

(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
2020 年 7 月 23 日 (23.07.2020)



(10) 国际公布号
WO 2020/147238 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2019/088544

(22) 国际申请日: 2019 年 5 月 27 日 (27.05.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910049180.5 2019 年 1 月 18 日 (18.01.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY(SHENZHEN)CO.,LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 金戈(JIN, Ge); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳众鼎专利商标代理事务所(普通合伙)(SHENZHEN ZHONGDING INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市龙岗区龙城街道中心城清林路 546 号城投商务中心 4 层/B, Guangdong 518172 (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,

(54) Title: KEYWORD DETERMINATION METHOD, AUTOMATIC SCORING METHOD, APPARATUS AND DEVICE, AND MEDIUM

(54) 发明名称: 关键词的确定方法、自动评分方法、装置、设备及介质

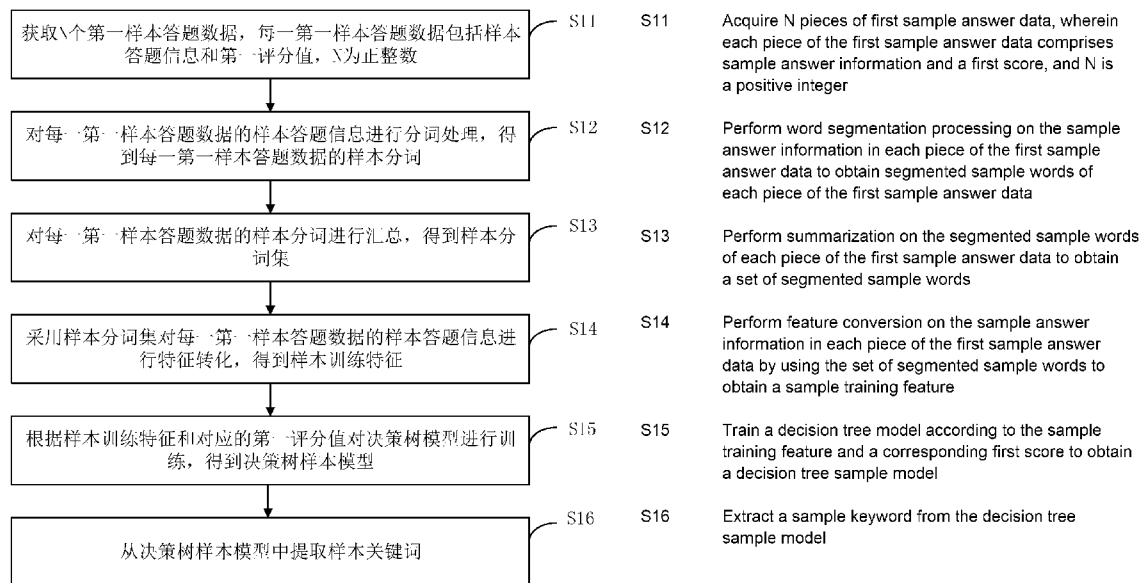


图 2

(57) Abstract: A keyword determination method, an automatic scoring method, apparatus and device, and a medium. The keyword determination method comprises: acquiring first sample answer data, and performing word segmentation processing and summarization on sample answer information in the first sample answer data to obtain a set of segmented sample words; performing feature conversion on the sample answer information to obtain a sample training feature; training a decision tree model according to the sample training feature and a first score to obtain a decision tree sample model; and extracting a sample keyword from the decision tree sample model. The automatic scoring method comprises: extracting a keyword from answer information to be scored to obtain a core keyword; and

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, 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))。

performing feature conversion on the core keyword by means of a target examination point to obtain an examination point feature to be scored, and then inputting same into a decision tree reference model to obtain an accurate score of the answer information to be scored. Thus, the keyword generalization capability and accuracy are improved; and efficient and accurate scoring of answer content of an examinee is realized.

(57) 摘要: 一种关键词的确定方法、自动评分方法、装置、设备及介质; 该关键词确定方法包括: 获取第一样本答题数据, 对第一样本答题数据中的样本答题信息进行分词处理和汇总, 得到样本分词集; 对样本答题信息进行特征转化, 得到样本训练特征; 根据样本训练特征和第一评分值对决策树模型进行训练, 得到决策树样本模型; 从决策树样本模型中提取样本关键词; 该自动评分方法包括: 从待评分答题信息中提取关键词, 得到核心关键词; 通过目标考点对核心关键词进行特征转化, 得到待评分考点特征后输入到决策树参考模型中, 得到待评分答题信息的准确分值; 提高了关键词的泛化能力和准确性, 并实现了对考生答题内容的高效准确评分。

关键词的确定方法、自动评分方法、装置、设备及介质

本申请以 2019 年 1 月 18 日提交的申请号为 201910049180.5，名称为“一种关键词的确定方法、自动评分方法、装置、设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请涉及智能决策领域，尤其涉及一种关键词的确定方法、自动评分方法、装置、设备及存储介质。

背景技术

随着社会的发展，竞争越来越大，而考试逐渐成为了衡量一个人学到多少知识掌握多少技能的一项常规手段，因此一系列对考生的答题内容进行评分的系统也随着考试的盛行而发展起来。随着计算机技术的发展，对于考生客观题的答题内容已经可实现全自动的计算机在线阅卷并实时评分，但是基于主观题存在一定的随机性以及记忆成分，若根据同样的评分方法，采用计算机对考生主观题的答题内容进行评分，则极易发生误判或者产生误差。此外，如果采用人工阅卷，当考生人群数量较大时，人工评分的工作量就会变得非常大，操作也会变得非常困难。目前，对考生主观题的答题内容进行评分的方法通常是预先通过评分规则信息，人工确立考点内容以及相关关键词，然后根据考点内容以及相关关键词，通过正则匹配方法识别答题内容，对考生的答题内容进行评分。然而，只通过评分规则信息而不考虑其它考生对同一主观题的答题情况，所确定的考点以及相关关键词不但泛化能力较低，且准确性不高。因此也会导致在后续对考生的答题内容进行评分时，得出的最终评分结果出现偏差，而无法体现考生的真实水平。

发明内容

本申请实施例提供一种关键词的确定方法、装置、设备及存储介质，以解决关键词泛化能力低、准确性不高的问题。

本申请实施例提供一种自动评分方法、装置、设备及存储介质，以解决无法对考生的答题内容进行高效准确评分的问题。

一种关键词的确定方法，包括：

获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

从所述决策树样本模型中提取样本关键词。

一种自动评分方法，包括：

获取待评分答题信息；

对所述待评分答题信息进行关键词提取，得到核心关键词；

采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

一种关键词的确定装置，包括：

第一样本答题数据获取模块，用于获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

分词处理模块，用于对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

分词汇总模块，用于对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

样本特征转化模块，用于采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

决策树样本模型训练模块，用于根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

样本关键词提取模块，用于从所述决策树样本模型中提取样本关键词。

一种自动评分装置，包括：

待评分答题信息获取模块，用于获取待评分答题信息；

关键词提取模块，用于对所述待评分答题信息进行关键词提取，得到核心关键词；

待评分考点特征转化模块，用于采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

输入模块，用于将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现上述关键词的确定方法的步骤，或所述处理器执行所述计算机可读指令时实现上述自动评分方法的步骤。

一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令被处理器执行时实现上述关键词的确定方法的步骤，或所述计算机可读指令被处理器执行时实现上述自动评分方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

附图说明

为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 是本申请一实施例中关键词的确定方法或自动评分方法的一应用环境示意图；

图 2 是本申请一实施例中关键词的确定方法的一示例图；

图 3 是本申请一实施例中关键词的确定方法的另一示例图；

图 4 是本申请一实施例中关键词的确定装置的一原理框图；

图 5 是本申请一实施例中关键词的确定装置的另一原理框图；

图 6 是本申请一实施例中自动评分方法的一示例图；

图 7 是本申请一实施例中自动评分方法的另一示例图；

图 8 是本申请一实施例中自动评分方法的另一示例图；

图 9 是本申请一实施例中自动评分装置的一原理框图；

图 10 是本申请一实施例中计算机设备的一示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请实施例提供一关键词的确定方法，该关键词的确定方法可应用如图 1 所示的应用环境中。具体地，该关键词的确定方法应用在关键词的确定系统中，该关键词的确定系统包括如图 1 所示的客户端和服务端，客户端与服务端通过网络进行通信，用于解决根据评分规则信息所确定的考点关键词泛化能力低、准确性不高的问题。其中，客户端又称为用户端，是指与服务端相对应，为客户提供本地服务的程序。客户端可安装在但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备上。服务端可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在一实施例中，如图 2 所示，提供一种关键词的确定方法，以该方法应用在图 1 中的服务端为例进行说明，包括如下步骤：

S11：获取 N 个第一样本答题数据，每一第一样本答题数据包括样本答题信息和第一评分值，N 为正整数。

其中，第一样本答题数据指考生的答题数据。每一第一样本答题数据包括样本答题信息和对应的第一评分值，即第一样本答题数据中包括样本答题信息和该样本答题信息进行初步评分后所得的对应的第一评分值。其中，样本答题信息指从评分系统的答题文本上获取的某一主观题的考生的答题信息。可选地，该第一样本答题数据可以从一评分系统中获取。该评分系统可以对样本答题信息进行初步评分，得到第一评分值。优选地，还可预先获取考生在纸质答卷上写入的答题信息，然后对考生在纸质答卷上写入的答题信息进行扫描识别，生成对应的答题文本提交到评分系统上，得到样本答题信息。第一评分值指通过人工评分或者计算机评分的方式，预先对样本答题信息进行初步评分后所得的评分值。该第一样本答题数据还可以是将纸质答卷上写入的答题信息以及人工评分进行扫描识别后获取得到。另外地，第一样本答题数据可以包括一个样本答题信息和对该样本答题信息进行初步评分后所得的一个第一评分值，还可以包括多个样本答题信息和对每一样本答题信息进行初步评分后所得的对应的多个第一评分值。

获取的第一样本答题数据的数量为 N 个，其中，N 为正整数。而 N 的具体数值可以根据实际需要进行设定。N 的数值越高，后续样本关键词提准的准确性会越高，然而提取效率会有所下降，可以在准确度和效率上进行综合考虑进行对 N 的选取。

S12：对每一第一样本答题数据的样本答题信息进行分词处理，得到每一第一样本答题数据的样本分词。

其中，样本分词指对每一第一样本答题数据的样本答题信息进行分词处理后，所得到的一个个独立的分词。具体地，对每一第一样本答题数据的样本答题信息进行分词处理包括：先采用分词算法对每一第一样本答题数据的样本答题信息进行词汇拆分。可选地，分词算法可以采用基于字符串匹配的分词算法，或者采用基于理解的分词算法，还可以采用基于统计的分词算法实现。优选地，还可通过 Java 语言的拆分函数，或者通过将样本答题信息导入计算机的 EXCEL 或 PPT 等具有自动拆分字符功能的软件实现对每一第一样本答题数据的样本答题信息的自动拆分。然后通过采用 Java 语言的正则表达式对拆分后的样本答题信息进行筛选，过滤掉一些不具有意义的特定词语，例如：助词、语气词或者连接

词等；最后得到每一第一样本答题数据的样本分词。

S13：对每一第一样本答题数据的样本分词进行汇总，得到样本分词集。

其中，样本分词集指对每一第一样本答题数据的样本分词进行统一汇总后所得到的分词集。具体地，获取每一第一样本答题数据的样本分词，然后对获取的每一第一样本答题数据的样本分词进行汇总，得到样本分词集。优选地，若每一第一样本答题数据中包含多个样本答题信息，则在每一第一样本答题数据的样本分词进行汇总时，需以每一第一样本答题数据的样本答题信息为单位进行汇总，即得到的样本分词集是对应每一样本答题信息的。

具体地，对每一第一样本答题数据的样本分词进行汇总包括：获取每一第一样本答题数据中样本答题信息的样本分词，然后按照从小到大的排列顺序给获取到的每一样本答题信息中的所有样本分词赋予对应的标识号，最后得到按照从小到大的排列顺序分布的样本分词集。例如：样本分词集为 $E = \{e_1, e_2, e_3, \dots, e_r\}$ ，其中， $e_1, e_2, e_3, \dots, e_r$ 表示该样本分词集所包含的样本分词， $1, 2, 3, \dots, r$ 表示每个样本分词所对应的标识号。

优选地，若获取的每一第一样本答题数据的样本分词中存在重复的样本分词，则在每一第一样本答题数据的样本分词进行汇总前，预先对第一样本答题数据的样本分词进行去重处理，然后再将去重后的每一第一样本答题数据的样本分词进行汇总，得到样本分词集。具体地，可采用 Count 函数、Editor 编辑器或者 R 语言对每一第一样本答题数据的样本分词进行去重处理。优选地，也可直接将每一第一样本答题数据的样本分词导入计算机的 EXCEL 表格中，通过 EXCEL 的高级筛选功能实现对样本分词的自动去重。

S14：采用样本分词集对每一第一样本答题数据的样本答题信息进行特征转化，得到样本训练特征。

其中，样本训练特征指每一第一样本答题数据的样本答题信息进行特征转化后所输出的结果。具体地，通过建立词袋模型的方式，采用样本分词集对每一第一样本答题数据的样本答题信息进行特征转化，得到样本训练特征。在本实施例，词袋模型指衡量每一第一样本答题数据的样本答题信息出现在样本分词集中的具体情况。具体地，词袋模型的建立可通过使用 SKLearn 中的 CountVectorizer 实现。其中，CountVectorizer 属于特征数值计算的常见方法。对于每一个训练文本，CountVectorizer 只考虑每种词汇在该训练文本中出现的频率，CountVectorizer 可通过计数来将一个文档转换为向量，并将提取的词汇进行训练，并生成一个 CountVectorizerModel 用于存储相应的词汇向量空间。

具体地，采用样本分词集对每一第一样本答题数据的样本答题信息进行特征转化包括：先以样本分词集的数量为基准，建立一组词向量，然后通过正则匹配法，将每一第一样本答题数据的样本答题信息与该样本分词集中的所有样本分词进行匹配；若第一样本答题数据的样本答题信息与样本分词集中的样本分词匹配成功，则词向量中对应的元素值为 1，若第一样本答题数据的样本答题信息与样本分词集中的样本分词不匹配，则词向量中对应的元素值为 0，最后得到一组由若干个 1 和 0 组成的词向量，即样本训练特征。

示例性地，若获有包含 B_1, B_2, C_1, C_2, C_3 五个样本分词的一样本分词集和 B、C 两个样本答题信息；样本答题信息 B 里面包含 B_1, B_2 两个分词，样本答题信息 C 里面包含 C_1, C_2, C_3 三个分词；则采用该样本分词集对样本答题信息 B 进行特征转化后，得到样本训练特征为 $[1, 1, 0, 0, 0]$ ；采用该样本分词集对样本答题信息 C 进行特征转化后，得到样本训练特征为 $[0, 0, 1, 1, 1]$ 。

其中，正则匹配法是用于测试正则表达式的应用，其中，正则表达式是对字符串或特殊字符操作的一种逻辑公式，指用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种文本模式，模式描述在搜索文本时要匹配的一个或多个字符串。

S15：根据样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样

本模型。

其中，决策树样本模型指以词袋模型的特征为基础，根据样本训练特征和对应的第一评分值对决策树模型进行训练后所生成的样本模型。具体地，决策树样本模型的建立过程包括：将样本训练特征和对应的第一评分值输入到决策树模型中，然后通过使用 C4.5 算法对决策树模型进行训练，生成训练后的决策树样本模型。C4.5 算法是一系列用在机器学习和数据挖掘的分类问题中的算法。C4.5 算法的目标是监督学习，给定一个数据集，其中的每一个元组都能用一组属性值来描述，每一个元组属于一个互斥的类别中的某一类。C4.5 算法可通过学习，找到一个从属性值到类别的映射关系，并且这个映射能用于对新的类别未知的实体进行分类。

进一步地，在建立决策树样本模型前，还需确认决策树样本模型的大小，其中，决策树样本模型的大小是由决策树的深度和节点样本数决定的。可选地，在本实施中，为了保证建立的决策树样本模型不会出现过渡拟合现象，以及保证决策树样本模型的精度，将决策树的最大深度设为 5，最小叶节点样本数设为 50，分类标准为熵。

S16：从决策树样本模型中提取样本关键词。

其中，样本关键词指决策树样本模型的每一输出节点上所对应的特征属性值。具体地，样本关键词的提取也称决策树样本模型的特征值提取。由于决策树样本模型的每一特征属于决策样本模型中的决策属性，因此决策树样本模型的每一特征值对应的是决策样本模型中决策属性的分支。可以理解地，决策树样本模型中每个分支的输出节点都有对应的样本关键词。

具体地，从决策树样本模型中提取样本关键词可通过先将决策树样本模型读作为一个 sourceable 对象，然后通过 tosource 方法把决策树样本模型代码化，再通过分析代码结构得到该决策树样本模型所输出的样本关键词，最后将样本关键词提取出来。

在本实施例中，通过获取 N 个第一样本答题数据，每一第一样本答题数据包括样本答题信息和第一评分值，对每一第一样本答题数据的样本答题信息进行分词处理，得到样本分词集，然后采用样本分词集对每一第一样本答题数据的样本答题信息进行特征转化，得到样本训练特征，再根据样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型，最后从决策树样本模型中提取样本关键词，不但可以提高考点关键词的泛化能力和准确性，保证关键词更加全面，也可以提高后续评分的准确性。

在一实施例中，如图 3 所示，在从决策树样本模型中提取样本关键词之后，该关键词的确定方法还包括如下步骤：

S17：获取评分规则信息，评分规则信息包括预设考点和每一预设考点对应的预设关键词。

其中，评分规则信息指业务方提供的基本评分依据，包括预设考点和每一预设考点对应的预设关键词。预设考点指业务方提供的评判考生的答题信息是否正确的知识点。预设考点包括评判考生答错的错误考点和评判考生答对的正确考点。可以理解地，该评分规则信息为一个初步的评分标准，可能存在关键词不够准确或者不够全面的问题。可选地，预设考点可以为一个词语、一句话、或者一段话。此外，在本实施中，为了便于区分不同的预设考点，还可预先给每一预设考点赋予不同的标识。具体地，每一预设考点对应的标识可以用阿拉伯数字、英文大写字母或英文小写字母中的至少一种表示。每一预设考点中都包含对应的预设关键词，预设关键词指从预设考点中提取出来的、可直接用于规则量化的词语。可以理解地，一个预设考点中至少包含一个预设关键词。例如：预设考点 1 为：杜甫是唐代伟大的现实主义诗人；则预设考点 1 所对应的预设关键词可以为“杜甫”、“唐代”、“现实主义”和“诗人”。

S18：从样本关键词中去除和预设关键词重复的关键词，得到目标关键词。

其中，目标关键词指从样本关键词中提取出来的不同于预设关键词的关键词。具体地，

从样本关键词中去除和预设关键词重复的关键词可通过使用 C++ 中的字符比较函数，将样本关键词与预设关键词进行一一比较，然后根据比较结果，去除掉与预设关键词相同的样本关键词，最后将剩下的与预设关键词不同的样本关键词提取出来，作为目标关键词。

S19: 发送目标关键词至客户端，获取客户端根据目标关键词返回的考点标签。

其中，考点标签指根据预设考点，给获取目标关键词赋予对应标识号的标签。具体地，将目标关键词发送至客户端后，用户可对获取的目标关键词进行分析，然后根据预设考点，给每一目标关键词赋予与对应的预设考点相同的标识号，得到考点标签发送至服务端。优选地，还可以将每一目标关键词对应的考点标签统一生成考点标签文本后再发送至服务端。

S20: 根据考点标签将每一目标关键词加入到对应的预设考点中，得到目标考点。

其中，目标考点指加入目标关键词后的考点。具体地，服务端接收从客户端发送的考点标签文本后，根据考点标签文本中每一目标关键词所对应的标识号，将每一目标关键词加入到对应的相同标识号的预设考点中。可以理解地，目标考点所包含的关键词比预设考点所包含的关键词更丰富更全面。

示例性地，若有预设考点 1 和预设考点 2，预设考点 1 包括三个预设关键词分别为 a_1, a_2, a_3 ，预设考点 2 包括三个预设关键词分别为 b_1, b_2, b_3 ，根据步骤 S18 获取的目标关键词为 a_4, a_5, b_4, b_5 ，将目标关键词 a_4, a_5 赋予考点标签为 1，将目标关键词 b_4, b_5 赋予考点标签为 2；则根据考点标签将 a_4, a_5 加入到预设考点 1 中，将 b_4, b_5 加入到预设考点 2 中；最后得到目标考点 1 所包含的关键词为 a_1, a_2, a_3, a_4, a_5 ，目标考点 2 所包含的关键词为 b_1, b_2, b_3, b_4, b_5 。

在本实施例中，通过获取评分规则信息，评分规则信息包括预设考点和每一预设考点对应的预设关键词，从样本关键词中去除和预设关键词重复的关键词，得到目标关键词，发送目标关键词至客户端，然后获取客户端根据目标关键词返回的考点标签，最后根据考点标签将每一目标关键词加入到对应的预设考点中，得到目标考点；进一步丰富了根据评分规则信息所确定的考点所包含的关键词。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

在一实施例中，提供一种关键词的确定装置，该关键词的确定装置与上述实施例中关键词的确定方法一一对应。如图 4 所示，该关键词的确定装置包括第一样本答题数据获取模块 11、分词处理模块 12、分词汇总模块 13、样本特征转化模块 14、决策树样本模型训练模块 15 和样本关键词提取模块 16。各功能模块详细说明如下：

第一样本答题数据获取模块 11，用于获取 N 个第一样本答题数据，每一第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

分词处理模块 12，用于对每一第一样本答题数据的样本答题信息进行分词处理，得到每一第一样本答题数据的样本分词；

分词汇总模块 13，用于对每一第一样本答题数据的样本分词进行汇总，得到样本分词集；

样本特征转化模块 14，用于采用样本分词集对每一第一样本答题数据的样本答题信息进行特征转化，得到样本训练特征；

决策树样本模型训练模块 15，用于根据样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

样本关键词提取模块 16，用于从决策树样本模型中提取样本关键词。

优选地，如图 5 所示，该关键词的确定装置，还包括：

评分规则信息获取模块 17，用于获取评分规则信息，评分规则信息包括预设考点和每一预设考点对应的预设关键词；

重复关键词去除模块 18，用于从样本关键词中去除和预设关键词重复的关键词，得到目标关键词；

考点标签获取模块 19，用于发送目标关键词至客户端，获取客户端根据目标关键词返回的考点标签；

目标关键词加入模块 20，用于根据考点标签将每一目标关键词加入到对应的预设考点中，得到目标考点。

关于关键词的确定装置的具体限定可以参见上文中对于关键词的确定方法的限定，在此不再赘述。上述关键词的确定装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

本申请实施例还提供一自动评分方法，该自动评分方法可应用如图 1 所示的应用环境中。具体地，该自动评分方法应用在自动评分系统中，该自动评分系统包括如图 1 所示的客户端和服务端，客户端与服务端通过网络进行通信，用于解决无法对考生的答题内容进行高效准确评分的问题。其中，客户端又称为用户端，是指与服务端相对应，为客户提供本地服务的程序。客户端可安装在但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备上。服务端可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在一实施例中，如图 6 所示，提供一种自动评分方法，以该方法应用在图 1 中的服务端为例进行说明，包括如下步骤：

S21：获取待评分答题信息。

其中，待评分答题信息指从考生的答题文本上获取的答题信息。具体地，获取待评分答题信息可以直接从评分系统的答题文本上获取任意一考生的答题信息，或者将任意一考生在纸质答卷上写入的答题信息进行扫描识别后获取得到。

S22：对待评分答题信息进行关键词提取，得到核心关键词。

其中，核心关键词指从获取的待评分答题信息中提取出来的一个个独立的关键词。具体地，对待评分答题信息进行关键词提取包括：先采用分词算法对待评分答题信息进行词汇拆分。可选地，分词算法可以采用基于字符串匹配的分词算法，或者采用基于理解的分词算法，还可以是采用基于统计的分词算法实现。优选地，还可通过 Java 语言的拆分函数，或者通过将待评分答题信息导入计算机的 EXCEL 或 PPT 等具有自动拆分字符功能的软件实现对待评分答题信息的自动拆分。然后通过采用 Java 语言的正则表达式对拆分后的待评分答题信息进行筛选，过滤掉一些不具有意义的特定词语，例如：助词、语气词或者连接词等。最后将筛选后所得的词语提取出来作为核心关键词。在本实施例中，核心关键词的个数应不少于一个。

S23：采用目标考点对核心关键词进行特征转化，得到待评分考点特征；其中，目标考点是采用上述关键词的确定方法所得的。

其中，待评分考点特征指衡量核心关键词与目标考点中的关键词之间的相似度的一种特征。在本步骤中，目标考点是采用上述实施例中的关键词的确定方法所得的。

采用目标考点对核心关键词进行特征转化包括：先以目标考点的数量为基准，建立一组考点向量，然后通过正则匹配法，将每一核心关键词与目标考点中的关键词进行匹配，根据匹配结果，判断核心关键词是否与目标考点相匹配。具体地，评判核心关键词是否与目标考点相匹配，可根据核心关键词与该目标考点所包含的关键词匹配程度判断。可以为只要核心关键词与目标考点中的任意一个关键词匹配，则认为该核心关键词与对应的目标考点相匹配，也可以为核心关键词与对应的目标考点中的至少两个关键词相匹配，才认为该核心关键词与对应的目标考点相匹配，具体可根据实际情况自定义设置。优选地，若核

心关键词与该目标考点中的任意一关键词匹配成功,则表示该核心关键词与目标考点相匹配,在考点向量中对应的元素值为 1,若核心关键词与目标考点中的所有关键词都不匹配,则表示该核心关键词与目标考点匹配失败,在考点向量中对应的元素值为 0。最后得到一组由若干个 1 和 0 组成的考点向量,即待评分考点特征。

S24: 将待评分考点特征输入到预设的决策树参考模型中,得到待评分答题信息的准确分值。

其中,准确分值指待评分答题信息经决策树参考模型训练后所得的分值。在本实施例中,决策树参考模型是预设建立好并保存在服务端的后台数据库的,当执行完步骤 S23,得到待评分考点特征后,可直接从服务端的数据库调取。

其中,决策树参考模型指是在已知各种情况发生概率的基础上,通过构成决策树来求取净现值的期望值大于等于零的概率,它属于一种树状结构,其中每个内部节点表示一个属性上的测试,每个分支代表一个测试输出,每个叶节点代表一种类别。

在本实施例中,通过获取待评分答题信息,对待评分答题信息进行关键词提取,得到核心关键词,然后采用目标考点对核心关键词进行特征转化,得到待评分考点特征;其中,目标考点是采用上述关键词的确定方法所得的;最后将待评分考点特征输入到预设的决策树参考模型中,得到待评分答题信息的准确分值;实现了对考生的答题信息进行高效准确的评分。

在一实施例中,如图 7 所示,采用目标考点对核心关键词进行特征转化,得到待评分考点特征,包括如下步骤:

S231: 获取目标考点对应的有效关键词。

其中,有效关键词指目标考点所包含的所有关键词。具体地,根据上述关键词的确定方法可知,每一目标考点所对应的关键词是已确定好的,因此获取目标考点所对应的有效关键词可直接根据每一目标考点,从每一目标考点中获取对应的有效关键词即可。

S232: 通过正则匹配法,将有效关键词与核心关键词进行一一匹配,得到关键词匹配信息。

具体地,通过正则匹配法,将有效关键词与核心关键词进行一一匹配指将有效关键词定义为特定字符,然后将这些由有效关键词组成的特定字符组成一个“规则字符串”,用来表达对核心关键词的一种过滤逻辑,从而匹配出与有效关键词相对应的核心关键词,得到关键词匹配信息。

其中,关键词匹配信息指将有效关键词与核心关键词进行匹配后所得的匹配结果,包括匹配成功和匹配失败。具体地,通过正则匹配法,将有效关键词与核心关键词进行一一匹配,根据匹配结果,得到对应的关键词匹配信息。例如:若获取到 10 个核心关键词和 5 个有效关键词,则提取任意一核心关键词,通过正则匹配法,逐一与获取的 5 个有效关键词进行匹配,在匹配过程中,只要该核心关键词与获取的 5 个有效关键词中的任意一个有效关键词相匹配,则表示匹配成功,若该核心关键词与获取的 5 个有效关键词都不匹配,则表示匹配失败;根据上述步骤逐一提取核心关键词,并将提取的核心关键词,通过正则匹配法,逐一与获取的 5 个有效关键词进行匹配,直至将获取的 10 个核心关键词与获取 5 个有效关键词都逐一匹配完成,最后得到关键词匹配信息。

S233: 根据关键词匹配信息,为每一核心关键词分配对应的匹配标识。

其中,匹配标识指根据关键词匹配信息,为每一核心关键词自定义赋予的一种标识号,可以为阿拉伯数字、大写字母或小写字母等。具体地,该匹配标识反映了核心关键词与目标关键词的匹配情况。另外地,由于核心关键词与有效关键词匹配成功后,还需明确该有效关键词所对应的考点,因此,在给与有效关键词匹配成功的核心关键词赋予匹配标识时,还需标识上有效关键词所对应的考点标识。此方案对具体的匹配标识不做任何限定。优选地,为了便于与后续待评分考点特征的标识区分,将与有效关键词匹配成功的核心关键词

赋予大写字母标识和对应的考点标识,例如 A1,大写字母 A 表示与有效关键词匹配成功,1 表示有效关键词所对应的考点标识;将与有效关键词匹配失败的核心关键词只赋予小写字母标识,例如 a,小写字母 a 表示与有效关键词匹配失败。

S234: 根据每一核心关键词的匹配标识,得到待评分考点特征。

具体地,根据每一核心关键词的匹配标识,判断核心关键词与对应的目标考点是否匹配成功,若核心关键词与该目标考点匹配成功,则在考点向量中对应的元素值为 1,若核心关键词与该目标考点匹配失败,则在考点向量中对应的元素值为 0,最后得到一组由若干个 1 和 0 组成的考点向量,即待评分考点特征。

示例性的,若获有 6 个目标考点,每个目标考点都至少包含 1 个有效关键词,和 5 个核心关键词;根据正则匹配法,将 5 个核心关键词与目标考点中的有效关键词进行一一匹配后得到,只有前三个核心关键词与目标考点匹配成功,则得到待评分考点特征为 [1, 1, 1, 0, 0, 0]。

在本实施例中,通过获取目标考点所对应的有效关键词;通过正则匹配法,将有效关键词与核心关键词进行一一匹配,得到关键词匹配信息;然后根据关键词匹配信息,为每一核心关键词分配对应的匹配标识,最后根据每一核心关键词的匹配标识,得到待评分考点特征;进一步保证了新增的考点关键词的准确性和有效性。

在一实施例中,如图 8 所示,在将待评分考点特征输入到预设的决策树参考模型中,得到待评分答题信息的输出分值之前,自动评分方法还包括:

S241: 获取 M 个第二样本答题数据,每一第二样本答题数据包括原始答题信息和第二评分值,M 为正整数。

其中,第二样本答题数据指考生的答题数据。每一第二样本答题数据包括原始答题信息和第二评分值;即第二样本答题数据中包括原始答题信息和对该原始答题信息进行初步评分后所得的第二评分值。可选地,该第二样本答题数据可以从一评分系统中获取。该评分系统可以对原始答题信息进行初步评分,得到第二评分值。其中,原始答题信息指从评分系统的答题文本上获取的某一主观题的考生的答题信息。第二评分值指通过人工评分或者计算机评分的方式,预先对原始答题信息进行初步评分后所得的评分值。

获取的第二样本答题数据的数量为 M 个,其中,M 为正整数。而 M 的具体数值可以根据实际需要进行设定。M 的数值越高,后续决策树参考模型的准确性会越高,然而提取效率会有所下降,可以在准确度和效率上进行综合考虑进行对 M 的选取。

S242: 采用目标考点对每一第二样本答题数据的原始答题信息进行特征转化,得到考点训练特征。

其中,考点训练特征指衡量目标考点与每一第二样本答题数据的原始答题信息之间的相似度的一种特征。目标考点是采用上述关键词的确定方法所得的。

具体地,采用目标考点对每一第二样本答题数据的原始答题信息进行特征转化包括:先以目标考点的数量为基准,建立一组空的考点向量,然后按照同义词词林语义码,将每一第二样本答题数据的原始答题信息与目标考点进行词义比较;若原始答题信息与任一目标考点匹配成功,则在考点向量中对应的元素值为 1,若原始答题信息与任一目标考点都不匹配,则在考点向量中对应的元素值为 0,最后得到一组由若干个 1 和 0 组成的考点向量,即考点训练特征。其中,同义词词林语义码是一种用来计算词语间的相似度的一种方法。

S243: 将考点训练特征和对应的第二评分值组成考点样本集。

其中,考点样本集指待输入到决策树模型中进行训练的样本数据;包括考点训练特征和对应的第二评分值。具体的,考点样本集是由若干个考点样本组成的数据集,考点样本包括考点训练特征和与该考点训练特征相对应的第二评分值。可以理解地,每一考点训练特征都与对应的第二评分值相关联。

S244: 根据考点样本集对决策树模型进行训练, 得到决策树参考模型。

其中, 决策树参考模型是一种预测模型, 它代表对象属性与对象值之间的一种映射关系, 决策树中每个节点表示某个对象, 而每个分叉路径则代表的某个可能的属性值, 而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。具体地, 根据考点样本集对决策树模型进行训练, 得到决策树参考模型指将考点训练特征和对应的第二评分值输入到决策树模型中后, 通过使用 C4.5 算法对决策树模型进行训练, 生成训练后的决策树样本模型。

优选地, 为了进一步验证决策树参考模型的准确性, 还将考点样本集分成用于建模的训练集和用于验证模型效果的测试集。其中, 训练集指用来建立决策树样本模型的数据集。测试集指用来验证建立后的决策树样本模型的效果的数据集。将考点样本集分为训练集和测试集可采用随机划分数据集或交叉检验的方法进行划分; 划分后训练集和测试集的比例值可以为: 训练集:测试集=6:4、训练集:测试集=7:3 或训练集:测试集=75:25 等。优选地, 为了提高决策树样本模型的精度, 在本步骤中, 将获取的考点样本集的 75%作为训练集, 把获取的考点样本集的 25%作为测试集。

在本实施例中, 通过获取 M 个第二样本答题数据, 每一第二样本答题数据包括原始答题信息和第二评分值, 然后采用目标考点对每一第二样本答题数据的原始答题信息进行特征转化, 得到考点训练特征, 最后根据考点训练特征和对应的第二评分值对决策树模型进行训练, 得到决策树参考模型; 进一步保证了通过决策树参考模型对考生的答题信息进行评分的准确性。

应理解, 上述实施例中各步骤的序号的大小并不意味着执行顺序的先后, 各过程的执行顺序应以其功能和内在逻辑确定, 而不应对本申请实施例的实施过程构成任何限定。

在一实施例中, 提供一种自动评分装置, 该自动评分装置与上述实施例中自动评分方法一一对应。如图 9 所示, 该自动评分装置包括待评分答题信息获取模块 21、关键词提取模块 22、待评分考点特征转化模块 23 和输入模块 24。各功能模块详细说明如下:

待评分答题信息获取模块 21, 用于获取待评分答题信息;

关键词提取模块 22, 用于对待评分答题信息进行关键词提取, 得到核心关键词;

待评分考点特征转化模块 23, 用于采用目标考点对核心关键词进行特征转化, 得到待评分考点特征; 其中, 目标考点是采用关键词的确定方法所得的;

输入模块 24, 用于将待评分考点特征输入到预设的决策树参考模型中, 得到待评分答题信息的准确分值。

优选地, 待评分考点特征转化模块 23, 包括:

有效关键词获取单元, 用于获取目标考点对应的有效关键词;

匹配单元, 用于通过正则匹配法, 将有效关键词与核心关键词进行一一匹配, 得到关键词匹配信息;

分配单元, 用于根据关键词匹配信息, 为每一核心关键词分配对应的匹配标识;

得到单元, 用于根据每一核心关键词的匹配标识, 得到待评分考点特征。

优选地, 输入模块 24, 包括:

第二样本答题数据获取单元, 用于获取 M 个第二样本答题数据, 每一第二样本答题数据包括原始答题信息和第二评分值, M 为正整数;

考点特征转化单元, 用于采用目标考点对每一第二样本答题数据的原始答题信息进行特征转化, 得到考点训练特征;

组成单元, 用于将考点训练特征和对应的第二评分值组成考点样本集;

决策树参考模型训练单元, 用于根据考点样本集对决策树模型进行训练, 得到决策树参考模型。

关于自动评分装置的具体限定可以参见上文中对于自动评分方法的限定, 在此不再赘

述。上述自动评分装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 10 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存。该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存为易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储上述关键词的确定方法和上述自动评分方法中使用到的数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种关键词的确定方法，或该计算机可读指令被处理器执行时以实现一种自动评分方法。

在一个实施例中，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机可读指令，处理器执行计算机可读指令时实现以下步骤：

获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

从所述决策树样本模型中提取样本关键词。

在一个实施例中，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机可读指令，处理器执行计算机可读指令时实现以下步骤：

获取待评分答题信息；

对所述待评分答题信息进行关键词提取，得到核心关键词；

采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

在一个实施例中，一个或多个存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

从所述决策树样本模型中提取样本关键词。

在一个实施例中，一个或多个存储有计算机可读指令的非易失性可读存储介质，所述

计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取待评分答题信息；

对所述待评分答题信息进行关键词提取，得到核心关键词；

采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性计算机可读存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程 ROM (PROM)、电可编程 ROM (EPROM)、电可擦除可编程 ROM (EEPROM) 或闪存。易失性存储器可包括随机存取存储器 (RAM) 或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，诸如静态 RAM (SRAM)、动态 RAM (DRAM)、同步 DRAM (SDRAM)、双数据率 SDRAM (DDRSDRAM)、增强型 SDRAM (ESDRAM)、同步链路 (Synchlink) DRAM (SLDRAM)、存储器总线 (Rambus) 直接 RAM (RDRAM)、直接存储器总线动态 RAM (DRDRAM)、以及存储器总线动态 RAM (RDRAM) 等。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。

以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围，均应包含在本申请的保护范围之内。

权 利 要 求 书

1. 一种关键词的确定方法，其特征在于，包括：

获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

从所述决策树样本模型中提取样本关键词。

2. 如权利要求 1 所述的关键词的确定方法，其特征在于，在所述从所述决策树样本模型中提取样本关键词之后，所述关键词的确定方法还包括：

获取评分规则信息，所述评分规则信息包括预设考点和每一所述预设考点对应的预设关键词；

从所述样本关键词中去除和所述预设关键词重复的关键词，得到目标关键词；

发送所述目标关键词至客户端，获取所述客户端根据所述目标关键词返回的考点标签；

根据所述考点标签将每一所述目标关键词加入到对应的所述预设考点中，得到目标考点。

3. 一种自动评分方法，其特征在于，包括：

获取待评分答题信息；

对所述待评分答题信息进行关键词提取，得到核心关键词；

采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

4. 如权利要求 3 所述的自动评分方法，其特征在于，所述采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征，包括：

获取所述目标考点对应的有效关键词；

通过正则匹配法，将所述有效关键词与所述核心关键词进行一一匹配，得到关键词匹配信息；

根据所述关键词匹配信息，为每一所述核心关键词分配对应的匹配标识；

根据每一所述核心关键词的匹配标识，得到待评分考点特征。

5. 如权利要求 3 所述的自动评分方法，其特征在于，在将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的输出分值之前，所述自动评分方法还包括：

获取 M 个第二样本答题数据，每一所述第二样本答题数据包括原始答题信息和第二评分值，M 为正整数；

采用所述目标考点对每一所述第二样本答题数据的所述原始答题信息进行特征转化，得到考点训练特征；

将所述考点训练特征和对应的所述第二评分值组成考点样本集；

根据所述考点样本集对决策树模型进行训练，得到决策树参考模型。

6. 一种关键词的确定装置，其特征在于，包括：

第一样本答题数据获取模块，用于获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值， N 为正整数；

分词处理模块，用于对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

分词汇总模块，用于对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

样本特征转化模块，用于采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

决策树样本模型训练模块，用于根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

样本关键词提取模块，用于从所述决策树样本模型中提取样本关键词。

7. 一种关键词的确定装置，其特征在于，还包括：

评分规则信息获取模块，用于获取评分规则信息，所述评分规则信息包括预设考点和每一所述预设考点对应的预设关键词；

重复关键词去除模块，用于从所述样本关键词中去除和所述预设关键词重复的关键词，得到目标关键词；

考点标签获取模块，用于发送所述目标关键词至客户端，获取所述客户端根据所述目标关键词返回的考点标签；

目标关键词加入模块，用于根据所述考点标签将每一所述目标关键词加入到对应的所述预设考点中，得到目标考点。

8. 一种自动评分装置，其特征在于，包括：

待评分答题信息获取模块，用于获取待评分答题信息；

关键词提取模块，用于对所述待评分答题信息进行关键词提取，得到核心关键词；

待评分考点特征转化模块，用于采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

输入模块，用于将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

9. 如权利要求 8 所述的自动评分装置，其特征在于，所述待评分考点特征转化模块包括：

有效关键词获取单元，用于获取所述目标考点对应的有效关键词；

匹配单元，用于通过正则匹配法，将所述有效关键词与所述核心关键词进行一一匹配，得到关键词匹配信息；

分配单元，用于根据所述关键词匹配信息，为每一所述核心关键词分配对应的匹配标识；

得到单元，用于根据每一所述核心关键词的匹配标识，得到待评分考点特征。

10. 如权利要求 9 所述的自动评分装置，其特征在于，所述输入模块包括：

第二样本答题数据获取单元，用于获取 M 个第二样本答题数据，每一所述第二样本答题数据包括原始答题信息和第二评分值， M 为正整数；

考点特征转化单元，用于采用所述目标考点对每一所述第二样本答题数据的所述原始答题信息进行特征转化，得到考点训练特征；

组成单元，用于将所述考点训练特征和对应的所述第二评分值组成考点样本集；

决策树参考模型训练单元，用于根据所述考点样本集对决策树模型进行训练，得到决策树参考模型。

11. 一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如

下步骤:

获取 N 个第一样本答题数据, 每一所述第一样本答题数据包括样本答题信息和第一评分值, N 为正整数;

对每一所述第一样本答题数据的所述样本答题信息进行分词处理, 得到每一所述第一样本答题数据的样本分词;

对每一所述第一样本答题数据的所述样本分词进行汇总, 得到样本分词集;

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化, 得到样本训练特征;

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练, 得到决策树样本模型;

从所述决策树样本模型中提取样本关键词。

12. 如权利要求 11 所述的计算机设备, 其特征在于, 在所述从所述决策树样本模型中提取样本关键词之后, 所述处理器执行所述计算机可读指令时还实现如下步骤:

获取评分规则信息, 所述评分规则信息包括预设考点和每一所述预设考点对应的预设关键词;

从所述样本关键词中去除和所述预设关键词重复的关键词, 得到目标关键词;

发送所述目标关键词至客户端, 获取所述客户端根据所述目标关键词返回的考点标签;

根据所述考点标签将每一所述目标关键词加入到对应的所述预设考点中, 得到目标考点。

13. 一种计算机设备, 包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令, 其特征在于, 所述处理器执行所述计算机可读指令时实现如下步骤:

获取待评分答题信息;

对所述待评分答题信息进行关键词提取, 得到核心关键词;

采用目标考点对所述核心关键词进行特征转化, 得到待评分考点特征; 其中, 所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的;

将所述待评分考点特征输入到预设的决策树参考模型中, 得到所述待评分答题信息的准确分值。

14. 如权利要求 13 所述的计算机设备, 其特征在于, 所述采用目标考点对所述核心关键词进行特征转化, 得到待评分考点特征, 包括:

获取所述目标考点对应的有效关键词;

通过正则匹配法, 将所述有效关键词与所述核心关键词进行一一匹配, 得到关键词匹配信息;

根据所述关键词匹配信息, 为每一所述核心关键词分配对应的匹配标识;

根据每一所述核心关键词的匹配标识, 得到待评分考点特征。

15. 如权利要求 14 所述的计算机设备, 其特征在于, 在将所述待评分考点特征输入到预设的决策树参考模型中, 得到所述待评分答题信息的输出分值之前, 所述处理器执行所述计算机可读指令时还实现如下步骤:

获取 M 个第二样本答题数据, 每一所述第二样本答题数据包括原始答题信息和第二评分值, M 为正整数;

采用所述目标考点对每一所述第二样本答题数据的所述原始答题信息进行特征转化, 得到考点训练特征;

将所述考点训练特征和对应的所述第二评分值组成考点样本集;

根据所述考点样本集对决策树模型进行训练, 得到决策树参考模型。

16. 一个或多个存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取 N 个第一样本答题数据，每一所述第一样本答题数据包括样本答题信息和第一评分值，N 为正整数；

对每一所述第一样本答题数据的所述样本答题信息进行分词处理，得到每一所述第一样本答题数据的样本分词；

对每一所述第一样本答题数据的所述样本分词进行汇总，得到样本分词集；

采用所述样本分词集对每一所述第一样本答题数据的所述样本答题信息进行特征转化，得到样本训练特征；

根据所述样本训练特征和对应的第一评分值对决策树模型进行训练，得到决策树样本模型；

从所述决策树样本模型中提取样本关键词。

17. 如权利要求 16 所述的非易失性可读存储介质，其特征在于，在所述从所述决策树样本模型中提取样本关键词之后，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器还执行如下步骤：

获取评分规则信息，所述评分规则信息包括预设考点和每一所述预设考点对应的预设关键词；

从所述样本关键词中去除和所述预设关键词重复的关键词，得到目标关键词；

发送所述目标关键词至客户端，获取所述客户端根据所述目标关键词返回的考点标签；

根据所述考点标签将每一所述目标关键词加入到对应的所述预设考点中，得到目标考点。

18. 一个或多个存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取待评分答题信息；

对所述待评分答题信息进行关键词提取，得到核心关键词；

采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征；其中，所述目标考点是采用权利要求 2 所述的关键词的确定方法所得的；

将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的准确分值。

19. 如权利要求 18 所述的非易失性可读存储介质，其特征在于，所述采用目标考点对所述核心关键词进行特征转化，得到待评分考点特征，包括：

获取所述目标考点对应的有效关键词；

通过正则匹配法，将所述有效关键词与所述核心关键词进行一一匹配，得到关键词匹配信息；

根据所述关键词匹配信息，为每一所述核心关键词分配对应的匹配标识；

根据每一所述核心关键词的匹配标识，得到待评分考点特征。

20. 如权利要求 19 所述的非易失性可读存储介质，其特征在于，在将所述待评分考点特征输入到预设的决策树参考模型中，得到所述待评分答题信息的输出分值之前，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行如下步骤：

获取 M 个第二样本答题数据，每一所述第二样本答题数据包括原始答题信息和第二评分值，M 为正整数；

采用所述目标考点对每一所述第二样本答题数据的所述原始答题信息进行特征转化，得到考点训练特征；

将所述考点训练特征和对应的所述第二评分值组成考点样本集；

根据所述考点样本集对决策树模型进行训练，得到决策树参考模型。

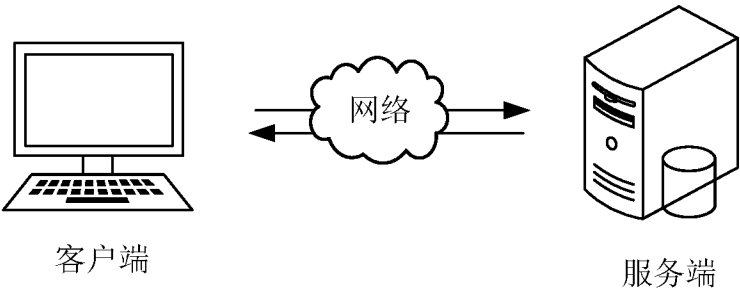


图 1

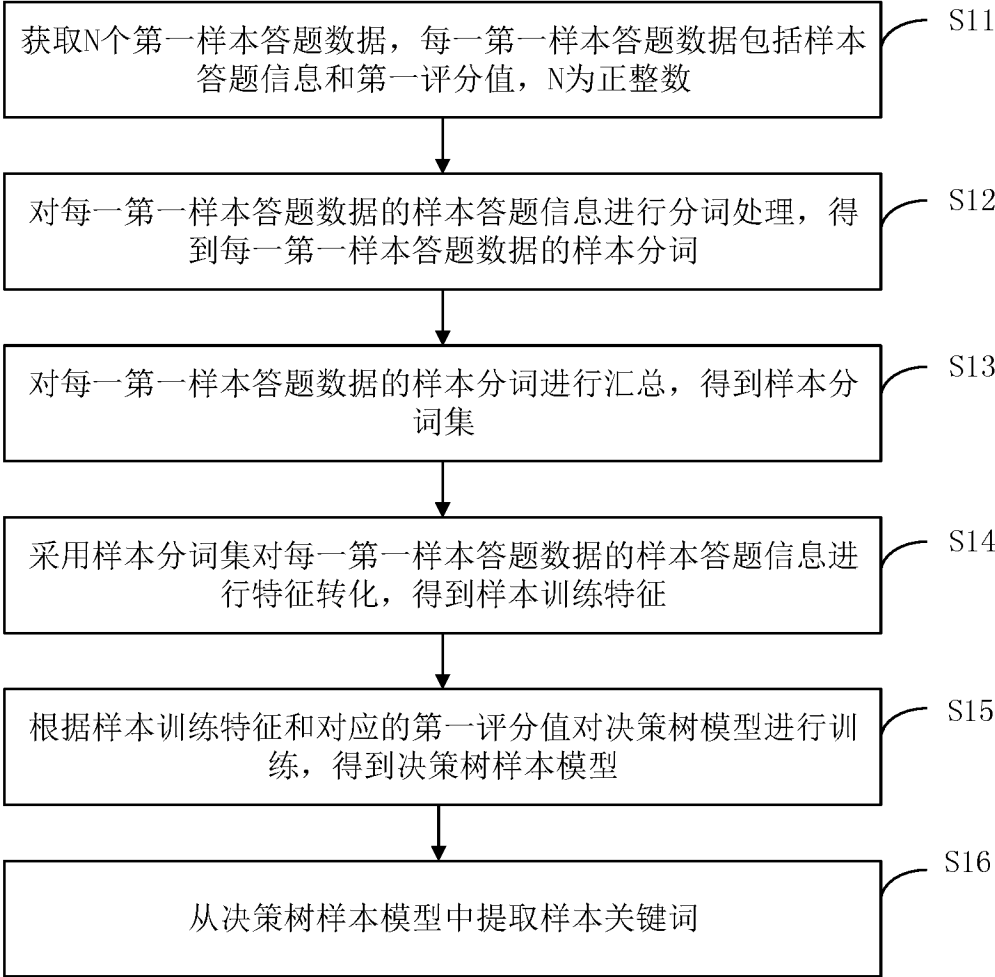


图 2

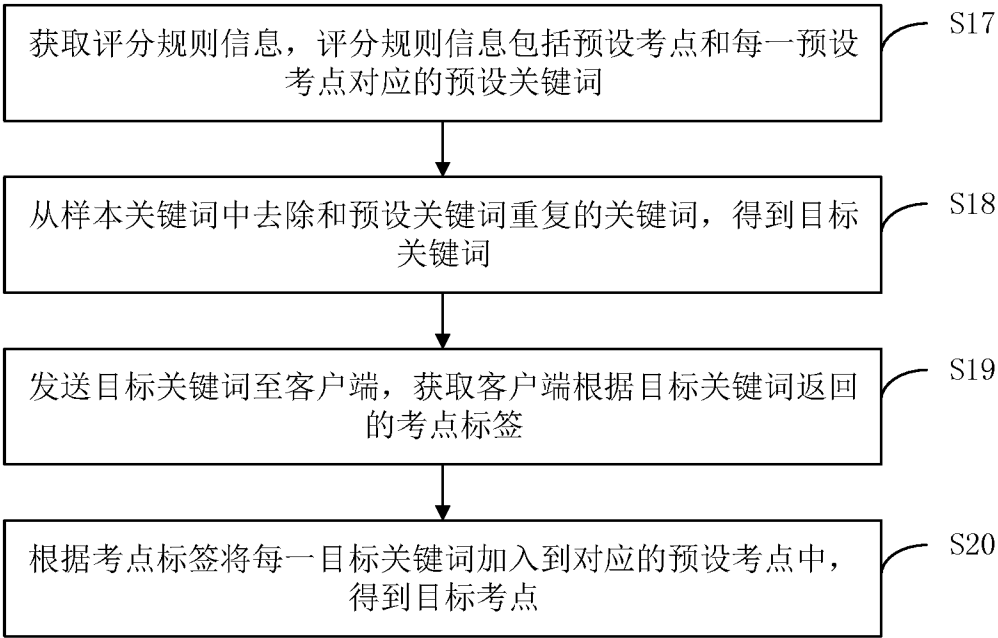


图 3



图 4

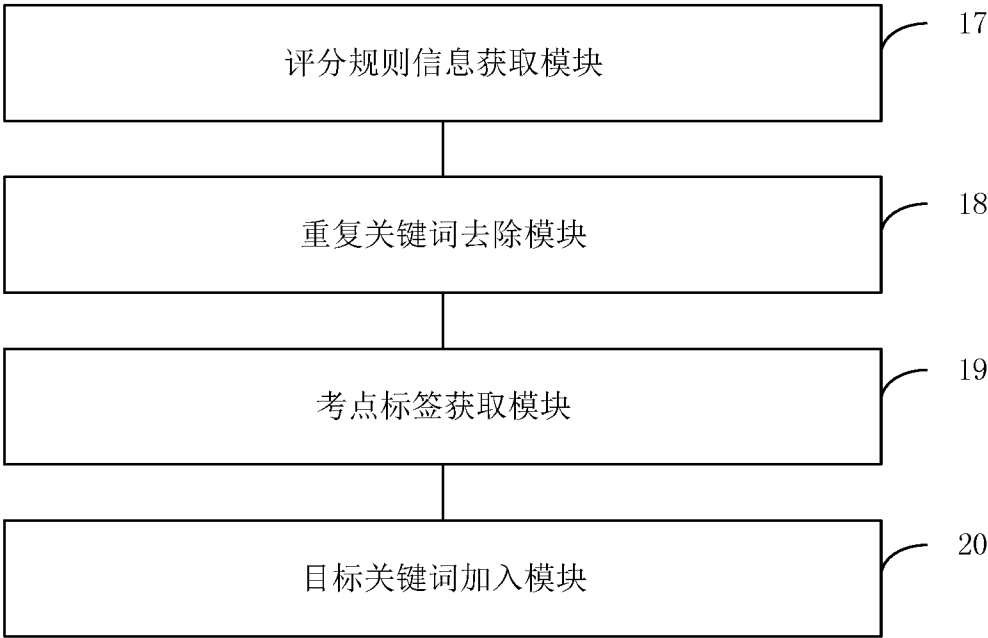


图 5

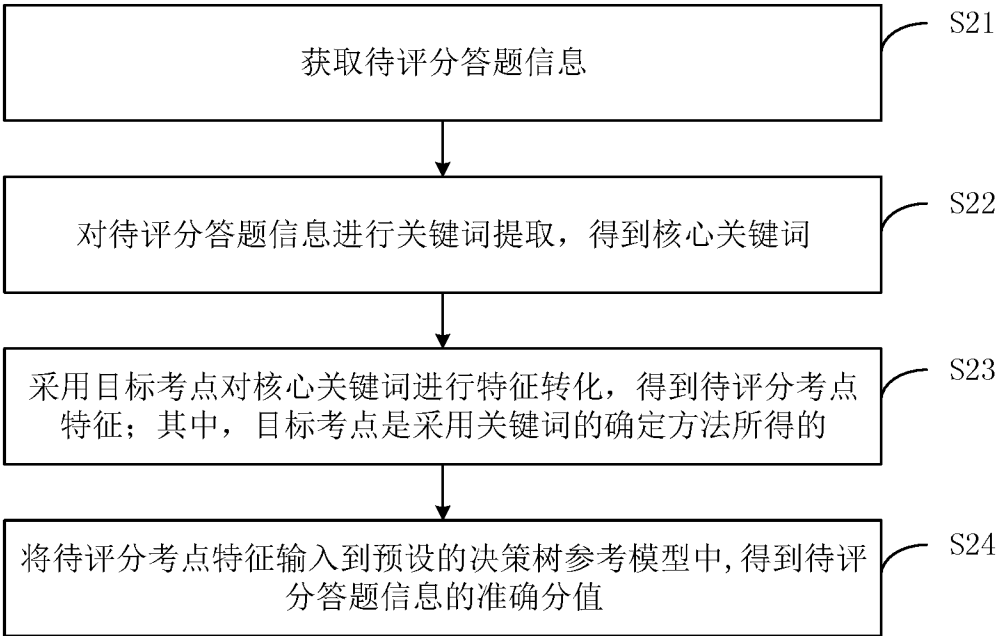


图 6

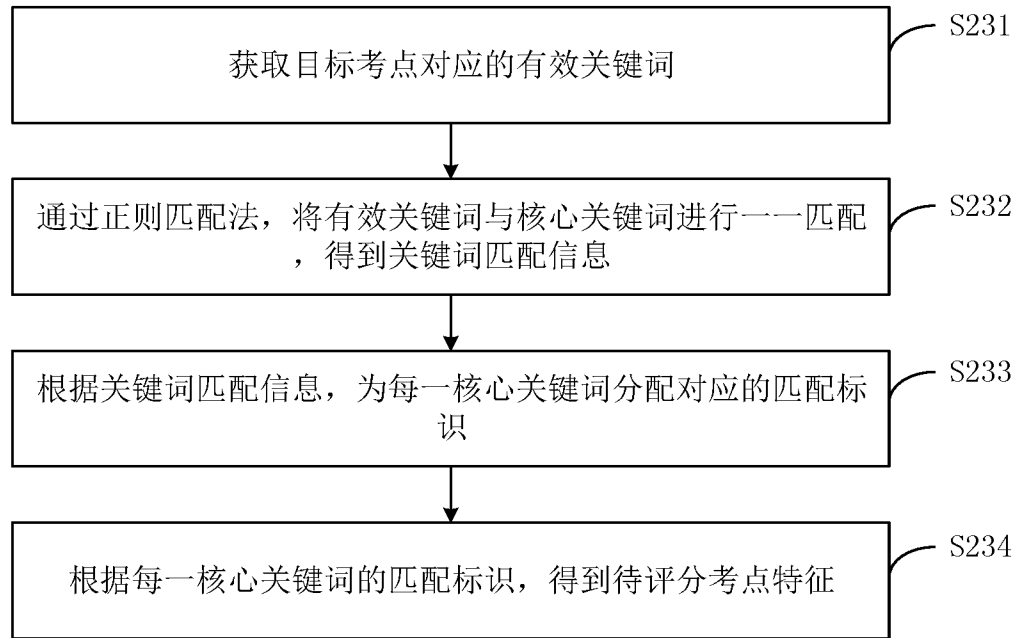


图 7

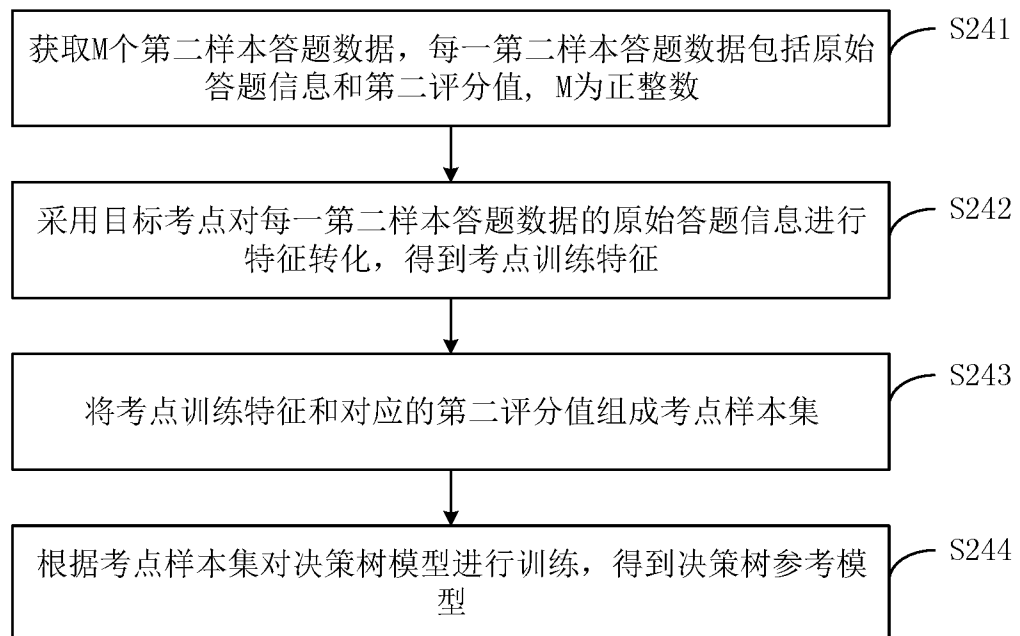


图 8

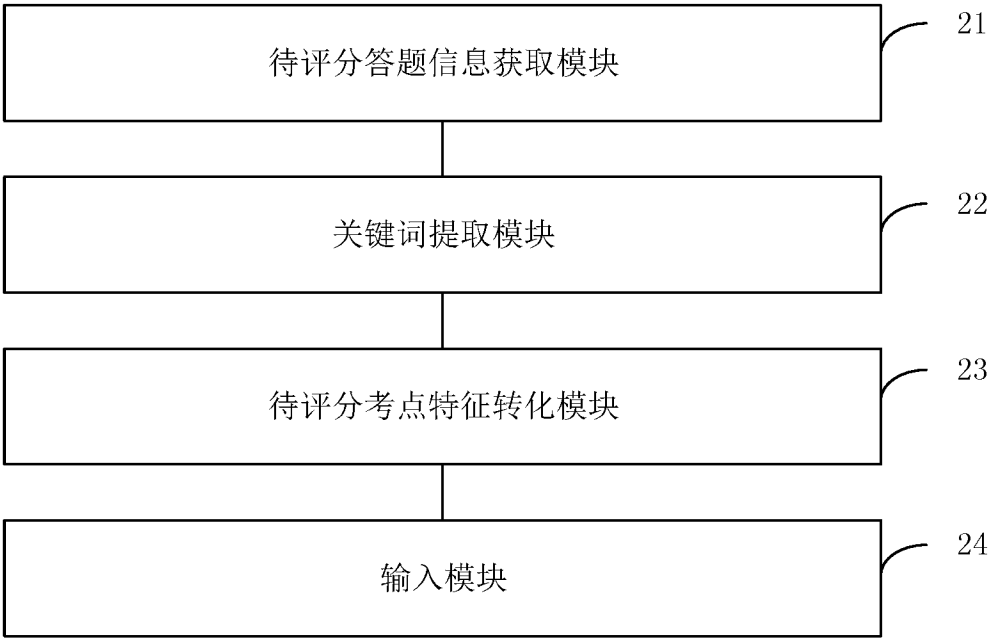


图 9

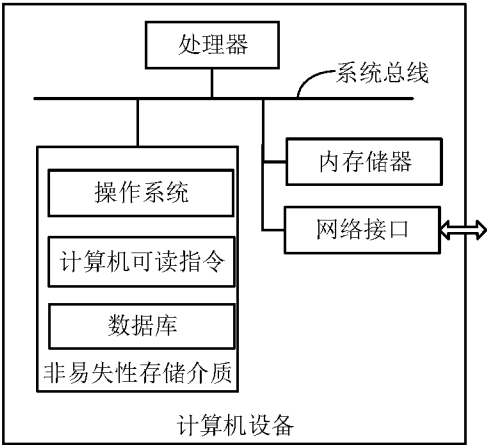


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/088544

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

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: 答题, 评分, 阅卷, 关键词, 训练, 模型, 决策树, 机器学习, 考点, model, train, keyword, tree, decision, feature, determin+, score, machine, study

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108363687 A (SHENZHEN NAODONG TECH CO., LTD.) 03 August 2018 (2018-08-03) description, paragraphs [0044]-[0076]	1, 6, 11, 16
Y	CN 108363687 A (SHENZHEN NAODONG TECH CO., LTD.) 03 August 2018 (2018-08-03) description, paragraphs [0044]-[0076]	2-5, 7-10, 12-15, 17-20
Y	CN 101859338 A (SHENZHEN HAIYUNTIAN TECHNOLOGY CO., LTD.) 13 October 2010 (2010-10-13) description, paragraphs [0028]-[0048]	2-5, 7-10, 12-15, 17-20
A	CN 105989040 A (ALIBABA GROUP HOLDING LIMITED) 05 October 2016 (2016-10-05) entire document	1-20
A	US 9342491 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 17 May 2016 (2016-05-17) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 September 2019

Date of mailing of the international search report

18 October 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/088544

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108363687	A	03 August 2018	None			
CN	101859338	A	13 October 2010	None			
CN	105989040	A	05 October 2016	HK	1229484	A0	17 November 2017
US	9342491	B2	17 May 2016	CN	103577452	A	12 February 2014
				US	9311282	B2	12 April 2016

国际检索报告

国际申请号

PCT/CN2019/088544

A. 主题的分类 G06F 17/27 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 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, EPDOC, CNPAT, CNKI: 答题, 评分, 阅卷, 关键词, 训练, 模型, 决策树, 机器学习, 考点, model, train, keyword, tree, decision, feature, determin+, score, machine, study																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段</td> <td>1, 6, 11, 16</td> </tr> <tr> <td>Y</td> <td>CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段</td> <td>2-5, 7-10, 12-15, 17-20</td> </tr> <tr> <td>Y</td> <td>CN 101859338 A (深圳市海云天科技股份有限公司) 2010年 10月 13日 (2010 - 10 - 13) 说明书第[0028]-[0048]段</td> <td>2-5, 7-10, 12-15, 17-20</td> </tr> <tr> <td>A</td> <td>CN 105989040 A (阿里巴巴集团控股有限公司) 2016年 10月 5日 (2016 - 10 - 05) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 9342491 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016年 5月 17日 (2016 - 05 - 17) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段	1, 6, 11, 16	Y	CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段	2-5, 7-10, 12-15, 17-20	Y	CN 101859338 A (深圳市海云天科技股份有限公司) 2010年 10月 13日 (2010 - 10 - 13) 说明书第[0028]-[0048]段	2-5, 7-10, 12-15, 17-20	A	CN 105989040 A (阿里巴巴集团控股有限公司) 2016年 10月 5日 (2016 - 10 - 05) 全文	1-20	A	US 9342491 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016年 5月 17日 (2016 - 05 - 17) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段	1, 6, 11, 16																		
Y	CN 108363687 A (深圳市脑洞科技有限公司) 2018年 8月 3日 (2018 - 08 - 03) 说明书第[0044]-[0076]段	2-5, 7-10, 12-15, 17-20																		
Y	CN 101859338 A (深圳市海云天科技股份有限公司) 2010年 10月 13日 (2010 - 10 - 13) 说明书第[0028]-[0048]段	2-5, 7-10, 12-15, 17-20																		
A	CN 105989040 A (阿里巴巴集团控股有限公司) 2016年 10月 5日 (2016 - 10 - 05) 全文	1-20																		
A	US 9342491 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2016年 5月 17日 (2016 - 05 - 17) 全文	1-20																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期		国际检索报告邮寄日期																		
2019年 9月 10日		2019年 10月 18日																		
ISA/CN的名称和邮寄地址		受权官员																		
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		田晶																		
		电话号码 86-(10)-53961334																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/088544

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	108363687	A	2018年 8月 3日	无			
CN	101859338	A	2010年 10月 13日	无			
CN	105989040	A	2016年 10月 5日	HK	1229484	A0	2017年 11月 17日
US	9342491	B2	2016年 5月 17日	CN	103577452	A	2014年 2月 12日
				US	9311282	B2	2016年 4月 12日