频(Radio Frequency,RF)模块,其用于通过无线方式与互联网进行通讯。

[0042] 在上述运行环境下,本申请提供了如图2所示的数据检测的方法。图2是根据本发明实施例1的数据检测方法的流程图,如图2所示,该方法包括以下步骤:、

[0043] 步骤S202,获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录。

[0044] 本申请上述步骤S202中,搜索行为信息可以表征用户搜寻行为,即用户在产生购买目标对象的行为之前所做的努力,如提交查询、点击浏览待选对象等;目标对象可以为用户最终购买的商品。

[0045] 例如,电商平台A禁止买卖各种药品。然而电商平台A上的卖家会使用其它词汇代替药品,以规避电商平台A的数据检测规则,获取报酬。用户为了在电商平台A买到一款廉价的感冒药,如果直接搜索感冒药必然查询不到任何商品,因此,搜索行为信息可以变换为感冒、着凉、头痛等。

[0046] 容易注意到,并非每条用户搜寻行为都能购买到目标对象,所以用户搜寻行为不仅包括成功查询到目标对象并购买的查询记录,也包括没有查询到目标对象的查询记录,甚至包括只提交了查询但是并未点击任何商品的查询记录。

[0047] 步骤S204,依据搜索行为信息和待选对象信息生成检测模型。

[0048] 本申请上述步骤S204中,检测模型可以为机器学习模型,例如递归神经网络。

[0049] 递归神经网络是一种具有树状阶层结构且网络节点按其连接顺序对输入信息进行递归的人工神经网络,其输入之间可能是存在联系的,所以在多次输入x_1,x_2,x_3,…中,每次的中间信息都保存下来传给下次输入的中间信息,每次输出的计算结果不一定是目标结果,可以不使用,只有最终的输出才是需要的预测结果。

[0050] 由于用户的前次搜索行为信息和下一次的搜索行为信息往往存在关联,因此,递归神经网络尤其适用于自然语言处理。

[0051] 具体地,平台在获取到用户针对目标对象的搜索行为信息后,可以将之以Berrypicking搜寻信息树的形式进行建模。对于每条购买记录,提取其对应的用户搜寻行为,以用户为根,将用户提交的所有查询序列作为树枝,每个查询对应点击的商品序列作为对应的叶子。在构建好Berrypicking搜寻信息树后,利用递归神经网络对树进行编码。

[0052] 其中,上述方法通过挖掘每条购买记录中的Berrypicking搜寻信息树来建模用户在购买前的行为,辅助检测隐蔽商品。

[0053] 需要说明的是,搜索行为信息可以来自平台已经成功检测到的隐蔽商品的查询记录,也可以随机抽取若干正常商品的查询记录,前者作为正样本、后者作为负样本来训练模型,以使待训练的模型更加准确。

[0054] 步骤S206,依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。

[0055] 本申请上述步骤S206中,预设周期可以根据目标对象的商品属性而设置,例如更新换代时间、交易数量、危害人群等,例如一周或一个月等。

[0056] 虽然卖家可以通过改变商品的描述信息来规避平台的防控机制,但是卖家无法控制买家的搜寻行为。买家的第一个查询通常是隐蔽商品的直接描述,在搜索结果不理想的情况下,才会不断修改自己的查询以找到自己的目标商品。所以买家的查询序列和点击的商品序列中蕴含着大量的信息可供挖掘。

[0057] 具体的,结合步骤S202至步骤S206,仍以电商平台A禁止买卖各种药品为例,如前所述,如果用户直接搜索感冒药必然查询不到任何商品,因此,搜索行为信息可以变换为着凉,但是着凉对应的待选商品为汗巾、空调挡风板、夏凉被等,此时用户不作任何点击操作,继续输入下一个查询词头痛、瞌睡等,每个查询词都对应若干待选商品。最终在瞌睡对应的商品信息里找到了感冒药。后台依据这些搜索行为信息和待选对象信息生成检测模型,并对一段时间内的类似感冒药和类似感冒药对应的用户的搜素行为信息进行预测,以获得这些类似感冒药的商品的检测结果,最终由人工判断究竟其是否为感冒药。

[0058] 在本发明实施例中,采用引入用户在搜索隐蔽商品时的搜索行为信息的方式,通过获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果,达到了避免卖家刻意规避交易平台数据检测规则导致检测疏漏问题产生的目的,从而实现了提高对隐蔽商品检测的技术效果,进而解决了现有技术在对隐藏商品进行数据检测的过程中,由于技术本身的缺陷导致的检测效率低的技术问题。

[0059] 可选的，步骤S202获取搜索行为信息包括：

[0060] 步骤S2021，获取每条用户确定目标对象的记录信息。

[0061] 本申请上述步骤S2021中，记录信息可以为用户购买到至少一个目标对象的信息。

[0062] 步骤S2022，提取记录信息中用户的查询记录。

[0063] 本申请上述步骤S2022中，查询记录可以为用户查询到目标对象前的一系列操作，例如输入查询词、点击查询词对应的商品中的某一个等。

[0064] 步骤S2023，获取查询记录中的查询序列以及查询序列中的待选对象序列。

[0065] 本申请上述步骤S2023中，查询序列可以为用户查询到目标对象前所有的查询词构成的序列；待选对象序列可以为每条查询词所对应的待选对象构成的序列。

[0066] 步骤S2024，依据查询序列和所待选对象序列得到搜索行为信息。

[0067] 其中，对于每条购买记录，提取其对应的用户搜寻行为，以用户为根，将用户提交的所有查询序列作为树枝，每个查询对应点击的商品序列作为对应的叶子。每条购买记录的用户搜索行为信息被组织成树的形状。

[0068] 可选的，步骤S204依据搜索行为信息和待选对象信息生成检测模型包括：

[0069] 步骤S20401，依据搜索行为信息中的查询数据进行向量计算，得到查询向量矩阵。

[0070] 步骤S20402，依据待选对象信息中的待选对象数据进行向量计算，得到待选对象向量矩阵。

[0071] 步骤S20403，依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0072] 本申请上述步骤S20401到步骤S20403中，将搜索行为信息中的查询数据和待选对象数据进行分词，统计所有出现的词做成词典，然后将每个分词映射成对应的ID，通过每个分词的ID找到对应的词嵌入向量，这样每个查询数据和待选对象数据都可以表示为词向量构成的矩阵，即查询向量矩阵和待选对象向量矩阵。

[0073] 需要说明的是，上述方法中的词嵌入向量可以通过预训练或者直接随机初始化获得。另外，除了传统词向量之外，还可以使用BERT用来增强每个字的表达。

[0074] BERT(Bidirectional Encoder Representation from Transformers)在本申请实施例中，BERT框架替换掉原来的tf.embedding_lookup(传统的词向量)；载入BERT的预训练模型，并将fine-tune参数设置为true，将整个模型重新训练。进一步地，可选的，本申请实施例提供的数据检测的方法还包括：

[0075] 步骤S20404，依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息。

[0076] 本申请上述步骤S20404中，语义信息可以为商品的隐藏语义信息，意图信息可以包括前向潜在的用户意图信息和后向潜在的用户意图信息。

[0077] 需要说明的是，隐藏语义信息和潜在的用户意图信息可以由递归神经网络进行编码和建模。

[0078] 步骤S20405，依据语义信息和意图信息获取向量。

[0079] 本申请上述步骤S20405中，获取向量的方法可以为平均池化法，也可以为基于查询信息的注意力池化法，其中，平均池化或者注意力池化(按照注意力加权平均)的根本目的在于压缩提取信息。将获得的是每一个树枝对应的向量，这些向量会组成一个矩阵。但是为了获得输出，需要对这些矩阵做进一步处理以重新压缩至一个向量。

[0080] 平均池化中所有树枝的重要性是一样的；注意力池化会先计算各个树枝的注意力

分布(即权重分布),然后再加权求平均,这里本申请实施例中可以优选注意力池化。

[0081] 其中,对查询向量矩阵和待选对象向量矩阵可以做平均池化操作,即将每个查询向量矩阵和待选对象向量矩阵在词的维度上做向量平均,得到一个新的向量,分别表示查询信息和商品信息。这样每个查询可以表示成一个向量,与之对应的被点击的商品序列可以表示成一个矩阵。对于每一个树枝,将被点击的商品序列做平均池化操作,也可以做基于查询信息的注意力池化操作。

[0082] 步骤S20406,依据语义信息和意图信息的向量进行拼接,得到标签。

[0083] 由于Berrypicking搜寻信息树的每个树枝由查询向量矩阵和待选对象向量矩阵组成。对于每个树枝,将查询向量和待选对象向量做集合,如可以直接拼接,或者拼接之后加全连接做映射,或者使用门机制来将之结合。

[0084] 其中,在本申请实施例中直接拼接是把两种特征简单的拼接在一起变成一个维度更大的序列;

[0085] 拼接后加全联接映射相当于是在直接拼接的技术上,做一次信息过滤和压缩;

[0086] 使用门机制来融合是对两个向量的每一维都采用不同的权重组合结合的融合方式,这种方式对两种信息的融合更加精细。

[0087] 这样整个树枝的序列可以表示为一个矩阵。由于传统的递归神经网络仅可以输出一个隐藏状态,即无法同时建模序列的隐藏语义信息和潜在的用户意图信息,所以本方案提出了一种原创的递归神经网络神经元,如下所示:

[0088] $$z_t = \sigma(W_z \cdot x_t + v_z \odot h_{t-1}^1 + b_z),$$

[0089] $$r_t = \sigma(W_r \cdot x_t + v_r \odot h_{t-1}^2 + b_r),$$

[0090] $$i_t = \sigma(W_i^1 \cdot h_{t-1}^1 + W_i^2 \cdot h_{t-1}^2 + b_i),$$

[0091] $$\tilde{h}_t^1 = i_t \odot h_{t-1}^2,$$

[0092] $$\tilde{h}_t^2 = i_t \odot h_{t-1}^1,$$

[0093] $$h_t^1 = tanh(z_t \odot h_{t-1}^1 + (1 - z_t) \odot (W_h^1 \cdot x_t) + \tilde{h}_t^1),$$

[0094] $$h_t^2 = tanh(r_t \odot h_{t-1}^2 + (1 - r_t) \odot (W_h^2 \cdot x_t) + \tilde{h}_t^2), \quad (1)$$

[0095] 式(1)中,W表示参数矩阵,v表示参数向量,b表示偏置向量,这三者是递归神经网络模型需要学习的值;z_t、r_t、i_t均为sigmoid的函数,代表三个门,分别表示语义融合门、意图融合门和交互门;h^1表示隐藏语义信息,h^2表示潜在的用户意图信息,x表示当前的输入,$\tilde{h}_t^1$代表准备增强隐藏语义状态的中间向量;同理$\tilde{h}_t^2$代表准备增强隐含意图的中间向量,tanh(0)函数是一种非线性激活函数,可以把输入压缩至-1至1之间的一个数。tanh(0)是非常标准的神经网络的激活函数。

[0096] 经过这一系列的操作,前一个状态的隐藏语义信息、潜在的用户意图信息与当前状态的输入得到了更好地融合。

[0097] 按照步骤S20404-步骤S20406的方法,可以将所有树枝的序列,即上述矩阵输入到

本申请的原创递归神经网络中，得到两个对应的输出矩阵。同理，使用双向的该递归神经网络，会得到四个对应的输出矩阵。

[0098] 可选的，本申请实施例提供的数据检测的方法还包括：

[0099] 步骤S20407，获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0100] 由于递归神经网络每次输出的计算结果不一定是目标结果，可以不使用，只有最终的输出结果才是需要的预测结果。因此，本方案取潜在的用户意图信息的前向和后向的最后一个时间步的拼接作为全联结的输入。

[0101] 容易注意到，上述方法既可以形式化的挖掘其行为的表面语义，还可以挖掘用户潜在的意图信息。将隐藏的语义信息和潜在的用户意图信息得到的向量进行拼接，接两个全连接神经网络得到最终的标签。整个从输入到输出是一个端到端的神经网络框架，将交叉熵作为目标函数，然后用随机梯度下降进行训练。

[0102] 可选的，本申请实施例提供的数据检测的方法还包括：

[0103] 步骤S20408，获取语义信息中的最后一个潜在语义状态。

[0104] 步骤S20409，依据最后一个潜在语义状态计算与剩余潜在语义状态的相似度。

[0105] 步骤S20410，依据相似度对所有语义状态进行池化。

[0106] 其中，将最后一个潜在语义状态作为依据，计算其与剩余潜在语义状态的相似度，然后利用逻辑函数，如sigmoid函数做激活，接着用softmax做激活，降低与最后一个潜在语义状态完全不相关的其它树枝的作用。然后利用这个相似度，对所有潜在语义状态做池化。具体地，上述树剪枝机制的公式如下所示：

[0107] $H^{11}=\mathrm{copy}(\mathrm{last}(H^{l}))$，

[0108] $pm=\mathrm{softmax}(\sigma(\mathrm{similar}(H^{11},H^{l})))$，

[0109] $h^{l*}=pm^{T}\cdot H^{l}$， (2)

[0110] 式(2)中，H^{l}为枝叶序列的表达，经过取最后一个枝叶以及复制之后，H^{11}为最后一个枝叶表达的和H1同样长度的矩阵；在第二个公式中，首先计算H^{l}和H^{11}的余弦相似度，然后接sigmoid函数激活，然后再接softmax函数激活，pm为最终得到的剪枝机制权重的分布；最后一个公式是利用剪枝机制对所有枝叶序列的表达做加权求和，得到所有枝叶的最终表达$h^{l}*$。通过上述方法，剪枝机制可以对递归神经网络的结果进行优化。

[0111] 可选的，步骤S206依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果包括：

[0112] 步骤S2061，获取预设周期内的目标对象信息。

[0113] 本申请上述步骤S2061中，预设周期也可以根据目标对象的商品属性而设置，例如更新换代时间、交易数量、危害人群等，例如一周或一个月等。例如，平台获取近一个月内所有感冒药的交易信息，

[0114] 步骤S2062，提取目标对象信息对应的用户的搜索行为信息。

[0115] 步骤S2063，依据检测模型对目标信息和搜索行为信息进行检测，获取目标对象信息中不满足预设检测条件的目标对象数量。

[0116] 本申请上述步骤S2063中，预设检测条件可以为目标对象和目标信息的语义相关度低于最小阈值。

[0117] 步骤S2064,将目标对象数量作为检测结果。

[0118] 其中,平台为了实时打击隐蔽商品的传播,首先确定当前时刻前若干天内所有存在交易的商品,然后提取与这些商品相关联的用户搜寻行为数据。由于每个商品一般会被多个用户购买,所以在预测时,统计每个商品被训练好的分类器检测到有问题用户搜索序列的数量,并将之排序提交给平台维护商。由于用户的搜寻行为数据中蕴含着大量的信息,所以相较于现有技术中仅考虑商品信息的模型,可以极大的提高检测的效率。

[0119] 需要说明的是,在预测或者商品实际上线的过程中,任何有交易记录的商品都可以作为模型的输入,用来判定是否为违规商品。至于选取的种类可以由平台根据检查需求设置。

[0120] 图3是根据本实施例的一种基于原创递归神经网络和树剪枝机制检测隐蔽商品的原理图。如图3所示,平台在获取到用户针对目标对象的搜索行为信息后,可以将之以Berrypicking搜寻信息树的形式进行建模。对于每条购买记录q,提取其对应的用户搜寻行为,以用户为根,将用户提交的所有查询作为树枝,每个查询对应点击的商品p作为对应的叶子。在树枝3中,用户仅提交了查询,但是没有点击任何商品,对于这种情况,对应的模型直接为空即可。经过树枝1至树枝4的查询,用户在q_4的查询对话库中找到了目标商品$\hat{p}$并将其购买。因此,本模型不再考虑树枝4之后的树枝。

[0121] 当每条购买记录的用户搜寻行为被组织成Berrypicking搜寻信息树后,利用双向递归神经网络bi-BPTRU对树进行编码。这样将所有树枝的序列输入到该双向递归神经网络中,得到四个对应的输出矩阵:前向隐藏语义信息、前向潜在的用户意图信息、后向隐藏语义信息和后向潜在的用户意图信息。本实施例取隐含用户意图的前向和后向的最后一个时间步的拼接作为最终的表达。同时,将最后一个潜在语义状态作为依据,计算与剩余潜在语义状态的相似度,然后利用逻辑函数做激活,接着用softmax做激活,降低与最后一个潜在语义状态完全不相关的其他树枝的作用。然后利用这个相似度,对所有隐含语义状态做池化,得到剪枝机制权重的分布pm。最后将隐藏语义信息和潜在用户意图得到的向量进行拼接,接两个全联接神经网络得到最终的标签。整个过程中,从输入到输出是一个端到端的神经网络框架,将交叉熵作为目标函数,然后用随机梯度下降进行训练。

[0122] 需要说明的是,图3中每个树枝由查询机器对应点击的商品信息构成;整个树为这些枝叶的序列,另外每个树枝中所有表达的颜色基本统一。

[0123] 在本发明实施例中,采用引入用户在搜索隐蔽商品时的搜索行为信息的方式,通过获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果,达到了避免卖家刻意规避交易平台数据检测规则导致检测疏漏问题产生的目的。上述方案将搜索行为信息建模为Berrypicking搜寻信息树,充实了商品的特征,然后用原创的递归神经网络对树进行编码,挖掘其中的隐藏语义信息和潜在的用户意图信息,并利用树剪枝机制对编码进行优化。同时,由于卖家无法直接改变买家的想法和搜索行为,所以买家行为蕴含的信息远大于商品本身的信息,从而实现了提高对隐蔽商品检测的技术效果,进而解决了现有技术在对隐藏商品进行数据检测的过程中,由于技术本身的缺陷导致的检测效率低的技术问题。

[0124] 此外,本申请实施例提供的数据检测的方法还包括:根据特定时间段内的流行语

进行商品推荐。

[0125] 具体的，根据时下流行的网络流行语，在电商交易平台中用户输入该网络流行语后(即，本申请实施例中目标对象的搜索行为信息)，通过检测模型基于该网络流行语进行商品推荐，以通过新的维度进行商品推荐，提升商品推荐效率。

[0126] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0127] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的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(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0128] 实施例2

[0129] 根据本发明实施例，还提供了一种数据检测的方法，可以应用在各线上购物平台中，图4是根据本发明实施例2的一种数据检测的方法的流程图，如图4所示，该方法可以包括如下步骤：

[0130] 步骤S402，获取在线交易平台上的所有搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录。

[0131] 本申请上述步骤S402中，搜索行为信息可以表征用户搜寻行为，即用户在产生购买目标对象的行为之前所做的努力，如提交查询、点击浏览待选对象等；目标对象可以为用户最终购买的商品。

[0132] 容易注意到，并非每条用户搜寻行为都能购买到目标对象，所以用户搜寻行为不仅包括成功查询到目标对象并购买的查询记录，也包括没有查询到目标对象的查询记录，甚至包括只提交了查询但是并未点击任何商品的查询记录。

[0133] 步骤S404，依据搜索行为信息和待选对象信息生成检测模型。

[0134] 本申请上述步骤S404中，检测模型可以为机器学习模型，例如递归神经网络。

[0135] 递归神经网络是一种具有树状阶层结构且网络节点按其连接顺序对输入信息进行递归的人工神经网络，其输入之间可能是存在联系的，所以在多次输入x_1，x_2，x_3，…中，每次的中间信息都保存下来传给下次输入的中间信息，每次输出的计算结果不一定是目标结果，可以不使用，只有最终的输出才是需要的预测结果。

[0136] 由于用户的前次搜索行为信息和下一次的搜索行为信息往往存在关联，因此，递归神经网络尤其适用于自然语言处理。

[0137] 具体地，平台在爬取到用户针对目标对象的搜索行为信息后，可以将之以Berrypicking搜寻信息树的形式进行建模。对于每条购买记录，提取其对应的用户搜寻行为，以用户为根，将用户提交的所有查询序列作为树枝，每个查询对应点击的商品序列作为对应的叶子。在构建好Berrypicking搜寻信息树后，利用递归神经网络对树进行编码。

[0138] 其中，上述方法通过挖掘每条购买记录中的Berrypicking搜寻信息树来建模用户在购买前的行为，辅助检测隐蔽商品。

[0139] 需要说明的是，搜索行为信息可以来自平台已经成功检测到的隐蔽商品的查询记录，也可以随机抽取若干正常商品的查询记录，前者作为正样本、后者作为负样本来训练模型，以使待训练的模型更加准确。

[0140] 步骤S406，依据检测模型对目标对象信息和目标对象信息对应的搜索行为信息进行处理，预测得到检测结果。

[0141] 虽然卖家可以通过改变商品的描述信息来规避平台的防控机制，但是卖家无法控制买家的搜寻行为。买家的第一个查询通常是隐蔽商品的直接描述，在搜索结果不理想的情况下，才会不断修改自己的查询以找到自己的目标商品。所以买家的查询序列和点击的商品序列中蕴含着大量的信息可供挖掘。

[0142] 可选的，步骤S402爬取在线交易平台上的所有搜索行为信息包括：

[0143] 步骤S4021，获取每条用户确定目标对象的记录信息。

[0144] 本申请上述步骤S4021中，记录信息可以为用户购买到至少一个目标对象的信息。

[0145] 步骤S4022，提取记录信息中用户的查询记录。

[0146] 本申请上述步骤S4021中，查询记录可以为用户查询到目标对象前的一系列操作，例如输入查询词、点击查询词对应的商品中的某一个等。

[0147] 步骤S4023，获取查询记录中的查询序列以及查询序列中的待选对象序列。

[0148] 本申请上述步骤S4023中，查询序列可以为用户查询到目标对象前所有的查询词构成的序列；待选对象序列可以为每条查询词所对应的待选对象构成的序列。

[0149] 步骤S4024，依据查询序列和所待选对象序列得到搜索行为信息。

[0150] 其中，对于每条购买记录，提取其对应的用户搜寻行为，以用户为根，将用户提交的所有查询序列作为树枝，每个查询对应点击的商品序列作为对应的叶子。每条购买记录的用户搜索行为信息被组织成树的形状。

[0151] 可选的，步骤S404依据搜索行为信息和待选对象信息生成检测模型包括：

[0152] 步骤S40401，依据搜索行为信息中的查询数据进行向量计算，得到查询向量矩阵。

[0153] 步骤S40402，依据待选对象信息中的待选对象数据进行向量计算，得到待选对象向量矩阵。

[0154] 步骤S40403，依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0155] 本申请上述步骤S40401到步骤S40403中，将搜索行为信息中的查询数据和待选对象数据进行分词，统计所有出现的词做成词典，然后将每个分词映射成对应的ID，通过每个分词的ID找到对应的词嵌入向量，这样每个查询数据和待选对象数据都可以表示为词向量构成的矩阵，即查询向量矩阵和待选对象向量矩阵。

[0156] 需要说明的是，上述方法中的词嵌入向量可以通过预训练或者直接随机初始化获得。另外，除了传统词向量之外，还可以使用BERT用来增强每个字的表达。

[0157] BERT(Bidirectional Encoder Representation from Transformers)在本申请实施例中，BERT框架替换掉原来的tf.embedding_lookup(传统的词向量)；载入BERT的预训练模型，并将fine-tune参数设置为true，将整个模型重新训练。

[0158] 进一步地，可选的，本申请实施例提供的数据检测的方法还包括：

[0159] 步骤S40404，依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息。

[0160] 本申请上述步骤S40404中，语义信息可以为商品的隐藏语义信息，意图信息可以包括前向潜在的用户意图信息和后向潜在的用户意图信息。

[0161] 需要说明的是，隐藏语义信息和潜在的用户意图信息可以由递归神经网络进行编码和建模。

[0162] 步骤S40405，依据语义信息和意图信息获取向量。

[0163] 本申请上述步骤S40405中，获取向量的方法可以为平均池化法，也可以为基于查询信息的注意力池化法，其中，平均池化或者注意力池化(按照注意力加权平均)的根本目的在于压缩提取信息。将获得的是每一个树枝对应的向量，这些向量会组成一个矩阵。但是为了获得输出，需要对这些矩阵做进一步处理以重新压缩至一个向量。

[0164] 平均池化中所有树枝的重要性是一样的；注意力池化会先计算各个树枝的注意力分布(即权重分布)，然后再加权求平均，这里本申请实施例中可以优选注意力池化。

[0165] 其中，对查询向量矩阵和待选对象向量矩阵可以做平均池化操作，即将每个查询向量矩阵和待选对象向量矩阵在词的维度上做向量平均，得到一个新的向量，分别表示查询信息和商品信息。这样每个查询可以表示成一个向量，与之对应的被点击的商品序列可以表示成一个矩阵。对于每一个树枝，将被点击的商品序列做平均池化操作，也可以做基于查询信息的注意力池化操作。

[0166] 步骤S40406，依据语义信息和意图信息的向量进行拼接，得到标签。

[0167] 由于Berrypicking搜寻信息树的每个树枝由查询向量矩阵和待选对象向量矩阵组成。对于每个树枝，将查询向量和待选对象向量做集合，如可以直接拼接，或者拼接之后加全连接做映射，或者使用门机制来将之结合。

[0168] 其中，在本申请实施例中直接拼接是把两种特征简单的拼接在一起变成一个维度更大的序列；

[0169] 拼接后加全联接映射相当于是在直接拼接的技术上，做一次信息过滤和压缩；

[0170] 使用门机制来融合是对两个向量的每一维都采用不同的权重组合结合的融合方式，这种方式对两种信息的融合更加精细。

[0171] 这样整个树枝的序列可以表示为一个矩阵。由于传统的递归神经网络仅可以输出一个隐藏状态，即无法同时建模序列的隐藏语义信息和潜在的用户意图信息，所以本方案提出了一种原创的递归神经网络神经元，如下所示：

[0172] $$z_t = \sigma(W_z \cdot x_t + v_z \odot h_{t-1}^1 + b_z),$$

[0173] $$r_t = \sigma(W_r \cdot x_t + v_r \odot h_{t-1}^2 + b_r),$$

[0174] $$i_t = \sigma(W_i^1 \cdot h_{t-1}^1 + W_i^2 \cdot h_{t-1}^2 + b_i),$$

[0175] $$\tilde{h}_t^1 = i_t \odot h_{t-1}^2,$$

[0176] $$\tilde{h}_t^2 = i_t \odot h_{t-1}^1,$$

[0177] $$h_t^1 = tanh(z_t \odot h_{t-1}^1 + (1 - z_t) \odot (W_h^1 \cdot x_t) + \tilde{h}_t^1),$$

[0178] $$h_t^2 = tanh(r_t \odot h_{t-1}^2 + (1 - r_t) \odot (W_h^2 \cdot x_t) + \tilde{h}_t^2), \quad (1)$$

[0179] 式(1)中,W表示参数矩阵,v表示参数向量,b表示偏置向量,这三者是递归神经网络模型需要学习的值;z_t、r_t、i_t均为sigmoid的函数,代表三个门,分别表示语义融合门、意图融合门和交互门;h^1表示隐藏语义信息,h^2表示潜在的用户意图信息,x表示当前的输入,$\tilde{h}_t^1$代表准备增强隐藏语义状态的中间向量;同理$\tilde{h}_t^2$代表准备增强隐含意图的中间向量,tanh(0)函数是一种非线性激活函数,可以把输入压缩至-1至1之间的一个数。tanh(0)是非常标准的神经网络的激活函数。经过这一系列的操作,前一个状态的隐藏语义信息、潜在的用户意图信息与当前状态的输入得到了更好地融合。

[0180] 按照步骤S40404-步骤S40406的方法,可以将所有树枝的序列,即上述矩阵输入到本申请的原创递归神经网络中,得到两个对应的输出矩阵。同理,使用双向的该递归神经网络,会得到四个对应的输出矩阵。

[0181] 可选的,本申请实施例提供的数据检测的方法还包括:

[0182] 步骤S40407,获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0183] 由于递归神经网络每次输出的计算结果不一定是目标结果,可以不使用,只有最终的输出结果才是需要的预测结果。因此,本方案取潜在的用户意图信息的前向和后向的最后一个时间步的拼接作为全联结的输入。

[0184] 容易注意到,上述方法既可以形式化的挖掘其行为的表面语义,还可以挖掘用户潜在的意图信息。将隐藏的语义信息和潜在的用户意图信息得到的向量进行拼接,接两个全连接神经网络得到最终的标签。整个从输入到输出是一个端到端的神经网络框架,将交叉熵作为目标函数,然后用随机梯度下降进行训练。

[0185] 可选的,本申请实施例提供的数据检测的方法还包括:

[0186] 步骤S40408,获取语义信息中的最后一个潜在语义状态。

[0187] 步骤S40409,依据最后一个潜在语义状态计算与剩余潜在语义状态的相似度。

[0188] 步骤S40410,依据相似度对所有语义状态进行池化。

[0189] 其中,将最后一个潜在语义状态作为依据,计算其与剩余潜在语义状态的相似度,然后利用逻辑函数,如sigmoid函数做激活,接着用softmax做激活,降低与最后一个潜在语义状态完全不相关的其它树枝的作用。然后利用这个相似度,对所有潜在语义状态做池化。具体地,上述树剪枝机制的公式如下所示:

[0190] H^{11}=copy(last(H^1)),

[0191] pm=softmax(σ(similar(H^{11},H^1))),

[0192] $h^{1*}=pm^T \cdot H^1$, (2)

[0193] 式(2)中,H^1为枝叶序列的表达,经过取最后一个枝叶以及复制之后,H^{11}为最后一个枝叶表达的和H1同样长度的矩阵;在第二个公式中,首先计算H^1和H^{11}的余弦相似度,然后接sigmoid函数激活,然后再接softmax函数激活,pm为最终得到的剪枝机制权重的分布;最后一个公式是利用剪枝机制对所有枝叶序列的表达做加权求和,得到所有枝叶的最终表达h^{1*}。通过上述方法,剪枝机制可以对递归神经网络的结果进行优化。

[0194] 可选的,步骤S406依据检测模型对预设周期内的目标对象信息和目标对象信息对

应的用户的搜索行为信息进行处理，预测得到检测结果包括：

[0195] 步骤S4061，获取预设周期内的目标对象信息。

[0196] 本申请上述步骤S4061中，预设周期也可以根据目标对象的商品属性而设置，例如更新换代时间、交易数量、危害人群等，例如一周或一个月等。例如，平台获取近一个月内所有感冒药的交易信息，

[0197] 步骤S4062，提取目标对象信息对应的用户的搜索行为信息。

[0198] 步骤S4063，依据检测模型对目标信息和搜索行为信息进行检测，获取目标对象信息中不满足预设检测条件的目标对象数量。

[0199] 本申请上述步骤S4063中，预设检测条件可以为目标对象和目标信息的语义相关度低于最小阈值。

[0200] 步骤S4064，将目标对象数量作为检测结果。

[0201] 其中，平台为了实时打击隐蔽商品的传播，首先确定当前时刻前若干天内所有存在交易的商品，然后提取与这些商品相关联的用户搜寻行为数据。由于每个商品一般会被多个用户购买，所以在预测时，统计每个商品被训练好的分类器检测到有问题用户搜索序列的数量，并将之排序提交给平台维护商。由于用户的搜寻行为数据中蕴含着大量的信息，所以相较于现有技术中仅考虑商品信息的模型，可以极大的提高检测的效率。

[0202] 需要说明的是，在预测或者商品实际上线的过程中，任何有交易记录的商品都可以作为模型的输入，用来判定是否为违规商品。至于选取的种类可以由平台根据检查需求设置。

[0203] 此外还需要说明的是，本实施例的可选或优选实施方式可以参见实施例1中的相关描述，但不仅限于实施例1所公开的内容，在此不再赘述。

[0204] 实施例3

[0205] 根据本申请实施例，还提供了一种数据检测的装置，图5是根据本发明实施例3的一种数据检测装置的示意图，如图5所示，该装置500包括：获取模块502、模型生成模块504和检测模块506。

[0206] 其中，获取模块502，用于获取搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录；模型生成模块504，用于依据搜索行为信息和待选对象信息生成检测模型；检测模块506，用于据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果。

[0207] 可选的，获取模块502包括：获取子模块，用于获取每条用户确定目标对象的记录信息；提取模块，用于提取记录信息中用户的查询记录；序列获取模块，用于获取查询记录中的查询序列以及查询序列中的待选对象序列；得到模块，用于依据查询序列和所待选对象序列得到搜索行为信息。

[0208] 可选的，模型生成模块504包括：第一计算模块，用于依据搜索行为信息中的查询数据进行向量计算，得到查询向量矩阵；第二计算模块，用于依据待选对象信息中的待选对象数据进行向量计算，得到待选对象向量矩阵；生成模块，用于依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0209] 进一步地，可选的，本申请实施例提供的数据检测的装置还包括：信息获取模块，用于依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息；向量获取模块，用

于依据语义信息和意图信息获取向量;拼接模块,用于依据语义信息和意图信息的向量进行拼接,得到标签。

[0210] 可选的,本申请实施例提供的数据检测的装置还包括:意图获取模块,用于获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0211] 可选的,本申请实施例提供的数据检测的装置还包括:状态获取模块,用于获取语义信息中的最后一个潜在语义状态;第三计算模块,用于依据最后一个潜在语义状态计算与剩余潜在语义状态的相似度;池化模块,用于依据相似度对所有语义状态进行池化。

[0212] 可选的,检测模块506包括:对象信息获取模块,用于获取预设周期内的目标对象信息;提取子模块,用于提取目标对象信息对应的用户的搜索行为信息;检测子模块,用于依据检测模型对目标信息和搜索行为信息进行检测,获取目标对象信息中不满足预设检测条件的目标对象数量;结果模块,用于将目标对象数量作为检测结果。

[0213] 此处需要说明的是,上述获取模块502、模型生成模块504和检测模块506对应于实施例1中的步骤S202至步骤S206,三个模块与对应的步骤所实现的实例和应用场景相同,但不限于上述实施例1所公开的内容。需要说明的是,上述模块作为装置的一部分可以运行在实施例1提供的计算机终端10中。

[0214] 实施例4

[0215] 根据本申请实施例,还提供了一种数据检测的装置,图6是根据本发明实施例4的一种数据检测装置的示意图,如图6所示,该装置600包括:爬取模块602、模型生成模块604和检测模块606。

[0216] 其中,爬取模块602,用于获取在线交易平台上的所有搜索行为信息,其中,所述搜索行为信息用于指示目标对象所输入的查询记录;模型生成模块604,用于依据搜索行为信息和待选对象信息生成检测模型;检测模块606,用于依据所述检测模型对目标对象信息和所述目标对象信息对应的搜索行为信息进行处理,预测得到检测结果。

[0217] 可选的,爬取模块602包括:获取子模块,用于获取每条用户确定目标对象的记录信息;提取模块,用于提取记录信息中用户的查询记录;序列获取模块,用于获取查询记录中的查询序列以及查询序列中的待选对象序列;得到模块,用于依据查询序列和所待选对象序列得到搜索行为信息。

[0218] 可选的,模型生成模块604包括:第一计算模块,用于依据搜索行为信息中的查询数据进行向量计算,得到查询向量矩阵;第二计算模块,用于依据待选对象信息中的待选对象数据进行向量计算,得到待选对象向量矩阵;生成模块,用于依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0219] 进一步地,可选的,本申请实施例提供的数据检测的装置还包括:信息获取模块,用于依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息;向量获取模块,用于依据语义信息和意图信息获取向量;拼接模块,用于依据语义信息和意图信息的向量进行拼接,得到标签。

[0220] 可选的,本申请实施例提供的数据检测的装置还包括:意图获取模块,用于获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0221] 可选的,本申请实施例提供的数据检测的装置还包括:状态获取模块,用于获取语义信息中的最后一个潜在语义状态;第三计算模块,用于依据最后一个潜在语义状态计算

与剩余潜在语义状态的相似度;池化模块,用于依据相似度对所有语义状态进行池化。

[0222] 可选的,检测模块606包括:对象信息获取模块,用于获取预设周期内的目标对象信息;提取子模块,用于提取目标对象信息对应的用户的搜索行为信息;检测子模块,用于依据检测模型对目标信息和搜索行为信息进行检测,获取目标对象信息中不满足预设检测条件的目标对象数量;结果模块,用于将目标对象数量作为检测结果。

[0223] 此处需要说明的是,上述爬取模块602、模型生成模块604和检测模块606对应于实施例2中的步骤S402至步骤S406,三个模块与对应的步骤所实现的实例和应用场景相同,但不限于上述实施例2所公开的内容。

[0224] 实施例5

[0225] 本发明的实施例可以提供一种计算机终端,该计算机终端可以是计算机终端群中的任意一个计算机终端设备。可选地,在本实施例中,上述计算机终端也可以替换为移动终端等终端设备。

[0226] 可选地,在本实施例中,上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

[0227] 在本实施例中,上述计算机终端可以执行应用程序的数据检测方法中以下步骤的程序代码:获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。

[0228] 可选的,上述计算机终端包括存储介质和处理器,处理器用于运行存储于存储介质中的程序,其中,程序运行时执行上述实施例1或2的数据检测的方法。

[0229] 可选地,图7是根据本发明实施例的一种计算机终端的结构框图。如图7所示,该计算机终端A可以包括:一个或多个(图中仅示出一个)处理器、存储器、以及传输模块。

[0230] 其中,存储器可用于存储软件程序以及模块,如本发明实施例中的数据检测方法和装置对应的程序指令/模块,处理器通过运行存储在存储器内的软件程序以及模块,从而执行各种功能应用以及数据处理,即实现上述的数据检测方法。存储器可包括高速随机存储器,还可以包括非易失性存储器,如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中,存储器可进一步包括相对于处理器远程设置的存储器,这些远程存储器可以通过网络连接至终端A。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0231] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:获取搜索行为信息,其中,搜索行为信息用于指示确定目标对象前所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。

[0232] 可选的,上述处理器还可以执行如下步骤的程序代码:获取搜索行为信息包括:获取每条用户确定目标对象的记录信息;提取记录信息中用户的查询记录;获取查询记录中的查询序列以及查询序列中的待选对象序列;依据查询序列和所待选对象序列得到搜索行为信息。

[0233] 可选的,上述处理器还可以执行如下步骤的程序代码:依据搜索行为信息和待选对象信息生成检测模型包括:依据搜索行为信息中的查询数据进行向量计算,得到查询向

量矩阵;依据待选对象信息中的待选对象数据进行向量计算,得到待选对象向量矩阵;依据查询向量矩阵和待选对象向量矩阵生成检测模型。

[0234] 可选的,上述处理器还可以执行如下步骤的程序代码:上述方法还包括:依据查询向量矩阵和待选对象向量矩阵获取语义信息和意图信息;依据语义信息和意图信息获取向量;依据语义信息和意图信息的向量进行拼接,得到标签。

[0235] 可选的,上述处理器还可以执行如下步骤的程序代码:上述方法还包括:获取意图信息的前向和后向的最后一个时间点的拼接作为潜在意图。

[0236] 可选的,上述处理器还可以执行如下步骤的程序代码:上述方法还包括:获取语义信息中的最后一个潜在语义状态;依据最后一个潜在语义状态计算与剩余潜在语义状态的相似度;依据相似度对所有语义状态进行池化。

[0237] 可选的,上述处理器还可以执行如下步骤的程序代码:上述方法还包括:依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果包括:获取预设周期内的目标对象信息;提取目标对象信息对应的用户的搜索行为信息;依据检测模型对目标信息和搜索行为信息进行检测,获取目标对象信息中不满足预设检测条件的目标对象数量;将目标对象数量作为检测结果。

[0238] 本领域普通技术人员可以理解,图7所示的结构仅为示意,计算机终端也可以是智能手机(如Android手机、iOS手机等)、平板电脑、掌声电脑以及移动互联网设备(Mobile Internet Devices,MID)、PAD等终端设备。图7其并不对上述电子装置的结构造成限定。例如,计算机终端10还可包括比图7中所示更多或者更少的组件(如网络接口、显示装置等),或者具有与图7所示不同的配置。

[0239] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:闪存盘、只读存储器(Read-Only Memory,ROM)、随机存取器(Random Access Memory,RAM)、磁盘或光盘等。

[0240] 实施例6

[0241] 本发明的实施例还提供了一种存储介质。可选地,在本实施例中,上述存储介质可以用于保存上述实施例1或2所提供的数据检测方法所执行的程序代码。

[0242] 可选地,在本实施例中,上述存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中,或者位于移动终端群中的任意一个移动终端中。

[0243] 可选地,在本实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:获取搜索行为信息,其中,搜索行为信息用于指示目标对象所输入的查询记录;依据搜索行为信息和待选对象信息生成检测模型;依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理,预测得到检测结果。

[0244] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0245] 在本发明的上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其他实施例的相关描述。

[0246] 在本申请所提供的几个实施例中,应该理解到,所揭露的技术内容,可通过其它的方式实现。其中,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。另一点,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,单元或模块的间接耦合或通信连接,可以是电性或其它的形式。

[0247] 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实施例方案的目的。

[0248] 另外,在本发明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。

[0249] 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,本发明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(可为个人计算机、服务器或者网络设备等)执行本发明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、只读存储器(ROM,Read-OnlyMemory)、随机存取存储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0250] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

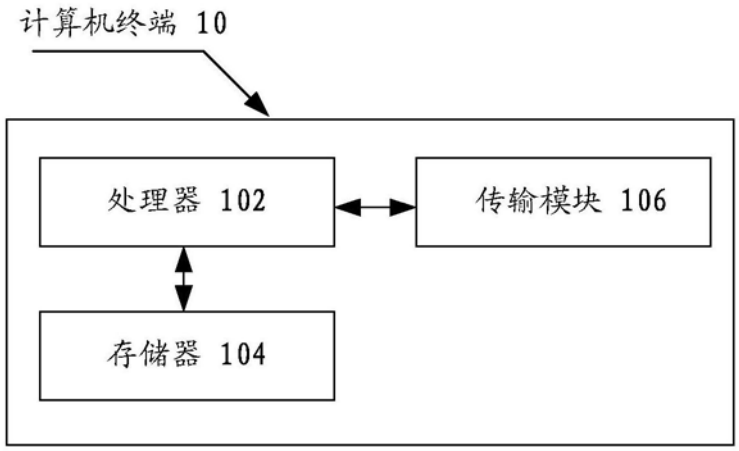

计算机终端 10
处理器 102
传输模块 106
存储器 104

图1

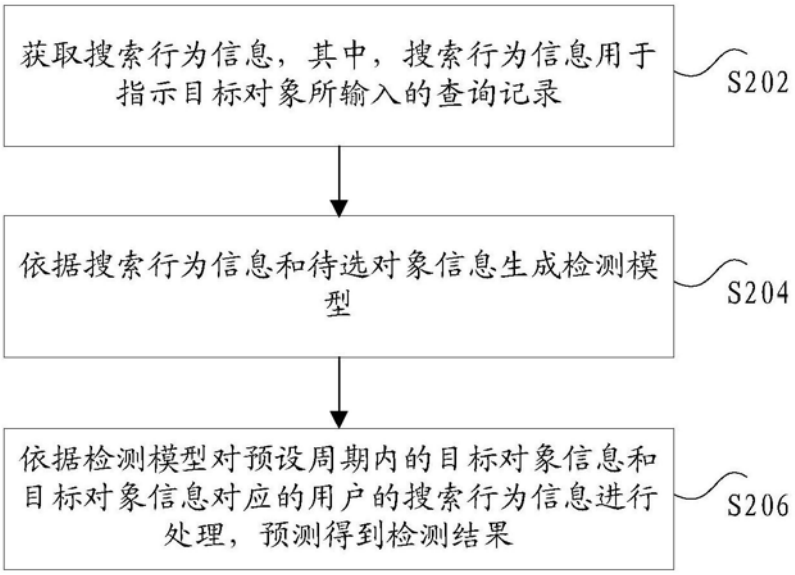

获取搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录
S202
依据搜索行为信息和待选对象信息生成检测模型
S204
依据检测模型对预设周期内的目标对象信息和目标对象信息对应的用户的搜索行为信息进行处理，预测得到检测结果
S206

图2

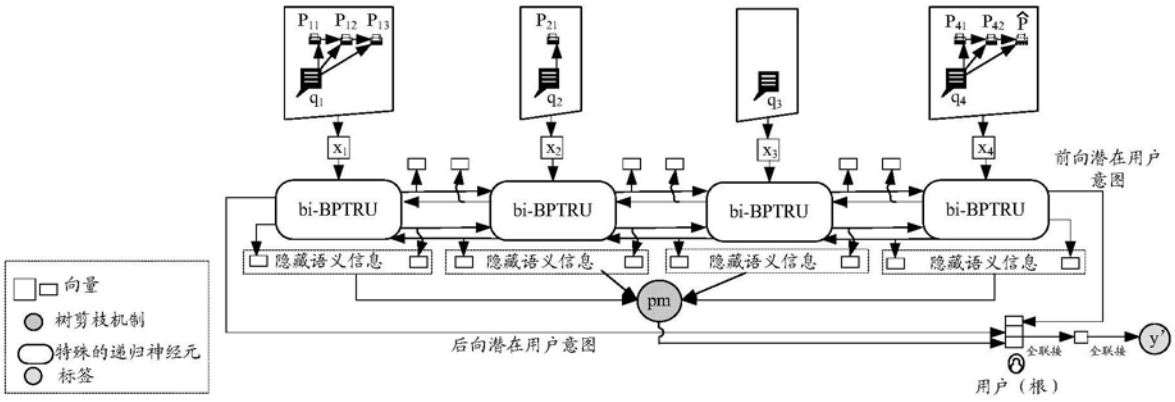

bi-BPTRU
bi-BPTRU
bi-BPTRU
bi-BPTRU
隐藏语义信息
隐藏语义信息
隐藏语义信息
隐藏语义信息
pm
前向潜在用户意图
后向潜在用户意图
用户（根）
向量
树剪枝机制
特殊的递归神经元
标签

图3

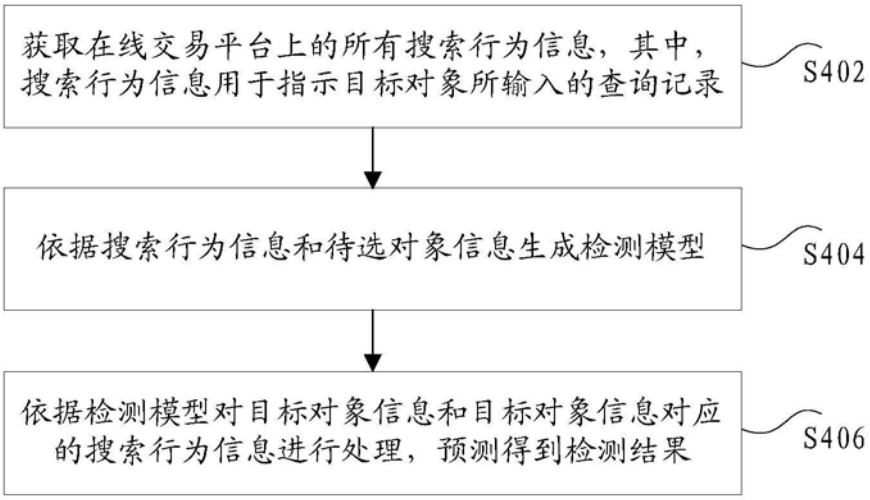

获取在线交易平台上的所有搜索行为信息，其中，搜索行为信息用于指示目标对象所输入的查询记录
S402
依据搜索行为信息和待选对象信息生成检测模型
S404
依据检测模型对目标对象信息和目标对象信息对应的搜索行为信息进行处理，预测得到检测结果
S406

图4

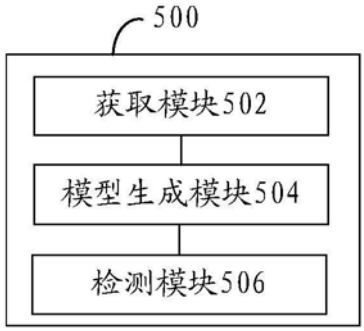

500
获取模块502
模型生成模块504
检测模块506

图5

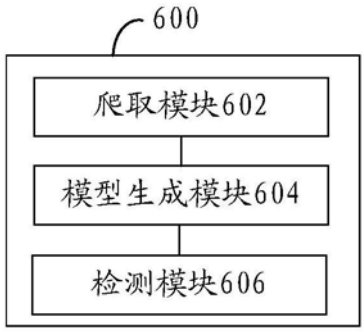

600
爬取模块602
模型生成模块604
检测模块606

图6

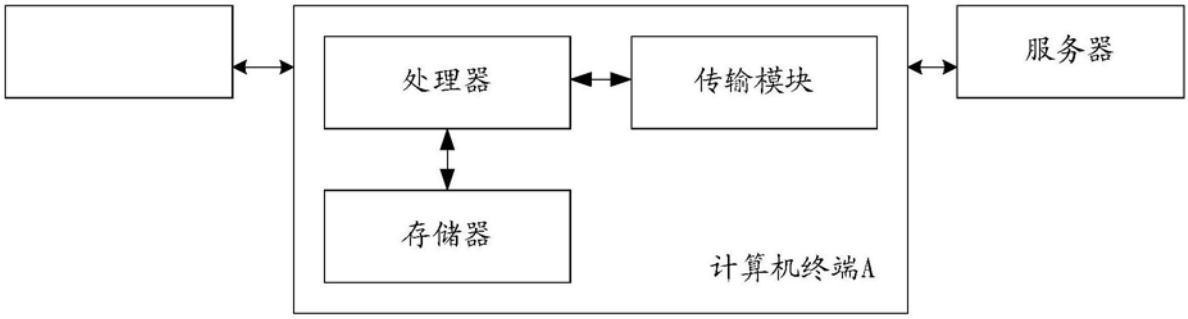


图7