(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 111095308 A
(43)申请公布日 2020.05.01

(21)申请号 201880046890.6

(22)申请日 2018.05.14

(30)优先权数据
62/505,936 2017.05.14 US

(85)PCT国际申请进入国家阶段日
2020.01.14

(86)PCT国际申请的申请数据
PCT/US2018/032607 2018.05.14

(87)PCT国际申请的公布数据
WO2018/213205 EN 2018.11.22

(71)申请人 数字推理系统有限公司
地址 美国田纳西州

(72)发明人 科里·休斯 提莫西·埃斯蒂斯 约翰·刘 布兰登·卡尔 乌黛·卡马斯

(74)专利代理机构 北京聿宏知识产权代理有限公司 11372
代理人 吴大建 张杰

(51)Int.Cl.
G06N 20/00(2019.01)
G06N 3/04(2006.01)
G06K 9/62(2006.01)
G06F 8/36(2018.01)

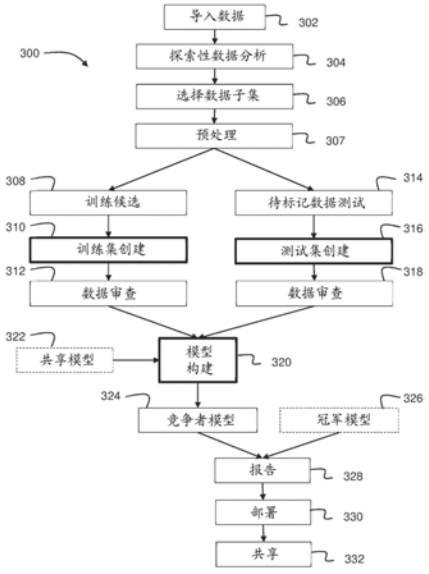
权利要求书6页 说明书25页 附图24页

(54)发明名称

用于快速构建、管理和共享机器学习模型的系统和方法

(57)摘要

在一些方面，提供用于快速构建、管理和共享机器学习模型的系统和方法。管理所述机器学习模型的生命周期可包括：接收未注释数据集；请求所述未注释数据的样本的注释以产生注释数据集；基于所述注释数据集构建机器学习模型；将所述机器学习模型部署到客户端系统，其中生成生产注释；收集所生成的生产注释并生成并入所述生产注释的新机器学习模型；以及选择基于所述注释数据集构建的所述机器学习模型或所述新机器学习模型中的一者。



CN 111095308 A

1.一种管理机器学习模型的生命周期的方法,所述方法包括:

接收未注释数据集;

请求所述未注释数据的样本的注释以产生注释数据集;

基于所述注释数据集构建机器学习模型;

将所述机器学习模型部署到客户端系统,其中生成生产注释;

收集所生成的生产注释并生成并入所述生产注释的新机器学习模型;以及

选择基于所述注释数据集构建的所述机器学习模型或所述新机器学习模型中的一者。

2.如权利要求1所述的方法,其还包括:

报告所述机器学习模型的质量的一个或多个量度,所述一个或多个量度包括精度、召回率、平均精度、接收者操作员特征分数或F-β分数。

3.如权利要求1所述的方法,其还包括:

与第三方共享所述模型。

4.如权利要求1所述的方法,其中请求样本的注释包括:

基于用户输入或自动化采样器选择来从所述未注释数据集选择样本。

5.如权利要求4所述的方法,其中所述用户输入包括语义搜索、类似样本的选择或所述未注释数据的可视图上的选择中的一者或多者。

6.如权利要求4所述的方法,其中所述自动化采样器选择来自进展中的多个采样器中的一个。

7.如权利要求6所述的方法,其中所述多个采样器中的每一个使用不同的采样算法。

8.如权利要求7所述的方法,其中所述相应的采样算法选自:密度采样算法;熵采样算法;所估计误差减小采样算法;详尽采样算法;标记预测算法;难例挖掘采样算法;高置信度采样算法;线性采样算法;图可视化采样算法;元数据搜索采样算法;最小裕度采样算法;委员会查询采样算法;随机采样算法;审查采样算法;搜索采样算法;相似性采样算法;对针对其所述输入为跳过样本类型算法的样本的采样;分层采样算法;最高置信度样本算法;或最不确定样本算法。

9.如权利要求7所述的方法,其中所述进展包括所述多个所述采样器中的采样器之间的连续改变。

10.如权利要求9所述的方法,其中所述多个采样器中的每个采样器具有预期的结果分布,所述预期的结果分布确定是移动到所述进展中的前一还是后一采样器。

11.如权利要求10所述的方法,其中在接收到具有不正确的模型预测的预定数量的样本注释后,所述进展在采样器之间改变为所述进展中的前一采样器。

12.如权利要求10所述的方法,其中在接收到具有一致的模型预测的预定数量的样本注释后,所述进展在采样器之间改变为所述进展中的后一采样器。

13.如权利要求1所述的方法,其中构建所述机器学习模型包括:接收共享模型,并且将中间模型的权重初始化为所述共享模型的权重并以不同的学习速率进行训练。

14.如权利要求1所述的方法,其中请求所述未注释数据的样本的注释包括:请求测试数据集的详尽注释。

15.如权利要求14所述的方法,其中所述测试数据集的所述详尽注释是通过包括密度采样、水平集树或随机采样中的一者或多者的远程监督来执行。

16.如权利要求1所述的方法，其中请求所述未注释数据的样本的注释包括：在多个采样器中的采样器的图形用户界面上呈现用于从所述未注释数据集中选择样本的推荐。

17.如权利要求16所述的方法，其还包括：在所述图形用户界面上呈现数据质量和数量指标。

18.如权利要求17所述的方法，其中所述数据数量指标包括多个训练的样本、多个正实例、多个负实例、或针对一类样本训练的多个样本中的一者或多者。

19.如权利要求17所述的方法，其中所述数据质量指标包括准确度、精度、召回率或F1分数中的一者或多者。

20.如权利要求1所述的方法，其还包括：在图形用户界面上呈现跨所述未注释数据集的注释的不一致性。

21.如权利要求1所述的方法，其中构建所述机器学习模型包括：选择建立所述机器学习模型的算法和损失函数。

22.如权利要求21所述的方法，其中选择所述算法是基于模型类型。

23.如权利要求21所述的方法，其还包括：

通过基于从所述未注释数据集进行注释的注释训练数据集对模型进行多次训练并测量跨运行的质量指标的分散度来测试收敛性。

24.如权利要求23所述的方法，其中所述质量指标包括学习曲线的斜率。

25.如权利要求21所述的方法，其中所述模型是使用针对给定模型类型选择的默认超参数和所述算法来训练。

26.如权利要求25所述的方法，其中所述超参数是使用随机选择、网格搜索或贝叶斯估计方法中的一者或多者来选择。

27.如权利要求25所述的方法，其中针对所述模型存储随机种子、算法选择、损失函数、超参数、数据集分割、数据集散列或类权重中的一者或多者。

28.如权利要求1所述的方法，其中对所述机器学习模型进行版本化、更改或回滚。

29.如权利要求1所述的方法，其还包括：

通过数据漂移或概念漂移来监测模型之间的变化。

30.如权利要求29所述的方法，其中概念漂移是通过基于对所述注释数据集与所述生产注释之间的多个变化预测的量化训练模型来计算。

31.如权利要求29所述的方法，其中数据漂移是基于所述注释数据集与所述生产注释之间的语料库统计和/或语料库对比来测量。

32.如权利要求29所述的方法，其中警示是在识别出数据漂移或概念漂移时生成。

33.如权利要求32所述的方法，其中所述数据漂移或所述概念漂移包括基于随时间推移的未注释数据的指标或基于随时间推移的模型预测的指标。

34.如权利要求3所述的方法，其中共享所述模型包括：执行特征散列、加密散列或随机投影中的一者或多者。

35.如权利要求3所述的方法，其中共享所述模型包括：共享所述模型的梯度更新。

36.如权利要求35所述的方法，其中所述梯度更新被添加到计算图形中的层。

37.如权利要求3所述的方法，其中共享所述模型包括：共享一个或多个模型资产。

38.如权利要求37所述的方法，其中所述一个或多个模型资产包括基于数据集、词向

量、注释集、关键字和短语列表、实例列表、语言模型、词典以及训练的模型和模型架构训练的词嵌入。

39.如权利要求38所述的方法，其中所述一个或多个模型资产被清除个人可识别信息。

40.如权利要求6所述的方法，其中所述进展包括从种子采样器到难例采样器、到分层采样器、到不确定性采样器的进展。

41.如权利要求1所述的方法，其中请求样本的注释包括：在图形用户界面上将问题呈现给用户以获得注释反馈。

42.如权利要求1所述的方法，其还包括：预测所述未注释数据的样本的一个或多个注释。

43.如权利要求42所述的方法，其中所述一个或多个注释的所述预测在请求所述未注释数据的样本的注释之前进行。

44.如权利要求42所述的方法，其还包括：基于采样分数将所预测的一个或多个注释存储在优先级队列中。

45.如权利要求44所述的方法，其中所述采样分数是所预测的一个或多个注释的置信度分数。

46.如权利要求44所述的方法，其还包括：在将所预测的一个或多个注释存储在所述优先级队列中之前确定所述采样分数是否大于阈值采样分数。

47.如权利要求46所述的方法，其还包括：丢弃所具有的采样分数被确定为小于所述阈值采样分数的预测。

48.如权利要求44所述的方法，其中所述优先级队列存储预定最大数量的预测。

49.如权利要求44所述的方法，其还包括：在将所述预测存储在所述优先级队列中之前确定存储在所述优先级队列中的预测数量小于所述预定最大预测数量。

50.如权利要求44所述的方法，其还包括：在将所述预测存储在所述优先级队列中之前确定所述采样分数大于所述优先级队列中的至少一个先前存储的预测。

51.如权利要求44所述的方法，其还包括：丢弃所述优先级队列中具有最低采样分数的先前存储的预测。

52.如权利要求44所述的方法，其中请求所述未注释数据的样本的注释包括：从多个优先级队列中选择所述优先级队列。

53.一种用于管理机器学习模型的生命周期的系统，其包括：

处理器；以及

非暂时性存储器装置，所述非暂时性存储器装置联接到所述处理器并存储计算机可读指令，所述计算机可读指令在由所述处理器执行时致使所述系统执行包括以下的功能：

接收未注释数据集；

请求所述未注释数据的样本的注释以产生注释数据集；

基于所述注释数据集构建机器学习模型；

将所述机器学习模型部署到客户端系统，其中生成生产注释；

收集所生成的生产注释并生成并入所述生产注释的新机器学习模型；以及

选择基于所述注释数据集构建的所述机器学习模型或所述新机器学习模型中的一者。

54.如权利要求53所述的系统，其中由所述系统执行的所述功能还包括：

报告所述机器学习模型的质量的一个或多个量度，所述一个或多个量度包括精度、召回率、平均精度、接收者操作员特征分数或F-β分数。

55.如权利要求53所述的系统，其中由所述系统执行的所述功能还包括：

与第三方共享所述模型。

56.如权利要求53所述的系统，其中请求样本的注释包括：

基于用户输入或自动化采样器选择来从所述未注释数据集选择样本。

57.如权利要求56所述的系统，其中所述用户输入包括语义搜索、类似样本的选择或所述未注释数据的可视图上的选择中的一者或多者。

58.如权利要求56所述的系统，其中所述自动化采样器选择来自进展中的多个采样器中的一个。

59.如权利要求58所述的系统，其中所述多个采样器中的每一个使用不同的采样算法。

60.如权利要求59所述的系统，其中所述相应的采样算法选自：密度采样算法；熵采样算法；所估计误差减小采样算法；详尽采样算法；标记预测算法；难例挖掘采样算法；高置信度采样算法；线性采样算法；图可视化采样算法；元数据搜索采样算法；最小裕度采样算法；委员会查询采样算法；随机采样算法；审查采样算法；搜索采样算法；相似性采样算法；对针对其所述输入为跳过样本类型算法的样本的采样；分层采样算法；最高置信度样本算法；或最不确定样本算法。

61.如权利要求58所述的系统，其中所述进展包括所述多个所述采样器中的采样器之间的连续改变。

62.如权利要求61所述的系统，其中所述多个采样器中的每个采样器具有预期的结果分布，所述预期的结果分布确定是移动到所述进展中的前一还是后一采样器。

63.如权利要求62所述的系统，其中在接收到具有不正确的模型预测的预定数量的样本注释后，所述进展在采样器之间改变为所述进展中的前一采样器。

64.如权利要求53所述的系统，其中在接收到具有一致的模型预测的预定数量的样本注释后，所述进展在采样器之间改变为所述进展中的后一采样器。

65.如权利要求53所述的系统，其中构建所述机器学习模型包括：接收共享模型，并且将中间模型的权重初始化为所述共享模型的权重并以不同的学习速率进行训练。

66.如权利要求53所述的系统，其中请求所述未注释数据的样本的注释包括：请求测试数据集的详尽注释。

67.如权利要求66所述的系统，其中所述测试数据集的所述详尽注释是通过包括密度采样、水平集树或随机采样中的一者或多者的远程监督来执行。

68.如权利要求53所述的系统，其中请求所述未注释数据的样本的注释包括：在多个采样器中的采样器的图形用户界面上呈现用于从所述未注释数据集中选择样本的推荐。

69.如权利要求68所述的系统，其中由所述系统执行的所述功能还包括：在所述图形用户界面上呈现数据质量和数量指标。

70.如权利要求69所述的系统，其中所述数据数量指标包括多个训练的样本、多个正实例、多个负实例、或针对一类样本训练的多个样本中的一者或多者。

71.如权利要求69所述的系统，其中所述数据质量指标包括准确度、精度、召回率或F1分数中的一者或多者。

72.如权利要求53所述的系统，其中由所述系统执行的所述功能还包括：在图形用户界面上呈现跨所述未注释数据集的注释的不一致性。

73.如权利要求53所述的系统，其中构建所述机器学习模型包括：选择建立所述机器学习模型的算法和损失函数。

74.如权利要求73所述的系统，其中选择所述算法是基于模型类型。

75.如权利要求73所述的系统，其中由所述系统执行的所述功能还包括：

通过基于从所述未注释数据集进行注释的注释训练数据集对模型进行多次训练并测量跨运行的质量指标的分散度来测试收敛性。

76.如权利要求69所述的系统，其中所述质量指标包括学习曲线的斜率。

77.如权利要求73所述的系统，其中所述机器学习模型是使用针对给定模型类型选择的默认超参数和所述算法来训练。

78.如权利要求77所述的系统，其中所述超参数是使用随机选择、网格搜索或贝叶斯估计系统中的一者或多者来选择。

79.如权利要求77所述的系统，其中针对所述模型存储随机种子、算法选择、损失函数、超参数、数据集分割、数据集散列或类权重中的一者或多者。

80.如权利要求53所述的系统，其中对所述机器学习模型进行版本化、更改或回滚。

81.如权利要求53所述的系统，其中由所述系统执行的所述功能还包括：

通过数据漂移或概念漂移来监测模型之间的变化。

82.如权利要求81所述的系统，其中概念漂移是通过基于对所述注释数据集与所述生产注释之间的多个变化预测的量化训练模型来计算。

83.如权利要求81所述的系统，其中数据漂移是基于所述注释数据集与所述生产注释之间的语料库统计和/或语料库对比来测量。

84.如权利要求81所述的系统，其中警示是在识别出数据漂移或概念漂移时生成。

85.如权利要求81所述的系统，其中所述数据漂移或所述概念漂移包括基于随时间推移的未注释数据的指标或基于随时间推移的模型预测的指标。

86.如权利要求55所述的系统，其中共享所述模型包括：执行特征散列、加密散列或随机投影中的一者或多者。

87.如权利要求55所述的系统，其中共享所述模型包括：共享所述模型的梯度更新。

88.如权利要求87所述的系统，其中所述梯度更新被添加到计算图形中的层。

89.如权利要求55所述的系统，其中共享所述模型包括：共享一个或多个模型资产。

90.如权利要求89所述的系统，其中所述一个或多个模型资产包括基于数据集、词向量、注释集、关键字和短语列表、实例列表、语言模型、词典以及训练的模型和模型架构训练的词嵌入。

91.如权利要求89所述的系统，其中所述一个或多个模型资产被清除个人可识别信息。

92.如权利要求58所述的系统，其中所述进展包括从种子采样器到难例采样器、到分层采样器、到不确定性采样器的进展。

93.如权利要求53所述的系统，其中请求样本的注释包括：在图形用户界面上将问题呈现给用户以获得注释反馈。

94.如权利要求53所述的系统，其还包括：预测所述未注释数据的样本的一个或多个注

释。

95.如权利要求94所述的系统，其中所述一个或多个注释的所述预测在请求所述未注释数据的样本的注释之前进行。

95.如权利要求94所述的系统，其中由所述系统执行的所述功能还包括：基于采样分数将所预测的一个或多个注释存储在优先级队列中。

96.如权利要求95所述的系统，其中所述采样分数是所预测的一个或多个注释的置信度分数。

97.如权利要求95所述的系统，其中由所述系统执行的所述功能还包括：在将所预测的一个或多个注释存储在所述优先级队列中之前确定所述采样分数是否大于阈值采样分数。

98.如权利要求97所述的系统，其中由所述系统执行的所述功能还包括：丢弃所具有的采样分数被确定为小于所述阈值采样分数的预测。

99.如权利要求95所述的系统，其中所述优先级队列存储预定最大数量的预测。

100.如权利要求99所述的系统，其中由所述系统执行的所述功能还包括：在将所述预测存储在所述优先级队列中之前确定存储在所述优先级队列中的预测数量小于所述预定最大预测数量。

101.如权利要求97所述的系统，其中由所述系统执行的所述功能还包括：在将所述预测存储在所述优先级队列中之前确定所述采样分数大于所述优先级队列中的至少一个先前存储的预测。

102.如权利要求95所述的系统，其中由所述系统执行的所述功能还包括：丢弃所述优先级队列中具有最低采样分数的先前存储的预测。

103.如权利要求95所述的系统，其中请求所述未注释数据的样本的注释包括：从多个优先级队列中选择所述优先级队列。

用于快速构建、管理和共享机器学习模型的系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求2017年5月14日提交的美国临时专利申请号62/505,936的优先权和权益,所述申请以引用的方式整体并入本文。

背景技术

[0003] 常规的机器学习技术分别处理有助于数据注释、数据探索和模型创建。在用于数据注释的一些界面中,用户可突出显示他们感兴趣的文本范围,并将注释分配给突出显示的文本。替代地,用户可突出显示图像中他们感兴趣的部分,并将注释分配给图像中突出显示的部分。这些方法常常采用手动"蛮力"注释数据,并且要求用户按顺序遍历数据,从而导致生成机器学习模型的大量成本和时间延迟。此外,此类现有工具可能需要关于数据预处理、特征提取和可视化类型的广泛知识来运行。

[0004] 在一些常规方法的其他缺点、短处和不利中,他们可能遭受以下问题:他们常常需要先前所注释数据,并且在此类数据不存在时不提供起点;他们常常不针对非结构化数据;模型训练常常缓慢并且需要大量的硬件资源;他们可能无法有效地处理不平衡的数据(即,在所需结果的发生率较低的情况下的数据,例如,低于时间的10%);并且他们可能不提供集成工作流。

发明内容

[0005] 本公开涉及用于快速构建、管理和共享机器学习模型的系统和方法。本公开提供一种管理机器学习模型的生命周期的方法。在一些方面,所述方法包括:管理所述机器学习模型的生命周期可包括:接收未注释数据集;请求所述未注释数据的样本的注释以产生注释数据集;基于所述注释数据集构建机器学习模型;将所述机器学习模型部署到客户端系统,其中生成生产注释;收集所生成的生产注释并生成并入所述生产注释的新机器学习模型;以及选择基于所述注释数据集构建的所述机器学习模型或所述新机器学习模型中的一者。

[0006] 根据本公开的上述方面中的任一个,所述方法还可包括:报告所述机器学习模型的质量的一个或多个量度,所述一个或多个量度包括精度、召回率、平均精度、接收者操作员特征分数或F-β分数。

[0007] 根据本公开的上述方面中的任一个,所述方法还可包括:与第三方共享所述模型。

[0008] 根据本公开的上述方面中的任一个,所述方法还可包括:与第三方共享所述模型。

[0009] 根据本公开的上述方面中的任一个,请求样本的注释可包括:基于用户输入或自动化采样器选择来从所述未注释数据集选择样本。

[0010] 根据本公开的上述方面中的任一个,所述用户输入可包括语义搜索、类似样本的选择或所述未注释数据的可视图上的选择中的一者或多者。

[0011] 根据本公开的上述方面中的任一个,所述自动化采样器选择可来自进展中的多个采样器中的一个。

[0012] 根据本公开的上述方面中的任一个，所述多个采样器中的每一个可使用不同的采样算法。

[0013] 根据本公开的上述方面中的任一个，所述相应的采样算法可选自：密度采样算法；熵采样算法；所估计误差减小采样算法；详尽采样算法；标记预测算法；难例挖掘采样算法；高置信度采样算法；线性采样算法；图可视化采样算法；元数据搜索采样算法；最小裕度采样算法；委员会查询采样算法；随机采样算法；审查采样算法；搜索采样算法；相似性采样算法；对针对其所述输入为跳过样本类型算法的样本的采样；分层采样算法；最高置信度样本算法；或最不确定样本算法。

[0014] 根据本公开的上述方面中的任一个，所述进展可包括所述多个所述采样器中的采样器之间的连续改变。

[0015] 根据本公开的上述方面中的任一个，所述多个采样器中的每个采样器可具有预期的结果分布，所述预期的结果分布确定是移动到所述进展中的前一还是后一采样器。

[0016] 根据本公开的上述方面中的任一个，在接收到具有不正确的模型预测的预定数量的样本注释后，所述进展在采样器之间可改变为所述进展中的前一采样器。

[0017] 根据本公开的上述方面中的任一个，在接收到具有一致的模型预测的预定数量的样本注释后，所述进展在采样器之间可改变为所述进展中的后一采样器。

[0018] 根据本公开的上述方面中的任一个，构建所述机器学习模型可包括：接收共享模型，并且将中间模型的权重初始化为所述共享模型的权重并以不同的学习速率进行训练。

[0019] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：请求测试数据集的详尽注释。

[0020] 根据本公开的上述方面中的任一个，所述测试数据集的所述详尽注释可以是通过包括密度采样、水平集树或随机采样中的一者或多者的远程监督来执行。

[0021] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：在多个采样器中的采样器的图形用户界面上呈现用于从所述未注释数据集中选择样本的推荐。

[0022] 根据本公开的上述方面中的任一个，所述方法还可包括：在所述图形用户界面上呈现数据质量和数量指标。

[0023] 根据本公开的上述方面中的任一个，所述数据数量指标可包括多个训练的样本、多个正实例、多个负实例、或针对一类样本训练的多个样本中的一者或多者。

[0024] 根据本公开的上述方面中的任一个，所述数据质量指标可包括准确度、精度、召回率或F1分数中的一者或多者。

[0025] 根据本公开的上述方面中的任一个，所述方法还可包括：在图形用户界面上呈现跨所述未注释数据集的注释的不一致性。

[0026] 根据本公开的上述方面中的任一个，构建所述机器学习模型可包括：选择建立所述机器学习模型的算法和损失函数。

[0027] 根据本公开的上述方面中的任一个，选择所述算法是基于模型类型。

[0028] 根据本公开的上述方面中的任一个，所述方法还可包括：通过基于从所述未注释数据集进行注释的注释训练数据集对模型进行多次训练并测量跨运行的质量指标的分散度来测试收敛性。

[0029] 根据本公开的上述方面中的任一个，所述质量指标可包括学习曲线的斜率。

[0030] 根据本公开的上述方面中的任一个，所述模型可以是使用针对给定模型类型选择的默认超参数和所述算法来训练。

[0031] 根据本公开的上述方面中的任一个，所述超参数可以是使用随机选择、网格搜索或贝叶斯估计方法中的一者或多者来选择。

[0032] 根据本公开的上述方面中的任一个，可针对所述模型存储随机种子、算法选择、损失函数、超参数、数据集分割、数据集散列或类权重中的一者或多者。

[0033] 根据本公开的上述方面中的任一个，可对所述机器学习模型进行版本化、更改或回滚。

[0034] 根据本公开的上述方面中的任一个，所述方法还可包括：通过数据漂移或概念漂移来监测模型之间的变化。

[0035] 根据本公开的上述方面中的任一个，概念漂移可以是通过基于对所述注释数据集与所述生产注释之间的多个变化预测的量化训练模型来计算。

[0036] 根据本公开的上述方面中的任一个，数据漂移可以是基于所述注释数据集与所述生产注释之间的语料库统计和/或语料库对比来测量。

[0037] 根据本公开的上述方面中的任一个，警示可以是在识别出数据漂移或概念漂移时生成。

[0038] 根据本公开的上述方面中的任一个，所述数据漂移或所述概念漂移可包括基于随时间推移的未注释数据的指标或基于随时间推移的模型预测的指标。

[0039] 根据本公开的上述方面中的任一个，共享所述模型可包括：执行特征散列、加密散列或随机投影中的一者或多者。

[0040] 根据本公开的上述方面中的任一个，共享所述模型可包括：共享所述模型的梯度更新。

[0041] 根据本公开的上述方面中的任一个，所述梯度更新可被添加到计算图形中的层。

[0042] 根据本公开的上述方面中的任一个，共享所述模型可包括：共享一个或多个模型资产。

[0043] 根据本公开的上述方面中的任一个，所述一个或多个模型资产可包括基于数据集、词向量、注释集、关键字和短语列表、实例列表、语言模型、词典以及训练的模型和模型架构训练的词嵌入。

[0044] 根据本公开的上述方面中的任一个，所述一个或多个模型资产可被清除个人可识别信息。

[0045] 根据本公开的上述方面中的任一个，所述进展可包括从种子采样器到难例采样器、到分层采样器、到不确定性采样器的进展。

[0046] 根据本公开的上述方面中的任一个，请求样本的注释可包括：在图形用户界面上将问题呈现给用户以获得注释反馈。

[0047] 根据本公开的上述方面中的任一个，所述方法还可包括：预测所述未注释数据的样本的一个或多个注释。

[0048] 根据本公开的上述方面中的任一个，所述一个或多个注释的所述预测可在请求所述未注释数据的样本的注释之前进行。

[0049] 根据本公开的上述方面中的任一个，所述方法还可包括：基于采样分数将所预测的一个或多个注释存储在优先级队列中。

[0050] 根据本公开的上述方面中的任一个，所述采样分数可以是所预测的一个或多个注释的置信度分数。

[0051] 根据本公开的上述方面中的任一个，在将所述预测的一个或多个注释存储在所述优先级队列中之前，可根据所述方法确定所述采样分数是否大于阈值采样分数。

[0052] 根据本公开的上述方面中的任一个，所述方法还可包括：丢弃所具有的采样分数被确定为小于所述阈值采样分数的预测。

[0053] 根据本公开的上述方面中的任一个，所述优先级队列可存储预定最大数量的预测。

[0054] 根据本公开的上述方面中的任一个，所述方法还可包括：在将所述预测存储在所述优先级队列中之前确定存储在所述优先级队列中的预测数量小于所述预定最大预测数量。

[0055] 根据本公开的上述方面中的任一个，所述方法还可包括：在将所述预测存储在所述优先级队列中之前确定所述采样分数大于所述优先级队列中的至少一个先前存储的预测。

[0056] 根据本公开的上述方面中的任一个，所述方法还可包括：丢弃所述优先级队列中具有最低采样分数的先前存储的预测。

[0057] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：从多个优先级队列中选择所述优先级队列。

[0058] 本公开还提供一种用于管理机器学习模型的生命周期的系统。在一些方面，所述系统包括处理器；以及非暂时性存储器装置，所述非暂时性存储器装置联接到所述处理器并存储计算机可读指令，所述计算机可读指令在由所述处理器执行时致使所述系统执行包括以下的功能：接收未注释数据集；请求所述未注释数据的样本的注释以产生注释数据集；基于所述注释数据集构建机器学习模型；将所述机器学习模型部署到客户端系统，其中生成生产注释；

[0059] 收集所生成的生产注释并生成并入所述生产注释的新机器学习模型；以及选择基于所述注释数据集构建的所述机器学习模型或所述新机器学习模型中的一者。

[0060] 根据本公开的上述方面中的任一个，其中由所述系统执行的所述功能还可包括：报告所述机器学习模型的质量的一个或多个量度，所述一个或多个量度包括精度、召回率、平均精度、接收者操作员特征分数或F-β分数。

[0061] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：与第三方共享所述模型。

[0062] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：与第三方共享所述模型。

[0063] 根据本公开的上述方面中的任一个，请求样本的注释可包括：基于用户输入或自动化采样器选择来从所述未注释数据集选择样本。

[0064] 根据本公开的上述方面中的任一个，所述用户输入可包括语义搜索、类似样本的选择或所述未注释数据的可视图上的选择中的一者或多者。

[0065] 根据本公开的上述方面中的任一个，所述自动化采样器选择可来自进展中的多个采样器中的一个。

[0066] 根据本公开的上述方面中的任一个，所述多个采样器中的每一个可使用不同的采样算法。

[0067] 根据本公开的上述方面中的任一个，所述相应的采样算法可选自：密度采样算法；熵采样算法；所估计误差减小采样算法；详尽采样算法；标记预测算法；难例挖掘采样算法；高置信度采样算法；线性采样算法；图可视化采样算法；元数据搜索采样算法；最小裕度采样算法；委员会查询采样算法；随机采样算法；审查采样算法；搜索采样算法；相似性采样算法；对针对其所述输入为跳过样本类型算法的样本的采样；分层采样算法；最高置信度样本算法；或最不确定样本算法。

[0068] 根据本公开的上述方面中的任一个，所述进展可包括所述多个所述采样器中的采样器之间的连续改变。

[0069] 根据本公开的上述方面中的任一个，所述多个采样器中的每个采样器可具有预期的结果分布，所述预期的结果分布确定是移动到所述进展中的前一还是后一采样器。

[0070] 根据本公开的上述方面中的任一个，在接收到具有不正确的模型预测的预定数量的样本注释后，所述进展在采样器之间可改变为所述进展中的前一采样器。

[0071] 根据本公开的上述方面中的任一个，在接收到具有一致的模型预测的预定数量的样本注释后，所述进展在采样器之间可改变为所述进展中的后一采样器。

[0072] 根据本公开的上述方面中的任一个，构建所述机器学习模型可包括：接收共享模型，并且将中间模型的权重初始化为所述共享模型的权重并以不同的学习速率进行训练。

[0073] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：请求测试数据集的详尽注释。

[0074] 根据本公开的上述方面中的任一个，所述测试数据集的所述详尽注释可以是通过包括密度采样、水平集树或随机采样中的一者或多者的远程监督来执行。

[0075] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：在多个采样器中的采样器的图形用户界面上呈现用于从所述未注释数据集中选择样本的推荐。

[0076] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：在所述图形用户界面上呈现数据质量和数量指标。

[0077] 根据本公开的上述方面中的任一个，所述数据数量指标可包括多个训练的样本、多个正实例、多个负实例、或针对一类样本训练的多个样本中的一者或多者。

[0078] 根据本公开的上述方面中的任一个，所述数据质量指标可包括准确度、精度、召回率或F1分数中的一者或多者。

[0079] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：在图形用户界面上呈现跨所述未注释数据集的注释的不一致性。

[0080] 根据本公开的上述方面中的任一个，构建所述机器学习模型可包括：选择建立所述机器学习模型的算法和损失函数。

[0081] 根据本公开的上述方面中的任一个，选择所述算法是基于模型类型。

[0082] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：通过

基于从所述未注释数据集进行注释的注释训练数据集对模型进行多次训练并测量跨运行的质量指标的分散度来测试收敛性。

[0083] 根据本公开的上述方面中的任一个，所述质量指标可包括学习曲线的斜率。

[0084] 根据本公开的上述方面中的任一个，所述模型可以是使用针对给定模型类型选择的默认超参数和所述算法来训练。

[0085] 根据本公开的上述方面中的任一个，所述超参数可以是使用随机选择、网格搜索或贝叶斯估计方法中的一者或多者来选择。

[0086] 根据本公开的上述方面中的任一个，可针对所述模型存储随机种子、算法选择、损失函数、超参数、数据集分割、数据集散列或类权重中的一者或多者。

[0087] 根据本公开的上述方面中的任一个，可对所述机器学习模型进行版本化、更改或回滚。

[0088] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：通过数据漂移或概念漂移来监测模型之间的变化。

[0089] 根据本公开的上述方面中的任一个，概念漂移可以是通过基于对所述注释数据集与所述生产注释之间的多个变化预测的量化训练模型来计算。

[0090] 根据本公开的上述方面中的任一个，数据漂移可以是基于所述注释数据集与所述生产注释之间的语料库统计和/或语料库对比来测量。

[0091] 根据本公开的上述方面中的任一个，警示可以是在识别出数据漂移或概念漂移时生成。

[0092] 根据本公开的上述方面中的任一个，所述数据漂移或所述概念漂移可包括基于随时间推移的未注释数据的指标或基于随时间推移的模型预测的指标。

[0093] 根据本公开的上述方面中的任一个，共享所述模型可包括：执行特征散列、加密散列或随机投影中的一者或多者。

[0094] 根据本公开的上述方面中的任一个，共享所述模型可包括：共享所述模型的梯度更新。

[0095] 根据本公开的上述方面中的任一个，所述梯度更新可被添加到计算图形中的层。

[0096] 根据本公开的上述方面中的任一个，共享所述模型可包括：共享一个或多个模型资产。

[0097] 根据本公开的上述方面中的任一个，所述一个或多个模型资产可包括基于数据集、词向量、注释集、关键字和短语列表、实例列表、语言模型、词典以及训练的模型和模型架构训练的词嵌入。

[0098] 根据本公开的上述方面中的任一个，所述一个或多个模型资产可被清除个人可识别信息。

[0099] 根据本公开的上述方面中的任一个，所述进展可包括从种子采样器到难例采样器、到分层采样器、到不确定性采样器的进展。

[0100] 根据本公开的上述方面中的任一个，请求样本的注释可包括：在图形用户界面上将问题呈现给用户以获得注释反馈。

[0101] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：预测所述未注释数据的样本的一个或多个注释。

[0102] 根据本公开的上述方面中的任一个，所述一个或多个注释的所述预测可在请求所述未注释数据的样本的注释之前进行。

[0103] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：基于采样分数将所预测的一个或多个注释存储在优先级队列中。

[0104] 根据本公开的上述方面中的任一个，所述采样分数可以是所预测的一个或多个注释的置信度分数。

[0105] 根据本公开的上述方面中的任一个，在将所述预测的一个或多个注释存储在所述优先级队列中之前，可根据所述方法确定所述采样分数是否大于阈值采样分数。

[0106] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：丢弃所具有的采样分数被确定为小于所述阈值采样分数的预测。

[0107] 根据本公开的上述方面中的任一个，所述优先级队列可存储预定最大数量的预测。

[0108] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：在将所述预测存储在所述优先级队列中之前确定存储在所述优先级队列中的预测数量小于所述预定最大预测数量。

[0109] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：在将所述预测存储在所述优先级队列中之前确定所述采样分数大于所述优先级队列中的至少一个先前存储的预测。

[0110] 根据本公开的上述方面中的任一个，由所述系统执行的所述功能还可包括：丢弃所述优先级队列中具有最低采样分数的先前存储的预测。

[0111] 根据本公开的上述方面中的任一个，请求所述未注释数据的样本的注释可包括：从多个优先级队列中选择所述优先级队列。

[0112] 这些和其他特征将从结合附图和权利要求书进行的以下具体实施方式而得以更清楚了解。

附图说明

[0113] 为了更完整的理解本公开，现参考结合附图和详细描述的以下简要描述，其中相似的附图标号表示相似的部分。未必按比例绘制的附图示出本公开的若干实施方案，并且与描述一起用于根据实施方案解释本公开技术的原理。

[0114] 图1示出根据本公开技术的示例性实施方案的用于创建机器学习模型的信息堆栈。

[0115] 图2示出根据本公开技术的示例性实施方案的有助于数据注释和机器学习模型创建的计算机架构。

[0116] 图3示出根据本公开技术的示例性实施方案的用于数据注释和模型构建的集成工作流过程的流程图。

[0117] 图4示出根据本公开技术的示例性实施方案的用于注释训练数据集的注释过程的流程图。

[0118] 图5示出根据本公开技术的示例性实施方案的有助于对未注释数据进行注释的采样技术的进展的序列图。

[0119] 图6示出根据本公开技术的示例性实施方案的用于注释数据的计算机架构的框图。

[0120] 图7示出根据本公开技术的示例性实施方案的加速预测并降低硬件需求所采用的优先级队列方法的框图。

[0121] 图8是根据本公开技术的示例性实施方案的用于注释测试数据集的注释过程的流程图。

[0122] 图9示出根据本公开技术的示例性实施方案的用于开始新机器学习模型创建的示例性图形用户界面。

[0123] 图10示出根据本公开技术的示例性实施方案的描绘用户能够管理多个数据集的方式的示例性图形用户界面。

[0124] 图11示出根据本公开技术的示例性实施方案的描绘用户能够管理多个注释集的方式的示例性图形用户界面。

[0125] 图12示出根据本公开技术的示例性实施方案的各种注释集的实例。

[0126] 图13示出根据本公开技术的示例性实施方案的描绘用户能够使用来建立注释的配置选项的示例性图形用户界面。

[0127] 图14示出根据本公开技术的示例性实施方案的描绘输入并管理关键字和短语列表的方法的示例性图形用户界面。

[0128] 图15示出根据本公开技术的示例性实施方案的描绘用户能够发现并管理相关词和短语的方式的示例性图形用户界面。

[0129] 图16示出根据本公开技术的示例性实施方案的描绘将本体并入单词列表管理中的示例性图形用户界面。

[0130] 图17示出根据本公开技术的示例性实施方案的描绘输入并管理实例列表的方法的示例性图形用户界面。

[0131] 图18示出根据本公开技术的示例性实施方案的描绘注释过程、用于管理注释过程的工具和关于进展的反馈的示例性图形用户界面。

[0132] 图19示出根据本公开技术的示例性实施方案的描绘允许用户对其响应的强度进行评分的注释过程的示例性图形用户界面。

[0133] 图20示出根据本公开技术的示例性实施方案的描绘对相邻条目进行注释的能力以及对注释的彩色反馈的示例性图形用户界面。

[0134] 图21示出根据本公开技术的示例性实施方案的描绘示出多种类型的候选采样的下拉菜单的示例性图形用户界面。

[0135] 图22示出根据本公开技术的示例性实施方案的描绘跨数据的一次性关键字搜索的示例性图形用户界面。

[0136] 图23示出根据本公开技术的示例性实施方案的描绘能够允许用户可视地探索其数据的数据图的示例性图形用户界面。

[0137] 图24示出根据本公开技术的示例性实施方案的描绘能够如何处理失效状态和如何将信息传递回用户的示例性图形用户界面。

[0138] 图25示出根据本公开技术的示例性实施方案的描绘先前所注释条目的列表和如何管理那些条目的示例性图形用户界面。

[0139] 图26示出根据本公开技术的示例性实施方案的示例性计算机系统。

具体实施方式

[0140] 在开始时应当理解，虽然下面示出一个或多个实施方案的例示性实现方式，但是可使用无论是当前已知的还是现有的任何数量的技术来实现公开的系统和方法。本公开绝不应限于下面所示的说明性实现方式、附图和技术，但是可在所附权利要求的范围内以及其等同物的全部范围内进行修改。

[0141] 创建机器学习模型可以是复杂且耗时的任务。常规地，这涉及到数据的聚合、预处理、注释和检查、特征提取和向量化以及模型训练和评估。因此，创建此类模型的能力常常受数据科学专家的限制。在本公开技术的实施方案所提供的其他优点和益处中，用户在没有广泛数据科学知识的情况下可创建强大的模型，同时还使得数据科学家能够更快地执行他们的工作。

[0142] 根据本公开技术的各种实施方案，用户可连接适当的数据源、建立注释过程、注释数据、根据那些注释构建机器学习模型、部署机器学习模型、收集生产反馈并将其并入模型的新版本中，并且共享模型和所学到的。

[0143] 图1示出根据本公开技术的示例性实施方案的用于创建机器学习模型的信息堆栈100。信息堆栈100包括未注释数据102、注释数据104、分析106和模型108。未注释数据102包括来自数据源的未处理数据。例如，未注释数据102可包括电子邮件通信、聊天记录、文档存储或其他来源的文本数据的集合。文本数据可来自纯文本文件，诸如来自通过电子邮件或聊天进行的电子通信、平面文件或其他类型的文档文件(例如，.pdf、.doc等)。未注释数据102还可包括图像库、视频库或其他来源的图像或视频数据。未注释数据102还可包括电话呼叫、播客和其他来源的音频数据。未注释数据102可从预先存在的数据存储库中提供，或者还包括任何所需格式的未注释数据直播流。在一些实现方式中，未注释数据102可包括文件目录并且可包括数据的图形格式。可使用其他电子数据来源。

[0144] 创建新机器学习模型108的瓶颈是将未注释数据注释成注释数据104。注释数据104可包括与注释或由用户例如通过本文所述的应用提供的注释相结合的一个或多个数据集。与使用数据的科学专家不同，领域的主题专家可参与本文所述的注释过程，以有助于他们的知识转移并且提高机器学习模型创建过程的速度并降低其成本。以语言不可知和领域不可知的方式执行本公开的注释过程。

[0145] 可执行分析106以确保在创建模型108之前已有充分注释。

[0146] 图2示出根据本公开技术的示例性实施方案的有助于数据注释和机器学习模型创建的计算机架构200。计算机架构200包括执行本文所述的注释过程的注释服务器202。注释服务器202与被配置为存储其中的信息堆栈100的数据库204通信。在显示为单个数据库时，一个或多个数据库可用于信息堆栈100的每个元素。注释服务器202可通过网络208从注释客户端206接收未注释数据102以存储在数据库204中。注释服务器202通过一个或多个图形用户界面与注释客户端206进行交互以有助于注释数据104的生成。如一个或多个注释训练准则(例如，每类20个注释)所指定的，在对未注释数据102进行充分注释后，注释服务器202被配置为生成一个或多个中间模型。

[0147] 这些中间模型生成对未注释数据的预测，所述未注释数据可通过网络208传达给

注释客户端206或另一个客户端计算机(未示出),以有助于生产注释。在客户端计算机206上的正常生产操作期间,生成另外的生产注释数据并将其存储在生产注释数据库210中。例如,当在客户端计算机206上输入或操纵新数据时,基线模型呈现对被接受或修改的新数据的注释的预测,以生成附加生产注释数据。生产注释周期性地反馈到注释服务器202并且用于生成考虑到附加生产注释数据的所更新模型。可通过导入具有生产注释的文件或通过暴露在注释服务器202上的标准API来将生产注释反馈到注释服务器202。可对API进行速率限制以防止攻击。

[0148] 图3示出根据本公开技术的示例性实施方案的用于数据注释和模型构建的集成工作流过程300的流程图。在302处,将未注释数据102导入注释服务器202以存储在数据库204上。

[0149] 在304处,对未注释数据102进行探索性数据分析。这使得能够将数据恰当地分层到子集中以进行注释。例如,对于文本,探索性数据分析可识别外语分布(使用诸如逻辑回归的方法)、文档类型分布(电子邮件、聊天、可移植文档格式文档、超文本标记语言等等)以及Flesch-Kincaid可读性分数的分布。对于图像数据,探索性数据分析可识别彩色图像与黑白图像的分布、图像的大小和分辨率以及图像中的熵分布。这些分布用于选择所分层子集以进行注释。例如,用户可选择对2018年4月2日至2018年4月7日的一周注释西班牙语聊天消息。

[0150] 在307处,使用预限定或用户指定的清洗管线对未注释数据102进行预处理。这是降维的形式,所述降维将数据规范化以进行分析以及将其分割成感兴趣的区域。例如,文本的预处理可包括执行诸如去除免责声明、无意义文本或电子邮件中的内联回复之类的任务。对于文本数据,这还可包括标记文本并将其拆分成句子、段落或文档、转换为小写字符、可在标点符号之间插入空格,以及可为类似日语的非空白语言插入空白。

[0151] 在306处,选择预处理数据的子集以提供在308处的训练候选集和在314处的测试数据集。在310处,通过下面结合图4至图7详细描述的引导注释过程来创建注释训练集。引导注释过程允许主题专家在短时间内以降低的成本和减少的计算机资源来注释大量的训练候选集。鉴于依赖于数据分析师对数据集的“蛮力”注释的现有注释方法通常需要人-年以对不平衡的数据集进行充分注释,本文所公开的引导注释过程可有助于在人-小时或人-天期间对数据集进行充分注释。

[0152] 在316处,如下文结合图8所详述,对为测试集创建而保留的数据进行注释以产生注释测试集,用于显式地或通过使用远程监控的代理来进行测试。在一些实现方式中,对测试数据集进行详尽注释。在一些情况下,主动学习产生不适合创建无偏测试集的偏置数据分布。相反,水平集树、利用随机采样的无监督聚类和基于密度的采样有助于对测试数据集进行充分且有效的注释。

[0153] 在312和318处,对注释训练集和注释测试集执行数据审查。数据审查包括注释“清洗”,其识别出跨多个审查者的注释之间的不一致性,即使底层样本在语义上相似但不相同。还可在用户内部(疲劳或判断力差)或跨用户检查注释一致性。注释一致性可使用用户注释和所分布表示上的相似性量度(例如,关于向量嵌入的余弦相似性)来测量。在此审查期间,监督员可建立“黄金标准”注释。在一些实现方式中,注释“清洗”可如Gardner等人共同拥有的美国9,058,317“Systemand Method for Machine Learning Management”所述来

执行，并且所述申请在此以引用方式整体并入本文。

[0154] 在320处，使用已清洗的注释训练集和注释测试集来构建机器学习模型。在一些情况下，可供应共享模型322以通知模型构建320。当提供共享模型322时，将正在构建的模型初始化为共享模型322的权重并以不同的学习速率进行训练。在一些实现方式中，使用逐渐降低的学习速率来对正在构建的模型进行训练。在一些实现方式中，由共享模型322提供的特定权重可保持未训练或轻度训练的。如果正在构建的模型具有未训练或轻度训练的权重，则可以选择性地保持高的学习速率以快速训练那些权重。

[0155] 在324处，作为模型构建320的结果，生成竞争者模型。在328处，可呈现关于所生成模型的报告。在一些实现方式中，可构建多个模型并使用共同质量量度来与注释测试集进行比较。例如，质量量度可包括精度、召回率、平均精度、接收者操作员特征分数和F-β分数。可使用其他质量量度。可通过报告328将在模型同意以及不同意的情况下的预测实例呈现给用户。可针对每个模型在不同阈值下的精度召回率曲线、ROC曲线和真/假阳性/阴性样本提供另外的可视化以便有助于模型选择。

[0156] 用户可在任何时候认为已收集足够的训练数据并准备继续进行模型构建。在模型构建期间，注释服务器202将以自动化方式引导用户完成一系列步骤。在一些实施方案中，用户将指定某些注释集用于训练机器学习，并且指定其他注释集用于测试机器学习模型的质量。在一些实施方案中，注释服务器202将针对给定概念的所有可用注释数据划分成训练数据集和测试数据集。

[0157] 给定训练数据及测试数据和模型类型(例如文本分类器、图像分类器、语义角色标记)，注释服务器202选择用于建立基线的适当的算法和损失函数。在大多数情况下，已针对模型类型和训练数据量预定具体算法。例如，可选择具有二元特征的逻辑回归作为文本分类的基线算法，而可选择具有光谱特征的隐马尔可夫模型作为自动语音识别的基线算法。除基线之外，每种模型类型具有由注释服务器202预定的适用算法的相关联列表。

[0158] 当已选择算法和损失函数时，注释服务器202测试收敛性、评估另外的训练数据的益处并建立基线模型。可通过基于训练数据对模型进行多次训练、基于测试数据测量质量指标并跨运行测量质量指标的分散性来测试收敛性，其中分散性是通过标准偏差来计算。通过学习曲线来评估另外的训练数据的益处并将其呈现回给用户以供反馈。下文更详细地描述使用学习曲线的评估。最后，“基线模型”是使用针对给定模型类型选择的默认超参数和算法来训练。多个指标是使用基线模型预测和测试集的参考注释来计算。这些指标与问题的类型相关，但是可包括下文更详细地描述的数据质量指标、数据数量指标和模型质量指标。

[0159] 在一些实施方案中，基于验证集运行指标。在其他实施方案中，不存在验证集并且训练数据通过典型的交叉验证方法用于训练和验证。

[0160] 如在基线选择过程中一样，注释服务器202使用模型类型来选择适当的搜索空间。搜索空间由一系列算法、其相关联损失函数和用于调整算法的潜在超参数组成。在单次超参数优化运行期间，选择算法和样本超参数、对模型进行训练并计算指标。

[0161] 算法和候选超参数选择是使用以下任意种方法来执行：随机选择、网格搜索或贝叶斯估计方法(例如Parzen估计器树)。在每次模型训练运行中，将重新创建实验所必需的参数和实验结果存储在数据库中。这些参数可包括随机种子、算法选择、损失函数、超参数、

数据集分割、数据集散列(例如,跨数据集确定是否发生任何变化的量度)和类权重。存储结果可包括基线以及在超参数优化期间执行的迭代。

[0162] 超参数评估在质量目标实现、质量变化变小时或在计算预算耗尽时停止。在一些实施方案中,向用户呈现所有算法和超参数运行的结果的图形列表,用户可从所述图形列表中选择模型。在其他实施方案中,自动选择最佳模型以使目标函数最大化或最小化。例如,在文本分类中,这可以是使接收者操作特征曲线下方面积最大化的模型。

[0163] 在一些实现方式中,在生成初始竞争者模型324时,模型可被认为是冠军并部署在330处。可从外部系统提供新注释,诸如注释客户端206或另一个客户端计算机(未示出)。例如,假设存在合规性监控系统,其中合规官员的日常活动是标记对公司具有潜在风险的消息。这些所标记消息是生产注释,其可以反馈回到注释服务器202以补充存储在数据库204中的注释训练集并且用于在324处生成新竞争者模型。基线模型或初始竞争者模型324可被认为是冠军模型326。如上所述,报告328可包括冠军模型326与新构建的竞争者模型324之间的对比,以便有助于在330处选择要部署的模型中的一个。

[0164] 当向注释服务器202供应新生产注释时,通过数据漂移和概念漂移计算来监控随后生成的模型的变化。例如,概念漂移可通过基于注释训练集的新版本和旧版本对模型进行训练并量化基于旧数据集和新数据集的多个变化预测来计算。数据漂移可基于语料库统计和/或注释训练集的新版本与旧版本之间的语料库对比来测量。例如,对于文本数据,语料库统计数据可包括文档固定百分比;HTML标记百分比;与参考词汇(例如,聊天词汇、标准词汇)相比的词汇表外的单词百分比;具有混合字母和/或数字的单词百分比;词类百分比;标点符号、字母(英文、西里尔字母等)、数字和/或其他文本符号的百分比;大写、小写、大小写和/或其他格式的单词的百分比;每个单词、句子、段落和/或文档中字符、单词、句子和/或段落的数量;每行新行中字符和/或单词分布;重复句子的分布;每封电子邮件或其他文档中句子数量的分布;形式;最常用的单词和二元词;和/或可读性分数。除此之外或替代地,语料库统计可包括基于随时间推移的未注释数据的指标或基于随时间推移的模型预测的指标。语料库对比包括基于以上语料库统计、斯皮尔曼等级相关系数和/或困惑度中的任一者或组合的对比。

[0165] 通过在客户端计算机206上显示的消息或屏幕来向用户警示其数据中的此类漂移。此外,基于通过客户端计算机206供应的用户输入,可根据需要对模型进行版本化、更改和回滚。

[0166] 在332处,除使得大型数据集能够进行快速探索和注释以及对应的模型创建之外,某些实现方式还使得能够购买、出售、共享和分配所生成的模型和/或所生成的模型资产。这些模型资产包括但不限于:基于数据集、词向量、注释集、关键字和短语列表、实例列表、语言模型、词典以及训练的模型和模型架构训练的词嵌入。在一些实现方式中,在对新模型进行之前,注释被“清除”个人可识别信息。可对特征进行安全散列以防止发现任何最初的原始特征。同态加密可用于简单模型。

[0167] 在一些实现方式中,这些模型的“学习”是在不共享模型本身的情况下发布或以其他方式共享的。例如,当“发布者”对底层模型进行调整时,将模型的梯度更新提交给托管的外部服务器,所述外部服务器将这些梯度更新重新分配给“订阅者”。“订阅者”可使用梯度更新来对他们的本地模型进行进一步训练。在一些实施方案中,梯度更新可以是加密的。在

一些实现方式中，梯度更新被添加到计算图形中的层。当对局部模型进行时，梯度更新可乘以局部学习速率。替代地，梯度更新可共享(X,y)，其中X是输入数据点，即输入数据或匿名数据的语义表示。

[0168] 图4示出根据本公开技术的示例性实施方案的用于注释训练候选集的注释过程400的流程图。注释过程400可在310处的注释训练集的创建期间发生，并且由注释服务器202或在注释客户端206上的本地安装上执行。

[0169] 在402处，接收未注释的训练候选集。训练候选集中的每个数据元素被称为未注释数据102的样本。例如，对于文本，样本包括预处理的标记化文本(例如，n元语法、句子、段落等)。在404处，由预测训练候选集或其子集中的样本的注释的模型406生成预测集。可以成批的预测流传输(例如，一次确定一个)或提供预测集中的预测。还可针对训练候选集中的未注释样本的一个或多个聚类中的样本进行预测集中的预测。可在未注释数据102的预处理期间识别所述聚类。模型406还针对每个预测提供预测向量分数。例如，对于分类任务，模型406可使用二元分类器算法或多类分类器算法来生成预测集。下文参考图9更详细地描述可使用的模型实例。在一些实现方式中，所述模型是具有线性分类器和可训练单词嵌入的连续词包模型。在其他实施方案中，所述模型可以是具有可训练或固定单词嵌入的深度学习模型(诸如卷积或递归神经网络)。本公开设想除文本以外的其他类型的数据模型。

[0170] 在408处，基于预测的预测向量评估预测集，并且确定是否请求样本中的一个或多个的注释。为了有助于通过注释过程对模型进行快速且有针对性的训练，通过根据多个采样算法中的一个对预测集进行采样并将所采样预测集中的每个样本以采样分数的顺序在队列中排列来生成采样预测集。采样分数可等于置信度分数或可从预测向量导出，以表示预测与采样算法的匹配如何。用于生成所采样预测集和其中排列采样预测集的队列的采样算法称为采样器或“实例候选生成引擎”。然后，可请求由采样器提供的样本的注释。

[0171] 采样算法包括基于以下的采样：密度采样、熵采样(例如，识别具有最高香农熵水平的预测)、所估计误差减小采样、详尽采样(例如，线性级数)、标记采样(例如，针对其提供用户输入以标记预测以供以后分析的预测)、难例挖掘采样、高置信度采样(例如，具有最高置信度分数的预测)、线性采样、图可视化采样(例如，下文结合图23更详细地描述的基于数据图接收到的用户输入)、元数据搜索采样、最小裕度采样、委员会查询采样、随机采样、审查采样、搜索采样(例如，下文结合图13至图16和图22更详细地描述的搜索参数和/或关键字的用户输入)、相似性采样、跳过采样(例如，对针对其用户输入为跳过对预测进行注释的预测)、分层采样、最不确定采样(例如，具有最低置信度分数的预测)。此采样算法列表并不旨在成为详尽列表，还可使用其他采样算法。

[0172] 在410处，可将采样器改变为使用不同采样算法的不同采样器。可基于接收改变采样器的用户选择或基于改变采样器的算法确定来改变采样器。如下文参考图6至图7更详细地描述的，改变采样器不需要重新训练模型。

[0173] 在412处，确定用户是否跳过对样本进行注释。如果是，则过程400循环以评估要请求注释的其他未注释数据402。反之，在414处，基于用户反馈对样本进行注释。在416处，基于注释样本更新模型。模型可以流传输方式更新，使得在每个新注释之后执行更新。替代地，模型可以成批的方式，诸如在预定数量的注释等等之后更新。作为另一种替代形式，模型可在接收到用于更新模型的用户输入时更新。作为另一种替代形式，模型可基于算法确

定更新，诸如基于周期性、基于跟踪多个正确预测，或加强学习。

[0174] 注释过程400可继续，直到所更新模型416满足停止准则。停止准则可提供有助于人类判断模型质量的信息。例如，可针对数据质量指标和数据数量指标的详尽注释的测试数据集来评估所更新模型416。数据质量指标可包括一致性指标。例如，对于多类分类算法，聚类一致性指标是基于基尼系数计数或最大熵的比例百分比来生成。

[0175] 数据数量指标可包括学习曲线指标或模型收敛性指标。例如，学习曲线指标可在多个预定的数据注释级别中的每一个(例如，在注释5％、10％、20％、50％、75％、100％数据时的每一个)测量针对测试数据集的所更新模型的迭代的预测的准确度。学习曲线的斜率是所更新模型正在学习的另外信息量的量度。如果学习曲线变平，则所更新模型的每个另外的迭代学习减少量的另外信息。因此，用于终止注释过程400的停止准则可以是在学习曲线的斜率低于预定阈值学习速率时。模型收敛性指标可以是跨运行、跨交叉验证折叠和/或跨交叉验证平均值指标的标准偏差。可使用注释过程400的其他停止准则。

[0176] 图5示出根据本公开技术的示例性实施方案的有助于训练候选集的注释的采样技术的算法采样进展500的序列图。一般来讲，对于所选采样器，如果模型以高置信度识别样本并通过注释客户端206上的注释输入接收确认反馈，则采样器进展500将继续将采样器改变为沿进展500进一步向下的采样器。也就是说，每个采样器具有预期的结果分布，所述预期的结果分布确定是移动到进展中的前一还是后一采样器。例如，如果所选采样器是难例采样器504，并且由用户提供的注释与模型预测一致，则可将采样器改变为分层采样器506。

[0177] 同样，如果由用户提供的注释与模型预测不同，则可将采样器改变为沿进展500位于更高处的采样器。也就是说，在接收到具有不正确的模型预测的预定数量的样本注释时，选择进展中的先前的采样器。例如，如果分层采样器506不能正确地提供预测，则可将采样器改变为难例采样器504。在一些实现方式中，进展500选择不同的采样器以使多个“意外”注释最大化或以其他方式加强使学习曲线(例如，在学习曲线上保持尽可能陡峭的斜率)最大化。

[0178] 最初，所选采样器是种子采样器502。种子采样器502基于用户所提供的输入来识别样本。如下文参考图13至图16更详细地描述的，输入可包括用户所输入的用户感兴趣分类的关键字、短语和/或实例。此外，可导入词典、本体或其他数据类型的其他此类数据库以补充并扩展用户所提供的输入。输入还可包括共享模型资产的输入，诸如上述共享模型资产。如下文所详述，种子采样器502还允许用户主动搜索训练候选集内的样本。在基于文本的注释的实例中，向种子采样器402提供的关键字和短语列表用于最初查找用户正在寻找的内容的实例，从而提供解决不平衡数据(例如，相较于数据集中样本的数量存在少量代表性样本的数据)问题的方法。

[0179] 在一些实施方案中，“种子”采样部分地通过预训练模型来完成。这减少了用户查找代表性种子实例的需要，并且使得能够更快地进展到难例采样器。在这种实施方案中，使用先前讨论的递增学习速率来递增地训练中间模型。

[0180] 进展500从难例采样器504前进到种子采样器/从种子采样器前进到难例采样器504。难例采样器504使用尝试识别“意外”注释的难例挖掘采样算法。也就是说，难例挖掘算法搜索这样的样本，其中模型具有对注释的具有高置信度分数的预测，但从用户接收注释不正确的注释(例如，通过分配不同的注释)。

[0181] 进展500从分层采样器506前进到难例采样器504/从难例采样器504前进到分层采样器506。分层采样器506使用分层采样算法。分层采样算法识别给定结果的分数在两个浮点数[A,B]之间的样本。

[0182] 进展500从不确定性采样器508前进到分层采样器504/从分层采样器504前进到不确定性采样器508。不确定性采样器508使用最大熵算法、最小裕度算法、委员会查询算法或其他此类不确定性采样算法中的一者或多者。不确定性采样器508特别有助于对不平衡数据集中唯一的或罕见的或以其他方式概率不相等的样本进行注释。

[0183] 如上面所提到的，可由用户从预先提供的采样器列表中手动选择采样器，每个采样器有其自己的采样算法。采样算法包括但不限于：信息量最大(最大熵)、最小裕度、具体类的随机样本、基于关键字的样本、随机样本或数据的线性进展。诸如“最大熵”的方法可有效地识别存在低置信度的预测以征求反馈。

[0184] 例如，在一些实现方式中，鼓励用户使用如图21所示的“自动采样”。如上面所讨论的，通过采样进展500的自动采样可使响应于用户所接收到的新注释而获得的信息值最大化。具体地，进展500最初可使用关键字和短语(以在不平衡数据中找到正例)，然后切换到更进一步的方法。进展500可响应于用户所提供的注释是否是“意外”(即，注释与模型的预测不同)而调整为一种或若干种可能的采样方法。例如，假设模型具有与预测相关联的高置信度分数。在人类注释者同意模型的预测的情况下，进展500可自动切换为提供存在更多不确定性的样本的采样器。然而，在人类注释者不同意模型的预测的情况下，进展500可继续示出其他“高置信度”样本，以便使预期的信息增益最大化。

[0185] 在一些实施方案中，为了保持注释质量，用户可标记不确定样本以供以后审查。在其他实施方案中，用户可指定待“存储”的样本——这将通过在哈希表中查找来覆写这些样本的机器学习模型。

[0186] 在提供关键字和短语的列表或提供其他此类输入以为模型播种的情况下，进展500可确保输入数据的适当“覆盖”。例如，给定十个关键字的列表，采样器可跟踪显示给用户的每个关键字的样本数。在确定特定关键字相对于其他关键字已“欠采样”的情况下，进展500可选择对此条目进行过采样，直到不平衡被修正为止。此方法改进相关联的学习模型的召回率。

[0187] 如果用户认为提供的“上下文”不足，则其可请求另外的上下文。例如，如果句子被认为是模棱两可的，则用户可请求看所述句子的前后。在这种情况下，将记录两个注释：需要更多上下文以及注释。

[0188] 用户可“提示”或以其他方式将训练实例的区域手动指定为最相关的。例如，这实现基于方面的情感分析。其他此类用户引导的采样方法包括“类似采样”和“数据图”。这些采样方法中的每一个利用以下表示：所述表示使用基于未注释数据102或训练候选集的无监督学习技术来发现。如果用户请求具体样本的“类似样本”，则采样器可使用通过无监督学习技术学习的信息来尝试查找接近的实例。类似地，在用户对具体术语或短语执行关键字搜索时，采样器可使用通过无监督学习技术学习的此信息来尝试查找具有关键字及其同义词的实例。在一些实施方案中，使用连续的词包模型计算句子向量，并且使用余弦距离计算附近的句子。

[0189] 非结构化表示还可用于使用诸如t-sne或PCA的技术来将数据维数降到二维或三

维。如图23所示，这些低维表示可以在视觉上呈现为“图”，用户可通过所述“图”导航自己的数据，并且找到具体实例。在一个此实施方案中，样本被表示为散点图，并且使用表示用户注释的颜色来呈现先前注释的样本。“图”表示可以使得用户能够在视觉上看到“未导航地区”，以及可能发生错误预测的区域。

[0190] 图6示出根据本公开技术的示例性实施方案的用于注释数据的计算机架构600的框图。如下所述，计算机架构600利用有限的计算资源近实时地对大型数据集进行操作。如上面结合图3和图4所讨论的，对存储在数据库204中的未注释数据102进行预处理，并且选择预处理数据的子集以便产生训练候选集。在计算机架构600中，以流传输方式处理未注释数据。在602处，从未注释数据102或训练候选集中检索样本并对其进行预处理。样本可以是从未注释数据102或训练候选集中抽出的随机选择样本。随机性的本质可通过随机种子来控制。检索到的文本是预处理的(例如，小写、标点之间插入的空格，以及针对类似日语的非空白语言插入的空白等)。

[0191] 在604处，将预处理数据(例如，文本)流经模型406，所述模型406将预处理数据(例如，文本)转换成分数向量(在分类器的情况下)或其他此类模型输出以产生预测集的预测。在606处，将(样本标识符、向量分数)的元组(或由模型输出的其他预测)流经采样存储写入器以选择在其中写入预测的一个或多个优先级队列608。在610处，采样选择逻辑选择优先级队列，在612处从所述优先级队列向用户呈现采样和预测以进行注释。在从用户接收到注释时，在614处对模型进行训练以产生所更新模型616，以便在604处继续过程以从预处理数据进行预测。

[0192] 如果系统存储针对每个样本所得的预测，则对内存和磁盘空间的需求将非常大。例如，对于未注释数据102或训练候选集典型地可能具有数百万或甚至数千万或数亿样本。产生并存储每个样本的预测所需的计算资源非常大。因此，优先级队列608针对每种类型的采样器各自提供有限长度的优先级队列。根据采样器所使用的采样算法，优先级队列608中的每一个可仅存储样本的前10、100、1000或10000个预测。在一些实现方式中，优先级队列608可一次存储2000个样本。如上面所讨论的，存在用于各种感兴趣的类以及各种目标函数的采样器，每个采样器具有对应的优先级队列608。在各种实现方式中，可存在2、5、10、20或更多个采样器。在一些实现方式中，采样器中的一个或多个可能不具有优先级队列608，而是依赖于蓄水池采样算法。例如，为了选择性地从具有在0.5和1.0之间的置信度水平的预测向量中为A类采样，蓄水池采样选择性地从符合这些要求的流传输样本中采样子集。优先级队列608可被持久化到客户端计算机206或注释服务器202上的磁盘。在一些实施方案中，使用分布式数据库技术(诸如通过存储在数据库204上)来存储优先级队列。如下面结合图7更详细地讨论的，优先级队列608的数据结构使得仅能够存储最上面的结果，而丢弃其他结果。

[0193] 采样器和优先级队列608中的每一个属于特定用户的单个注释过程。也就是说，不同用户可向同一未注释数据集提供注释，其中向每个用户提供单独的注释集。对于不同用户，优先级队列608与进展500中当前所选采样器可以不同。

[0194] 由于优先级队列608各自基于不同的采样算法保持不同的样本集，所以在注释过程中无明显延迟的情况下，注释过程400可诸如在410处改变采样器。此结果根据设计目标来具体实现以减少所需的用户认知负荷。注释可以反馈到系统中以便改进模型的当前迭

代，这继而通知采样器等等。

[0195] 例如，给定样本，系统可以进行分类预测。这些预测可用于计算必要的指标，诸如熵、最小裕度等等。这些分数可与针对每种类型的采样器存储的分数进行比较。在一些实施方案中，在预测满足某些准则的情况下，将其保存并将结果存储在优先级队列608中的一个或多个中；否则，将其丢弃。有利且有益的净影响是，优先级队列608所需的内存小而固定，对运行时的影响很小。在此类实施方案中，如上面所讨论的，可在小的固定内部、在用户的请求下或在算法重新训练确定时对模型进行重新训练。

[0196] 在一些实现方式中，模型可保持动态。当新注释到来时，模型可进行小调整。然后，其可继续预测，直到其遇到与采样器采样算法中的一个一致的样本。此时，模型可“暂停”等待用户的进一步反馈以对实例进行注释。一旦提供此注释，可重复所述过程。

[0197] 图7示出根据本公开技术的示例性实施方案的加速预测并降低硬件需求所采用的优先级队列方法700的框图。当预测702流经采样存储写入器606时，向多个优先级队列608提供预测。图7的实例中所示的优先级队列608包括针对待以“A类”注释的具有高置信度预测的样本的优先级队列704、针对待以“B类”注释的具有高置信度预测的样本的优先级队列706、针对具有高熵的样本的优先级队列708(例如，保持最高香农熵的顺序)，以及针对最小裕度样本的优先级队列710。可使用更多或更少的优先级队列608。以增大采样分数714的顺序将样本排列在优先级队列中。如上面所讨论的，采样分数可以是置信度分数或由预测向量以其他方式导出的值。

[0198] 当接收到新预测时，优先级队列608中的每一个评估新预测的采样分数。如果采样分数低于给定优先级队列608的阈值716，则优先级队列608可丢弃720预测。不同的优先级队列可使用不同的阈值716。如果采样分数高于给定优先级队列608的阈值716，则优先级队列评估是否保存718预测。例如，如果给定优先级队列608未满并且采样分数大于阈值716，则优先级队列608将保存预测。然而，如果给定优先级队列608是满的，则将采样分数与优先级队列608中先前保存的预测的采样分数中的一个或多个进行比较。在一些实施方案中，如果采样分数不大于先前存储的预测的采样分数中的任一个，则丢弃预测。否则，根据其优先级分数将预测保存在优先级队列608中的位置处，并且从优先级队列608中移除最低分数预测。如上面所讨论的，以此方式，优先级队列608保持基本上小于保存所有预测所需的固定内存需求。在其他实施方案中，诸如蓄水池采样的方法用于保持原始预测的子集，同时近似底层候选样本的分布。

[0199] 图8是根据本公开技术的示例性实施方案的用于注释测试数据集的注释过程800的流程图。注释过程800可在316处的注释测试集的创建期间发生，并且由注释服务器202或在注释客户端206上的本地安装上执行。在802处，接收未注释的测试数据集。在804处，由注释过程800或通过用户输入来确定是否需要测试集的详尽注释。如果是，则在806处，注释过程800提供供用户注释的测试集的线性进展以及线性进展的进度指示，诸如通过进度条等等。

[0200] 如果不需要详尽注释，则在808处，通过识别测试集数据的核心聚类来启动远程监督过程。例如，可通过与底层分布的分析相关联的各种技术来识别核心聚类。例如，具有分布模式的基于密度的聚类、具有分布平均值的基于概率的聚类或具有分布质心的基于分层的聚类。每种技术与对应的距离指标相关联(例如，基于层的聚类将使用欧氏距离)。在810

处，通过图形用户界面向用户呈现对从聚类中的一个或多个所获取的样本进行注释的请求。最初，可从聚类中随机抽取样本。在对数据进行注释时，将一致性指标分配给聚类，诸如平方距离和、基于样本的熵指标和基尼系数。在812处，将与其到注释样本的逆距离相关联的置信度分数分配给未注释数据点。所述系统在接近已知样本的开发点与新分布部分的探索之间交替进行。在一个此类实施方案中，通过诸如Bayesian bandits的强化学习方法来在探索与开发之间交替。在816处，由注释过程800确定最不确定预测的置信度分数是否超过阈值置信度分数。如果否，则注释过程800循环回以在810处请求测试数据集的另外样本的注释。否则，确定测试数据集被充分注释并将其在818处输出。在各种实施方案中，所有技术同时运行(基于密度、基于概率和基于分层)，并且通过强化学习来学习最成功的技术。

[0201] 图9至图25示出用于创建注释训练集以构建机器学习模型的各种示例性图形用户界面。将图9至图25的图形用户界面显示在注释客户端206的显示器上，并且通过注释客户端206的输入装置从注释用户接收输入。可从注释服务器202(诸如通过服务一个或多个网页的注释服务器202)来向注释客户端206提供图9至图25的图形用户界面，以在注释客户端206上的web浏览器上显示。替代地，注释客户端206上的本地安装可在注释客户端的显示器上呈现图9至图25的图形用户界面。本公开设想了其他配置。

[0202] 图9示出根据本公开技术的示例性实施方案的用于开始新机器学习模型创建的示例性图形用户界面900。模型名称字段902被配置为接收命名待创建的新模型的字母数字或其他字符串。模型选择部分904包括多个可选模型按钮，其中的每一个与不同类型的分类器相关联。例如，对于文本模型，可针对句子分类器、段落分类器、文档分类器、表分类器或表提取器中的每一个提供可选模型按钮。同样，对于图像模型，可针对目标检测模型或图像相似性模型中的每一个提供可选模型按钮。本文可使用其他类型的模型。上面所讨论的系统和过程对于所使用的数据或模型的类型是不可知的，并且对于文本数据，对于文本中使用的语言是不可知的。可显示用于识别当前登录以创建注释的用户帐户的用户指示符906。导航菜单908提供用于导航到本文所述的其他图形用户界面的可选按钮和/或菜单。在给定屏幕上提供所需输入时，图形用户界面之间的导航也可以自动化。例如，在命名新模型并在图9的图形用户界面上选择模型的类型时，可自动显示图10的图形用户界面。

[0203] 在各种实现方式中，待注释的未注释数据是未注释文本、图像、视频或音频数据。所述模型是单类分类器、二元分类器、多类分类器或语言分类器。所述模型可执行回归；信息提取；语义角色标记；文本摘要；句子、段落或文档分类；表格提取；机器翻译；蕴涵和矛盾；问答；音频标签；音频分类；说话人分类；语言模型调整；图像标签；目标检测；图像分割；图像相似性；逐像素注释；文本识别；或视频标签。上述未注释数据的模型和类型列表并不旨在详尽，并且仅作为实例提供。本公开设想任何其他类型的模型或类型的未注释数据。

[0204] 图10示出根据本公开技术的示例性实施方案的描绘用户能够管理多个数据集的方式的示例性图形用户界面1000。如图所示，在已导入到注释服务器202或注释客户端206或其相应的数据库204、208的可选数据集的列表1002中提供未注释数据102中的一个或多个数据集。

[0205] 图11示出根据本公开技术的示例性实施方案的描绘用户能够管理多个注释集的方式的示例性图形用户界面1100。一旦已导入一个或多个数据集，用户就可以创建“注释集”。图形用户界面1100允许用户管理多个注释集。在从未注释数据102生成的可选注释集

的列表1102中提供一个或多个注释集104，诸如注释或未注释的训练候选集或测试数据集。

[0206] 图12示出根据本公开技术的示例性实施方案的各种类别的注释集的实例。例如，可将注释集分类为情绪类别1202、行为类别1204、生活事件类别1206或顾客类别1208。本公开设想其他类别和类型的注释集。在每个类别中，可列出多个注释集。例如，对于情绪类别1202，注释集的列表包括喜爱、激动、愤怒、投诉、快乐、悲伤、团结和担忧。本公开设想了其他情绪。

[0207] 图13示出根据本公开技术的示例性实施方案的描绘用户能够使用来建立注释的配置选项的示例性图形用户界面1300。可选“编辑实例”按钮1302被提供用于导航到图17的图形用户界面以编辑实例。可选“编辑关键字”按钮1304被提供用于导航到图14的图形用户界面以编辑关键字。用户在数据集中查找的内容的实例和关键字向上述种子采样器502提供输入。可选注释按钮1306被提供用于导航到图18至图23的图形用户界面中的一个以根据上述过程300-800中的一个或多个注释样本。可选审查按钮1308被提供用于导航到图25的图形用户界面以审查并编辑注释。

[0208] 可选标记按钮1310有助于添加或改变用于注释样本的注释。所选注释显示在注释部分1312中。虽然只示出两个注释，但是可针对单类或多类分类器提供其他数量的注释。可选数据集按钮1314有助于添加另外的待注释数据集。可从图10保持的数据集中选择数据集。与数据集相关联的可选删除图标1316有助于移除要注释的数据集。为了解决关于不平衡数据的问题，频率选择1318有助于用户指示数据集中的实例或关键字的发生频率。在一些实现方式中，可从外部提供者(例如，外部服务器)下载、购买或出售关键字、实例和/或注释。进度指示符1320向用户显示已提供的输入和在开始样本注释之前仍需要的输入。和可选注释按钮1306一样，可选注释按钮1322被提供用于导航到图18至图23的图形用户界面中的一个以根据上述过程300-800中的一个或多个注释样本。同样，和可选审查按钮1308一样，可选审查按钮1324被提供用于导航到图25的图形用户界面以审查并编辑注释。

[0209] 图14示出根据本公开技术的示例性实施方案的描绘输入并管理关键字和短语列表的方法的示例性图形用户界面1400。文本输入框1402被提供用于添加关键字或短语以提供给种子采样器502。当添加关键字或短语时，更新关键字列表1412以显示输入的关键字或短语列表。可选按钮1404提供粘贴从另一个文档或程序复制的关键字或短语列表的选项。可选选项1406提供上传用于填充关键字列表的关键字或短语的文件的选项。利用所上传的关键字列表对外部关键字源列表1410进行更新。查找列表按钮1408有助于搜索并下载或购买一个或多个关键字列表。

[0210] 对于关键字列表1412中示出的每个关键字或短语，可选同义词按钮1414有助于扩展输入的关键字或短语以包括同义词。关键字或短语通过同义词典查找、通过近距离词嵌入以及通过外部本体扩展。词嵌入是指一组表示词的数字。这些词嵌入可提前提供，或从用户提供的数据集创建，或两者。可使用诸如跳跃图、负采样或移位正点互信息的非监督技术来学习词嵌入。本体是指开源的或用户提供的本体，诸如dbpedia。可以创建本体条目树，并且在给定关键字列表的情况下查找给定列表的最近共同祖先。然后，可在图形用户界面上呈现供用户选择的此祖先的后代，以包括在关键字列表1412中。在每一种情况下，诸如图15和图16所示，在给定词或短的短语的情况下，以可选方式向用户提供类似的词或短语以扩展关键字列表1412。在完成输入并扩展关键字和短语之后，可选结束按钮1416有助于导航

回到图13的图形用户界面。

[0211] 图15示出根据本公开技术的示例性实施方案的描绘用户能够发现并管理相关词和短语的方式的示例性图形用户界面1500。关键字指示符1502突出显示当前在考虑添加同义词或另外上下文的关键字。关键字的同义词列表1504被提供以组织到上下文聚类中。每个聚类提供用于选择聚类中的所有同义词的可选选项1506。此外，每个聚类内的同义词中的每一个提供有用于选择对应的同义词的可选选项1508。可选择取消按钮1510以丢弃任何所选的同义词，并返回到图14的图形用户界面。否则，接受按钮1512保存关键字列表1412中所选的同义词，并导航回图14的图形用户界面。

[0212] 图16示出根据本公开技术的示例性实施方案的描绘将本体并入单词列表管理中的示例性图形用户界面1600。文本输入框1402被提供用于向关键字列表1412添加关键字或短语。当关键字或短语被添加到关键字列表1412时，咨询以提供可选关键字列表1602的一个或多个本体来添加到关键字列表1412。如上面所讨论的，可以创建本体条目树，并且在给定关键字列表关键字的情况下查找给定列表的最近共同祖先。然后，可在图形用户界面1600上将此祖先的后代呈现为供用户选择的可选关键字1602，以包括在关键字列表1412中。

[0213] 图17示出根据本公开技术的示例性实施方案的描绘输入并管理实例列表的方法的示例性图形用户界面1700。通过选择图形用户界面1300上的编辑实例按钮1302，可以导航到图形用户界面1700。与关键字或短语类似，实例图形用户界面1700包括用于添加新实例的文本输入框1702。如图所示，实例通过一个或多个句子在上下文中提供关键字。每个实例是可选的，以为实例分配注释。例如，注释菜单1704被呈现用于指示所述实例是否是用户在数据集中查找的特定类或其他数据块的实例。示出先前输入的实例及对应注释的列表1706。结束按钮1708是可选的，以导航回图形用户界面1300。

[0214] 下面更详细地描述有助于注释过程的图形用户界面的各个方面。图18至图24的图形用户界面提供有助于人类判断模型质量的信息。一旦用户注释多个实例，就可以对初始模型进行训练并使得另外的抽样方法可用。

[0215] 图18示出根据本公开技术的示例性实施方案的描绘注释过程、用于管理注释过程的工具和关于进展的反馈的示例性图形用户界面1800。图形用户界面1800使认知负荷和训练机器学习模型所需的专业知识最小化。这可以通过显示来自用户所提供的数据集的一系列实例来实现。如上所述，实例由采样器选择，所述采样器可使用户创建强大模型所需的注释数量最小化。

[0216] 可向图形用户界面提供实例1802。提供有趣的单词或短语的突出显示1804，其中所述突出显示和颜色可表示对最终预测的影响的方向和大小。周围的上下文1806还可提供有实例，诸如前一个句子和后一个句子。此外，向用户显示预测1808以及该预测的置信度分数。数据注释可诸如通过选择“是”1810或“否”1812按钮来存储在问题的答案中。用户可通过键盘快捷键(诸如键入“Y”或“N”)来选择性地提供他们的响应。

[0217] 向用户提供关于注释数据的质量和数量的反馈1814。例如，关于注释范围的反馈可包括多个训练的实例，包括训练的多个正实例和多个负实例的分类。此外，还可示出模型的性能指标，诸如准确度、精度、召回率、F1分数或二元分类器的ROC下的面积。可示出其他指标。可提供一个或多个导航图标以诸如通过选择箭头1818来跳过实例的注释，或者诸如

通过选择箭头1816来返回到先前的实例。

[0218] 图19示出根据本公开技术的示例性实施方案的描绘允许用户对其响应的强度进行评分的注释过程的示例性图形用户界面1900。除了为注释提供“是”或“否”输入之外或替代地，用户还可对实例的好坏进行评分，诸如选择在一定范围的多个分数按钮1902(例如，从“1”至“5”个按钮)中的一个。还可以提供所述类型采样器1904的另外反馈，所述所述类型采样器目前用于选择要注释的实例。

[0219] 图20示出根据本公开技术的示例性实施方案的描绘对相邻条目进行注释的能力以及对注释的彩色反馈的示例性图形用户界面2000。例如，在呈现实例时，用户可以突出显示2002相邻条目并从菜单2004提供选择以例如将相邻条目注释为正实例、负实例，或者清除相邻条目的突出显示。

[0220] 图21示出根据本公开技术的示例性实施方案的描绘示出多种类型的候选采样的下拉菜单的示例性图形用户界面2100。如上面所讨论的，可通过用户输入来手动选择采样器。在图21所示的实例中，提供用于从不同类型的采样器进行选择以便提供实例的采样方法菜单2102。通过选择突出显示2104来指示当前所选的采样器。采样器可用的选择由选择图标2106指示。例如，在选择选择图标2106时，可将采样器从自动采样器(例如，进展500)改变为错误减少采样器。例如，可使用其他类型的采样方法菜单，诸如下拉列表。通过使用上述优先级队列608，虽然改变用于向图形用户界面提供实例的采样算法，但是用户在获得随后的实例时将不会注意到过程延迟。

[0221] 图22示出根据本公开技术的示例性实施方案的描绘跨数据的一次性关键字搜索的示例性图形用户界面2200。此外，如上面所讨论的，种子采样器502可使用用户所提供的输入来搜索另外实例。例如，用户可以在搜索框2202中输入关键字以在数据集内搜索，以便识别另外实例。响应于用户在搜索框2202中搜索关键字，用户界面2200可以呈现另外的可选关键字或短语2204以扩展用户的搜索。例如，如上所述，可使用同义词典、词典和/或本体来生成另外的关键字或短语。在选择可选关键字或短语2204中的一个时，使用所选关键字或短语可执行另外的搜索。

[0222] 图23示出根据本公开技术的示例性实施方案的描绘能够允许用户可视地探索其数据的数据图的示例性图形用户界面2300。如上面所提到的，数据集的非结构化表示可用于使用诸如t-sne或PCA的技术来将数据维数降到二维或三维。然后，这些低维表示可以在图形用户界面2300上显示为数据图2302。数据图2302可包括数据指示符2304，代表所识别的聚类或其他数据分组中的样本。每个样本可具有代表样本的注释或者指示样本是否未注释的颜色。此外，样本的置信度分数可以通过代表给定样本的图标的大小(例如，圆的直径基于置信度分数是不同的)来以图形方式表示。注释者之间的分歧/错误还可通过错误图标(例如，红色或指示错误的区别性形状或图案)来在数据图2302上指示。用户可以使用数据图2302来导航他们的数据，并且查找和选择要注释的具体实例2306。例如，用户可以在样本周围绘制边框或套索，他们为其提供注释。如上面所讨论的，种子采样器502可使用用户所提供的此输入来呈现另外实例。例如，用户可选择对尚未注释的所识别样本聚类进行注释。替代地，用户可选择使用与样本相关联的若干颜色来注释样本聚类，所述颜色代表用于注释聚类中的数据元素的多个不同注释。因此，用户可以使所需注释清晰，或以其他方式提供进一步的输入，以有助于对所选数据进行正确注释。

[0223] 图24示出根据本公开技术的示例性实施方案的描绘能够如何处理失效状态和如何将信息传递回用户的示例性图形用户界面2400。例如，反馈2402可被提供用于说明显示失败状态的原因，并且提供推荐2404说明如何解决失败状态。所推荐动作按钮2406可被提供用于自动启动推荐2404。

[0224] 图25示出根据本公开技术的示例性实施方案的描绘先前所注释条目的列表和如何管理那些条目的示例性图形用户界面2500。例如，用户界面2500可在选择审查按钮1308、1324时导航以审查注释。在执行注释的审查时，用户可选择注释中的任一个以改变与注释相关联的注释。例如，在选择注释时，注释菜单2502可呈现用于选择不同注释的选项。

[0225] 图26示出根据本公开技术的示例性实施方案的示例性计算机系统。

[0226] 关于系统与方法和/或应用、程序或其他计算机相关的实现方式与配置，本文描述本公开技术的某些方面。如本文所述的“系统”可以指计算机相关的系统和部件，其可以利用单台计算机或分布式计算架构。各种图的图示示出图形用户界面的方面，并且如本文所述是指可由一个或多个计算系统的输入和输出控制器和/或其他系统控制的显示数据以及功能交互元件和输出。所述一个或多个计算系统可包括用于实现本文提及的各种系统、方法和/或应用/程序的功能部件，例如，一个或多个计算机，其包括联接到一个或多个存储器装置和/或用于存储指令的其他存储装置的一个或多个处理器，所述指令在由所述一个或多个处理器实行时致使所述一个或多个计算机执行特定任务，以便实现本公开技术的所述实施方案的各个方面。

[0227] 如上面所简要提及的，所述一个或多个计算机的此类部件可联接到输入/输出控制器，用于从输入装置接收输入，例如从查看图形用户界面显示的计算机的用户的交互输入，以及用于控制将数据输出到一个或多个显示装置或其他输出外围装置。本文所指的“方法”可以是计算机实现的方法，其包括由一个或多个处理器和/或其他计算机系统部件执行的一系列操作。本文中对应用、程序等等的引用可以是计算机可执行指令，其可以存储在模块中、硬盘和/或可移动存储介质(又称为“计算机可读介质”或“计算机可读存储介质”或“非暂时性计算机可读存储介质”)上，并且所述指令在由一个或多个处理器执行时致使一个或多个计算机系统执行与本文所述的实施方案相关的特定功能。本文所述的各种计算机和/或系统的部件可包括网络接口部件，用于访问到诸如因特网或内部网络的网络的网络连接，以例如通过与一个或多个外部服务器交换数据来在此类网络上接收和传输数据。

[0228] 应当了解，本文相对于各图描述的逻辑操作可实施为(1)运行在计算装置(例如，图26中所述的计算装置)上的计算机实施的动作或程序模块(即，软件)的序列，(2)计算装置内的互连机器逻辑电路或电路模块(即，硬件)，和/或(3)计算装置的软件和硬件的组合。因此，本文所论述的逻辑操作并不限于硬件和软件的任何特定组合。实施方式是取决于计算装置的性能和其它要求的选择问题。因此，本文所描述的逻辑运算被不同地称为操作、结构设备、行为或模块。这些操作、结构装置、动作和模块可实施在软件、固件、专用数字逻辑以及其任何组合中。还应了解，可执行比附图中示出且本文中描述的更多或更少的操作。这些操作也可以与本文中描述的不同的顺序来执行。

[0229] 参考图26，示出了示例性计算装置2600，本发明的实施方案可在其上实现。例如，本文所述的注释服务器202或客户端计算机206中的每一个可作为计算装置来实现，诸如计算装置2600。应当理解，示例性计算装置2600仅是可在适当的计算环境上实现本发明的实

施方案的一个实例。可选地，计算装置2600可以是众所周知的计算系统，其包括但不限于个人计算机、服务器、手持或笔记本电脑装置、多处理器系统、基于微处理器的系统、网络个人计算机(PC)、小型计算机、大型计算机、嵌入式系统和/或包括多个任何上述系统或装置的分布式计算环境。分布式计算环境使得连接到通信网络或其他数据传输介质的远程计算装置能够执行各种任务。在分布式计算环境中，程序模块、应用和其他数据可存储在本地和/或远程计算机存储介质上。

[0230] 在一个实施方案中，计算装置2600可包括协作执行任务的相互通信的两台或更多台计算机。例如，但不是以限制的方式，可以允许对应用的指令进行并发和/或并行处理的方式来对应用进行分区。替代地，可以允许两台或更多台计算机对数据集的不同部分进行并发和/或并行处理的方式来对由应用处理的数据进行分区。在一个实施方案中，计算装置2600可利用虚拟化软件来提供不直接绑定到计算装置2600中的计算机数量的多个服务器的功能。例如，虚拟化软件可在四台物理计算机上提供二十台虚拟服务器。在一个实施方案中，可通过执行应用和/或在云计算环境中的应用来提供上面所公开的功能。云计算可包括使用动态可伸缩计算资源通过网络连接提供计算服务。云计算可至少部分地由虚拟化软件支持。云计算环境可由企业建立和/或根据需要从第三方提供商租用。一些云计算环境可包括企业拥有和操作的云计算资源，以及第三方提供商租用和/或租赁的云计算资源。

[0231] 在计算装置2600最基本的配置中，所述计算装置2600通常包括至少一个处理单元2620和系统存储器2630。根据计算装置的确切配置和类型，系统存储器2630可为易失性的(如随机存取存储器(RAM))、非易失性的(如只读存储器(ROM)、闪存等)、或这两者的某个组合。这种最基本的配置在图26中以虚线2610示出。处理单元2620可为执行对计算装置2600的操作必要的算术和逻辑操作的标准可编程处理器。虽然只示出一个处理单元2620，但可存在多个处理器。因此，虽然指令可被讨论为由处理器执行，但是指令可由一个或多个处理器同时、串行或以其他方式执行。计算装置2600还可包括用于在计算装置2600的各种部件之间传达信息的总线或其他通信机构。

[0232] 计算装置2600可具有另外的特征/功能性。例如，计算装置2600可包括另外的存储装置，诸如可移动存储装置2640和不可移动存储装置2650，包括但不限于磁性或光学盘或带。计算装置2600还可包含允许装置与其他装置通信的网络连接2680，诸如本文所述的通信路径。所述一个或多个网络连接2680可采用以下形式：调制解调器、调制解调器组、以太网卡、通用串行总线(USB)接口卡、串行接口、令牌环卡、光纤分布式数据接口(FDDI)卡、无线局域网(WLAN)卡、诸如码分多址(CDMA)的无线电收发器卡、全球移动通信系统(GSM)、长期演进(LTE)、全球微波接入互操作性(WiMAX)和/或其他空中接口协议无线电收发器卡以及其他众所周知的网络装置。计算装置2600还可具有一个或多个输入装置2670，诸如键盘、辅助键盘、开关、调节器、鼠标、轨迹球、触摸屏、语音识别器、读卡器、纸带读取器或其他众所周知的输入装置。还可以包括一个或多个输出装置2660，诸如打印机、视频监视器、液晶显示器(LCD)、触摸屏显示器、显示器、扬声器等。另外装置可连接到总线，以便有助于在计算装置2600的部件之间的数据通信。所有这些装置是本领域中众所周知的，并且无需在此详述。

[0233] 处理单元2620可配置成执行编码在有形计算机可读介质中的程序代码。有形的计算机可读介质是指能够提供致使计算装置2600(即，机器)以特定方式操作的数据的任何介

质。可利用各种计算机可读介质来向处理单元2620提供指令以供执行。示例性有形的计算机可读介质可包括但不限于在任何方法或技术中实施的易失性介质、非易失性介质、可移动介质和不可移动介质，用以存储信息如计算机可读指令、数据结构、程序模块或其它数据。系统存储器2630、可移动存储装置2640以及不可移动存储装置2650均为有形的计算机存储介质的实例。示例性有形计算机可读记录介质包括但不限于集成电路(例如，现场可编程门阵列或专用IC)、硬盘、光盘、磁光盘、软盘、磁带、全息存储介质、固态装置、RAM、ROM、电可擦除编程只读存储器(EEPROM)、闪存或其它存储器技术、CD-ROM、数字多功能盘(DVD)或其它光学存储装置、磁盒、磁带、磁盘存储装置或其它磁性存储装置。

[0234] 通过将可执行软件加载到计算机中所能够实现的功能可以通过众所周知的设计规则来转换为硬件实现，其对于电气工程和软件工程技术来说是至关重要的。在软件和硬件中实现概念之间的决定通常取决于设计的稳定性和待生产的单元的数量，而不是从软件域转变为硬件域所涉及的任何问题。一般来讲，仍然频繁变化的设计可优选地在软件中实现，因为重新旋转硬件实现比重新旋转软件设计昂贵。一般来讲，将大量生产的稳定设计可优选地在硬件中实现，例如在专用集成电路(ASIC)中实现，因为对于大型生产运行，硬件实现可以比软件实现便宜。常常可以软件的形式对设计进行开发和测试，然后根据众所周知的设计规则，将其转变为专用集成电路中的硬连接软件指令的等效硬件实现。与由新ASIC控制的机器是特定的机器或设备以同样的方式，同样地，被编程和/或加载有可执行指令的计算机可以被视为特定的机器或设备。

[0235] 在示例性实施方式中，处理单元2620可执行存储在系统存储器2630中的程序代码。例如，总线可将数据运载到系统存储器2630，处理单元2620从系统存储器2630接收指令并且执行指令。系统存储器2630接收到的数据可在由处理单元2620执行之前或之后任选地存储在可移动存储装置2640或不可移动存储装置2650上。

[0236] 应当理解，本文中描述的各种技术可结合硬件或软件、或在适当时结合其组合来实施。因此，目前所公开的标的物的方法和设备、或其某些方面或部分可采取体现在有形介质中的程序代码(即，指令)的形式，所述有形介质如软盘、CD-ROM、硬盘驱动器或任何其他机器可读存储介质，其中当将程序代码加载到机器(如计算装置)中并由所述机器执行时，所述机器变为一种用于实践目前所公开的标的物的设备。在程序代码在可编程计算机上执行的情况下，计算装置通常包括处理器、可由处理器读取的存储介质(包括易失性和非易失性存储器和/或存储元件)、至少一个输入装置以及至少一个输出装置。一个或多个程序可实施或利用结合目前所公开的标的物而描述的过程，例如，通过使用应用编程接口(API)、可再用控件等。此类程序可以高级程序或面向对象的编程语言来实施以与计算机系统通信。然而，如果期望的话，程序可以汇编或机器语言来实施。在任何情况下，语言可为编译或解释语言，并且其可与硬件实施方案组合。

[0237] 方法和系统的实施方案可参照方法、系统、设备和计算机程序产品的框图和流程图在本文进行描述。将理解，框图和流程图图解的每个方框以及框图和流程图图解中的方框的组合分别可以由计算机程序指令实现。这些计算机程序指令可加载到通用计算机、专用计算机或其他可编程数据处理设备的处理器上以便产生一种机器使得在计算机或其他可编程数据处理设备上执行的指令构建用于实施一个或多个流程图方框中所指定的功能的手段。

[0238] 还可以将这些计算机程序指令存储在可以引导计算机或其他可编程数据处理设备以特定方式起作用的计算机可读存储器中，使得存储在计算机可读存储器中的指令产生一种制品，所述制品包括可实现在一个或多个流程图方框中规定的功能的指令。计算机程序指令还可以加载到计算机或其他可编程数据处理设备上以使得在计算机或其他可编程设备上执行一系列操作步骤来产生计算机实现的过程，使得在计算机或其他可编程设备上执行的指令提供用于实现一个或多个流程图方框中规定的功能的步骤。

[0239] 因此，框图的方框和流程图图解支持用于执行指定功能的装置的组合、用于执行指定功能的步骤的组合以及用于执行指定功能的程序指令装置。还应当理解，框图和流程图图解的每个方框以及框图和流程图图解中的方框的组合可由执行指定功能或步骤的基于特定用途硬件的计算机系统或者特定用途硬件和计算机指令的组合来实现。

[0240] 短语“和/或”的使用指示可以使用选项列表的任一个或任何组合。例如，“A、B和/或C”意指“A”或“B”或“C”，或“A和B”或“A和C”或“B和C”，或“A和B和C”。如本说明书中使用，单数形式“一”和“所述”包含复数形式，除非上下文清楚地另外规定。此外，为了方便读者，本说明书中可使用标题或副标题，其不应影响本公开技术的范围。通过“包含”或“含有”或“包括”意指至少存在于组合物或制品或方法中所指定的化合物、元素、颗粒或方法步骤，但不排除其他化合物、材料、颗粒、方法步骤的存在，即使其他此类化合物、材料、颗粒、方法步骤与所指定的具有相同功能。

[0241] 在描述示例性实施方案时，为了清晰起见，将借助于术语。意图是，每个术语设想其如本领域技术人员所理解的最广泛的意义，并且包括以类似方式操作以达到类似目的的所有技术等价物。

[0242] 应当理解，提及的方法的一个或多个步骤并不排除附加方法步骤的存在或在明确标识的那些步骤之间插入方法步骤。方法的步骤可与本文中描述的不同的顺序来执行。类似地，还应当理解，提及的装置或系统中的一个或多个部件并不排除附加部件的存在或在明确标识的那些部件之间插入部件。

[0243] 虽然本公开提供若干实施方案，但是应当理解，所公开的系统和方法可在不背离本公开的精神或范围的情况下以许多其他具体形式体现。本实例应被认为是说明性的，而不是限制性的，并且目的不限于本文所给出的细节。例如，各种元件或部件可被组合或集成到另一个系统中，或者某些特征可被省略或不实现。

[0244] 同样，在各种实施方案中描述和说明的作为分立或分离的技术、系统、子系统和方法可在不背离本公开的范围的情况下与其他系统、模块、技术或方法组合或集成。显示或讨论为直接联接或相互通信的其他项目可以通过一些接口、装置或中间部件间接联接或通信，无论是以电的方式、机械的方式还是其他方式。变化、替换和变更的其他实例是由本领域技术人员确定的，并且可以在不背离本文所公开的精神和范围的情况下进行。

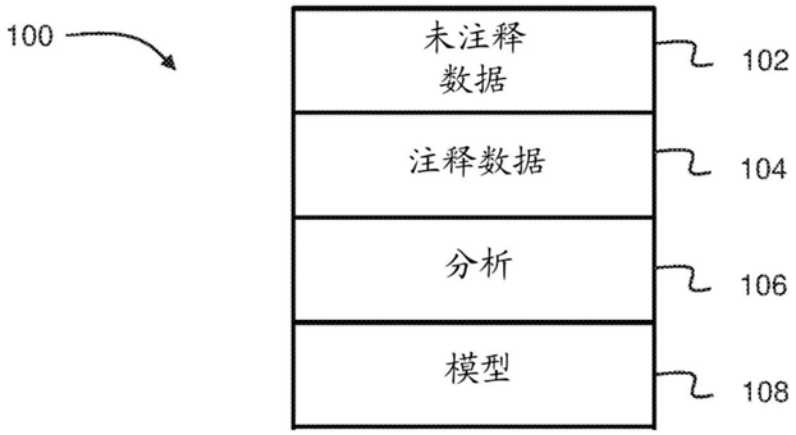


图1

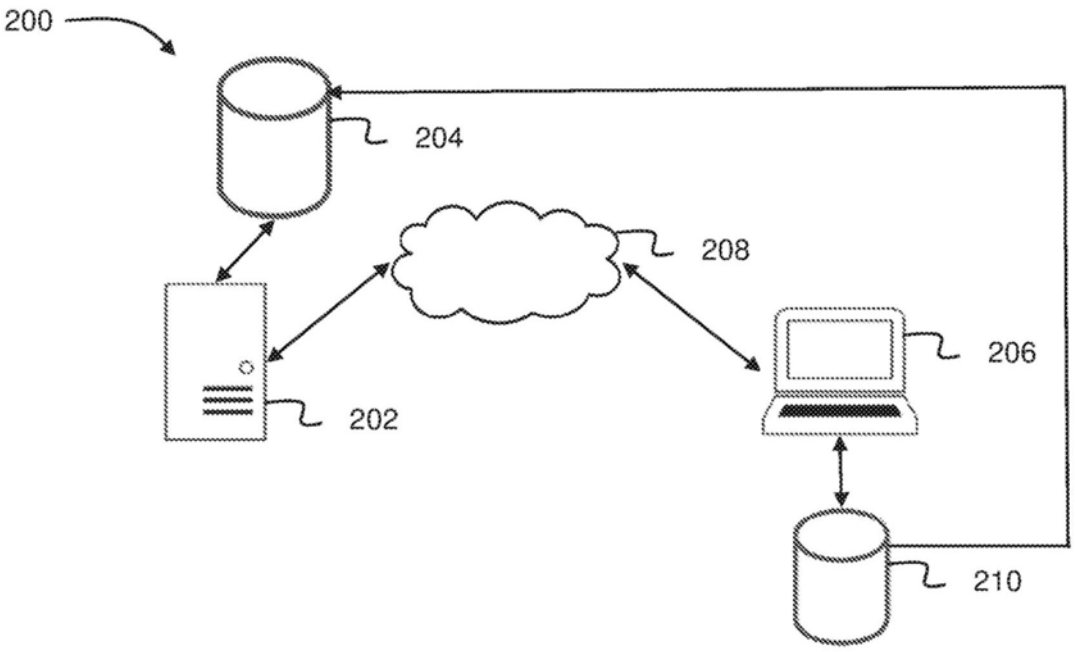


图2

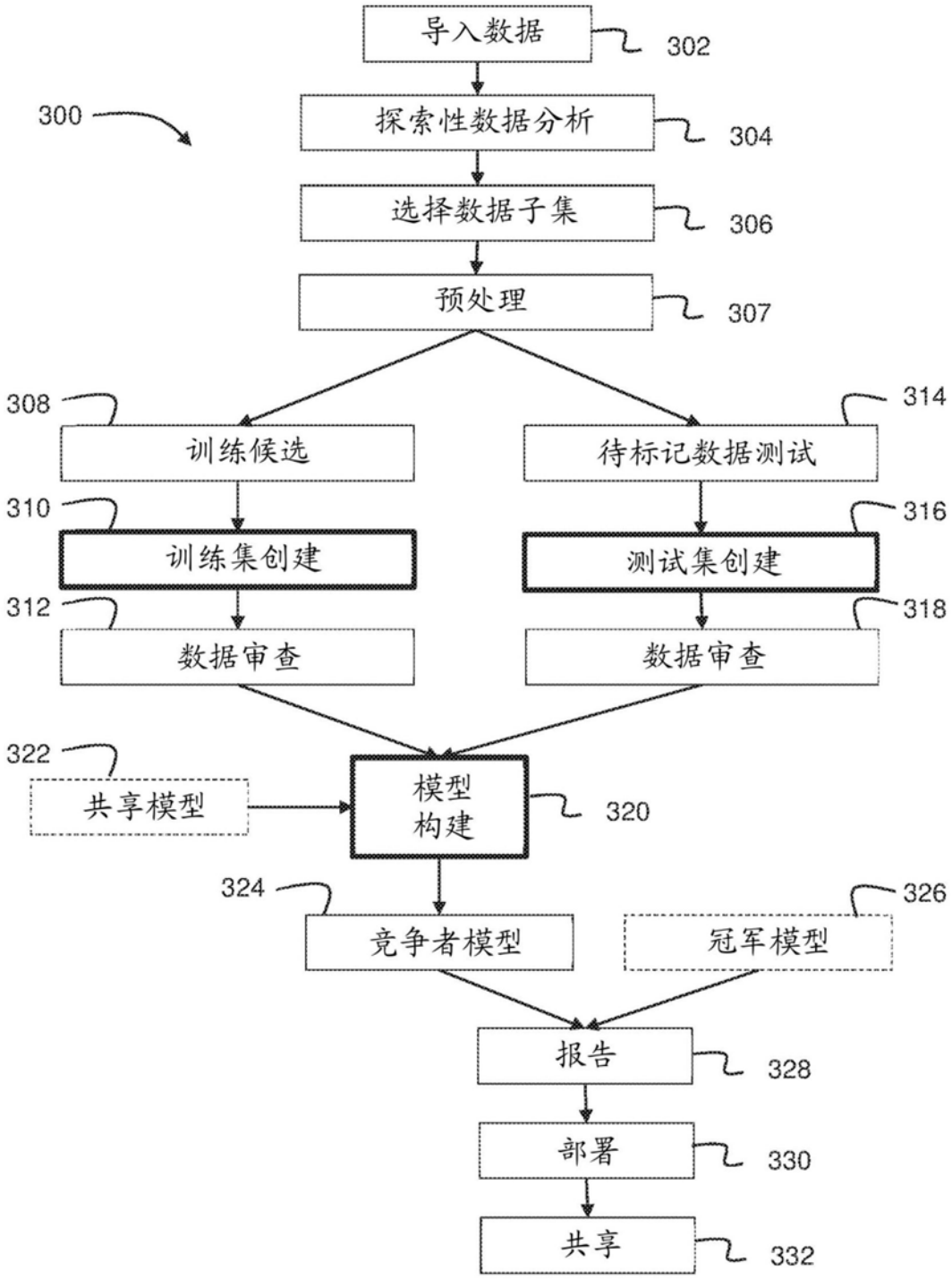

导入数据
302
300
探索性数据分析
304
选择数据子集
306
预处理
307
308
训练候选
314
待标记数据测试
310
训练集创建
316
测试集创建
312
数据审查
318
数据审查
322
共享模型
模型
构建
320
324
竞争者模型
326
冠军模型
报告
328
部署
330
共享
332

图3

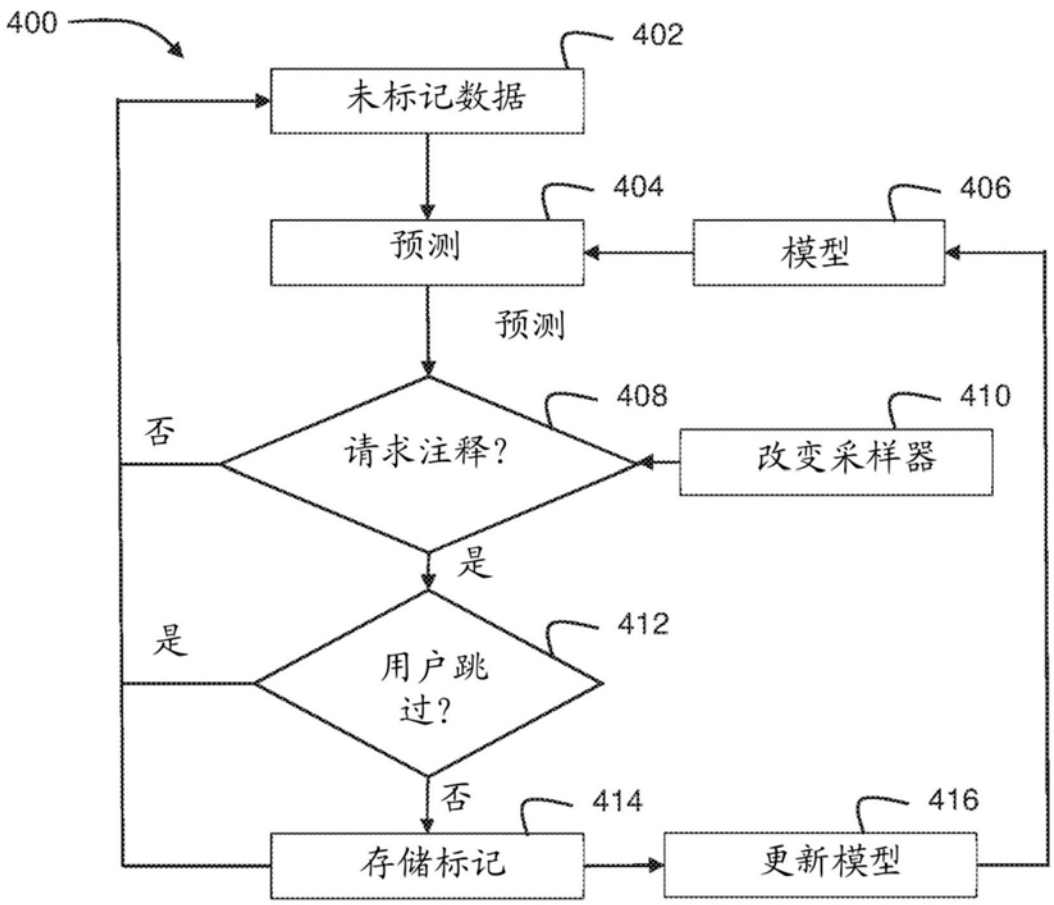

400
402
未标记数据
404
预测
406
模型
预测
408
请求注释?
410
改变采样器
否
是
是
412
用户跳过?
否
414
存储标记
416
更新模型

图4

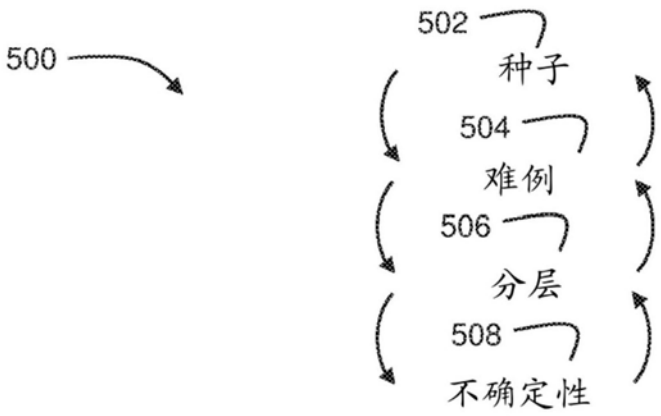

500
502
种子
504
难例
506
分层
508
不确定性

图5

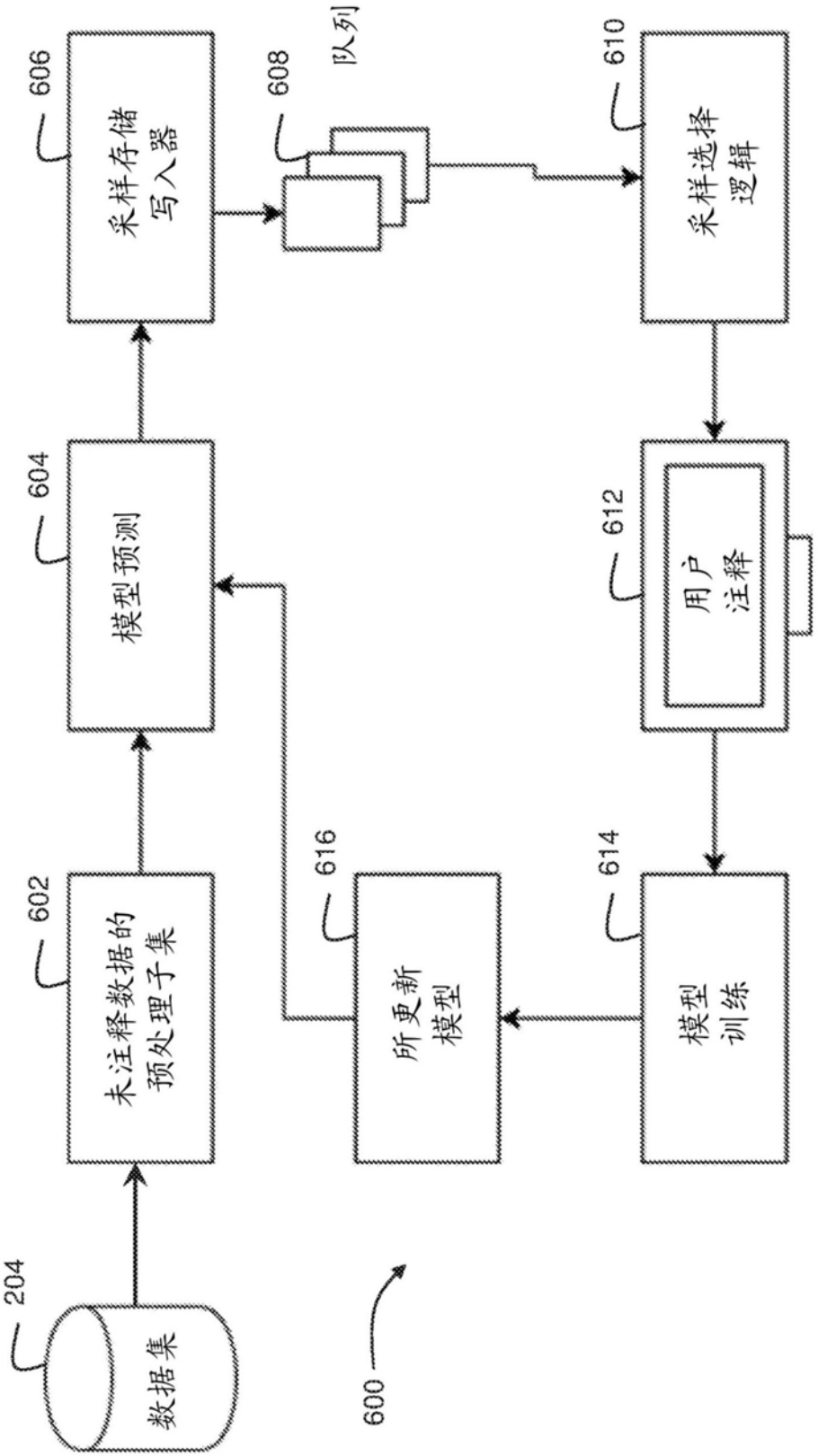

606
采样存储
写入器
608
队列
610
采样选择
逻辑
604
模型预测
612
用户
注释
602
未注释数据的
预处理子集
616
所更新
模型
614
模型
训练
204
数据集
600

图6

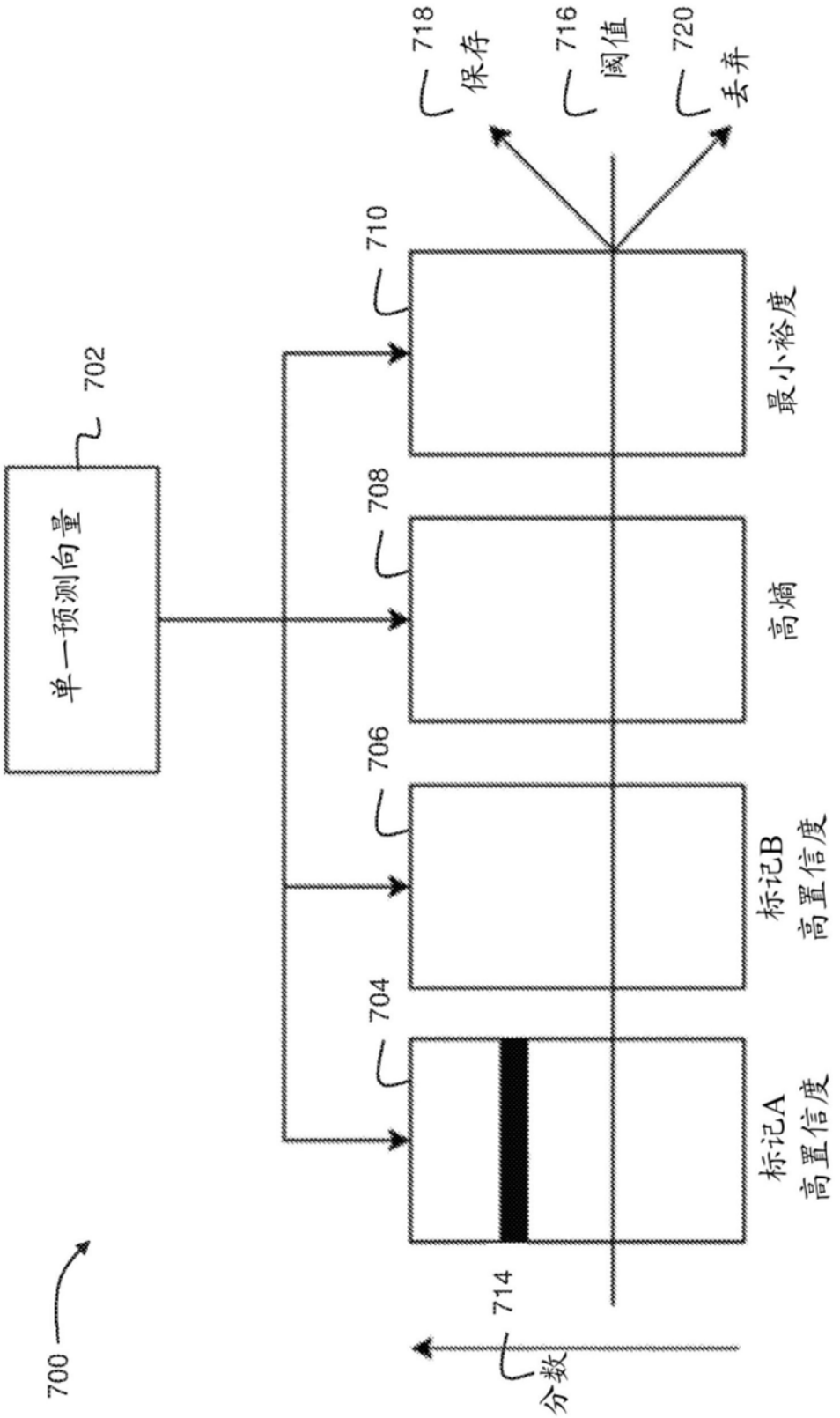

700
702
单一预测向量
704
标记A
高置信度
706
标记B
高置信度
708
高熵
710
最小裕度
714
分数
716
阈值
718
保存
720
丢弃

图7

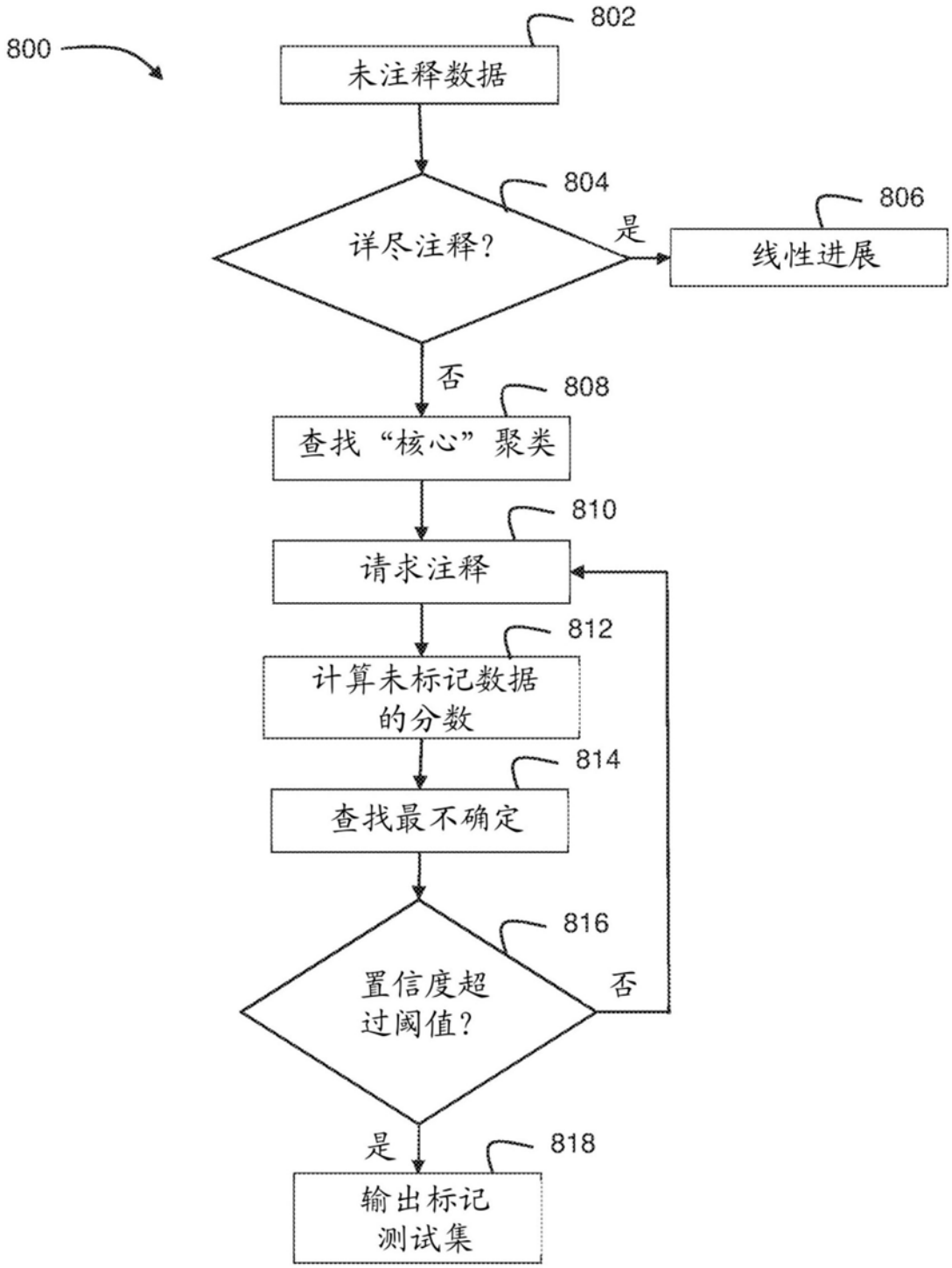

800
802
未注释数据
804
详尽注释?
是
806
线性进展
否
808
查找“核心”聚类
810
请求注释
812
计算未标记数据的分数
814
查找最不确定
816
置信度超过阈值?
否
是
818
输出标记测试集

图8

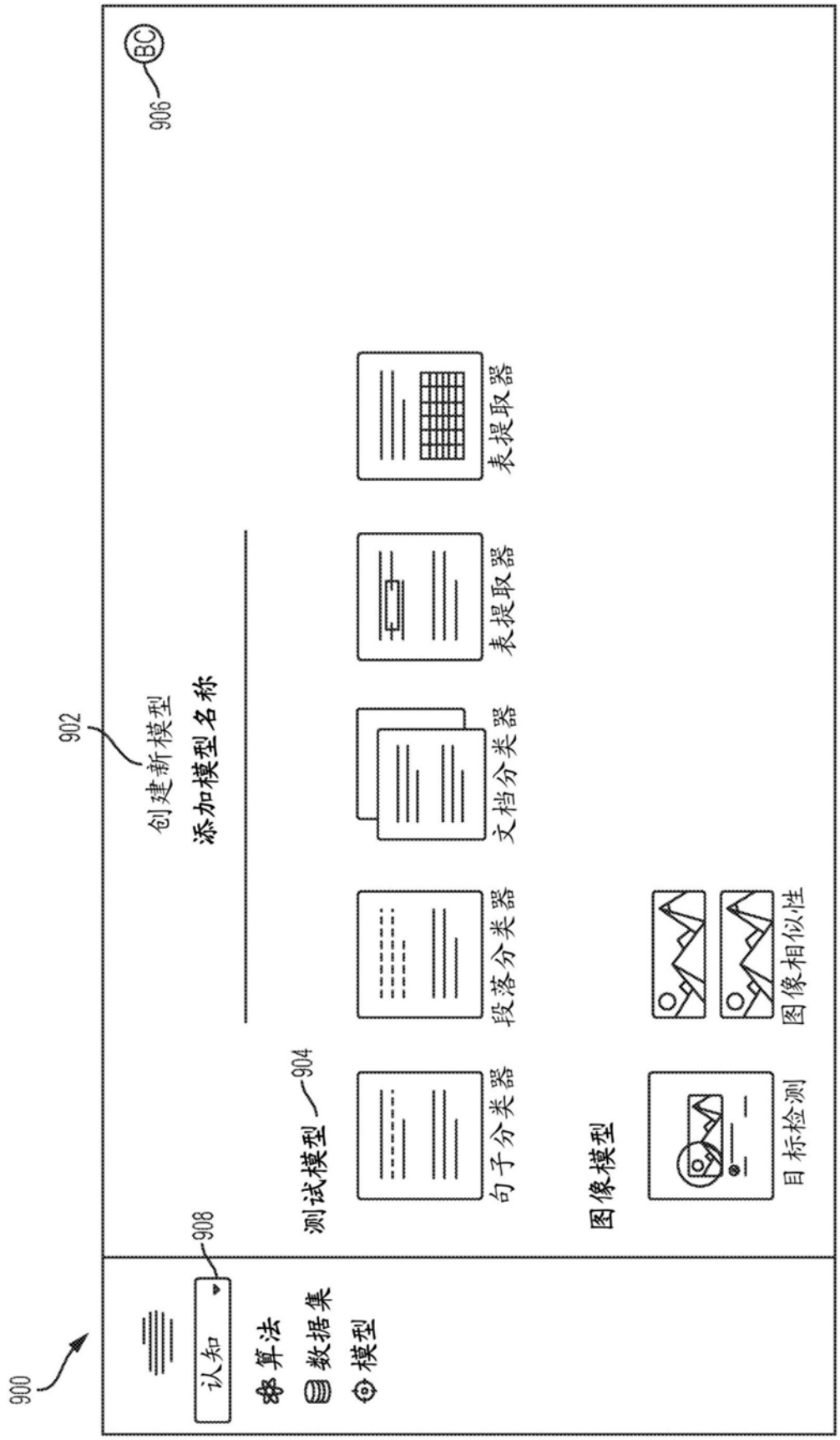

900
认知
算法
数据集
模型
908
测试模型
904
902
创建新模型
添加模型名称
906
BC
句子分类器
段落分类器
文档分类器
表提取器
表提取器
图像模型
目标检测
图像相似性

图9

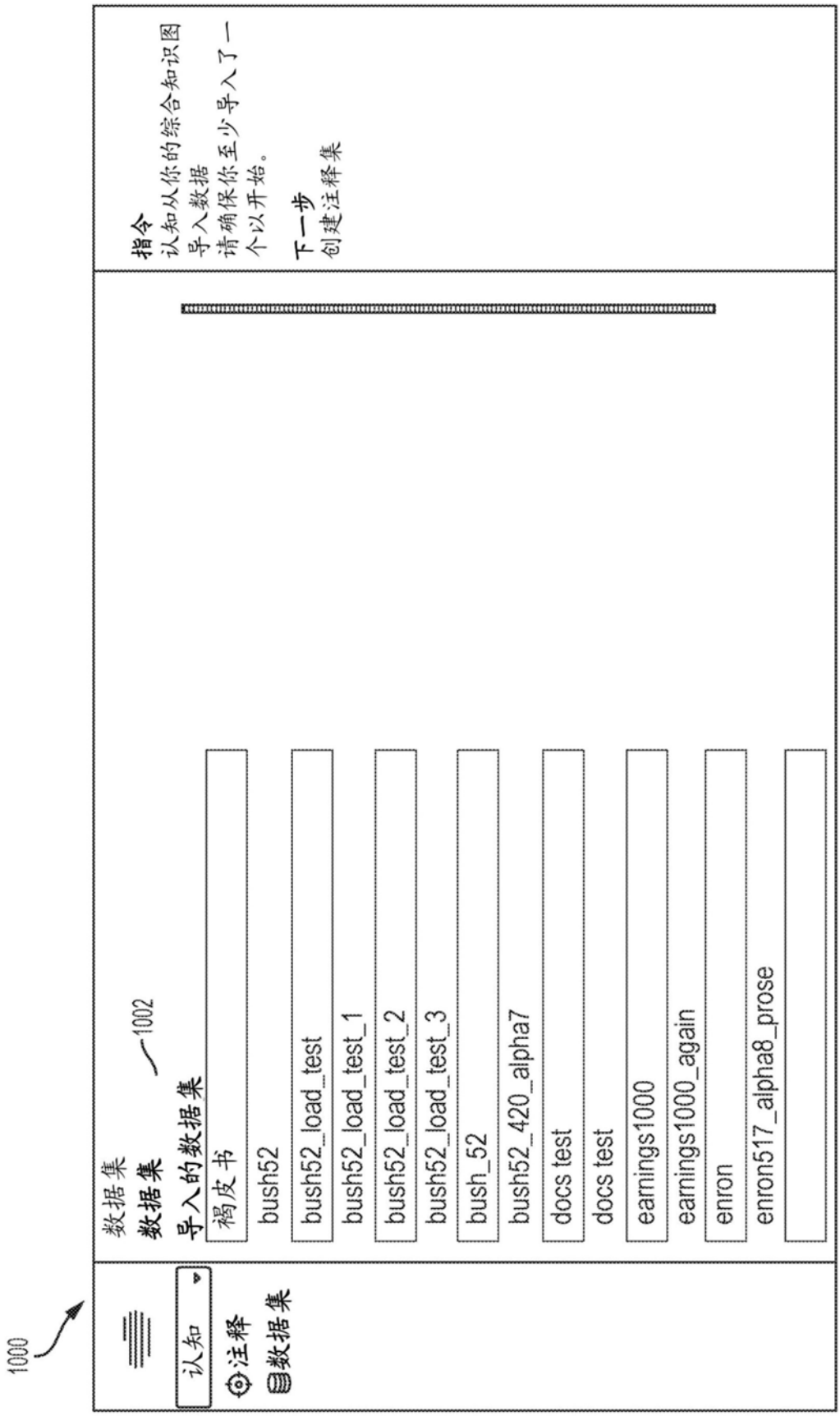

1000
认知
注释
数据集
数据集
数据集
1002
导入的数据集
褐皮书
bush52
bush52_load_test
bush52_load_test_1
bush52_load_test_2
bush52_load_test_3
bush_52
bush52_420_alpha7
docs test
docs test
earnings1000
earnings1000_again
enron
enron517_alpha8_prose
指令
认知从你的综合知识图
导入数据
请确保你至少导入了一
个以开始。
下一步
创建注释集

图10

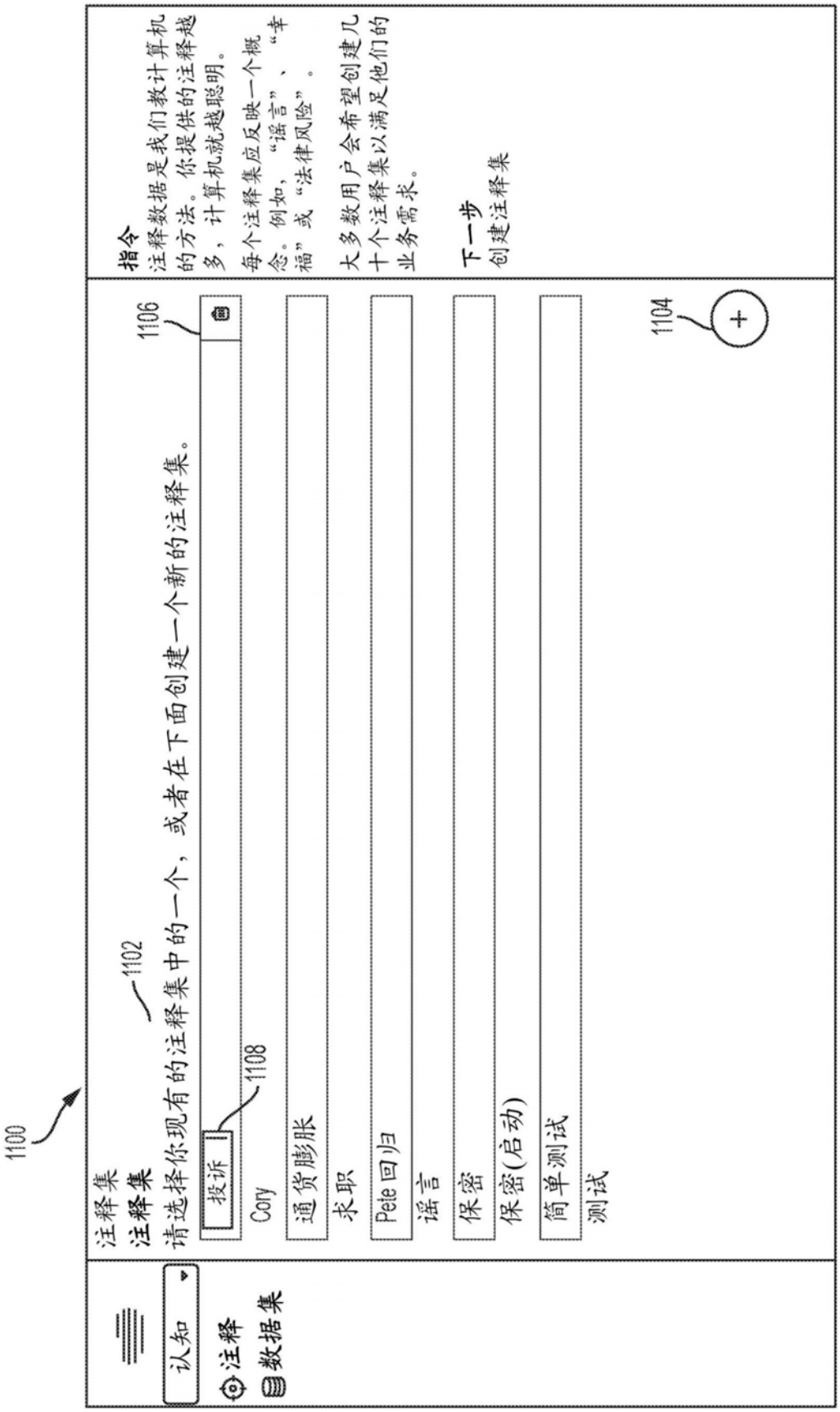

1100
注释集
注释集
1102
请选择你现有的注释集中的一个，或者在下面创建一个新的注释集。
1106
投诉
1108
Cory
通货膨胀
求职
Pete回归
谣言
保密
保密(启动)
简单测试
测试
1104
+
认知
注释
数据集
指令
注释数据是我们教计算机的方法。你提供的注释越多，计算机就越聪明。
每个注释集应反映一个概念。例如，“谣言”、“幸福”或“法律风险”。
大多数用户会希望创建几十个注释集以满足他们的业务需求。
下一步
创建注释集

图11

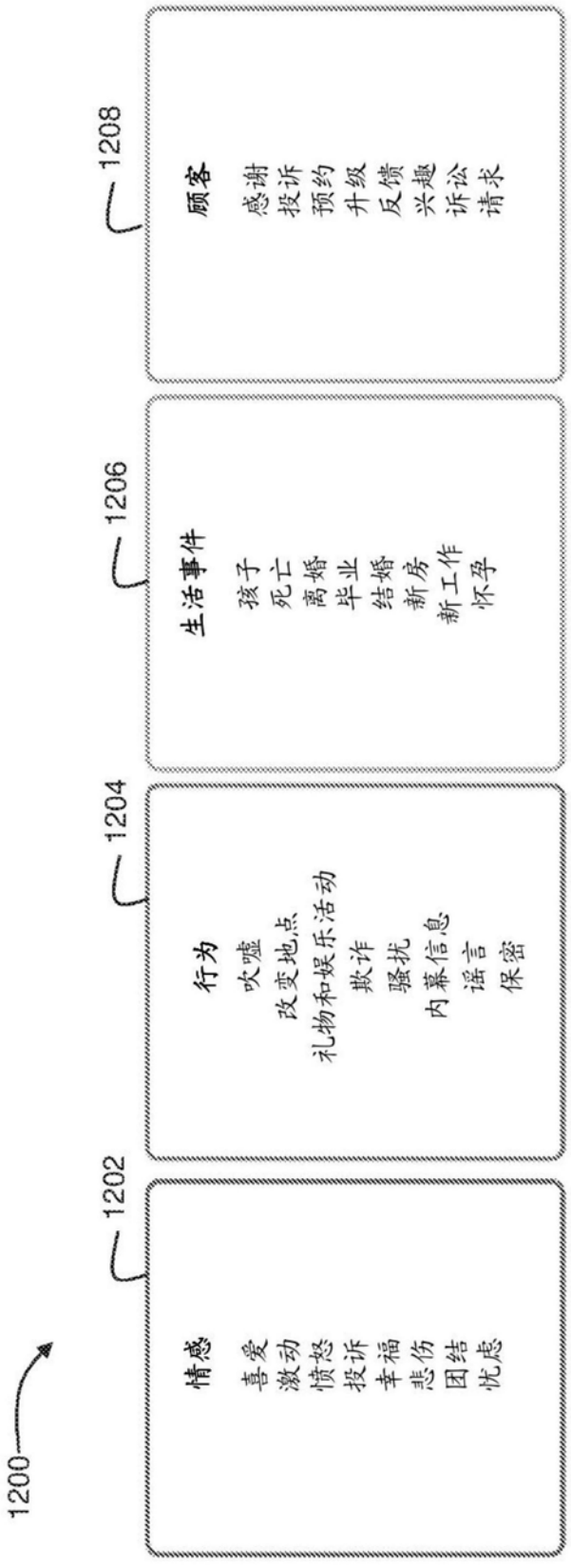

1200
1202
情感
喜爱
激动
愤怒
投诉
幸福
悲伤
团结
忧虑
1204
行为
吹嘘
改变地点
礼物和娱乐活动
欺诈
骚扰
内幕信息
谣言
保密
1206
生活事件
孩子
死亡
离婚
毕业
结婚
新房
新工作
怀孕
1208
顾客
感谢
投诉
预约
升级
反馈
兴趣
诉讼
请求

图12

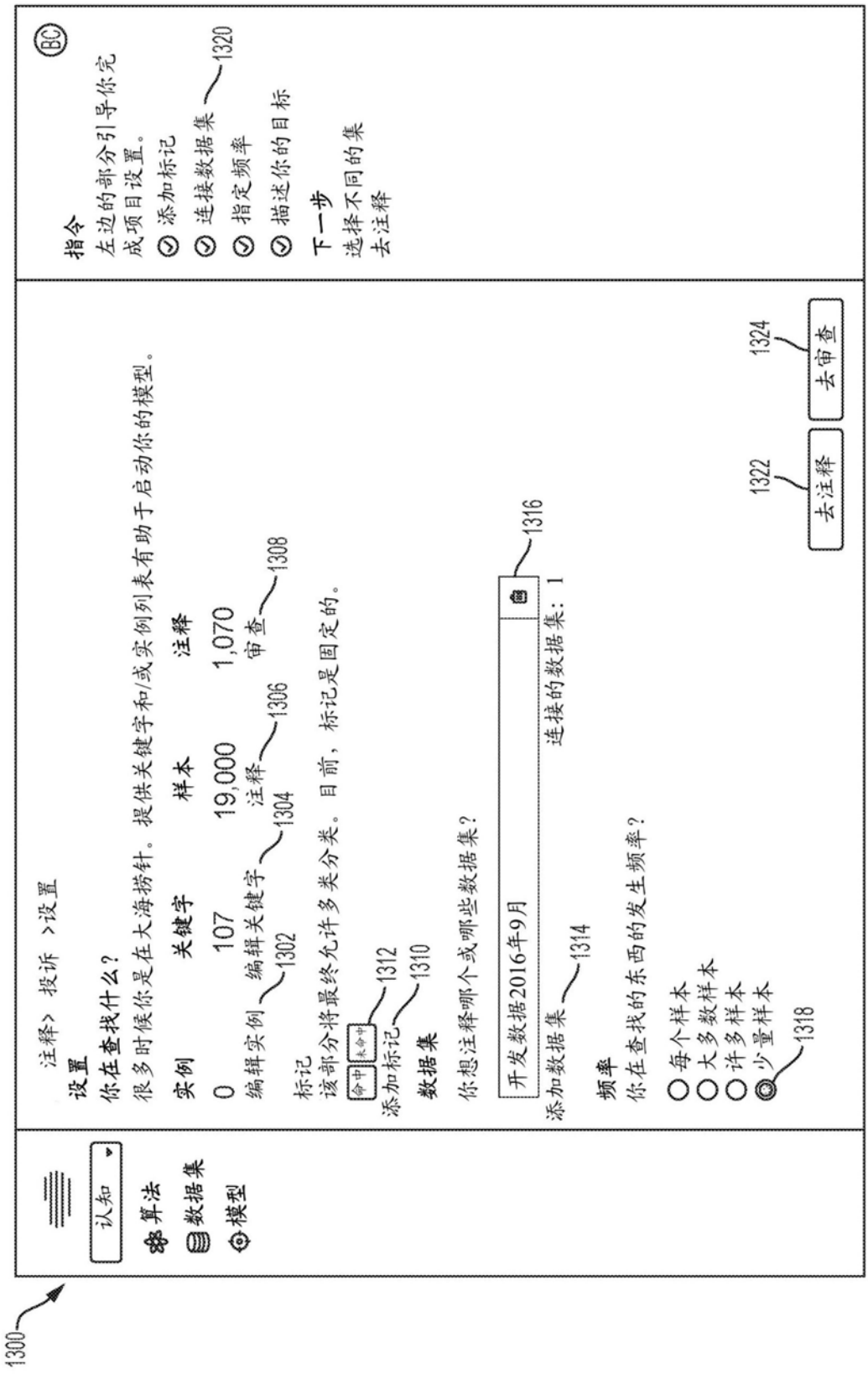

1300
认知
算法
数据集
模型
注释> 投诉 >设置
设置
你在查找什么?
很多时候你是在大海捞针。提供关键字和/或实例列表有助于启动你的模型。
实例
关键字
样本
注释
0
107
19,000
1,070
编辑实例
编辑关键字
注释
审查
1302
1304
1306
1308
标记
该部分将最终允许多类分类。目前，标记是固定的。
命中
未命中
添加标记
1312
数据集
1310
你想注释哪个或哪些数据集?
开发数据2016年9月
1316
添加数据集
1314
连接的数据集: 1
频率
你在查找的东西的发生频率?
每个样本
大多数样本
许多样本
少量样本
1318
1322
去注释
1324
去审查
BC
指令
左边的部分引导你完成项目设置。
添加标记
连接数据集
1320
指定频率
描述你的目标
下一步
选择不同的集
去注释

图13

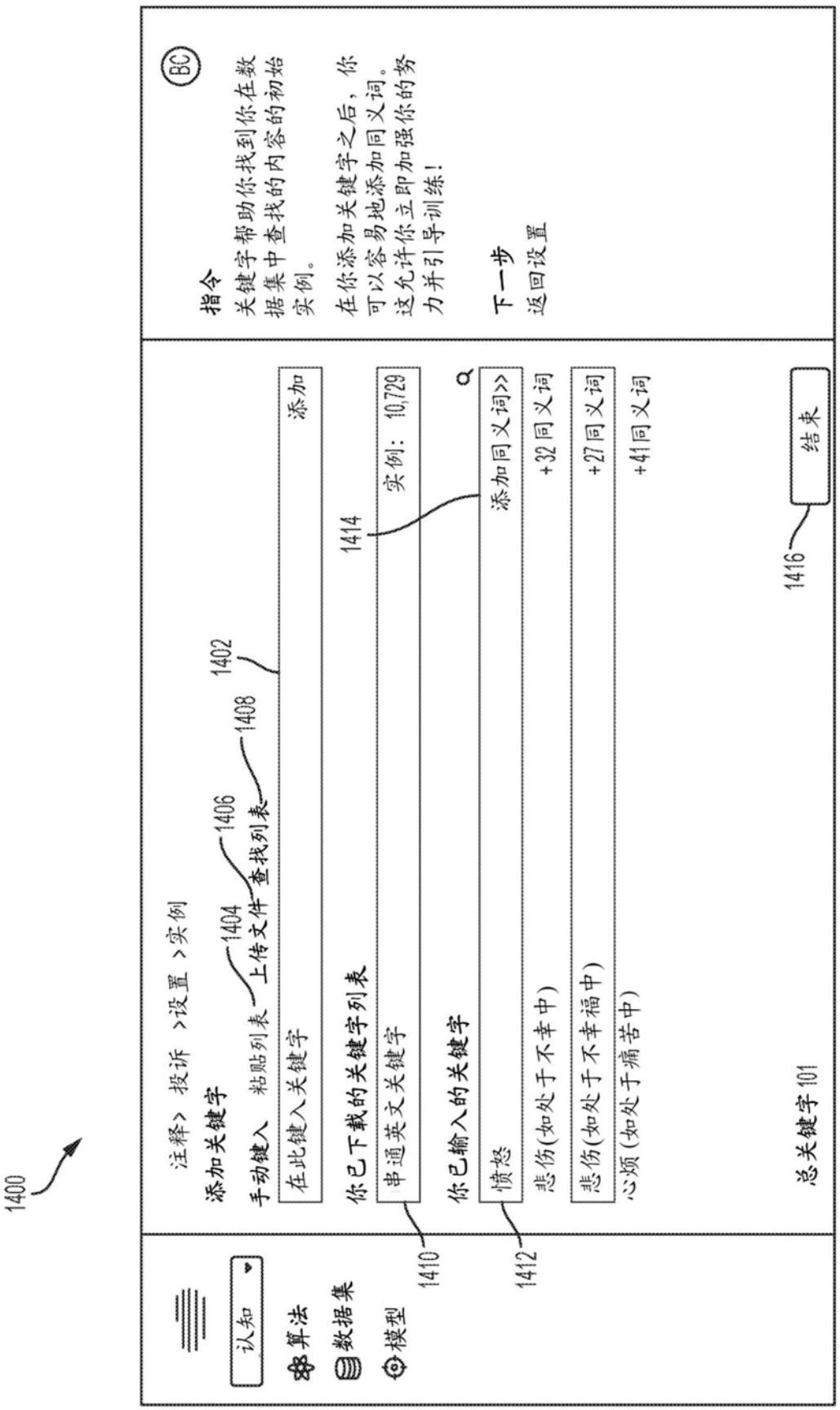

1400
注释> 投诉 >设置 >实例
添加关键字
手动键入
粘贴列表
上传文件
查找列表
1404
1406
1408
1402
在此键入关键字
添加
你已下载的关键字列表
1414
串通英文关键字
实例: 10,729
你已输入的关键字
添加同义词>>
愤怒
悲伤(如处于不幸中)
+32同义词
悲伤(如处于不幸福中)
+27同义词
心烦(如处于痛苦中)
+41同义词
总关键字101
1410
1412
1416
结束
认知
算法
数据集
模型
指令
关键字帮助你找到你在数据集中查找的内容的初始实例。
在你添加关键字之后，你可以容易地添加同义词。这允许你立即加强你的努力并引导训练！
下一步
返回设置
BC

图14

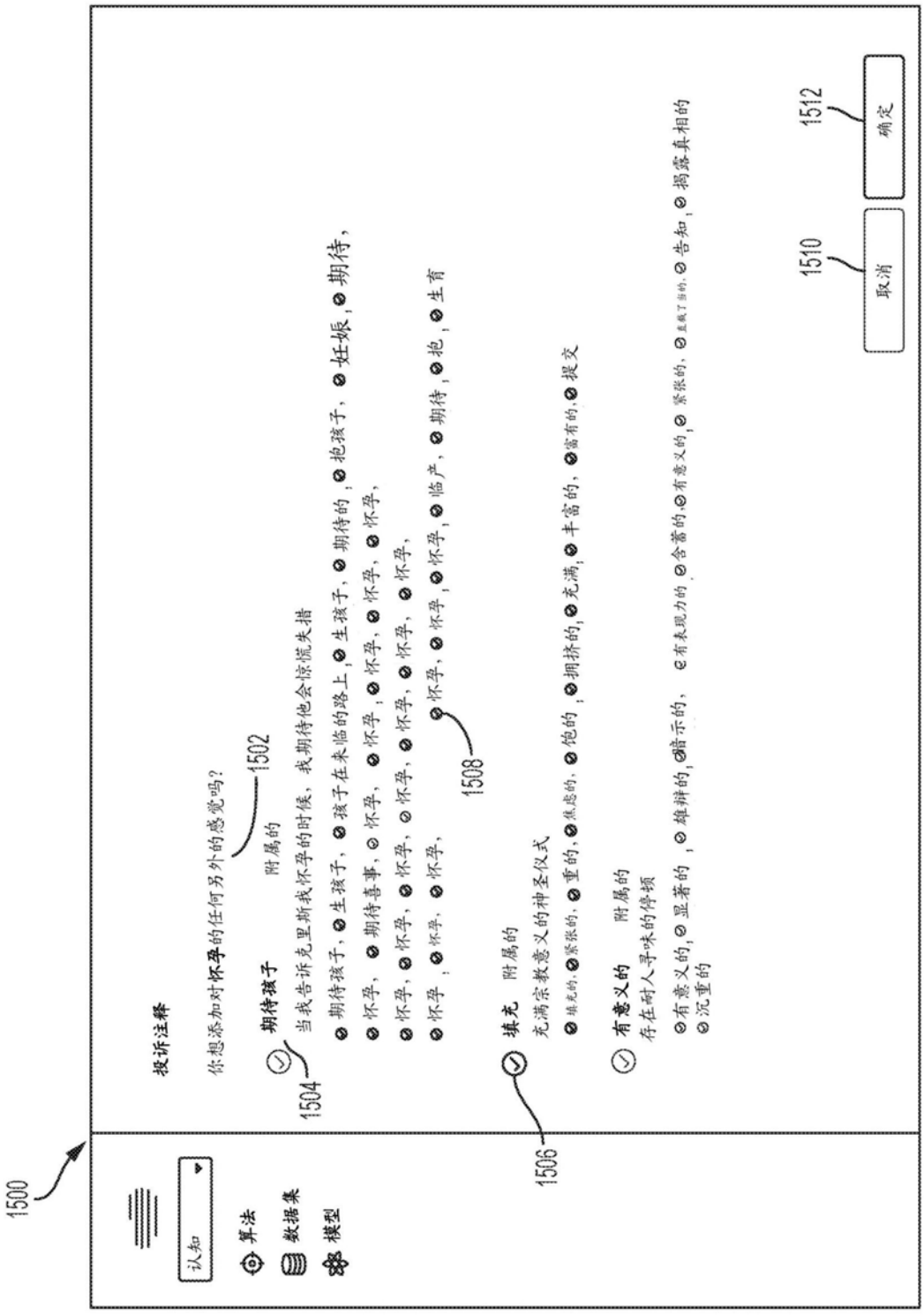

1500
认知
算法
数据集
模型
投诉注释
你想添加对怀孕的任何另外的感觉吗?
1502
期待孩子 附属的
当我告诉克里斯我怀孕的时候，我期待他会惊慌失措
1504
期待孩子，生孩子，孩子在未临的路上，生孩子，期待的，抱孩子，妊娠，期待，
怀孕，期待喜事，怀孕，怀孕，怀孕，怀孕，
怀孕，怀孕，怀孕，怀孕，怀孕，怀孕，怀孕，怀孕，
怀孕，怀孕，怀孕，
1508
怀孕，怀孕，怀孕，临产，期待，抱，生育
填充 附属的
充满宗教意义的神圣仪式
1506
重的，焦虑的，饱的，拥挤的，充满，丰富的，富有的，提交
有意义的 附属的
存在耐人寻味的停顿
有意义的，显著的，雄辩的，暗示的，有表现力的，含蓄的，有意义的，紧张的，告知，揭露真相的
沉重的
1510
取消
1512
确定

图15

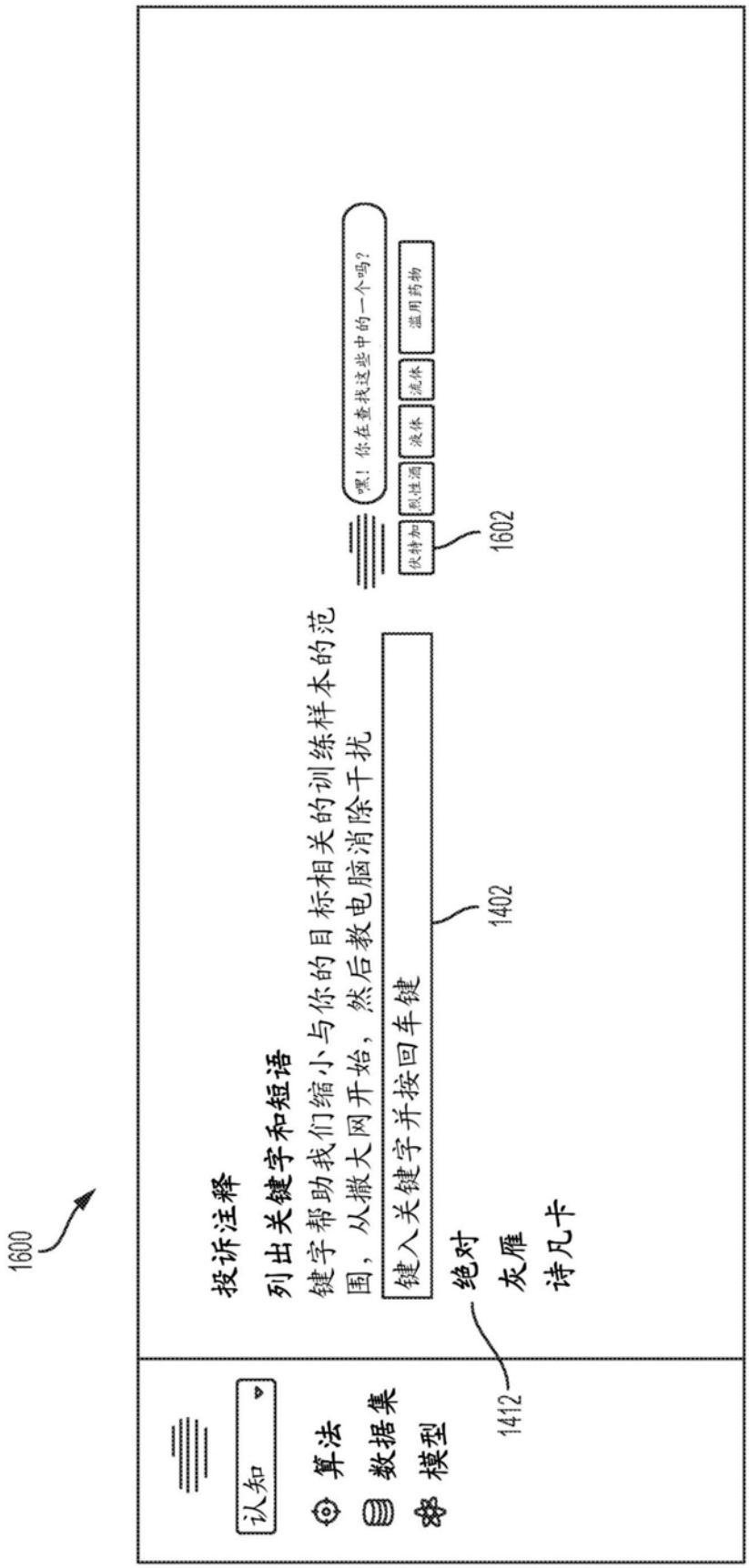

1600
投诉注释
列出关键字和短语
键字帮助我们缩小与你的目标相关的训练样本的范围，从撒大网开始，然后教电脑消除干扰
键入关键字并按回车键
1402
绝对
灰雁
诗凡卡
1412
认知
算法
数据集
模型
嘿！你在查找这些中的一个吗？
伏特加
烈性酒
液体
流体
滥用药物
1602

图16

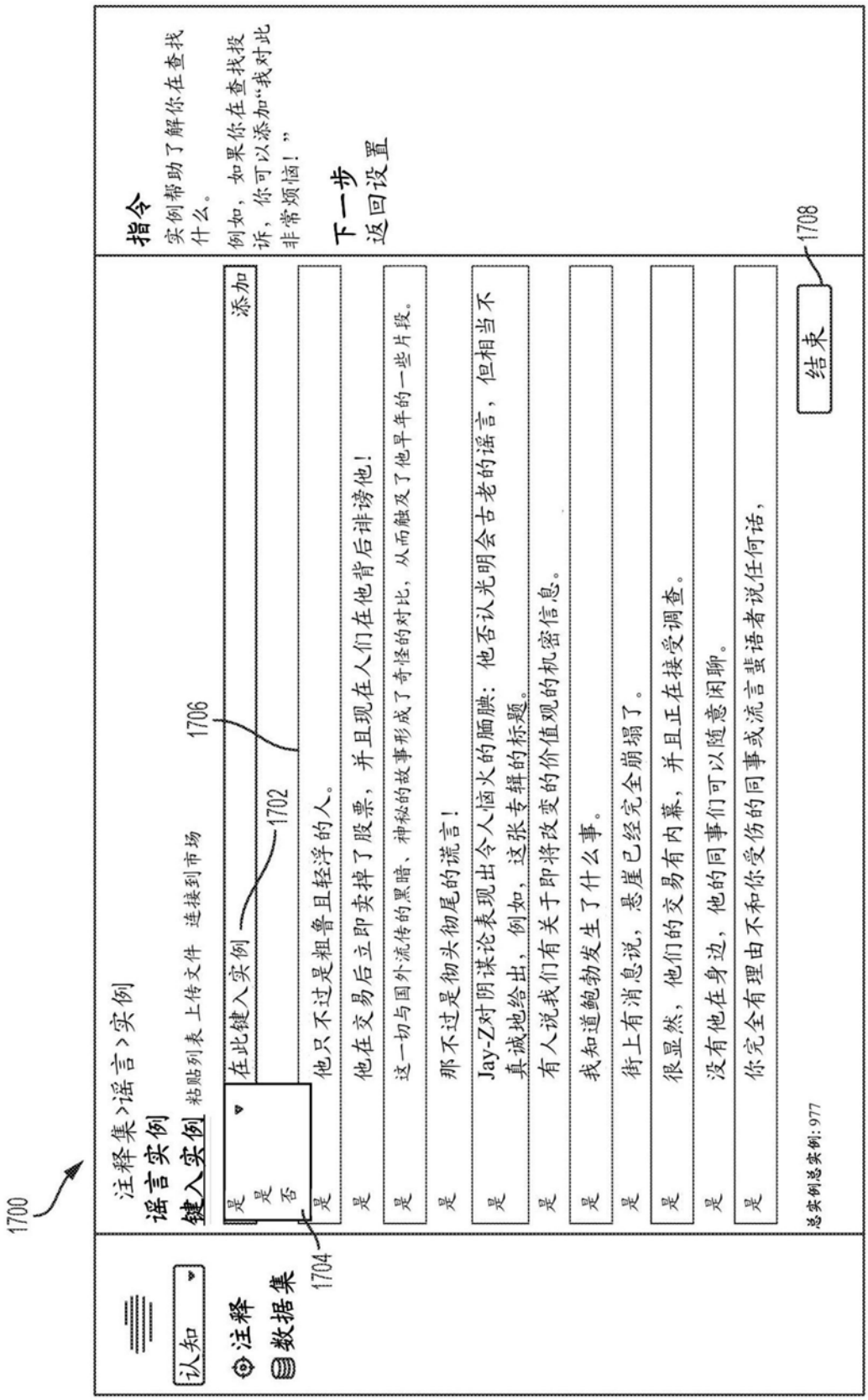

1700
注释集>谣言>实例
谣言实例
键入实例
粘贴列表 上传文件 连接到市场
1706
添加
1702
1704
是
是
否
是 他只不过是粗鲁且轻浮的人。
是 他在交易后立即卖掉了股票，并且现在人们在他背后诽谤他！
是 这一切与国外流传的黑暗、神秘的故事形成了奇怪的对比，从而触及了他早年的一些片段。
是 那不过是彻头彻尾的谎言！
是 Jay-Z对阴谋论表现出令人恼火的腼腆：他否认光明会古老的谣言，但相当不真诚地给出，例如，这张专辑的标题。
是 有人说我们有关于即将改变的价值观的机密信息。
是 我知道鲍勃发生了什么事。
是 街上有消息说，悬崖已经完全崩塌了。
是 很显然，他们的交易有内幕，并且正在接受调查。
是 没有他在身边，他的同事们可以随意闲聊。
是 你完全有理由不和你受伤的同事或流言蜚语者说任何话，
总实例总实例: 977
结束
1708
认知
注释
数据集
指令
实例帮助了解你在查找什么。
例如，如果你在查找投诉，你可以添加“我对此非常烦恼！”
下一步
返回设置

图17

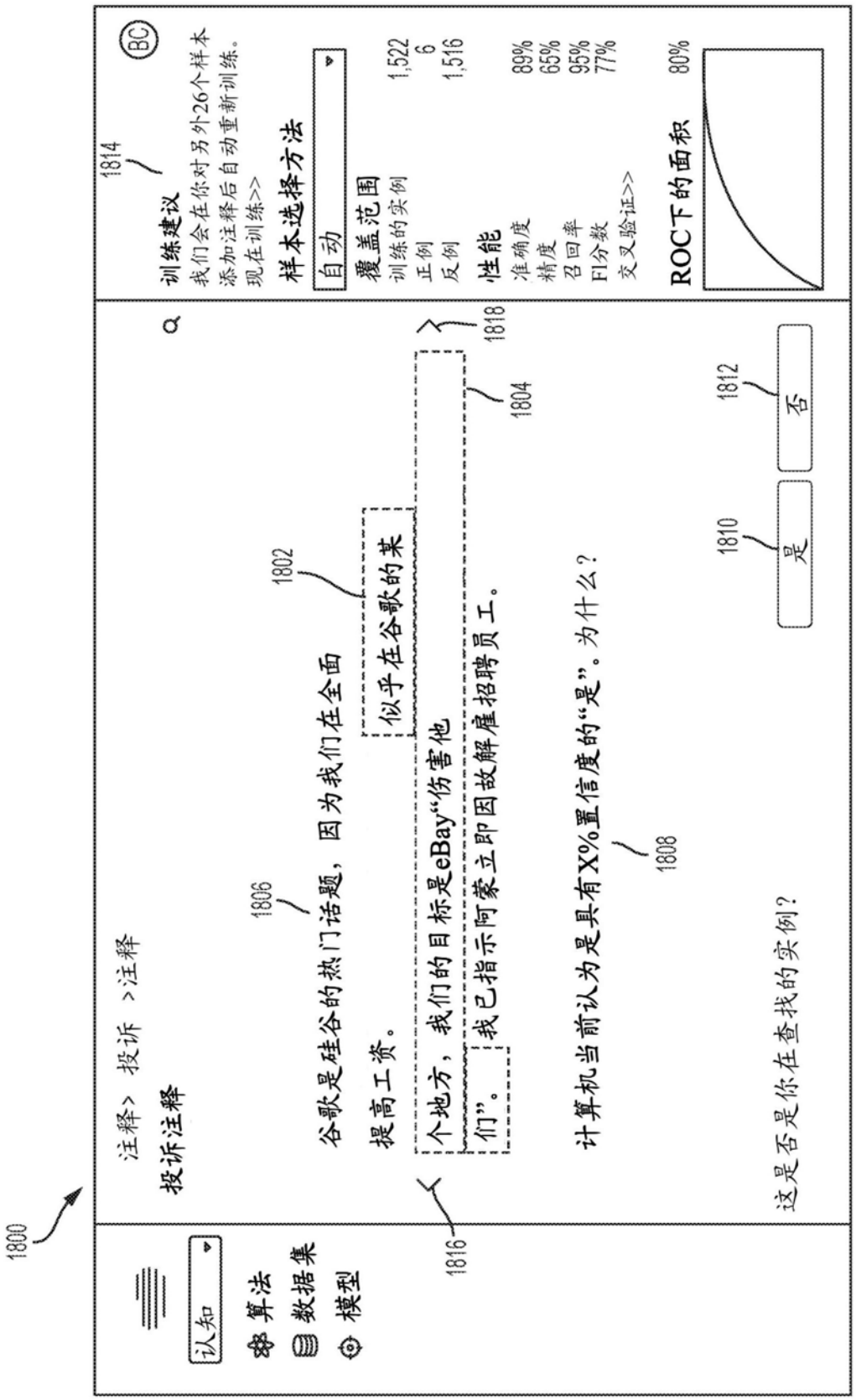

1800
注释> 投诉 >注释
投诉注释
认知
算法
数据集
模型
1816
谷歌是硅谷的热门话题，因为我们在全面
1806
提高工资。
1802
似乎在谷歌的某
个地方，我们的目标是eBay“伤害他
们”。 我已指示阿蒙立即因故解雇招聘员工。
1804
1818
计算机当前认为是具有X%置信度的“是”。为什么？
1808
这是否是你在查找的实例？
1810
是
1812
否
1814
训练建议
我们会在你对另外26个样本
添加注释后自动重新训练。
现在训练>>
样本选择方法
自动
覆盖范围
训练的实例 1,522
正例 6
反例 1,516
性能
准确度 89%
精度 65%
召回率 95%
F1分数 77%
交叉验证>>
ROC下的面积 80%
BC

图18

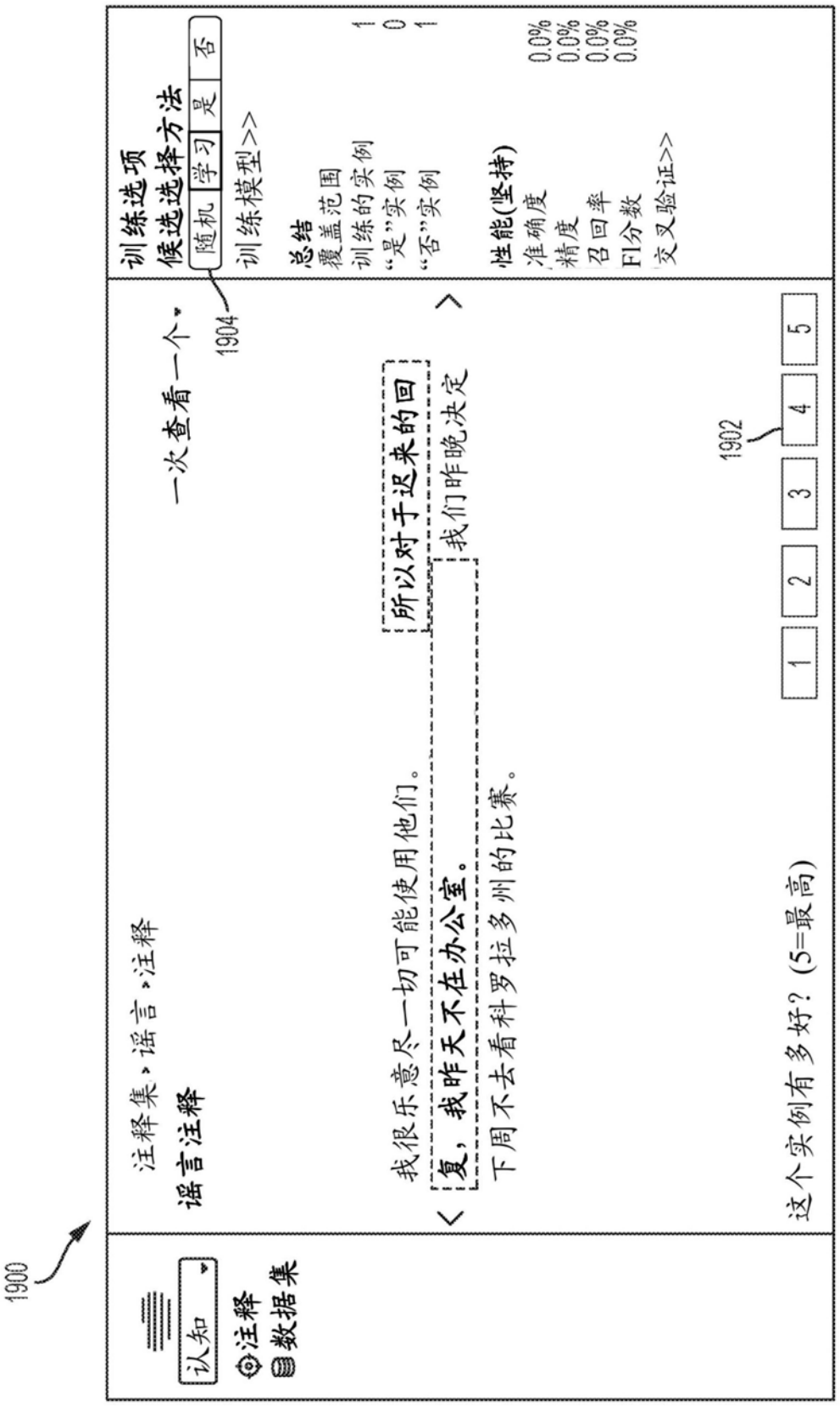

1900
认知
注释
数据集
注释集，谣言，注释
谣言注释
一次查看一个
1904
训练选项
候选选择方法
随机
学习
是
否
训练模型>>
总结
覆盖范围
训练的实例 1
“是”实例 0
“否”实例 1
性能(坚持)
准确度 0.0%
精度 0.0%
召回率 0.0%
F1分数 0.0%
交叉验证>>
我很乐意尽一切可能使用他们。
复，我昨天不在办公室。
下周不去看科罗拉多州的比赛。
所以对于迟来的回
我们昨晚决定
这个实例有多好？(5=最高)
1
2
3
4
5
1902

图19

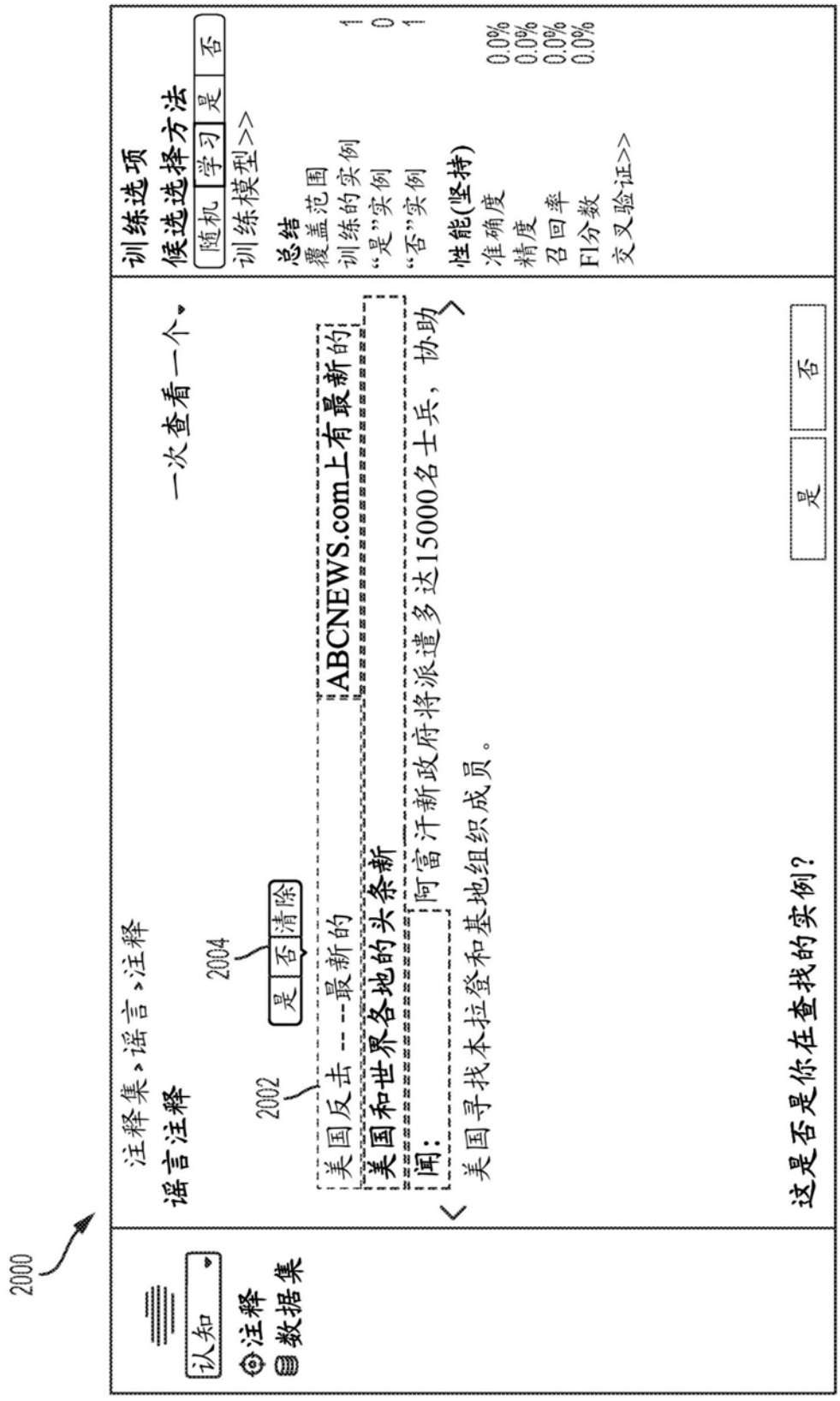

2000
认知
注释
数据集
注释集 › 谣言 › 注释
谣言注释
一次查看一个
2004
是
否
清除
2002
美国反击 -- --最新的
ABCNEWS.com上有最新的
美国和世界各地的头条新
闻：
阿富汗新政府将派遣多达15000名士兵，协助
美国寻找本拉登和基地组织成员。
这是否是你在查找的实例？
是
否
训练选项
候选选择方法
随机
学习
是
否
训练模型>>
总结
覆盖范围
训练的实例 1
“是”实例 0
“否”实例 1
性能(坚持)
准确度 0.0%
精度 0.0%
召回率 0.0%
F1分数 0.0%
交叉验证>>

图20

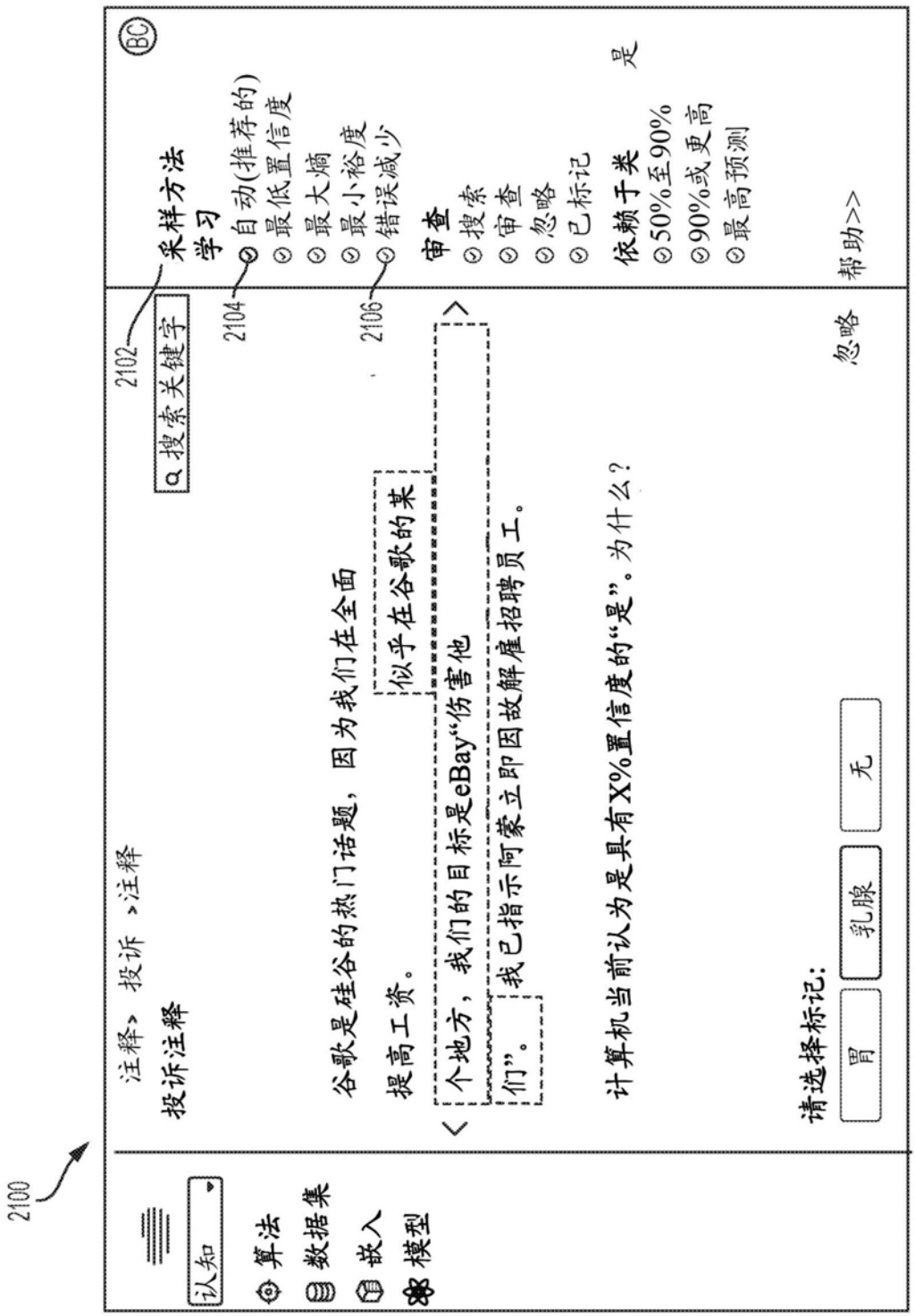

2100
认知
算法
数据集
嵌入
模型
注释› 投诉 ›注释
投诉注释
谷歌是硅谷的热门话题，因为我们在全面
提高工资。
个地方，我们的目标是eBay“伤害他
们”。
似乎在谷歌的某
我已指示阿蒙立即因故解雇招聘员工。
计算机当前认为是具有X%置信度的“是”。为什么？
请选择标记：
胃
乳腺
无
忽略
帮助>>
搜索关键字
2102
采样方法
学习
2104
自动(推荐的)
最低置信度
最大熵
最小裕度
2106
错误减少
审查
搜索
审查
忽略
已标记
依赖于类
是
50%至90%
90%或更高
最高预测
BC

图21

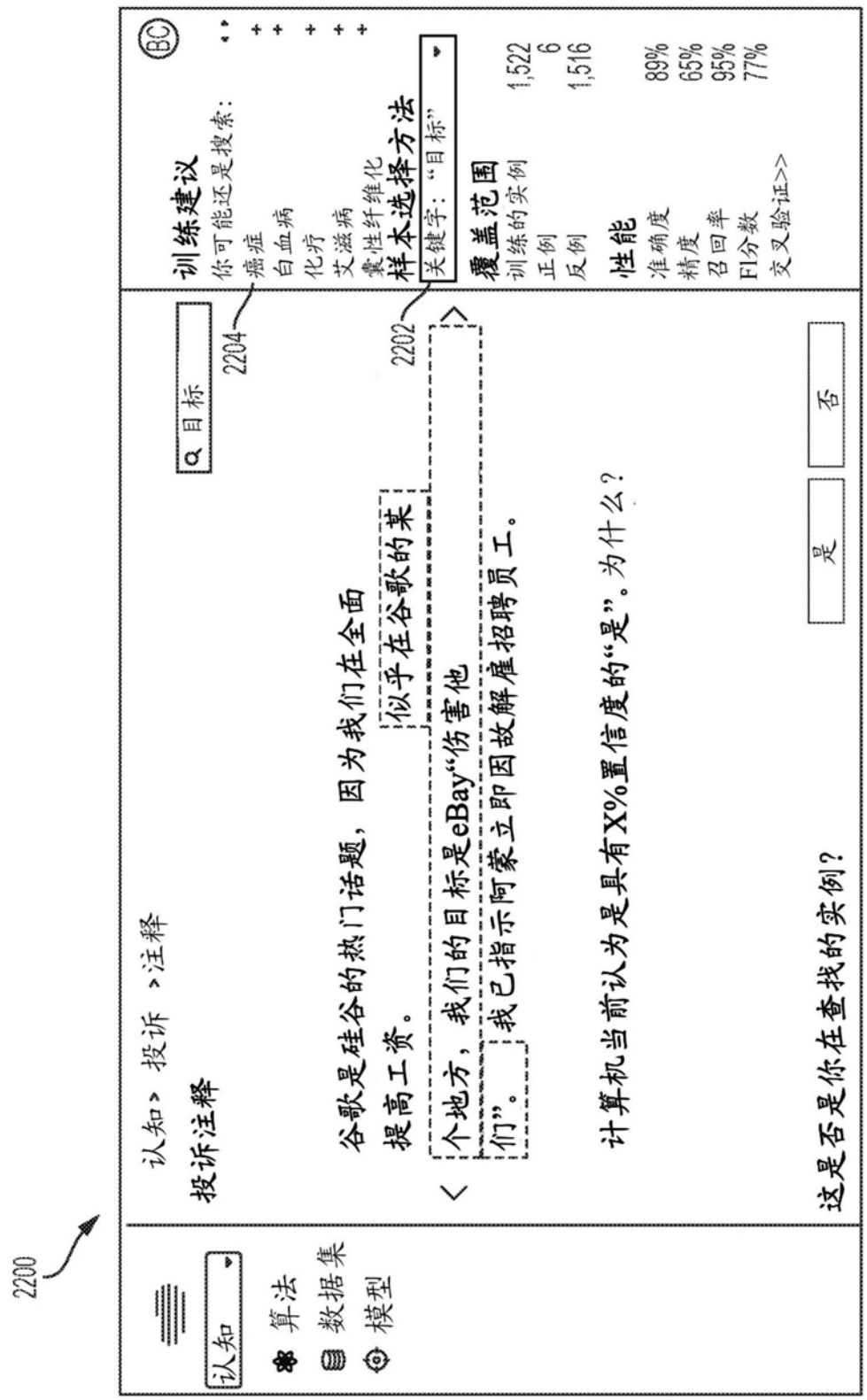

2200
认知
算法
数据集
模型
认知 › 投诉 › 注释
投诉注释
目标
2204
谷歌是硅谷的热门话题，因为我们在全面
提高工资。
似乎在谷歌的某
2202
个地方，我们的目标是eBay“伤害他
们”。 我已指示阿蒙立即因故解雇招聘员工。
计算机当前认为是具有X%置信度的“是”。为什么？
这是否是你在查找的实例？
是
否
BC
训练建议
你可能还是搜索：
癌症
白血病
化疗
艾滋病
囊性纤维化
样本选择方法
关键字：“目标”
覆盖范围
训练的实例 1,522
正例 6
反例 1,516
性能
准确度 89%
精度 65%
召回率 95%
F1分数 77%
交叉验证>>

图22

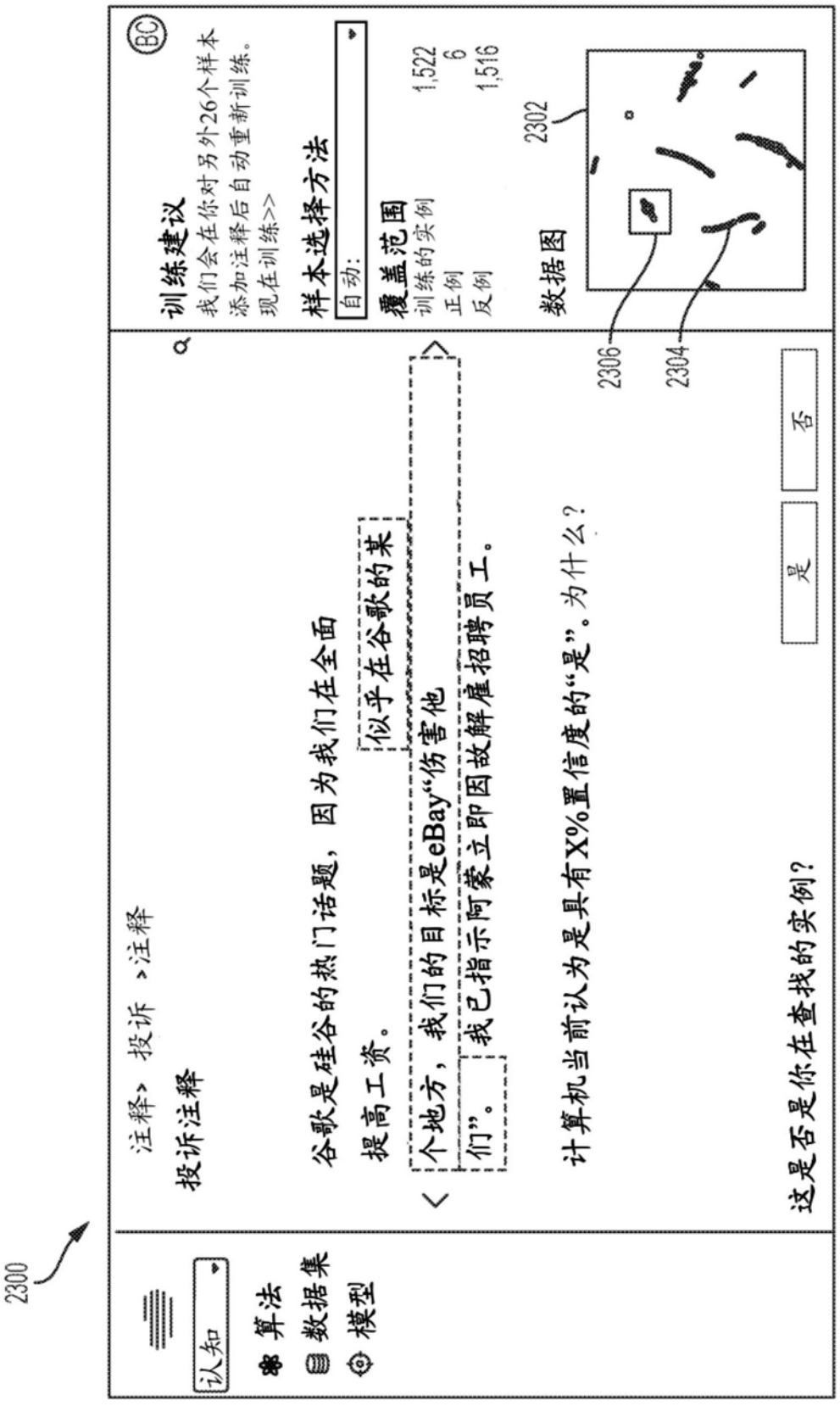

2300
认知
算法
数据集
模型
注释› 投诉 ›注释
投诉注释
谷歌是硅谷的热门话题，因为我们在全面
提高工资。
似乎在谷歌的某
个地方，我们的目标是eBay“伤害他
们”。 我已指示阿蒙立即因故解雇招聘员工。
计算机当前认为是具有X%置信度的“是”。为什么？
这是否是你在查找的实例？
是
否
BC
训练建议
我们会在你对另外26个样本
添加注释后自动重新训练。
现在训练>>
样本选择方法
自动:
覆盖范围
训练的实例 1,522
正例 6
反例 1,516
数据图
2302
2306
2304

图23

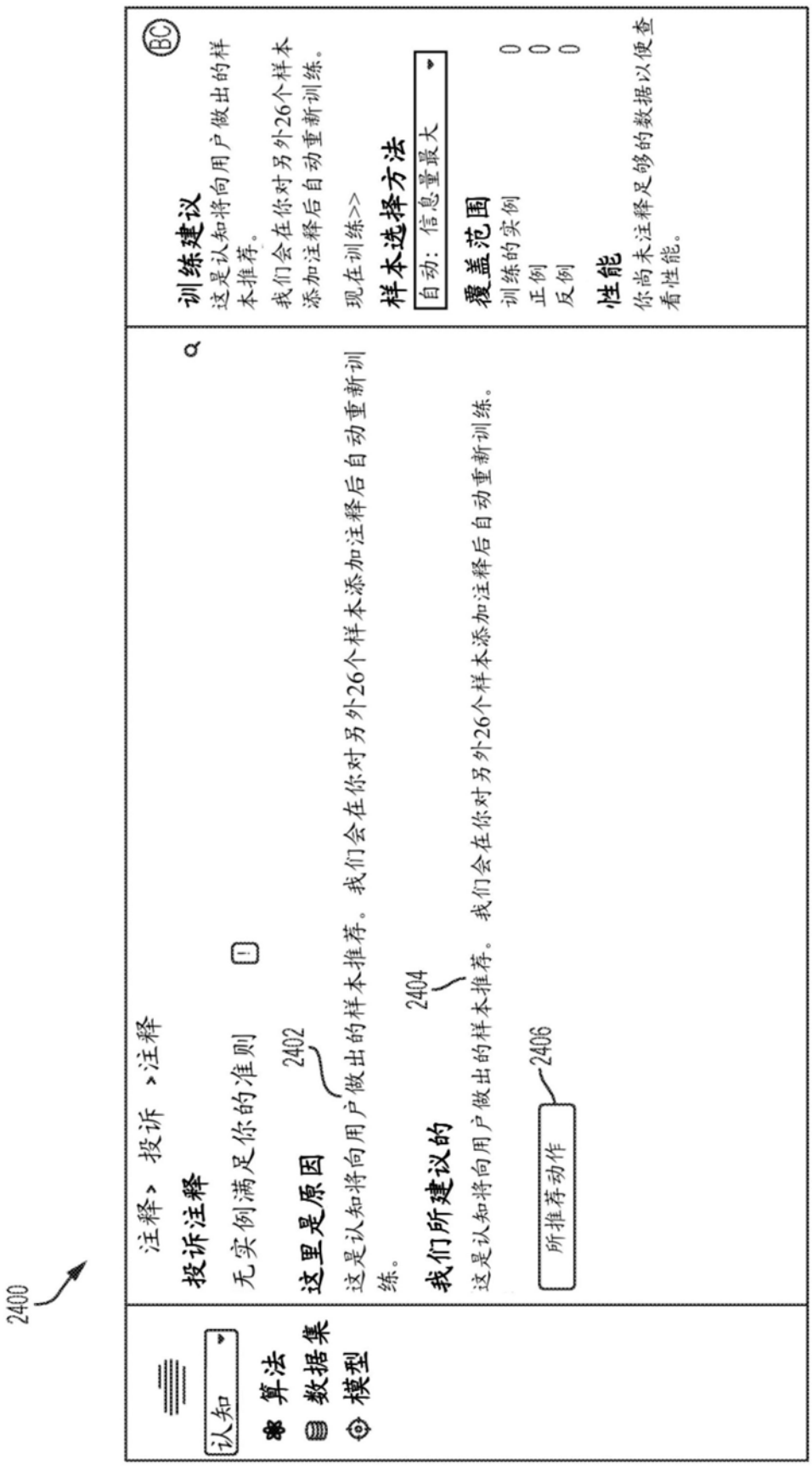

2400
注释› 投诉 ›注释
投诉注释
无实例满足你的准则
这里是原因
2402
这是认知将向用户做出的样本推荐。我们会在你对另外26个样本添加注释后自动重新训练。
我们所建议的
2404
这是认知将向用户做出的样本推荐。我们会在你对另外26个样本添加注释后自动重新训练。
所推荐动作
2406
认知
算法
数据集
模型
BC
训练建议
这是认知将向用户做出的样本推荐。
我们会在你对另外26个样本添加注释后自动重新训练。
现在训练>>
样本选择方法
自动：信息量最大
覆盖范围
训练的实例 0
正例 0
反例 0
性能
你尚未注释足够的数据以便查看性能。

图24

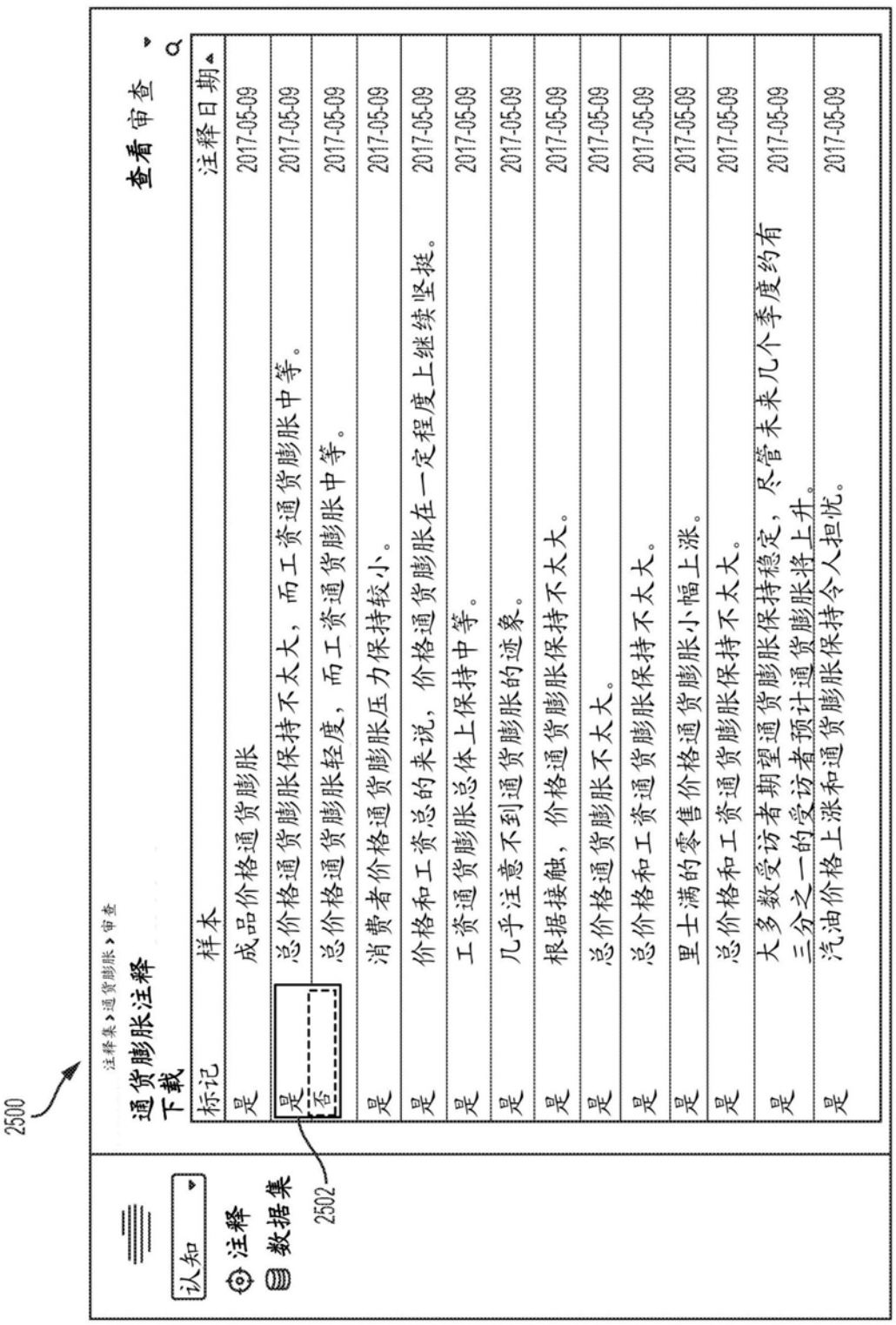

2500
注释集 › 通货膨胀 › 审查
通货膨胀注释
下载
查看 审查
认知
注释
数据集
2502
标记 样本 注释日期
是 成品价格通货膨胀 2017-05-09
是 总价格通货膨胀保持不太大，而工资通货膨胀中等。 2017-05-09
否 总价格通货膨胀轻度，而工资通货膨胀中等。 2017-05-09
是 消费者价格通货膨胀压力保持较小。 2017-05-09
是 价格和工资总的来说，价格通货膨胀在一定程度上继续坚挺。 2017-05-09
是 工资通货膨胀总体上保持中等。 2017-05-09
是 几乎注意不到通货膨胀的迹象。 2017-05-09
是 根据接触，价格通货膨胀保持不太大。 2017-05-09
是 总价格通货膨胀不太大。 2017-05-09
是 总价格和工资通货膨胀保持不太大。 2017-05-09
是 里士满的零售价格通货膨胀小幅上涨。 2017-05-09
是 总价格和工资通货膨胀保持不太大。 2017-05-09
是 大多数受访者期望通货膨胀保持稳定，尽管未来几个季度约有三分之一的受访者预计通货膨胀将上升。 2017-05-09
是 汽油价格上涨和通货膨胀保持令人担忧。 2017-05-09

图25

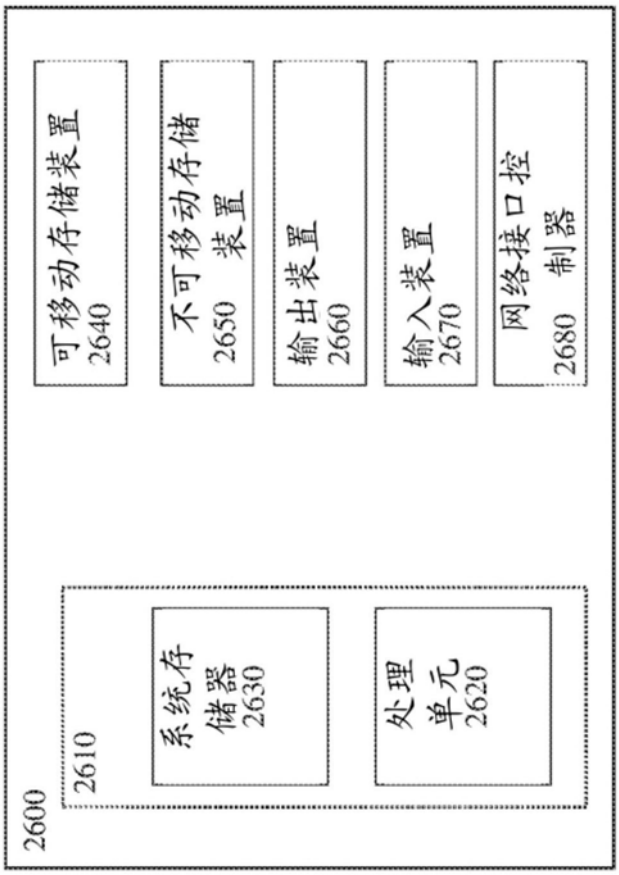

2600
2610
系统存储器 2630
处理单元 2620
可移动存储装置 2640
不可移动存储装置 2650
输出装置 2660
输入装置 2670
网络接口控制器 2680

图26