(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 4 月 18 日 (18.04.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/072128 A1

(51) 国际专利分类号:
G06Q 30/02 (2012.01)

(21) 国际申请号: PCT/CN2018/109020

(22) 国际申请日: 2018 年 9 月 30 日 (30.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710933661.3 2017年10月9日 (09.10.2017) CN

(71) 申请人: 北京京东尚科信息技术有限公司 (**BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD.**) [CN/CN]; 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。
北京京东世纪贸易有限公司(**BEIJING JINGDONG CENTURY TRADING CO., LTD.**) [CN/CN]; 中国北京市经济技术开发区科创十一街18号C座2层201室, Beijing 100176 (CN)。

(72) 发明人: 王颖帅(**WANG, Yingshuai**); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。 李晓霞(**LI, Xiaoxia**); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。 苗诗雨(**MIAO, Shiyu**); 中国北京市海淀区杏石口路65号西杉创意园四区11号楼东段1-4层西段1-4层, Beijing 100195 (CN)。

(74) 代理人: 中科专利商标代理有限责任公司 (**CHINA SCIENCE PATENT & TRADEMARK AGENT LTD.**); 中国北京市海淀区西三环北路87号4-1105室, Beijing 100089 (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, 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,

(54) **Title:** OBJECT IDENTIFICATION METHOD AND SYSTEM THEREFOR

(54) 发明名称: 对象识别方法及其系统

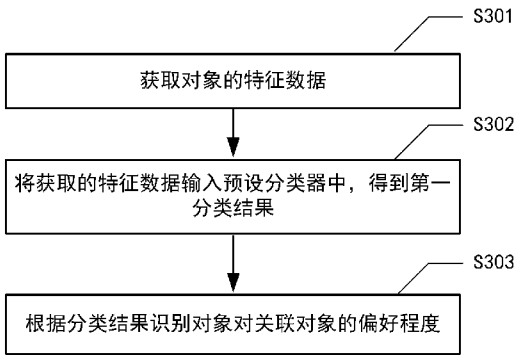

图 3

S301 Acquire feature data of an object
S302 Input the acquired feature data into a preset classifier to obtain a first classification result
S303 Identify a preference level of the object for a related object according to the first classification result

(57) **Abstract:** Provided by the present disclosure are an object identification method, the method comprising: acquiring feature data of an object, the feature data being used to reflect the preference level of the object for a related object; inputting the acquired feature data into a preset classifier to obtain a first classification result, the preset classifier being used to classify the object according to the inputted feature data; and identifying a preference level of the object for the related object according to the first classification result. Further provided by the present disclosure is an object identification system and a non-volatile storage medium.

(57) 摘要: 本公开提供了一种对象识别方法，该方法包括：获取对象的特征数据，其中，特征数据用于反映对象对其关联对象的偏好程度；将获取的特征数据输入预设分类器中，得到第一分类结果，其中，预设分类器用于依据输入的特征数据对对象进行分类；以及根据第一分类结果识别对象对关联对象的偏好程度。本公开还提供了一种对象识别系统以及一种非易失性存储介质。

WO 2019/072128 A1

[见续页]

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))。

对象识别方法及其系统

技术领域

本公开涉及数据处理领域，更具体地，涉及一种对象识别方法及其系统。

背景技术

随着网络技术的飞速发展，各行各业积累了海量的数据，如何对海量数据进行准确有效地分析显得非常重要。例如，用户对购物应用上展示的商品进行操作会产生大量的操作数据，这些数据往往会反映用户的个性化偏好，而能够根据用户对商品类型的偏好对用户进行准确识别，对个性化推荐有着非常重大的意义。

目前，相关技术中提供了多种用户识别方案。然而，在实现本公开构思的过程中，发明人发现相关技术中至少存在如下问题：相关技术中识别用户的准确度不高。

针对相关技术中的上述问题，目前还未提出有效的解决方案。

发明内容

有鉴于此，本公开提供了一种能够提高对象识别准确度的对象识别方法及其系统和系统、以及计算机可读存储介质。

本公开的一个方面提供了一种对象识别方法，包括：获取对象的特征数据，其中，上述特征数据用于反映上述对象对其关联对象的偏好程度；将获取的特征数据输入预设分类器中，得到第一分类结果，其中，上述预设分类器用于依据输入的特征数据对上述对象进行分类；以及根据上述第一分类结果识别上述对象对上述关联对象的偏好程度。

根据本公开的实施例，在将获取的特征数据输入预设分类器中，得到第一分类结果之前，上述方法还包括：获取训练数据集，其中，上述训练数据集中至少包含一类特征数据；以及对上述训练数据集中的特征数据进行训练，得到上述预设分类器。

根据本公开的实施例，获取训练数据集包括：获取样本数据集；计算包含在上述样本数据集中的各特征数据相对于上述样本数据集的

所有特征数据的条件信息熵；计算上述样本数据集中所有特征数据的信息熵；计算上述信息熵和条件信息熵的差值，得到上述各特征数据的信息增益；根据得到的信息增益的大小关系，从上述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据；以及将上述目标特征数据对应的数据集合作为上述训练数据集。

根据本公开的实施例，对上述训练数据集中的特征数据进行训练，得到上述预设分类器包括：根据上述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成上述第一决策树；以及根据上述第一决策树构建上述预设分类器。

根据本公开的实施例，在根据上述第一决策树构建上述预设分类器之后，上述方法还包括：获取包含有校验对象的特征数据的校验数据集；将上述校验对象的特征数据输入上述决策树，得到第二分类结果；判断是否能够根据上述第二分类结果识别上述校验对象对其关联对象的偏好程度；若否，则基于上述校验对象的特征数据确定第二决策树的根节点和叶子节点以生成上述第二决策树；根据上述第二决策树构建上述预设分类器；或者根据上述第一决策树和上述第二决策树构建上述预设分类器。

本公开的另一个方面提供了一种对象识别系统，包括：第一获取模块，用于获取对象的特征数据，其中，上述特征数据用于反映上述对象对其关联对象的偏好程度；第一处理模块，用于将获取的特征数据输入预设分类器中，得到第一分类结果，其中，上述预设分类器用于依据输入的特征数据对上述对象进行分类；以及识别模块，用于根据上述第一分类结果识别上述对象对上述关联对象的偏好程度。

根据本公开的实施例，上述系统还包括：第二获取模块，用于获取训练数据集，其中，上述训练数据集中至少包含一类特征数据；以及训练模块，用于对上述训练数据集中的特征数据进行训练，得到上述预设分类器。

根据本公开的实施例，上述第二获取模块包括：第一获取单元，用于获取样本数据集；第一计算单元，用于计算包含在上述样本数据集中的各特征数据相对于上述样本数据集的所有特征数据的条件信息熵；第二计算单元，用于计算上述样本数据集中所有特征数据的信息

熵；第三计算单元，用于计算上述信息熵和条件信息熵的差值，得到上述各特征数据的信息增益；选择单元，用于根据得到的信息增益的大小关系，从上述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据；以及第一确定单元，用于将上述目标特征数据对应的数据集合作为上述训练数据集。

根据本公开的实施例，上述训练模块包括：第二确定单元，用于根据上述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成上述第一决策树；以及构建单元，用于根据上述第一决策树构建上述预设分类器。

根据本公开的实施例，上述系统还包括：第三获取模块，用于获取包含有校验对象的特征数据的校验数据集；第二处理模块，用于将上述校验对象的特征数据输入上述决策树，得到第二分类结果；判断模块，用于判断是否能够根据上述第二分类结果识别上述校验对象对其关联对象的偏好程度；确定模块，用于若根据上述第二分类结果不能识别上述校验对象对其关联对象的偏好程度时，基于上述校验对象的特征数据确定第二决策树的根节点和叶子节点以生成上述第二决策树；第一构建模块，用于根据上述第二决策树构建上述预设分类器；或者第二构建模块，用于根据上述第一决策树和上述第二决策树构建上述预设分类器。

本公开的另一个方面提供了一种计算机可读存储介质，其上存储有可执行指令，上述指令被处理器执行时用于实现上述的对象识别方法。

本公开的另一个方面提供了一种对象识别系统，包括：上述计算机可读存储介质；以及上述处理器。

附图说明

为了更完整地理解本公开及其优势，现在将参考结合附图的以下描述，其中：

图1示意性示出了根据本公开的对象识别方法及其系统的示例性系统架构；

图2示意性示出了根据本公开实施例的对象识别方法及其系统的应

用场景图；

图 3 示意性示出了根据本公开实施例的对象识别方法的流程图；

图 4A 示意性示出了根据本公开另一实施例的对象识别方法的流程图；

图 4B 示意性示出了根据本公开另一实施例的对象识别方法的流程图；

图 4C 示意性示出了根据本公开另一实施例的对象识别方法的流程图；

图 4D 示意性示出了根据本公开另一实施例的对象识别方法的流程图；

图 5 示意性示出了根据本公开实施例的对象识别方法的识别结果图；

图 6 示意性示出了根据本公开实施例的对象识别系统的框图；以及

图 7 示意性示出了应用本公开实施例的对象识别方法的计算机系统的框图。

具体实施方式

以下，将参照附图来描述本公开的实施例。但是应该理解，这些描述只是示例性的，而并非要限制本公开的范围。此外，在以下说明中，省略了对公知结构和技术的描述，以避免不必要地混淆本公开的概念。

在此使用的术语仅仅是为了描述具体实施例，而并非意在限制本公开。这里使用的词语“一”、“一个（种）”和“该”等也应包括“多个”、“多种”的意思，除非上下文另外明确指出。此外，在此使用的术语“包括”、“包含”等表明了特征、步骤、操作和/或部件的存在，但是并不排除存在或添加一个或多个其他特征、步骤、操作或部件。

在此使用的所有术语（包括技术和科学术语）具有本领域技术人员通常所理解的含义，除非另外定义。应注意，这里使用的术语应解释为具有与本说明书的上下文相一致的含义，而不应以理想化或过于刻板的方式来解释。

附图中示出了一些方框图和/或流程图。应理解，方框图和/或流程

图中的一些方框或其组合可以由计算机程序指令来实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理系统的处理器，从而这些指令在由该处理器执行时可以创建用于实现这些方框图和/或流程图中所说明的功能/操作的系统。

因此，本公开的技术可以硬件和/或软件（包括固件、微代码等）的形式来实现。另外，本公开的技术可以采取存储有指令的计算机可读介质上的计算机程序产品的形式，该计算机程序产品可供指令执行系统使用或者结合指令执行系统使用。在本公开的上下文中，计算机可读介质可以是能够包含、存储、传送、传播或传输指令的任意介质。例如，计算机可读介质可以包括但不限于电、磁、光、电磁、红外或半导体系统、系统、器件或传播介质。计算机可读介质的具体示例包括：磁存储系统，如磁带或硬盘（HDD）；光存储系统，如光盘（CD-ROM）；存储器，如随机存取存储器（RAM）或闪存；和/或有线/无线通信链路。

本公开的实施例提供了一种对象识别方法及其系统。该方法包括数据获取过程和对象识别过程。在数据获取过程中，既需要获取能反映对象对其关联对象的特征数据，又需要获取用于依据输入的特征数据对对象进行分类的分类器。在完成数据获取之后，进入对象识别过程，此时可以将获取的对象的特征数据输入分类器中，得到分类结果，根据分类结果识别对象对关联对象的偏好程度，如识别一个或多个用户对商品分类的个性化偏好等。

图1示意性示出了根据本公开的对象识别方法及其系统的示例性系统架构。

如图1所示，系统架构100可以包括终端设备101、终端设备102、终端设备103，网络104和服务器105。网络104用以在终端设备101、终端设备102、终端设备103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备101、终端设备102、终端设备103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、终端设备102、终端设备103上可以安装有各种通讯客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客

户端、社交平台软件等，在此不再赘述。

终端设备 101、终端设备 102、终端设备 103 可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

服务器 105 可以是提供各种服务的服务器，例如对用户利用终端设备 101、终端设备 102、终端设备 103 所浏览的购物类网站提供支持的后台管理服务器（仅为示例）。后台管理服务器可以对接收到的产品信息查询请求等数据进行分析等处理，并将处理结果（例如目标推送信息、产品信息--仅为示例）反馈给终端设备。

需要说明的是，本公开实施例所提供的对象识别方法可以由服务器 105 执行，也可以由不同于服务器 105 的另外一个服务器或者一个服务器集群执行。相应地，用于对象识别的系统可以设置于服务器 105 中，也可以设置与服务器 105 以外的另一个服务器或者一个服务器集群中。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

目前，越来越多的用户会选择在电商平台或者其他交易平台上进行网购，系统会产生海量的交易数据，例如，在购物网站上，用户面对不同分类的商品，会对各商品分类（即品类）执行点击、浏览、关注、购买等操作，执行这些操作时，系统会产生海量的操作数据，这些数据往往会反映用户的个性化偏好，而能够根据用户对商品类型的偏好对用户进行准确识别，对个性化推荐有着非常重大的意义。

图 2 示意性示出了根据本公开实施例的对象识别方法及其系统的应用场景图。

如图 2 所示，在该应用场景中，对象可以是任意指定的用户，可以是一个也可以是多个，如可以是登录某交易平台的用户 A，也可以登录某交易平台的用户 A、用户 B、用户 C 和用户 D 等。对象的关联对象可以是显示在网页上的品类，例如，可以是服装品类，也可以是手机通讯品类等。用户会对不同品类的商品执行点击、浏览、关注、购买等操作，同时后台会产生大量的数据，这些数据用于反映用户对品类的偏好程度。电商为了给用户做个性化推荐，一般需要根据用户会对不同品类的商品

执行点击、浏览、关注、购买等操作时，从产生的海量数据中识别出用户 A、用户 B、用户 C 和用户 D 分别偏好哪个品类。

图 3 示意性示出了根据本公开实施例的对象识别方法的流程图。

如图 3 所示，该方法包括操作 S301~S303，其中：

操作 S301，获取对象的特征数据，其中，特征数据用于反映对象对其关联对象的偏好程度。

根据获取的与点击行为、浏览行为、关注行为、购物车行为相关的用户数据，得到用于反映用户对其品类的偏好程度的特征数据，如表1所示，其中，特征数据可以包括但不限于用户对品类的特征（编号为1-10的特征数据）、纯用户维度的特征（编号为11-13的特征数据）和纯品类维度的特征（编号为14-27的特征数据）。

表 1

编号	特征数据名称
1	用户三级品类浏览得分
2	用户三级品类关注得分
3	用户三级品类订单得分
4	用户三级品类分类页点击得分
5	用户三级品类下的品牌总得分
6	用户三级品类下的搜索总得分
7	用户三级品类下的购物车总得分
8	用户短期三级品类得分
9	用户偏好的三级品类价格
10	用户三级品类购买力等级
11	三级品类本身的平均价格
12	三级品类易购性因子
13	三级品类订单总数
14	用户购买商品数得分
15	用户购买品类数得分
16	用户下单频率得分
17	用户订单平均金额得分
18	用户订单最高金额得分
19	用户最近下单间隔得分
20	用户浏览页面数得分
21	用户浏览品类数得分
22	用户浏览频率得分
23	用户最近浏览间隔得分
24	用户浏览忠诚度得分
25	用户下单忠诚度得分
26	用户消费能力得分
27	用户价值总得分

需要说明的是，在获取对象的特征数据之前，根据过去用户在交易平台上对品类的点击行为、浏览行为、关注行为、购物车行为等，获取与行为相关的用户数据，如可以从浏览表中获取用户过去一个月浏览的商品；可以从关注表中获取用户过去三个月关注的商品；可以从订单表中获取用户过去一个年购买的商品；可以从点击表中获取用户过去一个月在某购物应用分类页点击的商品；可以从用户三级品类品牌得分表中获取用户三级品类品牌得分；可以从用户搜索表中获取用户三级品类搜索得分；可以从购物车表中获取用户三级品类加入购物车得分；可以从用户三级品类价格段表中获取用户偏好价格和购买力特征；可以从三级品类属性表中获取纯三级品类维度特征；还可以从用户价值表中获取纯用户维度特征，在此不做限定。

需要说明的是，在本公开实施例中，对商品的分类不作限定，它可以包括但不限于不同种类的商品，或者同一种类商品在不同维度上的分类，如三级品类。

在获取与点击行为、浏览行为、关注行为、购物车行为相关的用户数据之后，需要对数据进行清洗，如用户一天内浏览同一个三级品类超过100次，算作100次；关注数据中剔除取消关注的三级品类；一个用户同一天内购买多次三级品类，算作1次等。数据清洗的目的在于删除重复信息、纠正存在的错误，并提供数据的一致性。

在点击日志表中，用户点击过的三级品类，打标签1，用户没有点击过的三级品类打标签0，可以想象，若用户对该三级品类有偏好，则可以视为点击过，用户的标签为1，该标签可以作为识别用户偏好的重要指标。

操作 S302，将获取的特征数据输入预设分类器中，得到第一分类结果，其中，预设分类器用于依据输入的特征数据对对象进行分类。

需要说明的是，预设分类器用于依据输入的特征数据对对象进行分类。将对象的特征数据输入预设分类器中，可以得到该对象的分类结果。如将用户 A 对应的 100 个三级品类的特征数据输入到分类器后，得到用户 A 的 100 个对应的三级品类的分类结果；若将用户 A、用户 B、用户 C 和用户 D 的 100 个三级品类的特征数据输入到分类器后，

得到 4 个用户的 100 个对应的三级品类的分类结果。

需要说明的是，对象可以是一个用户也可以是多个用户，用户可以是网站上的所有用户，也可以是部分用户。特征数据可以是表 1 中的一类或多类，在此不做限定。

操作 S303，根据分类结果识别对象对关联对象的偏好程度。

分类结果可以是反映对象对其关联对象的偏好程度的预测得分，为了使各个结果达到等权重、等作用的效果，对预测分数做归一化处理，使预测分数分布在（0，1）之间。可以看出，预测得分越高，表明对象对其关联对象的偏好程度越高，在给用户推荐商品时，优先考虑用户偏好品类下的商品，这样可以更换的抓住用户的偏好心理，能够极大提升网购平台的用户体验，同时提升电商网站的点击率、商品交易总额（Gross Merchandise Volume，简称为 GMV）和订单等指标。

三级品类偏好的推荐数据格式如下图所示，其中，每一行以空格分割后的第一列是用户名，第二列首先以逗号分割，逗号分割后冒号前面的是三级品类编号，冒号后面的是用户对这个三级品类偏好的分数，并且本公开对用户偏好的三级品类分数进行了降序排列，使得分数结果一目了然，易于分析。

A 12354:0.04422928， 13783:0.04121400，12019:0.0318690
B 9724:0.04348646， 6907:0.0429534
C 655:0.47220629， 1350:0.12314458， 6907:0.120676905
D 6739：0.05305680， 9773:0.04604812， 1595:0.04556597

表 2

在表 2 中，第一行空格前面的 A 表示用户名，空格后面的先用逗号分割，12354 表示一个三级品类编号，0.04422928 表示用户 A 对 12354 这个三级品类的偏好分数，0.04121400 表示用户 A 对 13783 这个三级品类的偏好分数，可以看出，本公开推荐的数据格式对用户偏好的三级品类分数进行了降序排列，很容易可以得出结论：用户 A 对品类 12354 这个三级品类比较偏好，同理可以识别出其他用户比较偏好的三级品类。

需要说明的是，表 2 只是对用户三级品类偏好分数的推荐格式，仅为示意，并非对三级品类数量、用户数量及格式的限定。

在实现本公开构思的过程中，发明人发现，相关技术中提供了一种对象识别方法，即基于用户历史行为的统计用户偏好。这种方法参与识别的特征数据有限，维度比较单一，通常是用户的订单特征、用户的浏览特征、用户的关注特征和用户加入购物等特征，都是属于用户对品类的交互维度特征，且参与识别的特征数据是相互独立的，忽略了特征数据之间的交互关系，不能挖掘特征数据的潜在含义，同时每一种特征行为的权重依赖于分析师的经验，偏好得分不能根据线上得分及时更新，导致无法全面、有效的识别出对象对相关对象的偏好程度。

与相关技术相比，本公开实施例加入更多维度的特征数据，且考虑到特征数据之间的交互关系，克服了仅仅使用单一维度的特征数据进行识别而导致无法全面、有效的识别出对象对相关对象的偏好程度的缺陷，实现全面、有效地识别出对象对相关对象的偏好程度的技术效果。

在实现本公开构思的过程中，发明人发现，相关技术中还提供了一种对象识别方法，即基于时间权重计算的方法。这种方法一方面依赖于时间权重，而时间权重会在大促期间异常，另一方面依赖于购买周期模型，导致不足以准确识别出对象对相关对象的偏好程度。

与相关技术相比，本公开实施例将特征数据输入分类器，利用预设分类器对输入对象的特征数据进行分类，根据分类结果识别对象对关联对象的偏好程度，提高识别用户的准确度。这种机器学习的自动化模式，对推荐公司的智能化模型体系有重大意义。

本公开实施例，利用多维度的特征数据训练出具有高识别能力的的分类器，能够全面、有效、准确的识别出对象对其关联对象的偏好程度，向用户提供可以信赖的偏好商品，构造用于偏好图，对于识别用户搜索意向，提升用户体验，智能化平台服务，都具有重大意义。

图 4A 示意性示出了根据本公开另一实施例的对象识别方法的流程图。

如图 4A 所示，在将获取的特征数据输入预设分类器中，得到第一分类结果之前，该方法还包括操作 S401~S402，其中：

操作 S401，获取训练数据集，其中，训练数据集中至少包含一类特征数据。

根据本公开的实施例，在将获取的特征数据输入预设分类器中，得到第一分类结果之前，先需要获取用来生成预设分类器的训练数据集。

需要说明的是，训练数据集中的特征数据选取比较灵活，可以根据处理能力选择，也可以根据处理精度要求自主选择。训练数据集可以包含一个对象的特征数据，也可以包含多个对象的特征数据，特征数据有用户对品类的交互偏好特征、纯用户维度的特征和纯品类维度，可以实现对数据的大规模，如可以是表 2 中的一个或多个特征数据。

操作 S402，对训练数据集中的特征数据进行训练，得到预设分类器。

本公开的实施例，训练分类器可以包括多种方式/手段，在此不做限定。例如，可以使用随机森林和梯度提升树（Gradient Boosting Decision Tree，简称为GBDT）提供的机器学习算法进行训练，其中，本公开提供一种经过改进的GBDT，基本学习单元是回归决策树，对其中多个参数结合评估指标反馈，进行了最佳参数的调试。

本发明是改进的GBDT，通过调试最佳参数，模型本身就能发现特征之间的交互作用，比如分析师分析历史数据只想到了点击特征和订单特征，但是最新的数据显示把这两个特征进行某种合理的非线性组合就可以模拟出加入购物车特征，梯度提升树是可以做到挖掘更多的潜在特征，本发明用大数据技术Spark实现，每天做成订阅任务，数据更新及时，模型训练需要的时候也能随时自动化训练。

梯度提升树主要关注降低偏差，能基于泛化性能相当弱的学习器构建出很强的集成。梯度提升树是从弱学习算法出发，反复学习，得到一系列基本分类器，然后组合这些基本分类器，构成一个强分类器。提升方法实际采用加法模型与前向分布算法，以决策树为基函数的提升方法称为提升树，梯度提升算法是用最速下降法最小化损失函数，在回归预测问题上有比较好的效果。

在实现本公开构思的过程中，发明人发现，相关技术中提供的基

于历史行为的统计和基于时间权重的计算方法，依赖于分析师的业务经验，分析结果不能根据线上得分的变化而及时变化，导致识别结果不准确。

与相关技术相比，本公开实施例因为采用了预设分类器进行对象识别，预设分类器从训练数据集中训练出来，这种机器学习的自动化模式，可以对大数据进行识别，达到简化识别流程，提供识别效率和准确度的技术效果，对智能化模型体系有重大意义，同时训练过程可以实现自动化，训练结果随数据的变化而及时更新。

图 4B 示意性示出了根据本公开另一实施例的对象识别方法的流程图。

如图 4B 所示，获取训练数据集可以包括操作 S501~S506，其中：

操作 S501，获取样本数据集。

操作 S502，计算包含在样本数据集中的各特征数据相对于样本数据集的所有特征数据的条件信息熵。

操作 S503，计算样本数据集中所有特征数据的信息熵。

操作 S504，计算信息熵和条件信息熵的差值，得到各特征数据的信息增益。

操作 S505，根据得到的信息增益的大小关系，从样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据。

操作 S506，将目标特征数据对应的数据集合作为训练数据集。

需要说明的是，样本数据集可以是特征数据集的全部，也可以是特征数据集的部分。找出局部最优的特征，也就是按照这个特征进行分类后，数据集是否更加能被尽量分开。衡量局部最优特征的方法是计算样本集中所有特征数据的信息熵和样本数据集中的各特征数据相对于样本数据集的所有特征数据的条件信息熵，得出每个特征数据的信息增益值，值越大代表特征数据越重要。

例如按照如下公式计算特征数据代表的信息量。

$$I_e = -log_2 p_i$$

按照这个公式，认为小概率事件包含的信息量大，经常出现的事件

信息量小。

信息熵是信息量的期望，熵来衡量节点数据集合的有序性。

$$H(\mathrm{x}) = E(I(\mathrm{x})) = \sum_{i=1}^{n} p(x_i) I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

信息增益表示在某个特征参与下，整个数据的不确定性减少了多少，信息增益越大，特征就越重要。

$$g(\mathrm{D},\mathrm{A}) = \mathrm{H}(\mathrm{D}) - \mathrm{H}(\mathrm{D}|\mathrm{A})$$

$\mathrm{H}(\mathrm{D})$表示样本数据集 D 中所有特征数据的信息熵，$\mathrm{H}(\mathrm{D}|\mathrm{A})$在样本数据集中的特征数据 A 相对于样本数据集 D 的所有特征数据的条件信息熵，$g(\mathrm{D},\mathrm{A})$表示特征数据 A 给样本数据集 D 带来的信息增益。如样本数据集 D 中所有特征数据的信息熵是 80%，特征数据 A 相对于样本数据集 D 的所有特征数据的条件信息熵 60%，那么特征数据 A 给样本数据集 D 带来的信息增益是 20%，依此计算每个特征数据给样本数据集 D 带来的信息增益值。

在得到每个特征数据信息增益值之后，根据指定信息增益值的阈值，选择大于该阈值的特征数据对应的数据集作为训练数据集。

本公开实施例由于可以根据实际需要从特征数据集中按照每个特征数据对整个样本数据集带来的信息增益，筛选出重要性高的特征数据组成样本数据集，为后续分类器的提供了良好的数据基础。

图 4C 示意性示出了根据本公开另一实施例的对象识别方法的流程图。

如图 4C 所示，对训练数据集中的特征数据进行训练，得到预设分类器可以包括操作 S601~S602，其中：

操作 S601，根据目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成第一决策树。

操作 S602，根据第一决策树构建预设分类器。

需要说明的是，每次选择特征数据时，都挑选当前条件下最优的那个特征作为划分规则，即局部最优的特征。

图 5 给出了决策树的示意图。

如图 5 所示，在训练样本集中，当特征的信息增益大小排序为订单特征>浏览特征>关注特征时，确定出决策树叶子节点为关注特征，订单特征和浏览特征为根节点。

决策树的生成和决策树的修剪过程时，在每次选择特征数据时，也要挑选当前条件下最优的那个特征作为划分规则，即局部最优的特征，在此不再赘述。

本公开的实施例，根据信息增益的大小，每次选择特征数据时，都挑选出当前条件下最优的那个特征作为划分依据，即局部最优特征，以此构建出的决策树分类能力较强。

图 4D 示意性示出了根据本公开另一实施例的对象识别方法的流程图。

如图 4D 所示，在根据第一决策树构建预设分类器之后，该方法还可以包括操作 S701~S705，其中：

操作 S701，获取包含有校验对象的特征数据的校验数据集；

操作 S702，将校验对象的特征数据输入决策树，得到第二分类结果；

操作 S703，判断是否能够根据第二分类结果识别校验对象对其关联对象的偏好程度；

操作 S704，若否，则基于校验对象的特征数据确定第二决策树的根节点和叶子节点以生成第二决策树；

操作 S705，根据第二决策树构建预设分类器；或者

根据第一决策树和第二决策树构建预设分类器。

构建预设分类器有两种方案，方案一是根据第二决策树构建预设分类器；方案二是根据第一决策树和第二决策树构建预设分类器。图 4D 只是示意性示出方案一的流程图，方案二在此不再赘述。

需要说明的是，GBDT 里面的每一棵树看成是残差迭代树，残差最小化作为全局最优的方向，优化目标是让结果变成最好，后面每一棵回归树都在学习前面 N-1 棵树的残差，当采用平方误差损失函数时，每一棵回归树学习的是之前所有树的结论和残差，拟合得到一个当前

的残差回归树，残差的意义如公式：残差=真实值-预测值。提升树即是整个迭代过程生成的决策树的累加。

GBDT建立的梯度提升树残差用的是均方误差:

$$\mathrm{L}\left(\mathrm{Y},\mathrm{F}\left(\mathrm{X}\right)\right)=\left(\mathrm{Y}-\mathrm{F}\left(\mathrm{X}\right)\right)^{2}/2$$

GBDT的优化目标是最小化损失函数：

$$\mathrm{J}=\sum_{i}\mathrm{L}\left(Y_{i},F\left(X_{i}\right)\right)$$

求极小值的时候，把$F\left(X_{i}\right)$当做参数，对损失函数求偏导数,

$$\frac{\partial J}{\partial F\left(X_{i}\right)}=\frac{\partial\sum_{i}\mathrm{L}\left(Y_{i},F\left(X_{i}\right)\right)}{\partial F\left(X_{i}\right)}=\frac{\partial\mathrm{L}Y_{i},F\left(X_{i}\right)}{\partial F\left(X_{i}\right)}=F\left(X_{i}\right)-Y_{i}$$

所以在以均方误差作为损失函数的时候可以得出残差就是梯度的负方向：

$$Y_{i}-F\left(X_{i}\right)=-\frac{\partial J}{\partial F\left(X_{i}\right)}$$

梯度下降和预测的关系可以表示如下

$$F\left(X_{i}\right):=F\left(X_{i}\right)+h\left(X_{i}\right)$$
$$F\left(X_{i}\right):=F\left(X_{i}\right)+Y_{i}-F\left(X_{i}\right)$$
$$F\left(X_{i}\right):=F\left(X_{i}\right)+Y_{i}-1\frac{\partial J}{\partial F\left(X_{i}\right)}$$

本公开实施例中，根据第一决策树构建预设分类器之后，将校验对象的特征数据输入决策树，若能得到分类结果，则根据分类结果识别校验对象对其关联对象的偏好程度，若不能，则要学习第一决策树的残差，构建第二棵决策树，建立下一棵回归树减小残差就等价于建立下一棵决策树最小化梯度，更新到整个森林也是一样的，直到达到识别要求为止。

需要说明的是，分类器有多个决策树构成时，梯度提升树算法会根据每棵树的输出结果乘一个因子（$0<v<1$），其中$\sum_{j=1}^{J_{m}}\gamma_{jm}I\left(x\in R_{jm}\right)$是第 m 棵树的输出，而 f(m)是前 m 棵树的预测结果，

$$f_m(x) = f_{m-1}(x) + v \cdot \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

其中因子v可以看做是每棵树的权重，它可以控制提升过程的学习速率，对于GBDT来说，v和迭代轮数m(树的个数)是一个相互平衡的过程。

GBDT预测用户三级品类偏好，评估指标使用AUC，为了提高验证集中的AUC评估指标，通过观察训练集中AUC的变化，调试GBDT的特征标签组合和GBDT固有的参数，使AUC达到0.8以上。

通常影响梯度提升树GBDT的参数可以分为三类：

1.影响提升算法运行的参数：

a. LearningRate: 学习率。GBDT算法就是通过对初始模型进行一次次的调整来实现的，学习率就是衡量每次调整幅度的一个参数。理论上，这个参数越小，迭代出的结果往往越好，但需要的计算成本越大；

b. NumIterations: 弱分类器数量。就是生成的所有的弱学习器的数目，因为GBDT是串行的，就是迭代的次数，当然不是越多越好，提升算法也会有过拟合的风险；

c. 样本子集所占比例：用来训练弱学习器的样本子集占样本总体的比重，一般都是随机抽样以降低方差，默认是选总体80%的样本来训练。

2.单独影响每个弱学习器（回归树）的参数：

a. 分支最小样本量：一个节点想要继续分支所需要的最小样本数；

b. 叶节点最小样本量：一个节点要划分为叶节点所需最小样本数；

c. 树最大深度：树的层次，树越大越有过拟合的风险；

d. 最大叶节点量：叶节点的最大数目，和树最大深度可以相互替代；

e. 树的分类数量：NumClasses；

f. MaxBins 最大特征子集量。

3.学习目标参数，主要是下面两个：

a. 损失函数；b. 随机种子。

以上对提升树的效果评估不是本发明的重点，只要能实现相同目的的方法都在本发明保护范围内。

在本公开实施例中，梯度提升树主要关注降低偏差，能基于泛化性能相当弱的学习器构建出很强的集成。梯度提升树是从弱学习算法出发，反复学习，得到一系列基本分类器，然后组合这些基本分类器，构成一个强分类器。提升方法实际采用加法模型与前向分布算法，以决策树为基函数的提升方法称为提升树，梯度提升算法是用最速下降法最小化损失函数，在回归预测问题上有比较好的效果。

图 6 示意性示出了根据本公开实施例的对象识别系统的框图。

如图 6 所示，该对象识别系统包括：第一获取模块 601、第一处理模块 602、识别模块 603。

第一获取模块 601 用于获取对象的特征数据，其中，特征数据用于反映对象对其关联对象的偏好程度。

根据获取的与点击行为、浏览行为、关注行为、购物车行为相关的用户数据，得到用于反映用户对其品类的偏好程度的特征数据，如表1所示，其中，特征数据可以包括但不限于用户对品类的特征（编号为1-10的特征数据）、纯用户维度的特征（编号为11-13的特征数据）和纯品类维度的特征（编号为14-27的特征数据）。

需要说明的是，在获取对象的特征数据之前，根据过去用户在交易平台上对品类的点击行为、浏览行为、关注行为、购物车行为等，获取与行为相关的用户数据，如可以从浏览表中获取用户过去一个月浏览的商品；可以从关注表中获取用户过去三个月关注的商品；可以从订单表中获取用户过去一个年购买的商品；可以从点击表中获取用户过去一个月在某购物应用分类页点击的商品；可以从用户三级品类品牌得分表中获取用户三级品类品牌得分；可以从用户搜索表中获取用户三级品类搜索得分；可以从购物车表中获取用户三级品类加入购物车得分；可以从用户三级品类价格段表中获取用户偏好价格和购买力特征；可以从三级品类属性表中获取纯三级品类维度特征；还可以从用户价值表中获取纯用户维度特征，在此不做限定。

需要说明的是，在本公开实施例中，对商品的分类不作限定，它

可以包括但不限于不同种类的商品，或者同一种类商品在不同维度上的分类，如三级品类。

在获取与点击行为、浏览行为、关注行为、购物车行为相关的用户数据之后，需要对数据进行清洗，如用户一天内浏览同一个三级品类超过100次，算作100次；关注数据中剔除取消关注的三级品类；一个用户同一天内购买多次三级品类，算作1次等。数据清洗的目的在于删除重复信息、纠正存在的错误，并提供数据的一致性。

在点击日志表中，用户点击过的三级品类，打标签1，用户没有点击过的三级品类打标签0，可以想象，若用户对该三级品类有偏好，则可以视为点击过，用户的标签为1，该标签可以作为识别用户偏好的重要指标。

第一处理模块 602 用于将获取的特征数据输入预设分类器中，得到第一分类结果，其中，预设分类器用于依据输入的特征数据对对象进行分类。

需要说明的是，预设分类器用于依据输入的特征数据对对象进行分类。将对象的特征数据输入预设分类器中，可以得到该对象的分类结果。如将用户 A 对应的 100 个三级品类的特征数据输入到分类器后，得到用户 A 的 100 个对应的三级品类的分类结果；若将用户 A、用户 B、用户 C 和用户 D 的 100 个三级品类的特征数据输入到分类器后，得到 4 个用户的 100 个对应的三级品类的分类结果。

需要说明的是，对象可以是一个用户也可以是多个用户，用户可以是网站上的所有用户，也可以是部分用户。特征数据可以是表 1 中的一类或多类，在此不做限定。

识别模块 603 用于根据分类结果识别对象对关联对象的偏好程度。

分类结果可以是反映对象对其关联对象的偏好程度的预测得分，为了使各个结果达到等权重、等作用的效果，对预测分数做归一化处理，使预测分数分布在（0，1）之间。可以看出，预测得分越高，表明对象对其关联对象的偏好程度越高，在给用户推荐商品时，优先考虑用户偏好品类下的商品，这样可以更换的抓住用户的偏好心理，能够极大提升网购平台的用户体验，同时提升电商网站的点击率、商品交易总

额（Gross Merchandise Volume，简称为 GMV）和订单等指标。

三级品类偏好的推荐数据格式如下图所示，其中，每一行以空格分割后的第一列是用户名，第二列首先以逗号分割，逗号分割后冒号前面的是三级品类编号，冒号后面的是用户对这个三级品类偏好的分数，并且本公开对用户偏好的三级品类分数进行了降序排列，使得分数结果一目了然，易于分析。

在实现本公开构思的过程中，发明人发现，相关技术中提供了一种对象识别方法，即基于用户历史行为的统计用户偏好。这种方法参与识别的特征数据有限，维度比较单一，通常是用户的订单特征、用户的浏览特征、用户的关注特征和用户加入购物等特征，都是属于用户对品类的交互维度特征，且参与识别的特征数据是相互独立的，忽略了特征数据之间的交互关系，不能挖掘特征数据的潜在含义，同时每一种特征行为的权重依赖于分析师的经验，偏好得分不能根据线上得分及时更新，导致无法全面、有效的识别出对象对相关对象的偏好程度。

与相关技术相比，本公开实施例加入更多维度的特征数据，且考虑到特征数据之间的交互关系，克服了仅仅使用单一维度的特征数据进行识别而导致无法全面、有效的识别出对象对相关对象的偏好程度的缺陷，实现全面、有效地识别出对象对相关对象的偏好程度的技术效果。

在实现本公开构思的过程中，发明人发现，相关技术中还提供了一种对象识别方法，即基于时间权重计算的方法。这种方法一方面依赖于时间权重，而时间权重会在大促期间异常，另一方面依赖于购买周期模型，导致不足以准确识别出对象对相关对象的偏好程度。

与相关技术相比，本公开实施例将特征数据输入分类器，利用预设分类器对输入对象的特征数据进行分类，根据分类结果识别对象对关联对象的偏好程度，提高识别用户的准确度。这种机器学习的自动化模式，对推荐公司的智能化模型体系有重大意义。

本公开实施例，利用多维度的特征数据训练出具有高识别能力的的分类器，能够全面、有效、准确的识别出对象对其关联对象的偏好程度，向用户提供可以信赖的偏好商品，构造用于偏好图，对于识别用户搜索意向，提升用户体验，智能化平台服务，都具有重大意义。

作为一种可选的实施方式，系统还包括：第二获取模块，用于获取训练数据集，其中，训练数据集中至少包含一类特征数据；以及训练模块，用于对训练数据集中的特征数据进行训练，得到预设分类器。

作为一种可选的实施方式，第二获取模块包括：第一获取单元，用于获取样本数据集；第一计算单元，用于计算包含在样本数据集中的各特征数据相对于样本数据集的所有特征数据的条件信息熵；第二计算单元，用于计算样本数据集中所有特征数据的信息熵；第三计算单元，用于计算信息熵和条件信息熵的差值，得到各特征数据的信息增益；选择单元，用于根据得到的信息增益的大小关系，从样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据；以及第一确定单元，用于将目标特征数据对应的数据集合作为训练数据集。

作为一种可选的实施方式，训练模块包括：第二确定单元，用于根据目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成第一决策树；以及构建单元，用于根据第一决策树构建预设分类器。

作为一种可选的实施方式，系统还包括：第三获取模块，用于获取包含有校验对象的特征数据的校验数据集；第二处理模块，用于将校验对象的特征数据输入决策树，得到第二分类结果；判断模块，用于判断是否能够根据第二分类结果识别校验对象对其关联对象的偏好程度；确定模块，用于若根据第二分类结果不能识别校验对象对其关联对象的偏好程度时，基于校验对象的特征数据确定第二决策树的根节点和叶子节点以生成第二决策树；第一构建模块，用于根据第二决策树构建预设分类器；或者第二构建模块，用于根据第一决策树和第二决策树构建预设分类器。

需要说明的是，装置部分实施例中各模块/单元/子单元等的实施方式、解决的技术问题、实现的功能、以及达到的技术效果分别与方法部分实施例中各对应的步骤的实施方式、解决的技术问题、实现的功能、以及达到的技术效果相同或类似，在此不再赘述。

本公开的另一个方面提供了一种计算机可读存储介质，其上存储有可执行指令，指令被处理器执行时用于实现的上述对象识别方法。

本公开的另一个方面还提供了一种对象识别系统，包括：上述计算机可读存储介质；以及处理器。

图7示意性示出了应用本公开实施例的对象识别方法的计算机系统的框图。

如图7所示，根据本公开实施例的计算机系统700包括处理器701，其可以根据存储在只读存储器（ROM）702中的程序或者从存储部分708加载到随机访问存储器（RAM）703中的程序而执行各种适当的动作和处理。处理器701例如可以包括通用微处理器（例如CPU）、指令集处理器和/或相关芯片组和/或专用微处理器（例如，专用集成电路（ASIC）），等等。处理器701还可以包括用于缓存用途的板载存储器。处理器701可以包括用于执行参考图3、图4A~图4D描述的根据本公开实施例的方法流程的不同动作的单一处理单元或者是多个处理单元。

在RAM 703中，存储有系统700操作所需的各种程序和数据。处理器701、ROM 702以及RAM 703通过总线704彼此相连。处理器701通过执行ROM 702和/或RAM 703中的程序来执行以上参考图3、图4A~图4D描述的对象识别的各种操作。需要注意，所述程序也可以存储在除ROM 702和RAM 703以外的一个或多个存储器中。处理器701也可以通过执行存储在所述一个或多个存储器中的程序来执行以上参考图3、图4A~图4D描述的对象识别的各种操作。

根据本公开的实施例，系统700还可以包括输入/输出（I/O）接口705，输入/输出（I/O）接口705也连接至总线704。系统500还可以包括连接至I/O接口705的以下部件中的一项或多项：包括键盘、鼠标等的输入部分706；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分707；包括硬盘等的存储部分708；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分709。通信部分709经由诸如因特网的网络执行通信处理。驱动器710也根据需要连接至I/O接口705。可拆卸介质711，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器710上，以便于从其上读出的计算机程序根据需要被安装入存储部分708。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以

被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 709 从网络上被下载和安装，和/或从可拆卸介质 711 被安装。在该计算机程序被处理器 701 执行时，执行本公开的系统中限定的功能。

需要说明的是，本公开所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。根据本公开的实施例，计算机可读介质可以包括上文描述的 ROM 702 和/或 RAM 703 和/或 ROM 702 和 RAM 703 以外的一个或多个存储器。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码

的一部分，模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

本领域技术人员可以理解，本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合或/或结合，即使这样的组合或结合没有明确记载于本公开中。特别地，在不脱离本公开精神和教导的情况下，本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合。所有这些组合和/或结合均落入本公开的范围。

尽管已经参照本公开的特定示例性实施例示出并描述了本公开，但是本领域技术人员应该理解，在不背离所附权利要求及其等同物限定的本公开的精神和范围的情况下，可以对本公开进行形式和细节上的多种改变。因此，本公开的范围不应该限于实施例，而是应该不仅由所附权利要求来进行确定，还由所附权利要求的等同物来进行限定。

权利要求

1、一种对象识别方法，包括：

获取对象的特征数据，其中，所述特征数据用于反映所述对象对其关联对象的偏好程度；

将获取的特征数据输入预设分类器中，得到第一分类结果，其中，所述预设分类器用于依据输入的特征数据对所述对象进行分类；以及

根据所述第一分类结果识别所述对象对所述关联对象的偏好程度。

2、根据权利要求 1 所述的方法，其中，在将获取的特征数据输入预设分类器中，得到第一分类结果之前，所述方法还包括：

获取训练数据集，其中，所述训练数据集中至少包含一类特征数据；以及

对所述训练数据集中的特征数据进行训练，得到所述预设分类器。

3、根据权利要求 2 所述的方法，其中，获取训练数据集包括：

获取样本数据集；

计算包含在所述样本数据集中的各特征数据相对于所述样本数据集的所有特征数据的条件信息熵；

计算所述样本数据集中所有特征数据的信息熵；

计算所述信息熵和条件信息熵的差值，得到所述各特征数据的信息增益；

根据得到的信息增益的大小关系，从所述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据；以及

将所述目标特征数据对应的数据集合作为所述训练数据集。

4、根据权利要求 3 所述的方法，其中，对所述训练数据集中的特征数据进行训练，得到所述预设分类器包括：

根据所述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成所述第一决策树；以及

根据所述第一决策树构建所述预设分类器。

5、根据权利要求 4 所述的方法，其中，在根据所述第一决策树构建所述预设分类器之后，所述方法还包括：

获取包含有校验对象的特征数据的校验数据集；

将所述校验对象的特征数据输入所述决策树，得到第二分类结果；

判断是否能够根据所述第二分类结果识别所述校验对象对其关联对象的偏好程度；

若否，则基于所述校验对象的特征数据确定第二决策树的根节点和叶子节点以生成所述第二决策树；

根据所述第二决策树构建所述预设分类器；或者

根据所述第一决策树和所述第二决策树构建所述预设分类器。

6、一种对象识别系统，包括：

第一获取模块，用于获取对象的特征数据，其中，所述特征数据用于反映所述对象对其关联对象的偏好程度；

第一处理模块，用于将获取的特征数据输入预设分类器中，得到第一分类结果，其中，所述预设分类器用于依据输入的特征数据对所述对象进行分类；以及

识别模块，用于根据所述第一分类结果识别所述对象对所述关联对象的偏好程度。

7、根据权利要求 6 所述的系统，其中，所述系统还包括：

第二获取模块，用于获取训练数据集，其中，所述训练数据集中至少包含一类特征数据；以及

训练模块，用于对所述训练数据集中的特征数据进行训练，得到所述预设分类器。

8、根据权利要求 7 所述的系统，其中，所述第二获取模块包括：

第一获取单元，用于获取样本数据集；

第一计算单元，用于计算包含在所述样本数据集中的各特征数据相对于所述样本数据集的所有特征数据的条件信息熵；

第二计算单元，用于计算所述样本数据集中所有特征数据的信息熵；

第三计算单元，用于计算所述信息熵和条件信息熵的差值，得到所述各特征数据的信息增益；

选择单元，用于根据得到的信息增益的大小关系，从所述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据；以及

第一确定单元，用于将所述目标特征数据对应的数据集合作为所述训练数据集。

9、根据权利要求 8 所述的系统，其中，所述训练模块包括：

第二确定单元，用于根据所述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成所述第一决策树；以及

构建单元，用于根据所述第一决策树构建所述预设分类器。

10、根据权利要求 9 所述的系统，其中，所述系统还包括：

第三获取模块，用于获取包含有校验对象的特征数据的校验数据集；

第二处理模块，用于将所述校验对象的特征数据输入所述决策树，得到第二分类结果；

判断模块，用于判断是否能够根据所述第二分类结果识别所述校验对象对其关联对象的偏好程度；

确定模块，用于若根据所述第二分类结果不能识别所述校验对象对其关联对象的偏好程度时，基于所述校验对象的特征数据确定第二决策树的根节点和叶子节点以生成所述第二决策树；

第一构建模块，用于根据所述第二决策树构建所述预设分类器；或者

第二构建模块，用于根据所述第一决策树和所述第二决策树构建所述预设分类器。

11、一种计算机可读存储介质，其上存储有可执行指令，所述指令被处理器执行时用于实现权利要求 1 至 5 中任一项所述的对象识别系统。

12、一种对象识别系统，包括：

权利要求 11 所述的计算机可读存储介质；以及

处理器。

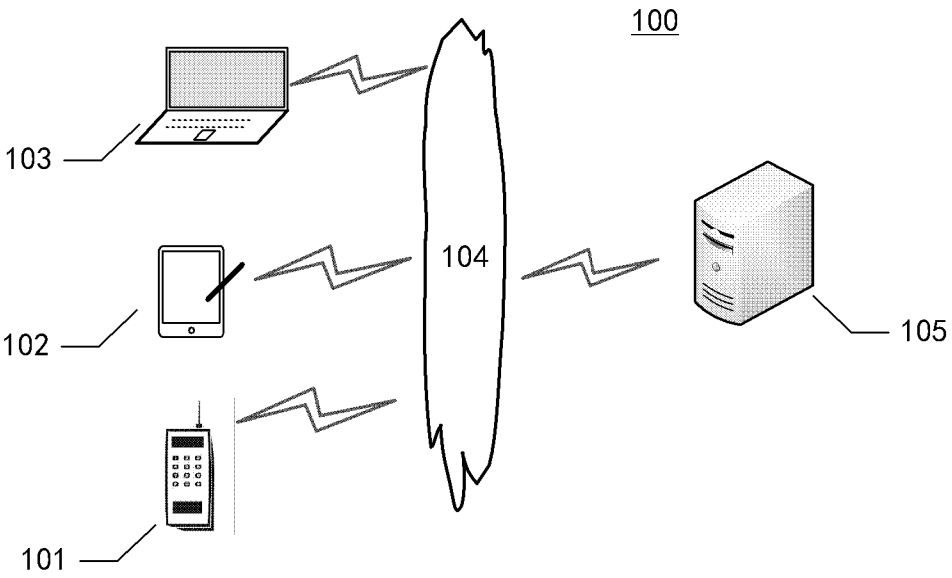
100
103
104
102
105
101

图 1

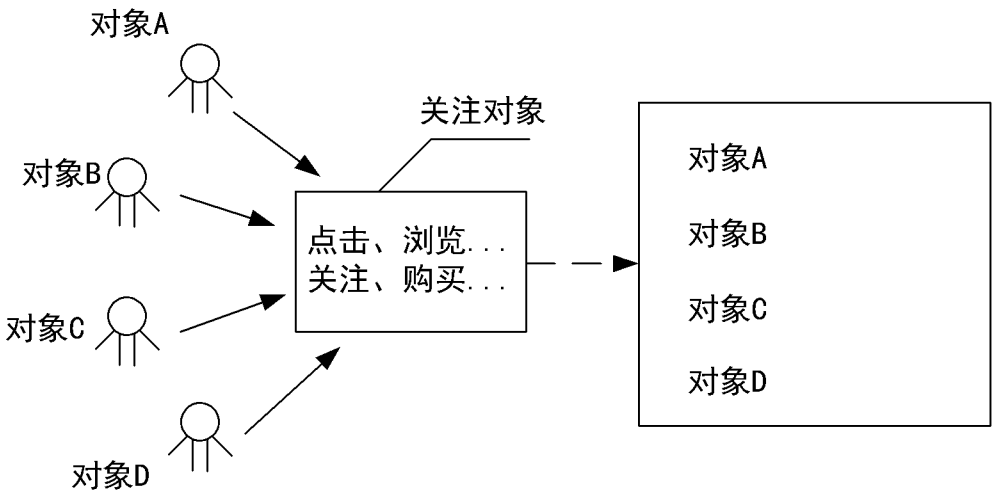
对象A
关注对象
对象B
点击、浏览...
关注、购买...
对象A
对象B
对象C
对象C
对象D
对象D

图 2

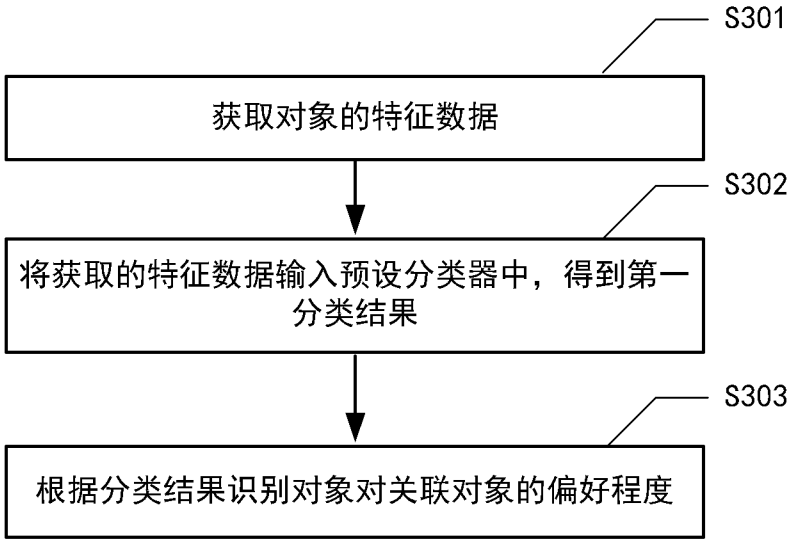
S301
获取对象的特征数据
S302
将获取的特征数据输入预设分类器中，得到第一分类结果
S303
根据分类结果识别对象对关联对象的偏好程度

图 3

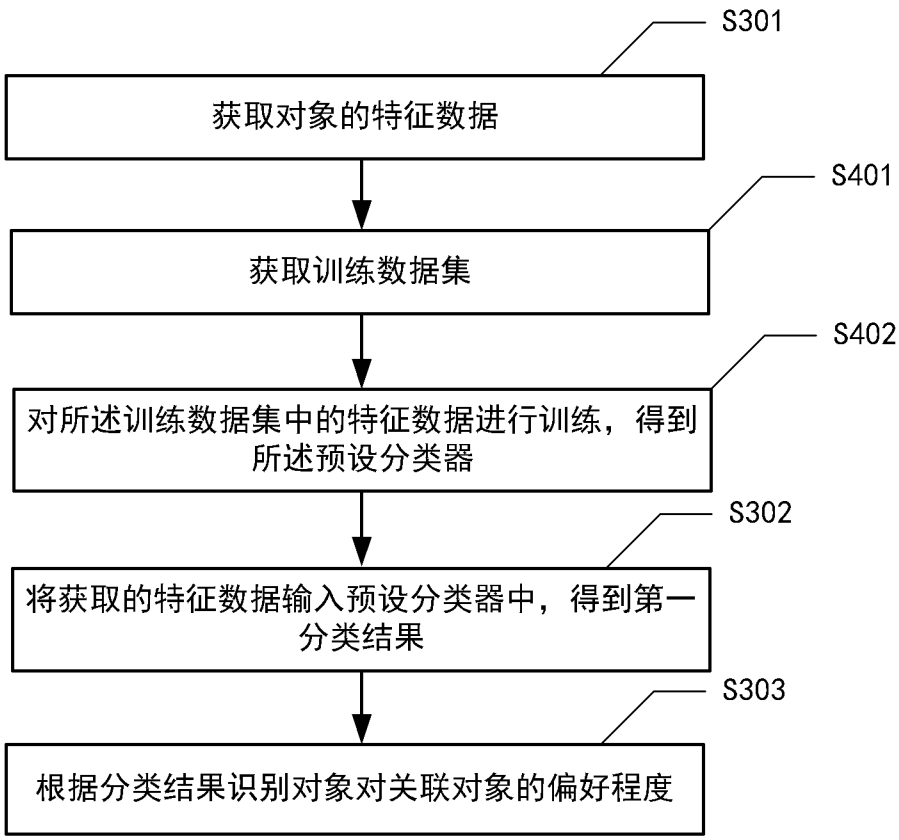
S301
获取对象的特征数据
S401
获取训练数据集
S402
对所述训练数据集中的特征数据进行训练，得到所述预设分类器
S302
将获取的特征数据输入预设分类器中，得到第一分类结果
S303
根据分类结果识别对象对关联对象的偏好程度

图 4A

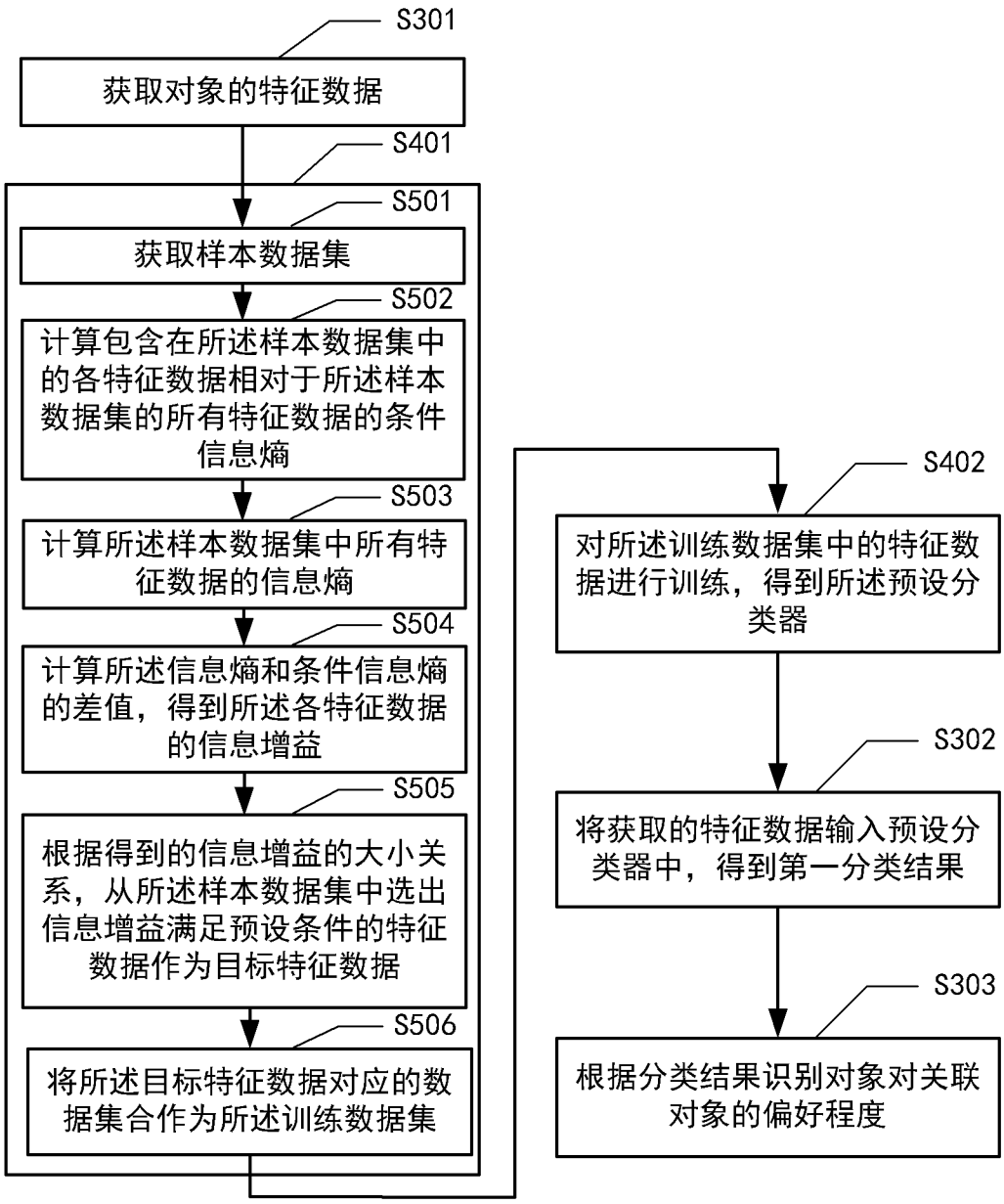
S301
获取对象的特征数据
S401
S501
获取样本数据集
S502
计算包含在所述样本数据集中的各特征数据相对于所述样本数据集的所有特征数据的条件信息熵
S503
计算所述样本数据集中所有特征数据的信息熵
S504
计算所述信息熵和条件信息熵的差值，得到所述各特征数据的信息增益
S505
根据得到的信息增益的大小关系，从所述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据
S506
将所述目标特征数据对应的数据集合作为所述训练数据集
S402
对所述训练数据集中的特征数据进行训练，得到所述预设分类器
S302
将获取的特征数据输入预设分类器中，得到第一分类结果
S303
根据分类结果识别对象对关联对象的偏好程度

图 4B

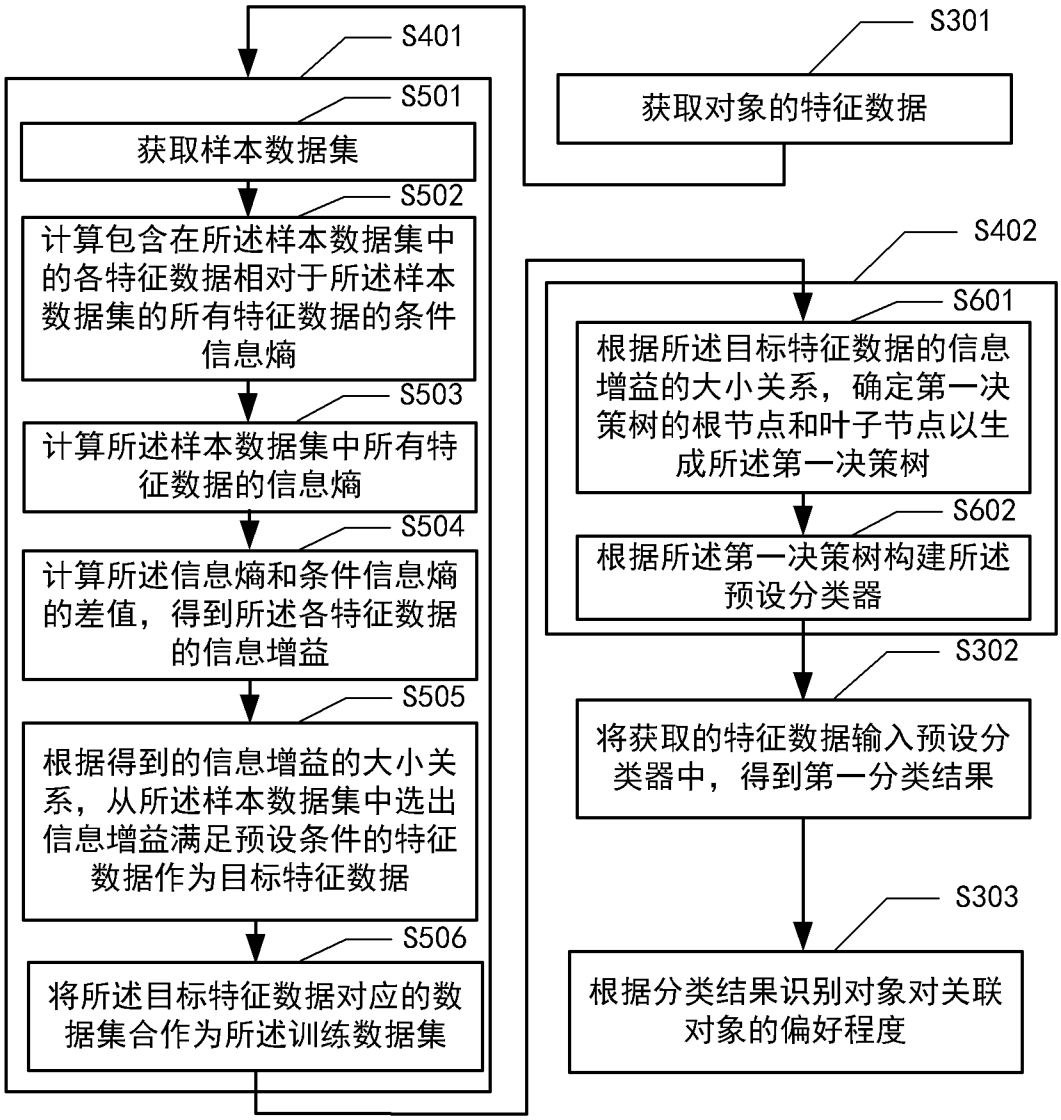
S401
S501
获取样本数据集
S502
计算包含在所述样本数据集中的各特征数据相对于所述样本数据集的所有特征数据的条件信息熵
S503
计算所述样本数据集中所有特征数据的信息熵
S504
计算所述信息熵和条件信息熵的差值，得到所述各特征数据的信息增益
S505
根据得到的信息增益的大小关系，从所述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据
S506
将所述目标特征数据对应的数据集合作为所述训练数据集
S301
获取对象的特征数据
S402
S601
根据所述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成所述第一决策树
S602
根据所述第一决策树构建所述预设分类器
S302
将获取的特征数据输入预设分类器中，得到第一分类结果
S303
根据分类结果识别对象对关联对象的偏好程度

图 4C

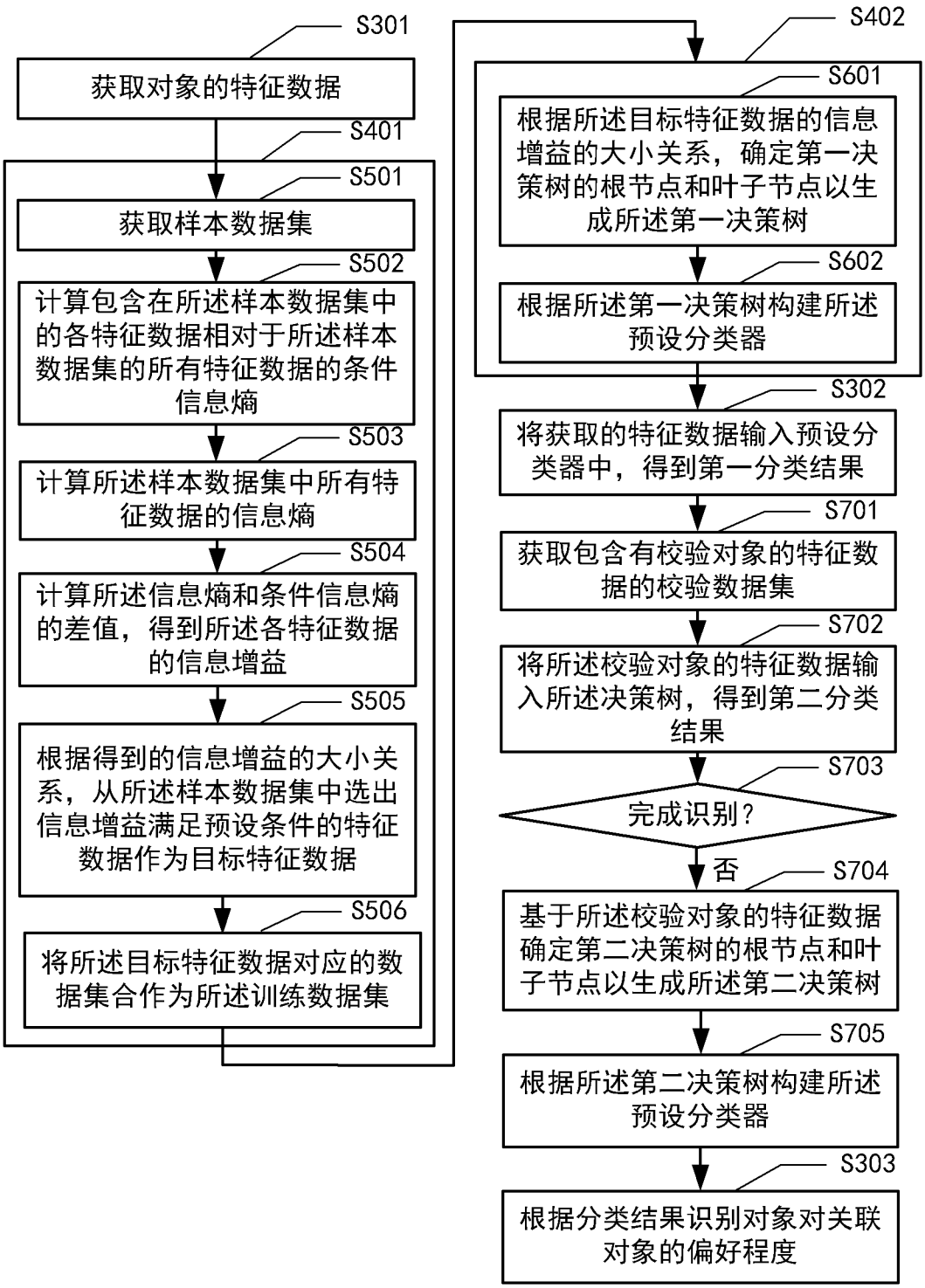
S301
获取对象的特征数据
S401
S501
获取样本数据集
S502
计算包含在所述样本数据集中的各特征数据相对于所述样本数据集的所有特征数据的条件信息熵
S503
计算所述样本数据集中所有特征数据的信息熵
S504
计算所述信息熵和条件信息熵的差值，得到所述各特征数据的信息增益
S505
根据得到的信息增益的大小关系，从所述样本数据集中选出信息增益满足预设条件的特征数据作为目标特征数据
S506
将所述目标特征数据对应的数据集合作为所述训练数据集
S402
S601
根据所述目标特征数据的信息增益的大小关系，确定第一决策树的根节点和叶子节点以生成所述第一决策树
S602
根据所述第一决策树构建所述预设分类器
S302
将获取的特征数据输入预设分类器中，得到第一分类结果
S701
获取包含有校验对象的特征数据的校验数据集
S702
将所述校验对象的特征数据输入所述决策树，得到第二分类结果
S703
完成识别？
否
S704
基于所述校验对象的特征数据确定第二决策树的根节点和叶子节点以生成所述第二决策树
S705
根据所述第二决策树构建所述预设分类器
S303
根据分类结果识别对象对关联对象的偏好程度

图 4D

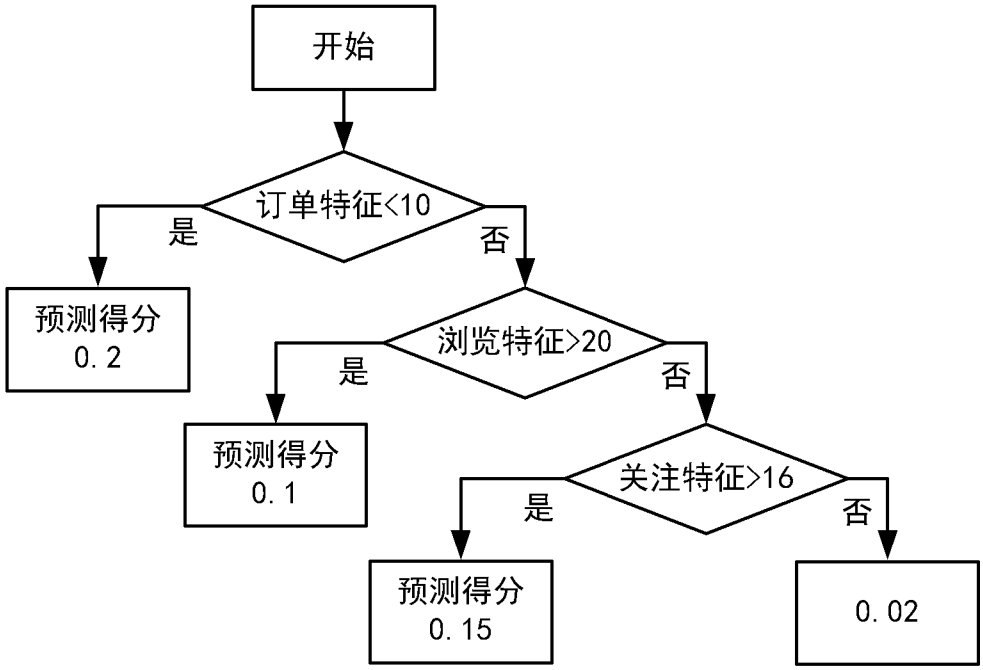

图 5

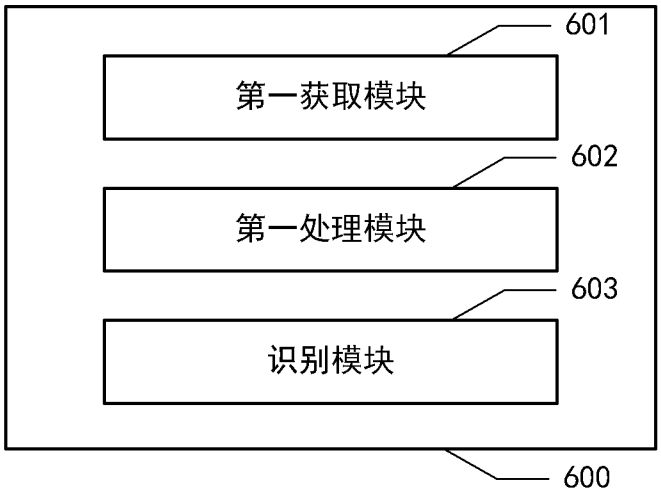
601
第一获取模块
602
第一处理模块
603
识别模块
600

图 6

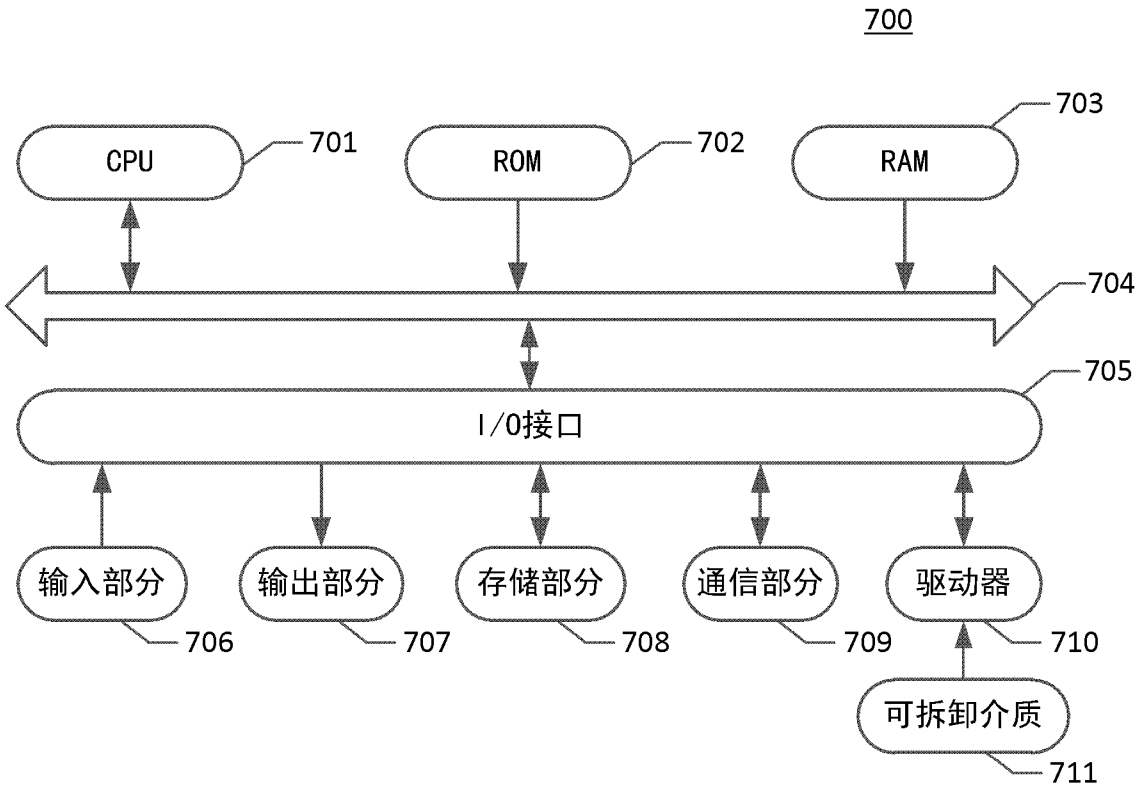
700
CPU
701
ROM
702
RAM
703
704
705
I/O接口
输入部分
706
输出部分
707
存储部分
708
通信部分
709
驱动器
710
可拆卸介质
711

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/109020

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 30/02(2012.01)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: 用户, 特征, 行为, 分类, 决策树, 爱好, 偏好, 兴趣, 样本, 信息熵, 条件信息熵, 信息增益, user, action, behavior, class+, GBDT, tree, interest+, hobby, fond, character+, sample, entropy, plus

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104951441 A (ALIBABA GROUP HOLDING LIMITED) 30 September 2015 (2015-09-30) claims 1-14, and description, paragraphs [0043]-[0045] and [0160]-[0162]	1-2, 6-7, 11-12
Y	CN 104951441 A (ALIBABA GROUP HOLDING LIMITED) 30 September 2015 (2015-09-30) claims 1-14, and description, paragraphs [0043]-[0045] and [0160]-[0162]	3-5, 8-10
Y	CN 105975929 A (PEKING UNIVERSITY SHENZHEN GRADUATE SCHOOL) 28 September 2016 (2016-09-28) claim 4, and description, paragraphs [0019]-[0025]	3-5, 8-10
Y	CN 106204122 A (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO., LTD. ET AL.) 07 December 2016 (2016-12-07) claim 5	5, 10
A	CN 104750789 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 01 July 2015 (2015-07-01) entire document	1-12
A	WO 2017107026 A1 (ZHENG, ZHAOHUI) 29 June 2017 (2017-06-29) entire document	1-12

☐ 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	Date of mailing of the international search report
07 December 2018	**04 January 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/109020

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 104951441 A	30 September 2015	HK 1211111 A1	13 May 2016
CN 105975929 A	28 September 2016	WO 2017190574 A1	09 November 2017
CN 106204122 A	07 December 2016	None	
CN 104750789 A	01 July 2015	None	
WO 2017107026 A1	29 June 2017	EP 3394769 A1	31 October 2018

国际检索报告

国际申请号
PCT/CN2018/109020

A. 主题的分类

G06Q 30/02(2012.01)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,EPODOC,CNPAT,CNKI:用户，特征，行为，分类，决策树，爱好，偏好，兴趣，样本，信息熵，条件信息熵，信息增益，user，action，behavior，class+，GBDT，tree，interest+，hobby，fond，character+，sample，entropy，plus

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
X	CN 104951441 A (阿里巴巴集团控股有限公司) 2015年 9月 30日 (2015-09-30) 权利要求1-14，说明书第[0043]-[0045]、[0160]-[0162]段	1-2，6-7，11-12
Y	CN 104951441 A (阿里巴巴集团控股有限公司) 2015年 9月 30日 (2015-09-30) 权利要求1-14，说明书第[0043]-[0045]、[0160]-[0162]段	3-5，8-10
Y	CN 105975929 A (北京大学深圳研究生院) 2016年 9月 28日 (2016-09-28) 权利要求4和说明书[0019]-[0025]段	3-5，8-10
Y	CN 106204122 A (北京京东尚科信息技术有限公司 等) 2016年 12月 7日 (2016-12-07) 权利要求5	5，10
A	CN 104750789 A (百度在线网络技术北京有限公司) 2015年 7月 1日 (2015-07-01) 全文	1-12
A	WO 2017107026 A1 (ZHENG ZHAOHUI) 2017年 6月 29日 (2017-06-29) 全文	1-12

☐ 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	国际检索报告邮寄日期
2018年 12月 7日	2019年 1月 4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 加玉 电话号码 86-10-53961785

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/109020

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	104951441	A	2015年 9月 30日	HK	1211111	A1	2016年 5月 13日
CN	105975929	A	2016年 9月 28日	WO	2017190574	A1	2017年 11月 9日
CN	106204122	A	2016年 12月 7日		无		
CN	104750789	A	2015年 7月 1日		无		
WO	2017107026	A1	2017年 6月 29日	EP	3394769	A1	2018年 10月 31日