(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107391760 B
(45)授权公告日 2018.05.25

(21)申请号 201710749088.0

(22)申请日 2017.08.25

(65)同一申请的已公布的文献号
申请公布号 CN 107391760 A

(43)申请公布日 2017.11.24

(73)专利权人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区八卦岭工业区平安大厦六楼

(72)发明人 王健宗 黄章成 吴天博 肖京

(74)专利代理机构 深圳市世纪恒程知识产权代理事务所 44287
代理人 胡海国

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)
G06K 9/62(2006.01)
G06Q 30/02(2012.01)

(56)对比文件
CN 103218436 A,2013.07.24,全文.
US 2017103343 A1,2017.04.13,全文.
CN 106157151 A,2016.11.23,全文.

审查员 刘芳

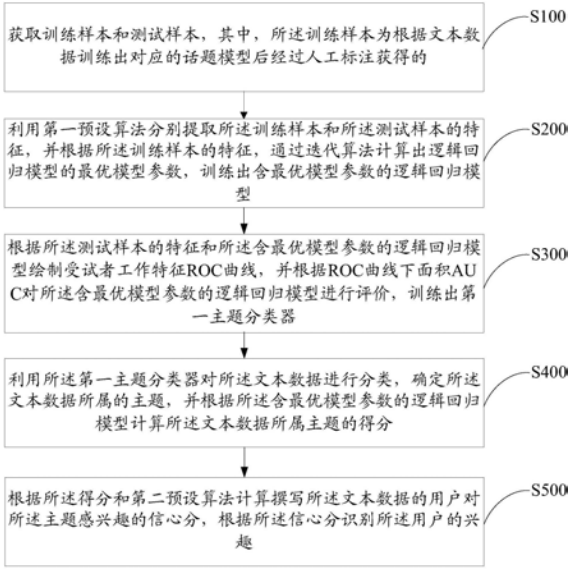
权利要求书3页 说明书14页 附图6页

(54)发明名称

用户兴趣识别方法、装置及计算机可读存储介质

(57)摘要

本发明公开了一种用户兴趣识别方法，该方法包括：获取训练样本和测试样本，其中训练样本为根据文本数据训练出对应话题模型后经人工标注获得的；利用第一预设算法提取训练样本和测试样本的特征，并根据训练样本的特征通过迭代算法计算逻辑回归模型的最优模型参数；根据测试样本的特征和ROC曲线下面积AUC对含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；利用第一主题分类器确定文本数据所属主题，根据含最优模型参数的逻辑回归模型计算文本数据所属主题的得分，并根据第二预设算法计算用户对所述主题感兴趣的信心分。本发明还公开了一种用户兴趣识别装置及计算机可读存储介质，可识别用户兴趣，帮助企业准确定位潜在客户。



CN 107391760 B

1.一种用户兴趣识别方法,其特征在于,所述用户兴趣识别方法包括以下步骤:

获取训练样本和测试样本,其中,所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的;

利用第一预设算法分别提取所述训练样本和所述测试样本的特征,并根据所述训练样本的特征,通过迭代算法计算出逻辑回归模型的最优模型参数,训练出含最优模型参数的逻辑回归模型;

根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线,并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价,训练出第一主题分类器;

利用所述第一主题分类器对所述文本数据进行分类,确定所述文本数据所属的主题,并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分;

根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,根据所述信心分识别所述用户的兴趣;

其中,所述第一预设算法为字节4元语法Byte 4-gram算法;

所述第二预设算法的计算公式为:

$$k_j=\frac{10}{\max(TN_j)-\min(TN_j)},\quad x_{j0}=median(TN_j),$$

$$s(u_i,topic_j,TN_{ij})=\frac{TN_{ij}*Avg(u_i,topic_j)}{1+e^{-k_j*(TN_{ij}-x_{jo})}},$$

其中,TN_j为所有用户对主题$topic_j$感兴趣的文本数,x_{j0}为TN_j的中位数,TN_{ij}为用户u_i发表的关于主题$topic_j$的微博数,$s(u_i,topic_j,TN_{ij})$为用户u_i对所述主题$topic_j$感兴趣的信心分,$Avg(u_i,topic_j)$为用户u_i在主题$topic_j$上的平均得分。

2.如权利要求1所述的用户兴趣识别方法,其特征在于,所述根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,根据所述信心分识别所述用户的兴趣,包括:

根据所述得分和第三预设算法计算所述文本数据所属主题的平均得分;

根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,根据所述信心分识别所述用户的兴趣;

其中,所述第三预设算法的计算公式为:

$$Avg(u_i,topic_j)=\frac{\sum_{m=1}^{n}s(u_i,tweet_m,topic_j)}{n},$$

其中,$s(u_i,tweet_m,topic_j)$为用户u_i的文本数据$tweet_m$分类之后属于主题$topic_j$的得分,n为用户u_i的文本数据信息$tweet_m$中关于主题$topic_j$的总数量。

3.如权利要求1所述的用户兴趣识别方法,其特征在于,所述获取训练样本和测试样本,其中,所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的,包括:

采集文本数据,并对所述文本数据进行预处理,获得对应的第一关键词集;

根据所述第一关键词集和预设数量的话题,利用预设主题模型计算得到所述文本数据

在所述话题上的分布，并根据所述文本数据在所述话题上的分布情况进行聚类，训练出所述文本数据对应的话题模型；

根据基于所述话题模型对所述文本数据的人工标注结果，从所述文本数据中筛选出与目标主题分类器对应的训练样本，并将除所述训练样本之外的文本数据作为测试样本。

4.如权利要求1所述的用户兴趣识别方法，其特征在于，所述利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型，包括：

利用第一预设算法分别提取所述训练样本和所述测试样本的特征，对应建立第一哈希散列表和第二哈希散列表；

将所述第一哈希散列表代入逻辑回归模型，并通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。

5.如权利要求4所述的用户兴趣识别方法，其特征在于，所述根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器，包括：

将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP；

根据所述TP，TN，FN和FP绘制ROC曲线；

计算ROC曲线下面积AUC，根据AUC值对所述含最优模型参数的逻辑回归模型进行评价；

当所述AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

当所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器；

其中，所述根据所述TP，TN，FN和FP绘制ROC曲线，包括：

根据所述TP，TN，FN和FP计算出伪阳性率FPR和真阳性率TPR，对应的计算公式分别为FPR＝FP/(FP+TN)，TPR＝TP/(TP+FN)；

以所述FPR为横坐标，所述TPR为纵坐标，绘制ROC曲线。

6.如权利要求5所述的用户兴趣识别方法，其特征在于，所述根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器的步骤之后，包括：

将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率；

调整所述预设AUC阈值，并根据所述TP，FP和FN计算准确率p和召回率r；

当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值时，训练出第二主题分类器；

利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题的步骤包括：

利用所述第二主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题。

7.如权利要求3所述的用户兴趣识别方法，其特征在于，所述采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集，包括：

采集文本数据，并对所述文本数据进行分词；

根据预设停用词表去除分词后的文本数据中的停用词，得到第二关键词集；

计算所述第二关键词集中各关键词的词频TF和逆向文件频率IDF；

根据所述TF和IDF计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

8.一种用户兴趣识别装置，其特征在于，所述用户兴趣识别装置包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的用户兴趣识别程序，所述用户兴趣识别程序被所述处理器执行时实现如权利要求1至7中任一项所述的用户兴趣识别方法的步骤。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有用户兴趣识别程序，所述用户兴趣识别程序被处理器执行时实现如权利要求1至7中任一项所述的兴趣识别方法的步骤。

用户兴趣识别方法、装置及计算机可读存储介质

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种用户兴趣识别方法、装置及计算机可读存储介质。

背景技术

[0002] 近年来，随着互联网的快速发展，尤其是社会化媒体的异军突起，人们越来越体会到其对自身和信息传播环境的各种影响。以往人们一直是被动的从互联网上获取信息，但是现在越来越多的人主动地参与社会化媒体上信息的产生与传播，随之产生了海量的用户信息和社交关系信息。

[0003] 然而，目前很多企业的内部数据中通常以交易记录为主，所包含的客户信息不够全面，无法准确定位潜在客户，了解用户需求。因此，如何通过互联网数据信息识别用户兴趣，全方面了解用户，从而帮助企业准确定位潜在客户已成为目前亟待解决的问题。

发明内容

[0004] 本发明的主要目的在于提供一种用户兴趣识别方法、装置及计算机可读存储介质，旨在通过互联网的数据信息识别用户兴趣，全方面了解用户，帮助企业快速准确地定位潜在客户，从而提高营销效率。

[0005] 为实现上述目的，本发明提供一种用户兴趣识别方法，所述用户兴趣识别方法包括以下步骤：

[0006] 获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的；

[0007] 利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0008] 根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；

[0009] 利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分；

[0010] 根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。

[0011] 可选地，所述根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣，包括：

[0012] 根据所述得分和第三预设算法计算所述文本数据所属主题的平均得分；

[0013] 根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣；

[0014] 其中，所述第三预设算法的计算公式为：

[0015] $$Avg(u_i, topic_j) = \frac{\sum_{m=1}^{n} s(u_i, tweet_m, topic_j)}{n},$$

[0016] 其中，$Avg(u_i, topic_j)$为用户u_i在主题$topic_j$上的平均得分，$s(u_i, tweet_m, topic_j)$为用户u_i的文本数据$tweet_m$分类之后属于主题$topic_j$的得分，n为用户u_i的文本数据信息$tweet_m$中关于主题$topic_j$的总数量。

[0017] 可选地，所述第二预设算法的计算公式为：

[0018] $$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = median(TN_j),$$

[0019] $$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{j0})}},$$

[0020] 其中，TN_j为所有用户对主题$topic_j$感兴趣的文本数，x_{j0}为TN_j的中位数，TN_{ij}为用户u_i发表的关于主题$topic_j$的微博数，$s(u_i, topic_j, TN_{ij})$为用户u_i对所述主题$topic_j$感兴趣的信心分。

[0021] 可选地，所述获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的，包括：

[0022] 采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集；

[0023] 根据所述第一关键词集和预设数量的话题，利用预设主题模型计算得到所述文本数据在所述话题上的分布，并根据所述文本数据在所述话题上的分布情况进行聚类，训练出所述文本数据对应的话题模型；

[0024] 根据基于所述话题模型对所述文本数据的人工标注结果，从所述文本数据中筛选出与目标主题分类器对应的训练样本，并将除所述训练样本之外的文本数据作为测试样本。

[0025] 可选地，所述利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型，包括：

[0026] 利用第一预设算法分别提取所述训练样本和所述测试样本的特征，对应建立第一哈希散列表和第二哈希散列表；

[0027] 将所述第一哈希散列表代入逻辑回归模型，并通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。

[0028] 可选地，所述根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器，包括：

[0029] 将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP；

[0030] 根据所述TP，TN，FN和FP绘制ROC曲线；

[0031] 计算ROC曲线下面积AUC，根据AUC值对所述含最优模型参数的逻辑回归模型进行评价；

[0032] 当所述AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0033] 当所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器；

[0034] 其中，所述根据所述TP，TN，FN和FP绘制ROC曲线，包括：

[0035] 根据所述TP，TN，FN和FP计算出伪阳性率FPR和真阳性率TPR，对应的计算公式分别为FPR＝FP/(FP+TN)，TPR＝TP/(TP+FN)；

[0036] 以所述FPR为横坐标，所述TPR为纵坐标，绘制ROC曲线。

[0037] 可选地，所述根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器的步骤之后，包括：

[0038] 将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率；

[0039] 调整所述预设AUC阈值，并根据所述TP，FP和FN计算准确率p和召回率r；

[0040] 当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值时，训练出第二主题分类器；

[0041] 利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题的步骤包括：

[0042] 利用所述第二主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题。

[0043] 可选地，所述采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集，包括：

[0044] 采集文本数据，并对所述文本数据进行分词；

[0045] 根据预设停用词表去除分词后的文本数据中的停用词，得到第二关键词集；

[0046] 计算所述第二关键词集中各关键词的词频TF和逆向文件频率IDF；

[0047] 根据所述TF和IDF计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0048] 此外，为实现上述目的，本发明还提供一种用户兴趣识别装置，所述用户兴趣识别装置包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的用户兴趣识别程序，所述用户兴趣识别程序被所述处理器执行时实现上述的用户兴趣识别方法的步骤。

[0049] 此外，为实现上述目的，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有用户兴趣识别程序，所述用户兴趣识别程序被处理器执行时实现上述的兴趣识别方法的步骤。

[0050] 本发明通过获取训练样本和测试样本，其中训练样本为根据文本数据训练出对应话题模型后经人工标注获得的；利用第一预设算法提取训练样本和测试样本的特征，并根据训练样本的特征通过迭代算法计算逻辑回归模型的最优模型参数；根据测试样本的特征

和ROC曲线下面积AUC对含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；利用第一主题分类器对文本数据进行分类，确定文本数据所属的主题，并根据含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分；根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。通过上述方式，本发明利用第一预设算法对训练样本和测试样本进行特征提取，缩短了特征提取和模型训练的时间，提高了分类效率。本发明采用人工标注的方式筛选训练样本，并采用ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价训练出主题分类器，可提高主题分类的准确率，同时，本发明利用了含最优模型参数的逻辑回归模型和第二预设算法，可提高信心分计算的准确性，从而根据计算得到的用户对所述主题感兴趣的信心分识别出用户兴趣，帮助企业全方面了解用户，从而快速准确地定位潜在客户，提高营销效率。

附图说明

[0051] 图1为本发明实施例方案涉及的用户兴趣识别装置结构示意图；

[0052] 图2为本发明用户兴趣识别方法第一实施例的流程示意图；

[0053] 图3为本发明实施例中利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型的细化流程示意图；

[0054] 图4为本发明实施例中根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣的细化流程示意图；

[0055] 图5为本发明实施例中获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的的细化流程示意图；

[0056] 图6为本发明实施例中根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器的细化流程示意图；

[0057] 图7为本发明用户兴趣识别方法第二实施例的流程示意图；

[0058] 图8为本发明实施例中采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集的细化流程示意图；

[0059] 图9为本发明实施例中计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集的细化流程示意图。

[0060] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0061] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0062] 由于现有分类技术的分类效率和准确率较低，导致用户面对海量的信息资源时，难以准确快捷地获取自身所需的相关主题信息。

[0063] 为了解决上述技术问题，本发明提供一种用户兴趣识别方法，通过获取训练样本和测试样本，其中训练样本为根据文本数据训练出对应话题模型后经人工标注获得的；利

用第一预设算法提取训练样本和测试样本的特征，并根据训练样本的特征通过迭代算法计算逻辑回归模型的最优模型参数；根据测试样本的特征和ROC曲线下面积AUC对含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；利用第一主题分类器对文本数据进行分类，确定文本数据所属的主题，并根据含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分；根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。通过上述方式，本发明利用第一预设算法对训练样本和测试样本进行特征提取，缩短了特征提取和模型训练的时间，提高了分类效率。本发明采用人工标注的方式筛选训练样本，并采用ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价训练出主题分类器，可提高主题分类的准确率，同时，本发明利用了含最优模型参数的逻辑回归模型和第二预设算法，可提高信心分计算的准确性，根据计算得到的用户对所述主题感兴趣的信心分识别出用户兴趣，帮助企业全方面了解用户，从而快速准确地定位潜在客户，提高营销效率。

[0064] 请参阅图1，为本发明实施例方案涉及的用户兴趣识别装置结构示意图。

[0065] 本发明实施例终端可以是PC，也可以是智能手机、平板电脑、便携计算机等具有显示功能的终端设备。

[0066] 如图1所示，该终端可以包括：处理器1001，例如CPU，网络接口1004，用户接口1003，存储器1005，通信总线1002。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard)，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如WI-FI接口)。存储器1005可以是高速RAM存储器，也可以是稳定的存储器(non-volatile memory)，例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[0067] 可选地，终端还可以包括摄像头、RF(Radio Frequency，射频)电路，传感器、音频电路、Wi-Fi模块等等。其中，传感器比如光传感器、运动传感器以及其他传感器。具体地，光传感器可包括环境光传感器及接近传感器，其中，环境光传感器可根据环境光线的明暗来调节显示屏的亮度，接近传感器可在移动终端移动到耳边时，关闭显示屏和/或背光。作为运动传感器的一种，重力加速度传感器可检测各个方向上(一般为三轴)加速度的大小，静止时可检测出重力的大小及方向，可用于识别移动终端姿态的应用(比如横竖屏切换、相关游戏、磁力计姿态校准)、振动识别相关功能(比如计步器、敲击)等；当然，移动终端还可配置陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器，在此不再赘述。

[0068] 本领域技术人员可以理解，图1中示出的终端结构并不构成对终端的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

[0069] 如图1所示，作为一种计算机可读存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及用户兴趣识别程序。

[0070] 在图1所示的终端中，网络接口1004主要用于连接后台服务器，与后台服务器进行数据通信；用户接口1003主要用于连接客户端(用户端)，与客户端进行数据通信；而处理器1001可以用于调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0071] 获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的；

[0072] 利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0073] 根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；

[0074] 利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分；

[0075] 根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。

[0076] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0077] 根据所述得分和第三预设算法计算所述文本数据所属主题的平均得分；

[0078] 根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。

[0079] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0080] 所述第三预设算法的计算公式为：

[0081] $$Avg(u_i, topic_j) = \frac{\sum_{m=1}^{n} s(u_i, tweet_m, topic_j)}{n},$$

[0082] 其中，$Avg(u_i, topic_j)$为用户u_i在主题$topic_j$上的平均得分，$s(u_i, tweet_m, topic_j)$为用户u_i的文本数据$tweet_m$分类之后属于主题$topic_j$的得分，n为用户u_i的文本数据信息$tweet_m$中关于主题$topic_j$的总数量。

[0083] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0084] 所述第二预设算法的计算公式为：

[0085] $$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = median(TN_j),$$

[0086] $$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{jo})}},$$

[0087] 其中，TN_j为所有用户对主题$topic_j$感兴趣的文本数，x_{j0}为TN_j的中位数，TN_{ij}为用户u_i发表的关于主题$topic_j$的微博数，$s(u_i, topic_j, TN_{ij})$为用户u_i对所述主题$topic_j$感兴趣的信心分。

[0088] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0089] 采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集；

[0090] 根据所述第一关键词集和预设数量的话题，利用预设主题模型计算得到所述文本数据在所述话题上的分布，并根据所述文本数据在所述话题上的分布情况进行聚类，训练

出所述文本数据对应的话题模型；

[0091] 根据基于所述话题模型对所述文本数据的人工标注结果，从所述文本数据中筛选出与目标主题分类器对应的训练样本，并将除所述训练样本之外的文本数据作为测试样本。

[0092] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0093] 利用第一预设算法分别提取所述训练样本和所述测试样本的特征，对应建立第一哈希散列表和第二哈希散列表；

[0094] 将所述第一哈希散列表代入逻辑回归模型，并通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。

[0095] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0096] 将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP；

[0097] 根据所述TP，TN，FN和FP绘制ROC曲线；

[0098] 计算ROC曲线下面积AUC，根据AUC值对所述含最优模型参数的逻辑回归模型进行评价；

[0099] 当所述AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0100] 当所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器。

[0101] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0102] 根据所述TP，TN，FN和FP计算出伪阳性率FPR和真阳性率TPR，对应的计算公式分别为FPR＝FP/(FP+TN)，TPR＝TP/(TP+FN)；

[0103] 以所述FPR为横坐标，所述TPR为纵坐标，绘制ROC曲线。

[0104] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0105] 将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率；

[0106] 调整所述预设AUC阈值，并根据所述TP，FP和FN计算准确率p和召回率r；

[0107] 当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值时，训练出第二主题分类器；

[0108] 利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题的步骤包括：

[0109] 利用所述第二主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题。

[0110] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0111] 采集文本数据，并对所述文本数据进行分词；

[0112] 根据预设停用词表去除分词后的文本数据中的停用词，得到第二关键词集；

[0113] 计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0114] 进一步地，处理器1001可以调用存储器1005中存储的用户兴趣识别程序，以实现以下步骤：

[0115] 计算所述第二关键词集中各关键词的词频TF和逆向文件频率IDF；

[0116] 根据所述TF和IDF计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0117] 请参阅图2，为本发明用户兴趣识别方法第一实施例的流程示意图。

[0118] 在本发明实施例中，所述用户兴趣识别方法包括：

[0119] 步骤S100，获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的；

[0120] 在本实施例中，获取训练主题分类器所需的训练样本和测试样本，其中，训练样本为根据文本数据训练处对应的话题模型后经过人工标注获得的，用于优化模型的参数，而测试样本为除训练样本之外的文本数据，用于对建立的模型进行性能评价。在具体实施例中，训练样本和测试样本的获得还可以通过程序直接从互联网中查找到的微博进行抽样，例如数学软件Matlab的Svmtrain函数。

[0121] 步骤S200，利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0122] 具体地，请参阅图3，图3为本发明实施例中利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型的细化流程示意图，步骤S200包括：

[0123] 步骤S210，利用第一预设算法分别提取所述训练样本和所述测试样本的特征，对应建立第一哈希散列表和第二哈希散列表；

[0124] 在本实施例中，利用第一预设算法分别提取训练样本和测试样本的特征，在本实施例中，采用二进制哈希散列表的字节4元语法Byte 4-gram算法分别提取所述训练样本和测试样本的特征，把每一个训练样本或测试样本对应地表示为一个由一组特征组成的特征向量。该方法抽取每一训练样本或测试样本数据中所有连续的4个字节为键(key)，将字符串转换成字符串的UTF-8编码所对应的byte数组，值为32bit的整数。进一步地，通过除留余数法构造出哈希函数，并分别对应建立第一哈希散列表和第二哈希散列表。其中，需要说明的是，对于散列表长为m的散列函数公式为：$f(key)=key \bmod p$，$(p\leq m)$。其中，mod表示求余数。在具体实施方式中，为例减小冲突的发生，避免哈希散列表分布过于稀疏，p通常取小于散列表长的最大素数。

[0125] 步骤S220，将所述第一哈希散列表代入逻辑回归模型，并通过迭代算法计算出逻

辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。

[0126] 进一步地，将所述第一哈希散列表代入逻辑回归模型，并通过优化方法迭代计算出最优的模型参数，训练出逻辑回归模型，其中逻辑回归模型用于估计某种事物的可能性，或者说判断一个样本属于某种类别的概率是多少。逻辑回归模型为：

[0127] $$h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}, \quad \theta^T x^{(i)} = \sum_{j=0}^{n} \theta_i x_j^{(i)},$$

[0128] 其中，x_j表示第j个训练样本的特征向量，$x^{(i)}$表示第i次取样，θ表示模型参数。

[0129] 此外，还需说明的是迭代算法包括梯度下降，共轭梯度法和拟牛顿法等。在具体实施例中，可以通过上述任一迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。当然，在具体实施例中，还可以采用其他方法分别提取训练样本和测试样本的特征，例如向量空间模型VSM、信息增益方法、期望交叉熵等。

[0130] 步骤S300，根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器；

[0131] 在本实施例中，将根据测试样本建立的第二哈希散列表代入所述含最优模型参数的逻辑回归模型，从而得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP，其中TP是利用逻辑回归模型对训练样本中正类进行判断后属于仍是正类的数目，TN利用逻辑回归模型对训练样本中负类进行判断后属于仍是负类的数目，FN利用逻辑回归模型对训练样本中负类进行判断后属于是正类的数目和FP利用逻辑回归模型对训练样本中正类进行判断后属于是负类的数目，正类和负类是指人工对训练样本标注的两种类别，即人工标注某个样本属于特定的类，则该样本属于正类，不属于该特定类的样本则属于负类。并根据所述TP，TN，FN和FP计算出伪阳性率FPR和真阳性率TPR，以FPR为横坐标，TPR为纵坐标，绘制出ROC曲线，ROC曲线是获得的各指标的特征曲线，用于展示各指标之间的关系，并进一步计算出ROC曲线下面积AUC，ROC曲线是获得的各指标的特征曲线，用于展示各指标之间的关系，AUC即ROC曲线下面积，AUC越大越好，提示该试验的诊断价值越高，对所述含最优模型参数的逻辑回归模型进行评价，当所述AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型，直至所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器。

[0132] 步骤S400，利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分；

[0133] 步骤S500，根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣。

[0134] 具体地，请参阅图4，图4为本发明实施例中根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣的细化流程示意图，步骤S500包括：

[0135] 步骤S510，根据所述得分和第三预设算法计算所述文本数据所属主题的平均得分；

[0136] 步骤S520,根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,根据所述信心分识别所述用户的兴趣。

[0137] 在本实施例中,利用训练出的第一主题分类器对文本数据进行分类,确定所述文本数据所属的主题,从中挑选出目标主题所对应的文本数据,并根据所述含最优模型参数的逻辑回归模型计算所述目标主题所对应的文本数据在所属目标主题上的得分,并根据该得分和第三预设算法计算所述目标主题所对应的文本数据所属主题的平均得分,第三预设算法的计算公式如下:

[0138] $$Avg(u_i, topic_j) = \frac{\sum_{m=1}^{n} s(u_i, tweet_m, topic_j)}{n},$$

[0139] 其中,$Avg(u_i, topic_j)$为用户u_i在主题$topic_j$上的平均得分,$s(u_i, tweet_m, topic_j)$为用户u_i的文本数据$tweet_m$分类之后属于主题$topic_j$的得分,n为用户u_i的文本数据信息$tweet_m$中关于主题$topic_j$的总数量。

[0140] 进一步地,根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,第二预设算法的计算公式如下:

[0141] $$k_j = \frac{10}{\max(TN_j) - \min(TN_j)}, \quad x_{j0} = median(TN_j),$$

[0142] $$s(u_i, topic_j, TN_{ij}) = \frac{TN_{ij} * Avg(u_i, topic_j)}{1 + e^{-k_j * (TN_{ij} - x_{jo})}},$$

[0143] 其中,TN_j为所有用户对主题$topic_j$感兴趣的文本数,x_{j0}为TN_j的中位数,TN_{ij}为用户u_i发表的关于主题$topic_j$的微博数,$s(u_i, topic_j, TN_{ij})$为用户u_i对所述主题$topic_j$感兴趣的信心分。

[0144] 进一步地,根据计算得到的信心分识别所述用户的兴趣。例如,根据第一主题分类器确定出用户的微博文本数据属于金融类话题,而且计算出的信心分大于预设信心分阈值,则说明该用户对金融产品感兴趣,从而帮助金融企业定位到该潜在用户,可以向其推荐相关的金融产品。

[0145] 此外,需要说明的是,在本实施例中,利用所述第一主题分类器对所述文本数据进行分类时,所述文本数据可以为步骤S100中的文本数据,也可以为从其他网络社交平台或信息资源数据库获得的文本数据。

[0146] 本发明实施例通过获取训练样本和测试样本,其中训练样本为根据文本数据训练出对应话题模型后经人工标注获得的;利用第一预设算法提取训练样本和测试样本的特征,并根据训练样本的特征通过迭代算法计算逻辑回归模型的最优模型参数;根据测试样本的特征和ROC曲线下面积AUC对含最优模型参数的逻辑回归模型进行评价,训练出第一主题分类器;利用第一主题分类器对文本数据进行分类,确定文本数据所属的主题,并根据含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分;根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分,根据所述信心分识别所述用户的兴趣。通过上述方式,本发明利用第一预设算法对训练样本和测试样本进行特征提取,缩短了特征提取和模型训练的时间,提高了分类效率。本发明采用人工标注的方式筛选训练样本,并采用ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价训

练出主题分类器，可提高主题分类的准确率，同时，本发明利用了含最优模型参数的逻辑回归模型和第二预设算法，可提高信心分计算的准确性，从而根据计算得到的用户对所述主题感兴趣的信心分识别出用户兴趣，帮助企业全方面了解用户，从而快速准确地定位潜在客户，提高营销效率。

[0147] 基于图2所示的第一实施例，请参阅图5，为本发明实施例中获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的的细化流程示意图，步骤S100包括：

[0148] 步骤S110，采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集；

[0149] 在本发明实施例中，文本数据可以从各大网络社交平台获得，例如微博、QQ空间、知乎、百度贴吧等，也可以从各大信息资源数据库获得，例如腾讯视频，知网，电子报等。本实施例以微博文本为例进行说明，具体地，微博文本数据的采集可以通过新浪API(Application Programming Interface)获取新浪微博文本数据，所述文本数据包括微博正文和评论。

[0150] 在本发明实施例中，对所述文本数据进行预处理的过程包括对所述文本数据进行分词，并进行词性标注，再根据预设停用词表去除分词后的文本数据中的停用词表，得到第二关键词集。进一步地，计算所述第二关键词集中各关键词的词频TF，逆向文件频率IDF及词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0151] 步骤S120，根据所述第一关键词集和预设数量的话题，利用预设主题模型计算得到所述文本数据在所述话题上的分布，并根据所述文本数据在所述话题上的分布情况进行聚类，训练出所述文本数据对应的话题模型；

[0152] 在本发明实施例中，预设主题模型采用LDA主题模型，该模型是一种非监督机器学习技术，可用于识别大规模文档集或语料库中潜藏的主题信息，将文档集中的每一篇文档用潜在主题的概率分布进行表示，而每一个潜在主题由词项的概率分布进行表示。具体地，本实施例在终端接收到输入的第一关键词集和设定的话题数量时，LDA主题模型会根据关键词在文档中的分布，计算得到所述话题在关键词上的分布，及文本数据在所述话题上的分布。进一步地，根据所述文本数据在所述话题上的分布情况进行聚类，训练出所述文本数据对应的话题模型。

[0153] 步骤S130，根据基于所述话题模型对所述文本数据的人工标注结果，从所述文本数据中筛选出与目标主题分类器对应的训练样本，并将除所述训练样本之外的文本数据作为测试样本。

[0154] 在本实施例中，由于LDA模型是一种话题生成模型，无法控制所得到的话题的种类，因此，需要对得到的话题进行人工标注，从而筛选出与目标主题相对应的文本数据，以此作为主题分类器的训练样本，有利于提高主题分类器的分类准确率。此外，将除训练样本之外的文本数据作为测试样本，用于对训练出的逻辑回归模型进行评价。

[0155] 基于图2所示的第一实施例，请参阅图6，为本发明实施例中根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器的细

化流程示意图，步骤S300包括：

[0156] 步骤S310，将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP；

[0157] 步骤S320，根据所述TP，TN，FN和FP绘制ROC曲线；

[0158] 步骤S330，计算ROC曲线下面积AUC，根据AUC值对所述含最优模型参数的逻辑回归模型进行评价；

[0159] 步骤S340，当所述AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型；

[0160] 步骤S350，当所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器。

[0161] 在本实施例中，将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，对测试样本进行分析，会出现以下四种情况：如果一个文本数据属于某一话题，同时被含最优模型参数的逻辑回归模型预测为属于该话题，则为真阳性TP；如果一个文本数据不属于某一话题，同时被预测为不属于该话题，则为真阴性TN；如果一个文本数据属于某一话题，却被预测为不属于该话题，则为伪阴性FN；如果一个文本数据不属于某一话题，却被预测为属于该话题，则为伪阳性FP。

[0162] 进一步，根据所述TP，TN，FN和FP绘制ROC曲线，具体地，ROC曲线以伪阳性率FPR为横坐标，以真阳性率TPR为纵坐标，具体计算公式如下：

[0163] FPR＝FP/(FP+TN)，TPR＝TP/(TP+FN)。

[0164] 进一步地，计算ROC曲线下面积AUC，计算公式如下：

[0165] $$\mathrm{AUC}=\int_0^1 TPR\,dFPR=\frac{1}{(TP+FN)(TN+FP)}\int_0^1 TP\,dFP$$

[0166] 在本实施例中，AUC值越大表示该含最优模型参数的逻辑回归模型的性能越好。当计算得到的AUC值小于或等于预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型。直至所述AUC值大于所述预设AUC阈值时，则判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器。

[0167] 基于图2所示的第一实施例，请参阅图7，为本发明用户兴趣识别方法第二实施例的流程示意图，该用户兴趣识别方法还包括：

[0168] 步骤S600，将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率；

[0169] 步骤S700，调整所述预设AUC阈值，并根据所述TP，FP和FN计算准确率p和召回率r；

[0170] 步骤S800，当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值时，训练出第二主题分类器；

[0171] 步骤S900，利用所述第二主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分。

[0172] 需要说明的是，相对于图2所示的第一实施例，图7所示第二实施例的区别在于：在实际使用过程中，由于文本数据过多，人工标注样本劳动力过大，可能无法涵盖所有可能的文本数据，导致使用效果不佳。此外，在使用ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价时，默认使用0.5作为预设AUC阈值，大于0.5则逻辑回归模型的预测结果为1，即表示属于该话题；小于或等于0.5时则逻辑回归模型的预测结果为0，即表示不属于该话题。因此，在第二实施例中，通过调整所述预设AUC阈值，在保证准确率p和召回率r的同时，进一步提高所述第二主题分类器的分类准确率。

[0173] 在本发明实施例中，将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率。进一步地，调整所述预设AUC阈值，并根据所述TP，FP和FN计算出准确率p和召回率r，计算公式如下：

[0174] $$p=\frac{TP}{TP+FP},\quad r=\frac{TP}{TP+FN},$$

[0175] 当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，继续进行调整，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值，训练出第二主题分类器，并利用所述第二主题分类器对所述文本数据进行分类。

[0176] 基于图5所示的实施方式，请参阅图8，为本发明实施例中采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集的细化流程示意图，步骤S110包括：

[0177] 步骤S111，采集文本数据，并对所述文本数据进行分词；

[0178] 步骤S112，根据预设停用词表去除分词后的文本数据中的停用词，得到第二关键词集；

[0179] 步骤S113，计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0180] 具体地，请参阅图9，图9为本发明实施例中计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集的细化流程示意图，步骤S113包括：

[0181] 步骤S1131，计算所述第二关键词集中各关键词的词频TF和逆向文件频率IDF；

[0182] 步骤S1132，根据所述TF和IDF计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集。

[0183] 在本发明实施例中，文本数据可以从各大网络社交平台获得，例如微博、QQ空间、知乎、百度贴吧等，也可以从各大信息资源数据库获得，例如腾讯视频，知网，电子报等。本实施例以微博文本为例进行说明，具体地，微博文本数据的采集可以通过新浪API(Application Programming Interface)获取新浪微博文本数据，所述文本数据包括微博正文和评论。

[0184] 进一步地，对所述文本数据进行预处理，预处理过程包括对所述文本数据进行分词，并进行词性标注。需要说明的是，分词处理可以通过分词工具实施，例如汉语词法分析系统ICTCLAS，清华大学中文词法分析程序THULAC，语言技术平台LTP等。分词主要是根据中文语言的特点，将所述样本数据中的每条中文文本切割成一个一个的单词，并进行词性标注。

[0185] 进一步地，预处理过程还包括根据预设停用词表去除分词后的文本数据中的停用词。停用词的去除有利于提高关键词的密度，从而有利于文本数据所属话题的确定。需要说明的是，停用词主要包括两类：第一类是使用过于频繁的一些单词，例如“我”，“就”等，这类词几乎在每个文档中均会出现；第二类是在文本中出现频率很高，但无实际意义的单词，这类词只有将其放入一个完整的句子中才有一定作用，包括语气助词、副词、介词、连接词等，如“的”、“在”，“接着”等。

[0186] 进一步地，预处理过程还包括计算所述第一关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第二关键词集。具体地，首先计算词频IF和逆向文件频率IDF，其中，TF表示某个关键词在当前文档中出现的频率，IDF表示该关键词在所有文本数据的文档中的分布情况，是一个词语普遍重要性的度量。TF和IDF的计算公式为：

[0187] $$\mathrm{TF}=\frac{n_i}{n},\quad \mathrm{IDF}=\log\frac{N}{1+N_i}$$

[0188] 其中，n_i表示该关键词在当前文档中出现的次数，n表示当前文档中的关键词总数，N表示数据集的文档总数，N_i表示在文本数据集在该关键词i的文档数。

[0189] 进一步地，根据公式TF-IDF＝TF×IDF计算TF-IDF值，去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的关键词集。

[0190] 此外，本发明实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有用户兴趣识别程序，所述用户兴趣识别程序被处理器执行时实现如上所述的用户兴趣识别方法的步骤。

[0191] 其中，在所述处理器上运行的用户兴趣识别程序被执行时所实现的方法可参照本发明用户兴趣识别方法的各个实施例，在此不作赘述。

[0192] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0193] 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。

[0194] 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上述实施例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如上所述的一个存储介质(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，空调器，或者网络设备等)执行本发明各个实施例所述的方法。

[0195] 以上仅为本发明的优选实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

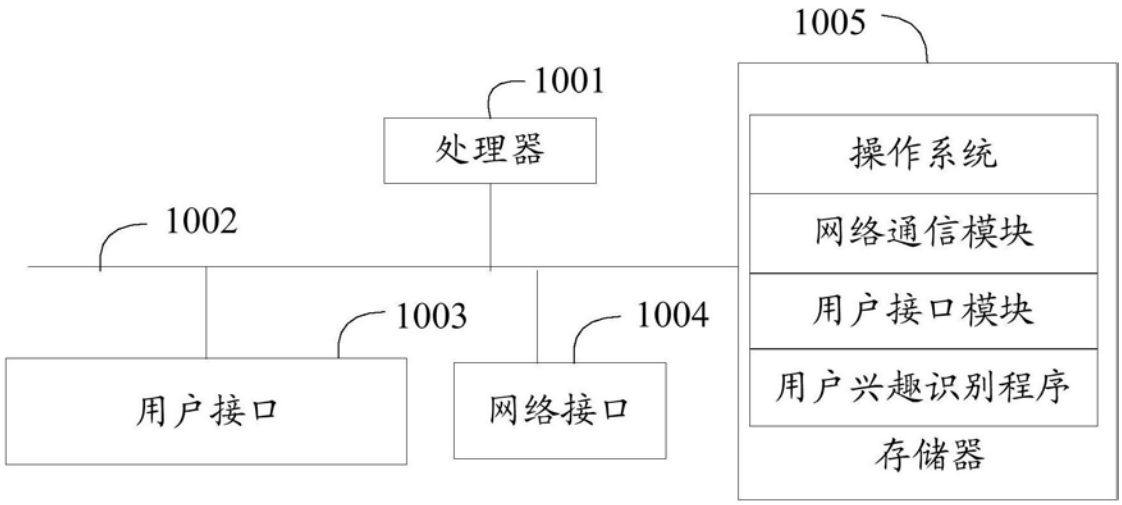

1005
1001
处理器
操作系统
网络通信模块
1002
用户接口模块
1003
1004
用户接口
网络接口
用户兴趣识别程序
存储器

图1

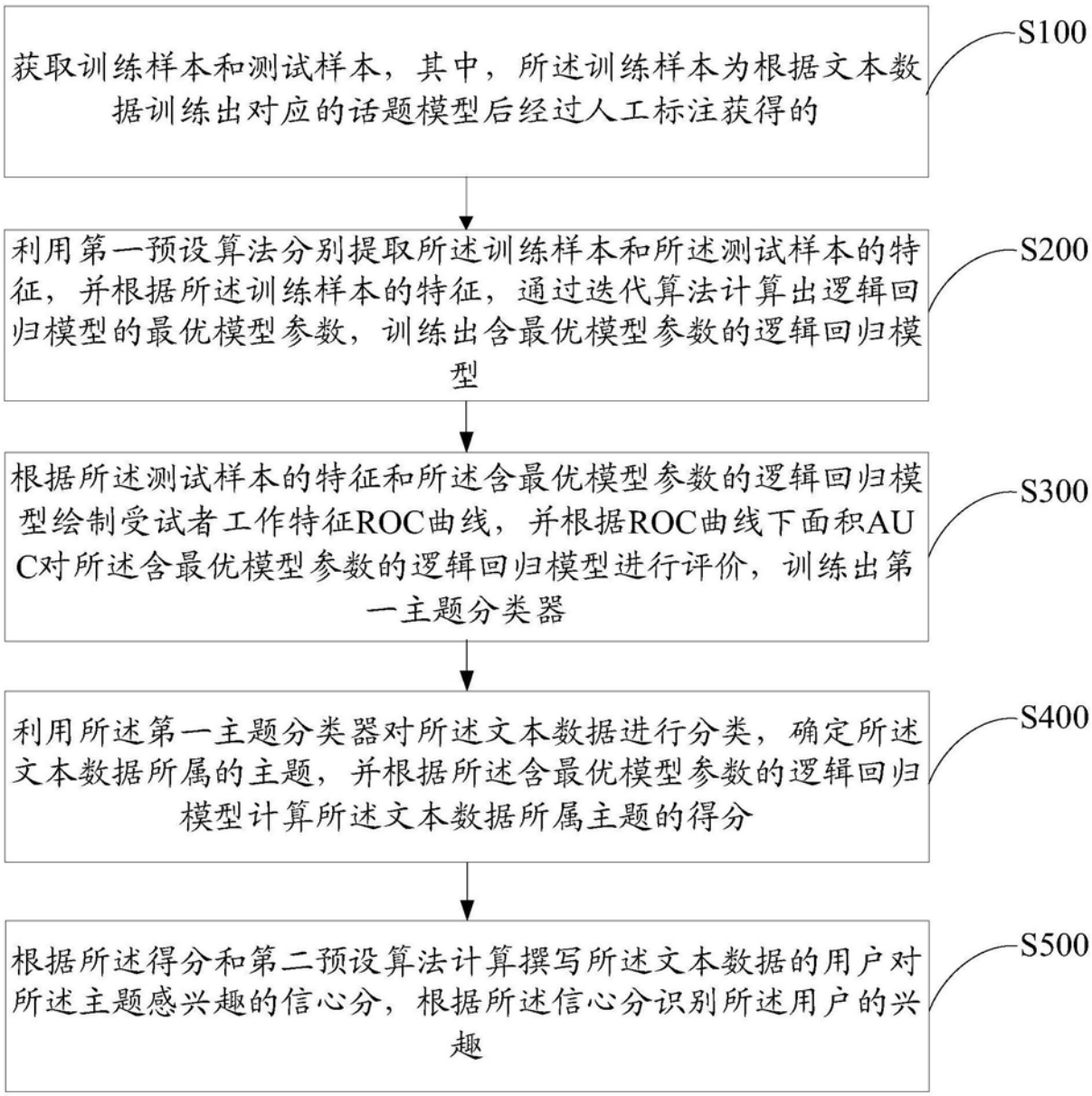

获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的
S100
利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型
S200
根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器
S300
利用所述第一主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分
S400
根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣
S500

图2

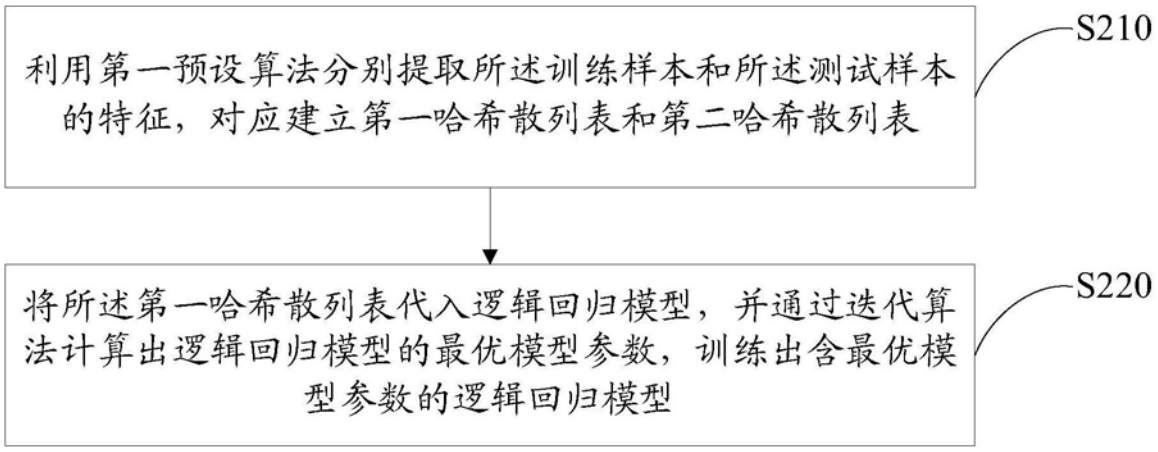

利用第一预设算法分别提取所述训练样本和所述测试样本的特征，对应建立第一哈希散列表和第二哈希散列表
S210
将所述第一哈希散列表代入逻辑回归模型，并通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型
S220

图3

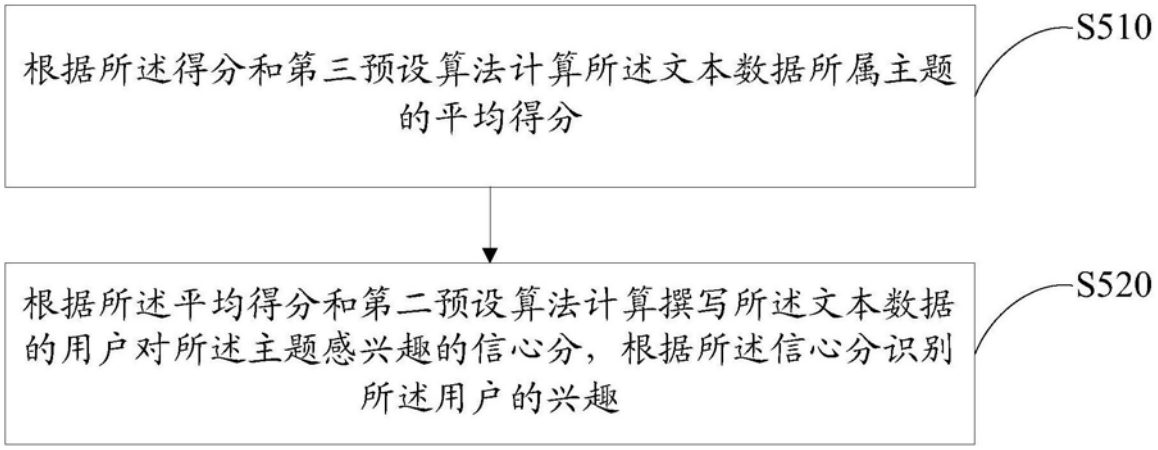

根据所述得分和第三预设算法计算所述文本数据所属主题的平均得分
S510
根据所述平均得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣
S520

图4

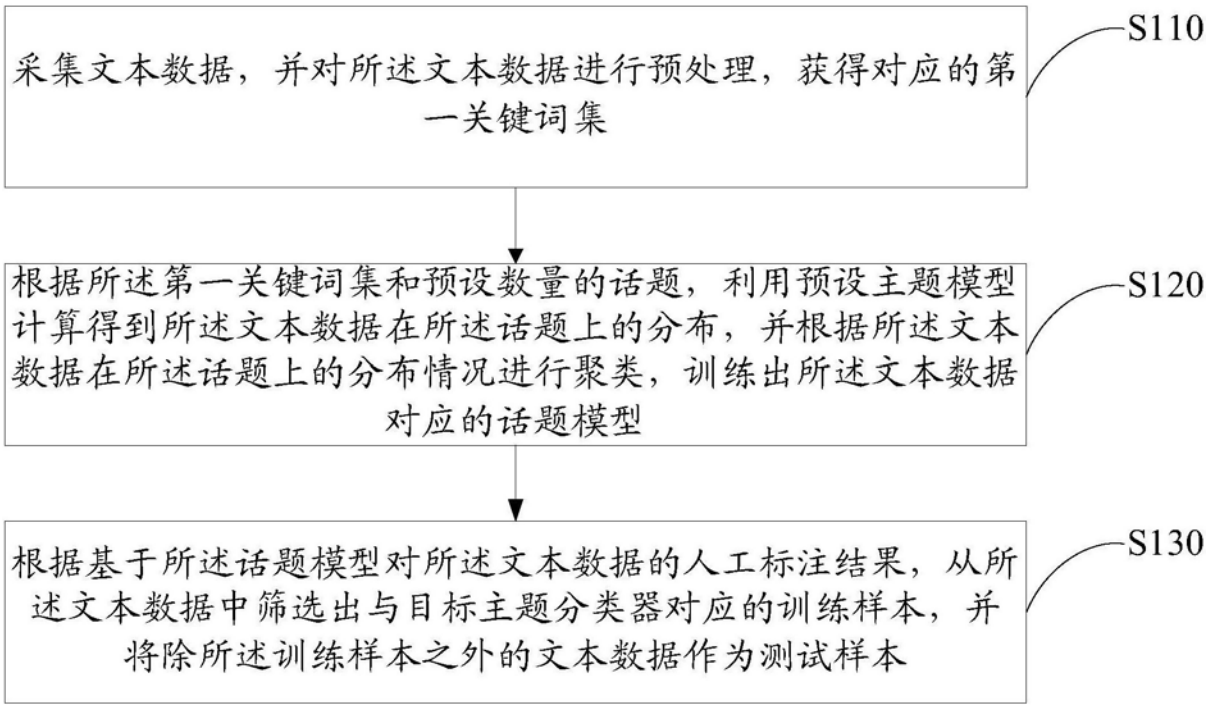

采集文本数据，并对所述文本数据进行预处理，获得对应的第一关键词集
S110
根据所述第一关键词集和预设数量的话题，利用预设主题模型计算得到所述文本数据在所述话题上的分布，并根据所述文本数据在所述话题上的分布情况进行聚类，训练出所述文本数据对应的话题模型
S120
根据基于所述话题模型对所述文本数据的人工标注结果，从所述文本数据中筛选出与目标主题分类器对应的训练样本，并将除所述训练样本之外的文本数据作为测试样本
S130

图5

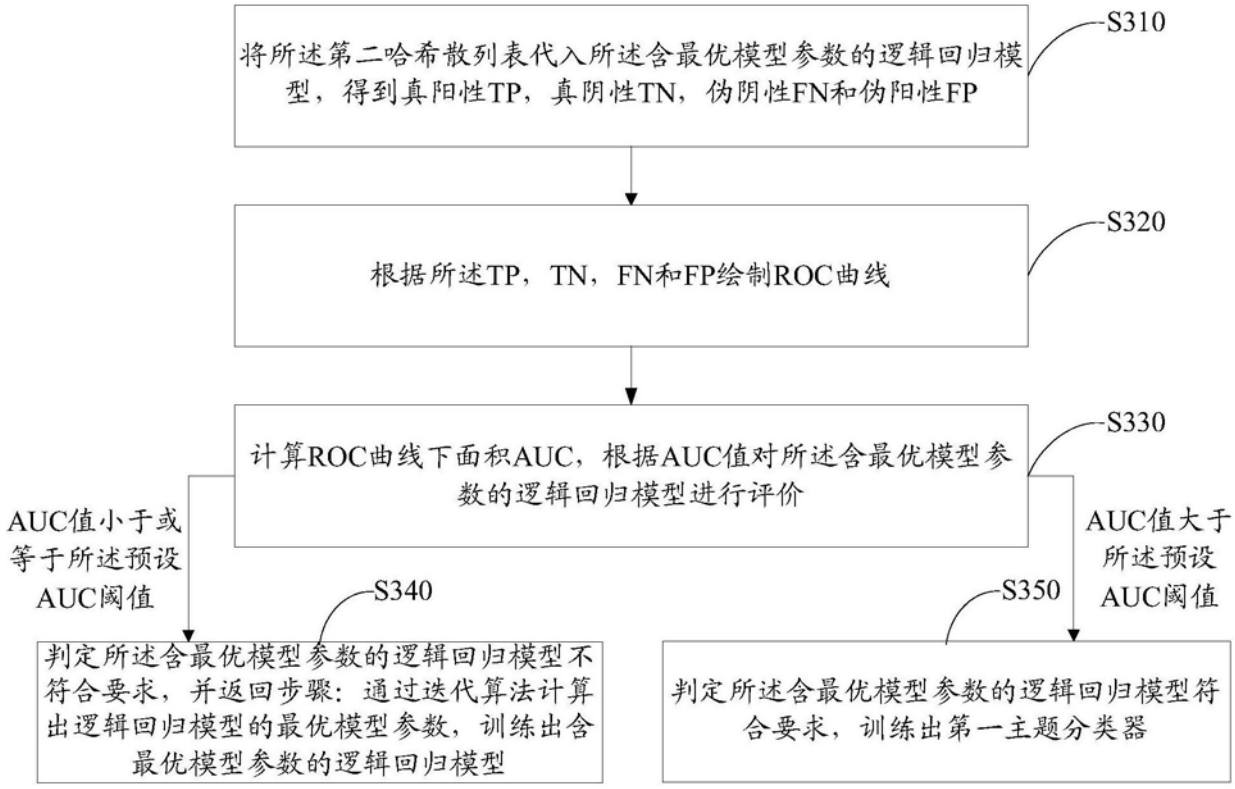

将所述第二哈希散列表代入所述含最优模型参数的逻辑回归模型，得到真阳性TP，真阴性TN，伪阴性FN和伪阳性FP
S310
根据所述TP，TN，FN和FP绘制ROC曲线
S320
计算ROC曲线下面积AUC，根据AUC值对所述含最优模型参数的逻辑回归模型进行评价
S330
AUC值小于或等于所述预设AUC阈值
AUC值大于所述预设AUC阈值
S340
判定所述含最优模型参数的逻辑回归模型不符合要求，并返回步骤：通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型
S350
判定所述含最优模型参数的逻辑回归模型符合要求，训练出第一主题分类器

图6

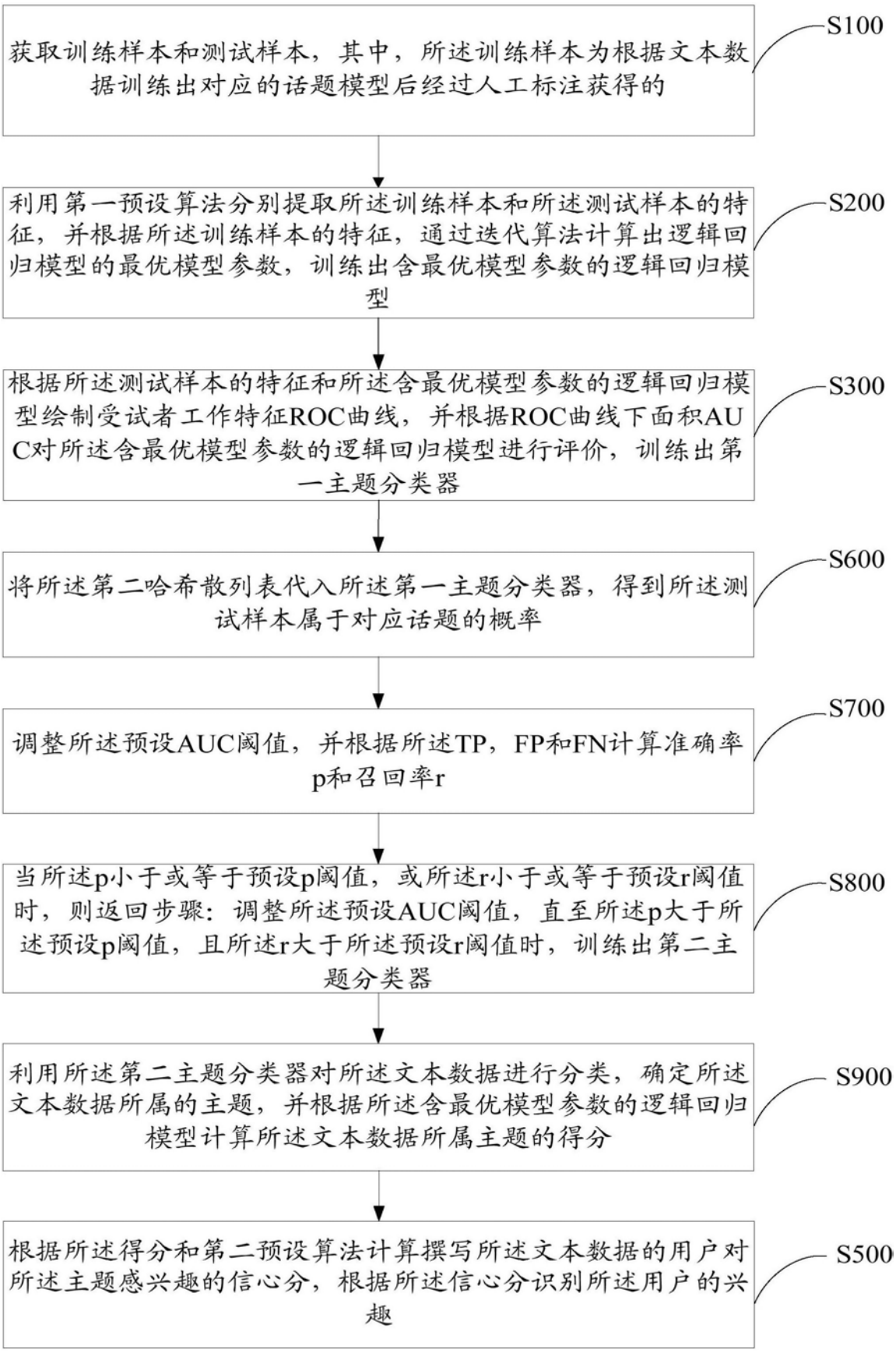

获取训练样本和测试样本，其中，所述训练样本为根据文本数据训练出对应的话题模型后经过人工标注获得的
S100
利用第一预设算法分别提取所述训练样本和所述测试样本的特征，并根据所述训练样本的特征，通过迭代算法计算出逻辑回归模型的最优模型参数，训练出含最优模型参数的逻辑回归模型
S200
根据所述测试样本的特征和所述含最优模型参数的逻辑回归模型绘制受试者工作特征ROC曲线，并根据ROC曲线下面积AUC对所述含最优模型参数的逻辑回归模型进行评价，训练出第一主题分类器
S300
将所述第二哈希散列表代入所述第一主题分类器，得到所述测试样本属于对应话题的概率
S600
调整所述预设AUC阈值，并根据所述TP，FP和FN计算准确率p和召回率r
S700
当所述p小于或等于预设p阈值，或所述r小于或等于预设r阈值时，则返回步骤：调整所述预设AUC阈值，直至所述p大于所述预设p阈值，且所述r大于所述预设r阈值时，训练出第二主题分类器
S800
利用所述第二主题分类器对所述文本数据进行分类，确定所述文本数据所属的主题，并根据所述含最优模型参数的逻辑回归模型计算所述文本数据所属主题的得分
S900
根据所述得分和第二预设算法计算撰写所述文本数据的用户对所述主题感兴趣的信心分，根据所述信心分识别所述用户的兴趣
S500

图7

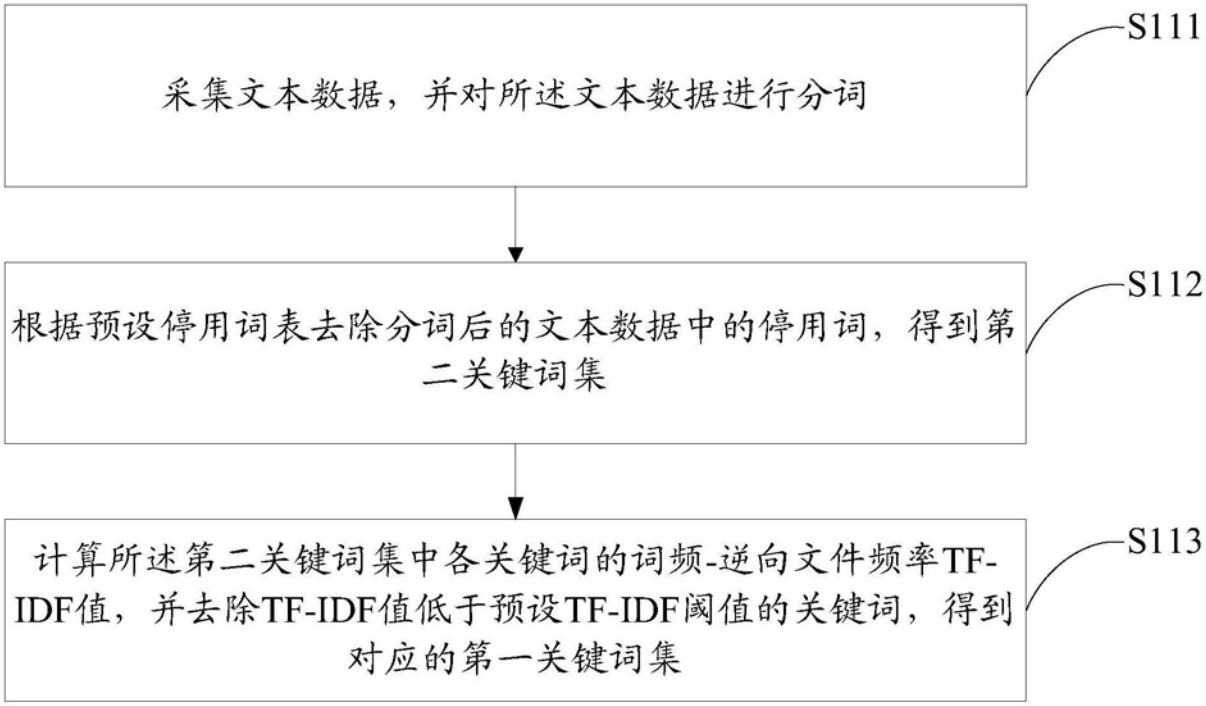

采集文本数据，并对所述文本数据进行分词
S111
根据预设停用词表去除分词后的文本数据中的停用词，得到第二关键词集
S112
计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集
S113

图8

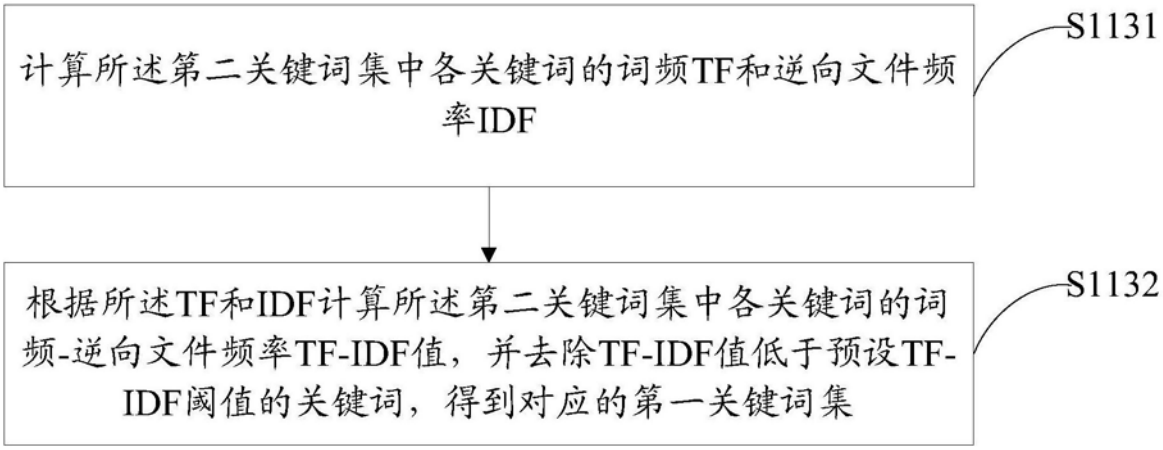

计算所述第二关键词集中各关键词的词频TF和逆向文件频率IDF
S1131
根据所述TF和IDF计算所述第二关键词集中各关键词的词频-逆向文件频率TF-IDF值，并去除TF-IDF值低于预设TF-IDF阈值的关键词，得到对应的第一关键词集
S1132

图9