(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
2021 年 7 月 15 日 (15.07.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/139265 A1

(51) 国际专利分类号:
G06F 40/205 (2020.01) *G06F 40/194* (2020.01)

(21) 国际申请号: PCT/CN2020/118737

(22) 国际申请日: 2020 年 9 月 29 日 (29.09.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010687313.4 2020年7月16日 (16.07.2020) 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: 阮鸿涛 (RUAN, Hongtao); 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)。 郑立颖 (ZHENG, Liying); 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)。 阮晓雯 (RUAN, Xiaowen); 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: 深圳市精英专利事务所 (SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路6009号绿景广场B栋20层B, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, 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,

(54) **Title:** COMPOSITION SCORING METHOD AND APPARATUS, COMPUTER DEVICE, AND COMPUTER READABLE STORAGE MEDIUM

(54) **发明名称**: 作文评分方法、装置、计算机设备及计算机可读存储介质

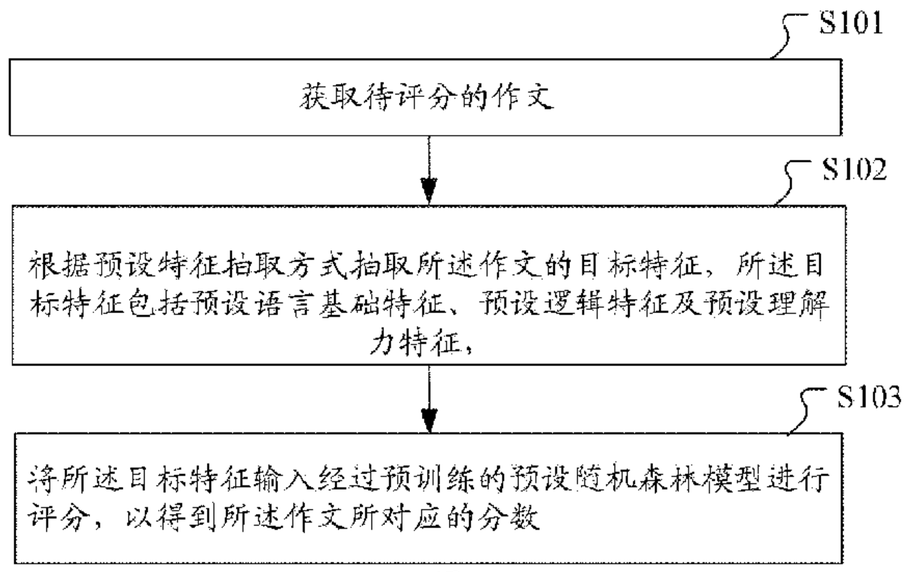


图 1

(57) **Abstract:** A feature extraction and random forest based composition scoring method and apparatus, a computer device, and a computer readable storage medium, being related to the technical field of artificial intelligence. The method comprises: obtaining the composition to be scored (S101); extracting, according to a pre-set feature extraction method, target features of the composition, said target features including pre-set language basic features, pre-set logic features, and pre-set understandability features (S102); inputting the target features into a pre-trained pre-set random forest model for scoring so as to obtain the score corresponding to the composition (S103).

(57) **摘要**: 一种基于特征抽取及随机森林的作文评分方法、装置、计算机设备及计算机可读存储介质，属于人工智能技术领域，该方法包括：获取待评分的作文（S101），根据预设特征抽取方式抽取作文的目标特征，目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征（S102），将目标特征输入经过预训练的预设随机森林模型进行评分，以得到作文所对应的分数（S103）。

[见续页]

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))。

— 在修改权利要求的期限届满之前进行，在收到该修改后将重新公布(细则48.2(h))。

— 根据申请人的请求，在条约第21条(2)(a)所规定的期限届满之前进行。

作文评分方法、装置、计算机设备及计算机可读存储介质

本申请要求于2020年07月16日提交中国专利局、申请号为202010687313.4、申请名称为“作文评分方法、装置、计算机设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种基于特征抽取及随机森林的作文评分方法、装置、计算机设备及计算机可读存储介质。

背景技术

作文自动评分（Automated Essay Scoring，AES）近年来被广泛应用于教师批改作文、学生自助评测、家长辅助检查和考试自动评分等场景。运用作文自动评分较为广泛的系统有E-rater 、Project Essay Grader、Intelligent Essay Assessor、IntelliMetric、批改网及腾讯云作文批改等。

目前工业上已经应用的作文自动评分系统主要利用线性回归方法建模进行得分预测，根据不同的考试类型选择十几个甚至更多和作文分数相关性最高的指标来获取最终评分。但是，发明人发现，传统技术中，每个作文自动评分系统采用的指标不一样，导致每个作文自动评分系统得出的评分结果较侧重于某一方面的特征，从而对作文进行评分的准确性较低。

发明内容

本申请提供了一种基于特征抽取及随机森林的作文评分方法、装置、计算机设备及计算机可读存储介质，能够解决传统技术中对对作文进行自动评分由于评价不全面而导致评分准确性较低的问题。

第一方面，本申请提供了一种基于特征抽取及随机森林的作文评分方法，所述方法包括：获取待评分的作文；根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

第二方面，本申请还提供了一种基于特征抽取及随机森林的作文评分装置，包括：获取单元，用于获取待评分的作文；抽取单元，用于根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；评分单元，用于将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

第三方面，本申请还提供了一种计算机设备，其包括存储器及处理器，所述存储器上存储有计算机程序，所述处理器执行所述计算机程序时执行如下步骤：获取待评分的作文；根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

第四方面，本申请还提供了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程序，所述计算机程序被处理器执行时使所述处理器实现如下步骤：获取待评分的作文；根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、

预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。执行所述基于特征抽取及随机森林的作文评分方法的步骤。

本申请通过获取待评分的作文，根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数，由于从语言基础能力、逻辑能力、理解力三大维度进行实施特征工程，涵盖了词汇、语法、句法、作文篇幅、可读性、文章的组织与结构、与优质文章的对比等多方面、全方位的特征，从而抽取全面的具有评价意义的特征指标，可以针对打分结果及不同特征维度给出解释和评价，能够提高作文自动评分所得到结果的准确性。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例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的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法的一个流程示意图；

图 2 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法中一个子流程的示意图；

图 3 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法的另一个子流程示意图；

图 4 为本申请实施例提供的基于特征抽取及随机森林的作文评分装置的一个示意性框图；以及

图 5 为本申请实施例提供的计算机设备的示意性框图。

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做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。

请参阅图 1，图 1 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法的一个流程示意图。如图 1 所示，该方法包括以下步骤 S101-S103:

S101、获取待评分的作文。

S102、根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征。

具体地，由于作文的语言基础能力包括作文中采取的词汇、语法、句法、作文篇幅及作文可读性等特征指标来体现，作文的逻辑能力涉及作文的组织与结构等特征指标，作文的理解力包括作文与优质文章的对比及跑题程度等特征指标，因此，本申请实施例所提供的基于特征抽取及随机森林的作文自动评分方法，从语言基础特征（用于描述作文的语言基础能力）、逻辑特征（用于描述作文的逻辑能力）、理解力特征（用于描述作文的理解力）三大维度实施特征工程，抽取对于作文具有评价意义的上述三大维度所包含的特征指标，用于对作文进行自动评分。

在对作文进行自动评分前，首先对作文语料数据进行预处理以对作文语料进行分类，划分为不同的分组类别，可以将作文语料按照不同年龄段作者的作文划分为不同的语料数据集，

例如划分为高中组、初中组及小学组等不同的分组类别，并在每一分组组中剔除作文单词数目过长或者作文单词数目过短等极端情形的作文，然后将作文按照主题划分为 K 个作文主题，针对不同年龄段及不同主题的作文分别训练基于随机森林模型的作文自动评分模型。其中，对作文进行特征抽取涉及的特征工程，具体包括如下：

（1）语言基础特征，用于描述作文的语言基础能力，包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征。具体分为以下特征内容：

1）词汇，用词汇特征进行描述，包含以下各个子特征：

①Unigram 单词（字）占比，用于描述词汇丰富度，Unigram 单词个数占全文单词总数目的比例，其中，重复使用单词只计算为一个，Unigram，是自然语言处理（NLP）中的概念，用于描述一元模型 ，比如，西安交通大学所对应的 Unigram 形式为：西/安/交/通/大/学。②平均笔画数，用于描述单个单词笔画数的平均数目，可以采用统计笔画的方式获得。③笔画数的标准方差，用于描述单个单词笔画数的方差。④单词类别数（含标点或不含标点），用于描述 Unigram 单词个数。⑤笔画数大于 n 的单词百分比（n=6，7，8，···）。⑥Mechanics 错误率，用于描述错别字、标点使用等错误次数占总单词数目的百分比。⑦词汇水平指标，通过将每个单词标记词汇难易水平，并选定作文中 Index（即特征）排名倒数第五的单词的水平值作为特征。⑧成语使用数，用于描述作文中使用的成语个数。

2）语法及句法，通过预设语法特征及预设句法特征进行描述，包括以下特征指标： ①词性标注占比，例如名词、动词、形容词、代词、借词、连词、虚词分别占总单词数目的百分比，可以通过预设语言工具进行分词处理，然后进行词性标注，再统计词性标注占比。②语法使用错误率（Grammar），用于描述语法（词性及、词组等）使用错误次数占总单词数目的百分比。③过度重复使用词，用于描述过度重复使用的词或词组的使用次数占总单词数目的百分比。④句法结构解析树，通过树的深度等树结构所对应的指标以描述句法结构。⑤从句关系，包括主语从句、宾语从句、定语从句、状语从句等数目。

3）作文篇幅，通过预设篇幅特征来描述，包括以下内容：①句长分布，用于描述句子单词数目的分布。②平均句长，用于描述句子单词数目的平均数。③从句总数，用于描述全文从句使用数量。④平均段落句子数，用于描述平均每个段落的句子数目。⑤句子总数，用于描述全文句子数目。⑥段落数量，用于描述全文段落数目。⑦全文单词数，用于描述全文总单词数目。

4）可读性，通过预设可读性特征来描述，包括以下内容： ①可读性指标，用于描述关于中文文本可读性的评价指标，计算公式为：N×中文平均笔画数+(1-N)×中文难词频度，其中 N 调整中文平均笔画数指标以及中文难词频度指标之间的权重配比。

（2）逻辑能力，通过预设逻辑特征进行描述，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，包括以下内容：

1）文章的组织与结构，采用预设文章结构特征来描述，包括以下内容：①文章结构指标，：将每个句子分类为背景、论点、主要观点、支撑观点、结论中的一类，将最大连续出现的某个类别的句子集合定义为该类别的一个元素，例如 3 句主要观点+4 句支撑观点+2 句主要观点，则对应 3 个元素，计算元素个数作为文章结构指标。②文章连贯性，用于描述相邻句子的 Overlap （即重叠）情况。③文章元素平均单词数，用于描述所有元素中单词数目的均值。

（3）作文理解力，通过所述预设理解力特征，所述预设理解力特征包括与预设文章的相似度特征，包括以下内容： 1）作文与优质文章的对比。①最相似文章得分：将语料库文章按得分分为 n 类，取与待评价文章最相似的文章类别的得分为指标。②与最优质论文相似度：待评价文章与得分最高的文章类别的相似度。2）作文跑题程度。①与同主题文章对比相似度：待评价文章与同主题文章类别的相似度。

获取待评分的作文后，根据预设特征抽取方式抽取所述作文的目标特征，针对上述每一项所述特征，可以预先针对每一项特征设置对应的抽取方式，进而在进行特征抽取时，根据预先设置的上述特征各自所对应的特征抽取方式抽取上述各项特征，从而得到所述作文的目

标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征。本申请实施例提供的基于特征抽取及随机森林的作文自动评分方法，针对自动评分用到的维度及特征，筛选从语言基础能力、逻辑能力、理解力三大维度进行实施特征工程，涵盖了词汇、语法/句法、作文篇幅、可读性、文章的组织与结构、与优质文章的对比、跑题程度等多方面的特征，抽取具有评价意义的指标，可以针对打分结果及不同特征维度给出解释和评价，作文自动评分模型可解释性高。

S103、将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

其中，随机森林，英文为 Random forest，是利用多棵树对样本进行训练并预测的一种分类器。

具体地，首先通过预训练作文样本对用于进行作文自动评分的预设随机森林模型进行预训练，所述预训练作文样本包括预训练作文及对所述预训练作文进行评分所对应得到的预训练作文分值，其中，对预训练作文样本进行评分以得到预训练作文分值，可以通过人工对预训练作文样本进行评分，其中，所述预训练作文分值还可以是对预训练作文进行多次评分得到的预训练作文分值的作文分值均值。比如，若通过人工对预训练作文进行评分时，所述预训练作文分值可以为多位不同的人工各自根据预设随机森林模型预设提取的目标特征作为作文评分指标对预训练作文进行评分后，根据得到多个预训练作文分值计算得到的作文分值均值，以使人工对该预训练作文的评分尽可能客观和全面，由于预训作文样本数据的准确性得到提升，从而使训练出的预设随机森林模型也更具备客观性，也能提升预设随机森林模型对作文自动评分的评分准确性。对预设随机森林模型进行预训练的过程包括以下步骤：获取预训练作文样本，所述预训练作文样本包括预训练作文及所述预训练作文所对应的作文分值；将所述预训练作文样本输入至预设随机森林模型；提取所述预训练作文的预设目标特征及所述作文分值进行学习，并对所述预训练作文样本进行分类，以实现预训练所述预设随机森林模型。

具体地，由于随机森林模型是以决策树为基学习器的集成学习算法，是利用多棵树对样本进行训练并预测的一种分类器。因此，在对预设随机森林模型进行预训练时，所述预设随机森林模型会提取所述预训练作文的预设目标特征及作文分值进行学习，并根据提取的所述预训练作文的预设目标特征及作文分值对所述预训练作文样本进行分类，以得到不同的作文类别，从而实现预训练预设随机森林模型。后续，预设随机森林模型对待评分作文进行自动评分时，也先提取待评分作文的预设目标特征，并根据提取的待评分作文的预设目标特征将所述待评分作文根据预训练时自动进行学习的学习结果对待评分作文进行分类，即将提取的待评分作文的预设目标特征与预训练时将预训练作文进行分类时对于各个作文类别学习的预设目标特征进行相似性比对，将待评分作文分类至相似度最高的作文类别中，并通过得到的相似度获取与待评分最相似的最相似作文，将最相似作文的作文分值作为待评分作文评分的基础，根据待评分作文与最相似作文的相似度，得到待评分作文的作文分值，以实现对待评分作为进行自动评分。

进一步地，对预设随机森林模型进行预训练时，还可以针对不同的目标群体进行分类，以训练不同的预设随机森林模型。例如，在对作文语料数据进行预处理以对作文语料进行分类，划分为不同的分组类别，可以将作文语料按照不同年龄段作者的作文划分为不同的语料数据集，例如划分为高中组、初中组及小学组等不同的分组类别等，采用所述作文语料及每篇作文所对应的评分训练预设随机森林模型，以使预设随机森林模型能够对作文进行自动评分。

进一步地，在训练所述随机森林模型时，以广泛应用于衡量有序量表评估一致性的统计指标 QWK(Quadratic Weighted Kappa)作为自动评分模型的评价指标，同时增加了 Correlation

（衡量了模型预测结果与实际结果的趋势一致性）、Accuracy（模型预测结果的准确度）、RMSE（Root Mean Squared Error）三种常用指标作为辅助输出指标值，对所述随机森林模型进行调试，以提供所述随机森林模型进行作文自动评分的准确性。

对预设随机森林模型预训练完成，预设随机森林模型满足预定目标后，完成所述预设随机森林模型的训练，获取待评分的作文，并在对所述作文完成目标特征抽取后，将目标特征输入随机森林模型，通过以上三个维度对作文进行总分评估。

本申请实施例通过获取待评分的作文，根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数，由于本申请实施例从语言基础能力、逻辑能力、理解力三大维度进行实施特征工程，涵盖了词汇、语法、句法、作文篇幅、可读性、文章的组织与结构、与优质文章的对比等多方面、全方位的特征，从而抽取全面的具有评价意义的特征指标，可以针对打分结果及不同特征维度给出解释和评价，能够提高作文自动评分所得到结果的准确性。

在一实施例中，所述根据预设特征抽取方式抽取所述作文的目标特征的步骤包括：根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征；根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征；根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征。

具体地，由于在本申请实施例中，为了提高对作文进行评分的准确性，采用了三大维度的较为全面的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，针对作文在每一方面的特征均采用所对应的预设特征抽取方式进行特征抽取，即根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征，根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征，根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征，从得到作文所对应的包含预设语言基础特征、预设逻辑特征及预设理解力特征的目标特征，并将将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数，从而对作文进行自动评分，由于针对作文抽取了全面的具有评价意义的特征指标，可以通过随机森林模型对作文的打分结果及不同特征维度给出解释和评价，能够提高作文自动评分所得到结果的准确性。

在一实施例中，所述根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征的步骤包括：采用第一预设语言工具对所述作文进行分词，以得到所述作文所包含的短语；基于所述短语，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征；将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征。

其中，第一预设语言工具可以为 Stanford CoreNLP 或者 HanLP 等支持进行分词的语言工具。

具体地，由于语言工具（例如 Stanford CoreNLP 或者 HanLP）支持包括标记化、浅层分析（句字分块）、分词、分句、分块、词性标注、命名实体识别及语法解析等 NLP 任务，可以通过预设语言工具对输入语句进行分词得到短语划分，再对短语进行词性标注，即采用第一预设语言工具对所述单语句进行分词，以得到所述单语句所包含的若干个短语。再基于所述短语，根据所述语言基础特征所包含的预设目标语言基础子特征，根据上述对子特征所对应内容的描述，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征，将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征。例如，若预设目标语言基础子特征为 Unigram 单词（字）占比，通过统计 Unigram 单词个数占全文单词总数目的比例以得到预设目标语言基础子特征，若预设目标语言基础子特征为平均笔画数，通过统计单个单词笔画数的平均数目以得到目标语言基础子特征，对于其它的预设目标语言基础子特征，均根据每个预设目标语言基础子特征抽取方式进行所对应的预设目标语言基础子特征抽取，以得到预设目标语言基础子特征，将得到的所有目标语言基础

子特征组成集合以得到所述预设语言基础特征。

在一实施例中，所述根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征的步骤包括：根据所述作文中所包含的标点符号将所述作文切割为若干个句子；基于所述句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素。

具体地，作文中一般是以标点符号作为句子的划分，因此通过对标点符号的识别，根据识别出来的标点符号，可以将作为切割为若干个独立的句子，基于上述对预设文章结构特征的描述，结合切割获得的句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素，从而根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征。比如，将每个句子分类为背景、论点、主要观点、支撑观点、结论中的一类，将最大连续出现的某个类别的句子集合定义为该类别的一个元素，例如 3 句主要观点+ 4 句支撑观点+ 2 句主要观点，则对应 3 个元素，计算元素个数作为作文结构指标，统计作文中相邻句子的 Overlap （即重叠）情况，以得到所述预设文章连贯性特征所对应的相邻句子之间的重叠数量作为作文结构指标，统计所述预设文章元素特征所对应的所有所述元素中单词数目的均值作为作文的结构指标等。

请参阅图 2，图 2 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法中一个子流程的示意图。在该实施例中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤包括：

S201、获取预设语料库中所包含的预设文章；S202、将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度。

其中，预设特征指标包括预设语言基础特征、预设逻辑特征及预设理解力特征中的全部或者部分子特征。

具体地，获取预设语料库中所包含的预设文章，预设特征指标包括预设语言基础特征、预设逻辑特征及预设理解力特征中的全部或者部分子特征。

请继续参阅图 2，如图 2 所示，进一步地，所述预设文章为若干篇文章，所述将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度的步骤之后，还包括：S203、从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章；S204、获取与所述最相似文章同属一个文章类别的得分指标为对比指标；S205、基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度。

具体地，预设语料库中所包含的预设文章为若干篇文章，每篇文章对应有相应的得分，将语料库中的若干篇文章可以按照得分分为 n 类，从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章，进而获取与待评价文章最相似的文章类别的得分指标为对比指标，基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度，从而实现将待评分作文与最优质论文进行对比以得到所述作文与最优质论文之间的相似度，从而可以得到待评价作文与得分最高的文章类别的相似度，从而实现对所述作文进行较为准确的评分。

请参阅图 3，图 3 为本申请实施例提供的基于特征抽取及随机森林的作文评分方法的另一个子流程示意图。在该实施例中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤还包括：

S301、获取所述作文所对应的主题；S302、获取所述预设语料库中所述主题所包含的同

主题文章，其中， 所述同主题文章为与所述作文的主题相同的文章；S303、将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度。

具体地，获取所述作文所对应的主题，例如从所述作文的标题获得所述作文的主题，获取所述预设语料库中所述主题所包含的同主题文章，其中，所述同主题文章为与所述作文的主题相同的文章，将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度，从而实现将所述作文与同主题文章进行对比，通过所述作文与同主题文章的进行对比的类别相似度，可进而判断得出所述作文的跑题程度，从而从较为全面的维度对作文进行自动评分，提高对作文自动评分的准确性。

需要说明的是，上述各个实施例所述的基于特征抽取及随机森林的作文评分方法，可以根据需要将不同实施例中包含的技术特征重新进行组合，以获取组合后的实施方案，但都在本申请要求的保护范围之内。

请参阅图 4，图 4 为本申请实施例提供的基于特征抽取及随机森林的作文评分装置的一个示意性框图。对应于上述所述基于特征抽取及随机森林的作文评分方法，本申请实施例还提供一种基于特征抽取及随机森林的作文评分装置。如图 4 所示，该基于特征抽取及随机森林的作文评分装置包括用于执行上述所述基于特征抽取及随机森林的作文评分方法的单元，该基于特征抽取及随机森林的作文评分装置可以被配置于计算机设备中。具体地，请参阅图 4，该基于特征抽取及随机森林的作文评分装置 400 包括获取单元 401、抽取单元 402 及评分单元 403。

其中，获取单元 401，用于获取待评分的作文；抽取单元 402，用于根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；评分单元 403，用于将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

在一实施例中，所述抽取单元 402 包括：语言基础特征抽取子单元，用于根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征；逻辑特征抽取子单元，用于根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征；理解力特征抽取子单元，用于根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征。

在一实施例中，所述语言基础特征抽取子单元包括：分词子单元，用于采用第一预设语言工具对所述作文进行分词，以得到所述作文所包含的短语；子特征抽取单元，用于基于所述短语，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征；集合子单元，用于将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征；

在一实施例中，所述逻辑特征抽取子单元包括：切割子单元，用于根据所述作文中所包含的标点符号将所述作文切割为若干个句子；统计子单元，用于基于所述句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素。

在一实施例中，所述理解力特征抽取子单元包括：获取子单元，用于获取预设语料库中所包含的预设文章；对比子单元，用于将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度。

在一实施例中，所述预设文章为若干篇文章，所述理解力特征抽取子单元还包括：最相似文章获得子单元，用于从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章；对比指标获取单元，用于获取与所述最相似文章同属一个文章类别的得分指标为对比

指标；对比子单元，用于基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度。

在一实施例中，所述理解力特征抽取子单元还包括：主题获取子单元，用于获取所述作文所对应的主题；同主题文章获取子单元，用于获取所述预设语料库中所述主题所包含的同主题文章，其中，所述同主题文章为与所述作文的主题相同的文章；类别相似度获取子单元，用于将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度。

需要说明的是，所属领域的技术人员可以清楚地了解到，上述基于特征抽取及随机森林的作文评分装置和各单元的具体实现过程，可以参考前述方法实施例中的相应描述，为了描述的方便和简洁，在此不再赘述。

同时，上述基于特征抽取及随机森林的作文评分装置中各个单元的划分和连接方式仅用于举例说明，在其他实施例中，可将基于特征抽取及随机森林的作文评分装置按照需要划分为不同的单元，也可将基于特征抽取及随机森林的作文评分装置中各单元采取不同的连接顺序和方式，以完成上述基于特征抽取及随机森林的作文评分装置的全部或部分功能。

上述基于特征抽取及随机森林的作文评分装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 5 所示的计算机设备上运行。

请参阅图 5，图 5 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 500 可以是台式机电脑或者服务器等计算机设备，也可以是其他设备中的组件或者部件。

参阅图 5，该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505，其中，存储器可以包括非易失性存储介质 503 和内存储器 504。

该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 被执行时，可使得处理器 502 执行一种上述基于特征抽取及随机森林的作文评分方法。

该处理器 502 用于提供计算和控制能力，以支撑整个计算机设备 500 的运行。

该内存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境，该计算机程序 5032 被处理器 502 执行时，可使得处理器 502 执行一种上述基于特征抽取及随机森林的作文评分方法。

该网络接口 505 用于与其它设备进行网络通信。本领域技术人员可以理解，图 5 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备 500 的限定，具体的计算机设备 500 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。例如，在一些实施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 5 所示实施例一致，在此不再赘述。

其中，所述处理器 502 用于运行存储在存储器中的计算机程序 5032，以实现本申请实施例所描述的作文评分方法。

应当理解，在本申请实施例中，处理器 502 可以是中央处理单元 (Central Processing Unit，CPU)，该处理器 502 还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor，DSP)、专用集成电路 (Application Specific Integrated Circuit，ASIC)、现成可编程门阵列 (Field-Programmable Gate Array，FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

本领域普通技术人员可以理解的是实现上述实施例的方法中的全部或部分流程，是可以通过计算机程序来完成，该计算机程序可存储于一计算机可读存储介质。该计算机程序被该计算机系统中的至少一个处理器执行，以实现上述方法的实施例的步骤。

因此，本申请实施例还提供一种计算机可读存储介质。该计算机可读存储介质可以为非易失性的计算机可读存储介质，也可以为易失性的计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序被处理器执行时使处理器执行以上各实施例中所描述

的所述基于特征抽取及随机森林的作文评分方法的步骤。

所述存储介质为实体的、非瞬时性的存储介质，例如可以是U盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、磁碟或者光盘等各种可以存储计算机程序的实体存储介质。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

以上所述，仅为本申请的具体实施方式，但本申请明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1. 一种基于特征抽取及随机森林的作文评分方法，包括:

获取待评分的作文;

根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征;

将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

2. 根据权利要求 1 所述基于特征抽取及随机森林的作文评分方法，其中，所述根据预设特征抽取方式抽取所述作文的目标特征的步骤包括:

根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征;

根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征;

根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征。

3. 根据权利要求 2 所述基于特征抽取及随机森林的作文评分方法，其中，所述根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征的步骤包括:

采用第一预设语言工具对所述作文进行分词，以得到所述作文所包含的短语;

基于所述短语，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征;

将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征。

4. 根据权利要求 2 所述基于特征抽取及随机森林的作文评分方法，其中，所述根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征的步骤包括:

根据所述作文中所包含的标点符号将所述作文切割为若干个句子;

基于所述句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素。

5. 根据权利要求 2 所述基于特征抽取及随机森林的作文评分方法，其中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤包括:

获取预设语料库中所包含的预设文章;

将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度。

6. 根据权利要求 5 所述基于特征抽取及随机森林的作文评分方法，其中，所述预设文章为若干篇文章，所述将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度的步骤之后，还包括:

从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章;

获取与所述最相似文章同属一个文章类别的得分指标为对比指标;

基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度。

7. 根据权利要求 5 所述基于特征抽取及随机森林的作文评分方法，其中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤还包括:

获取所述作文所对应的主题;

获取所述预设语料库中所述主题所包含的同主题文章，其中， 所述同主题文章为与所述作文的主题相同的文章;

将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度。

8. 一种基于特征抽取及随机森林的作文评分装置，包括：

获取单元，用于获取待评分的作文；

抽取单元，用于根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；

评分单元，用于将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

9. 一种计算机设备，所述计算机设备包括存储器以及与所述存储器相连的处理器；所述存储器用于存储计算机程序；所述处理器用于运行所述计算机程序，以执行如下步骤：

获取待评分的作文；

根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；

将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

10. 根据权利要求 9 所述计算机设备，其中，所述根据预设特征抽取方式抽取所述作文的目标特征的步骤包括：

根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征；

根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征；

根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征。

11. 根据权利要求 10 所述计算机设备，其中，所述根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征的步骤包括：

采用第一预设语言工具对所述作文进行分词，以得到所述作文所包含的短语；

基于所述短语，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征；

将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征。

12. 根据权利要求 10 所述计算机设备，其中，所述根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征的步骤包括：

根据所述作文中所包含的标点符号将所述作文切割为若干个句子；

基于所述句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素。

13. 根据权利要求 10 所述计算机设备，其中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤包括：

获取预设语料库中所包含的预设文章；

将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度。

14. 根据权利要求 13 所述计算机设备，其中，所述预设文章为若干篇文章，所述将所述

作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度的步骤之后，还包括：

从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章；

获取与所述最相似文章同属一个文章类别的得分指标为对比指标；

基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度。

15. 根据权利要求 13 所述计算机设备，其中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤还包括：

获取所述作文所对应的主题；

获取所述预设语料库中所述主题所包含的同主题文章，其中， 所述同主题文章为与所述作文的主题相同的文章；

将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度。

16. 一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现如下步骤：

获取待评分的作文；

根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，所述预设语言基础特征包括预设词汇特征、预设语法特征、预设句法特征、预设篇幅特征及预设可读性特征，所述预设逻辑特征包括预设文章结构特征、预设文章连贯性特征及预设文章元素特征，所述预设理解力特征包括与预设文章的相似度特征；

将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数。

17. 根据权利要求 16 所述计算机可读存储介质，其中，所述根据预设特征抽取方式抽取所述作文的目标特征的步骤包括：

根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征；

根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征；

根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征。

18. 根据权利要求 17 所述计算机可读存储介质，其中，所述根据预设语言基础特征抽取方式抽取所述作文的所述预设语言基础特征的步骤包括：

采用第一预设语言工具对所述作文进行分词，以得到所述作文所包含的短语；

基于所述短语，根据每个预设目标语言基础子特征抽取方式进行特征抽取，以得到每个目标语言基础子特征；

将得到的若干个目标语言基础子特征组成集合以得到所述预设语言基础特征。

19. 根据权利要求 17 所述计算机可读存储介质，其中，所述根据预设逻辑特征抽取方式抽取所述作文的所述预设逻辑特征的步骤包括：

根据所述作文中所包含的标点符号将所述作文切割为若干个句子；

基于所述句子，统计出所述作文的所述预设文章结构特征所对应的所述作文本所包含的元素个数、所述预设文章元素特征所对应的所有所述元素中单词数目的均值、所述预设文章连贯性特征所对应的相邻句子之间的重叠数量，其中，所述元素为所述作文中所包含的背景类、论点类、主要观点类、支撑观点类或者结论类别中，最大连续出现的类别的句子集合称为该类别的元素。

20. 根据权利要求 17 所述计算机可读存储介质，其中，所述根据预设理解力特征抽取方式抽取所述作文的所述预设理解力特征的步骤包括：

获取预设语料库中所包含的预设文章；

将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文

章的预设相似度。

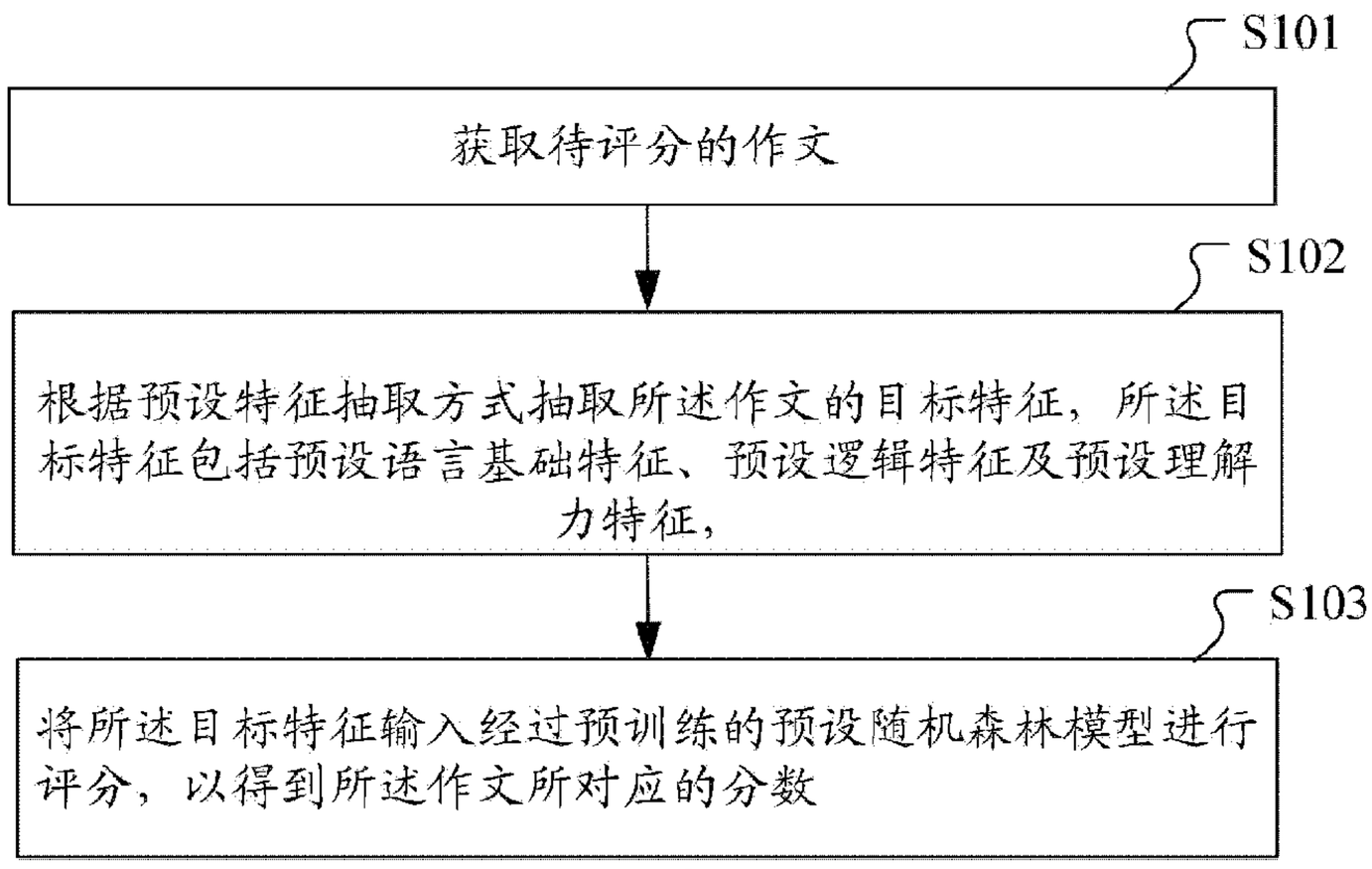

S101
获取待评分的作文
S102
根据预设特征抽取方式抽取所述作文的目标特征，所述目标特征包括预设语言基础特征、预设逻辑特征及预设理解力特征，
S103
将所述目标特征输入经过预训练的预设随机森林模型进行评分，以得到所述作文所对应的分数

图 1

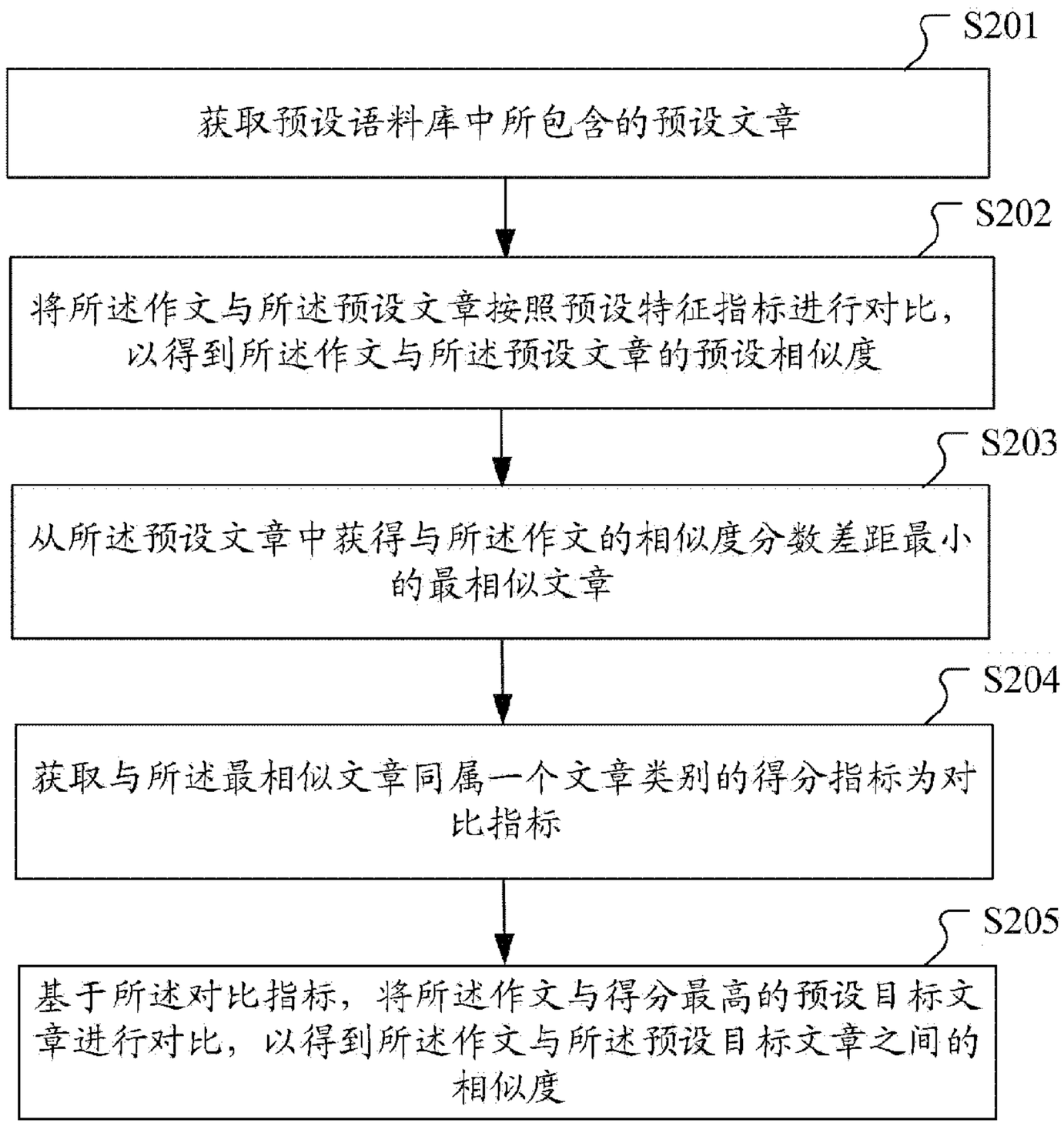

S201
获取预设语料库中所包含的预设文章
S202
将所述作文与所述预设文章按照预设特征指标进行对比，以得到所述作文与所述预设文章的预设相似度
S203
从所述预设文章中获得与所述作文的相似度分数差距最小的最相似文章
S204
获取与所述最相似文章同属一个文章类别的得分指标为对比指标
S205
基于所述对比指标，将所述作文与得分最高的预设目标文章进行对比，以得到所述作文与所述预设目标文章之间的相似度

图 2

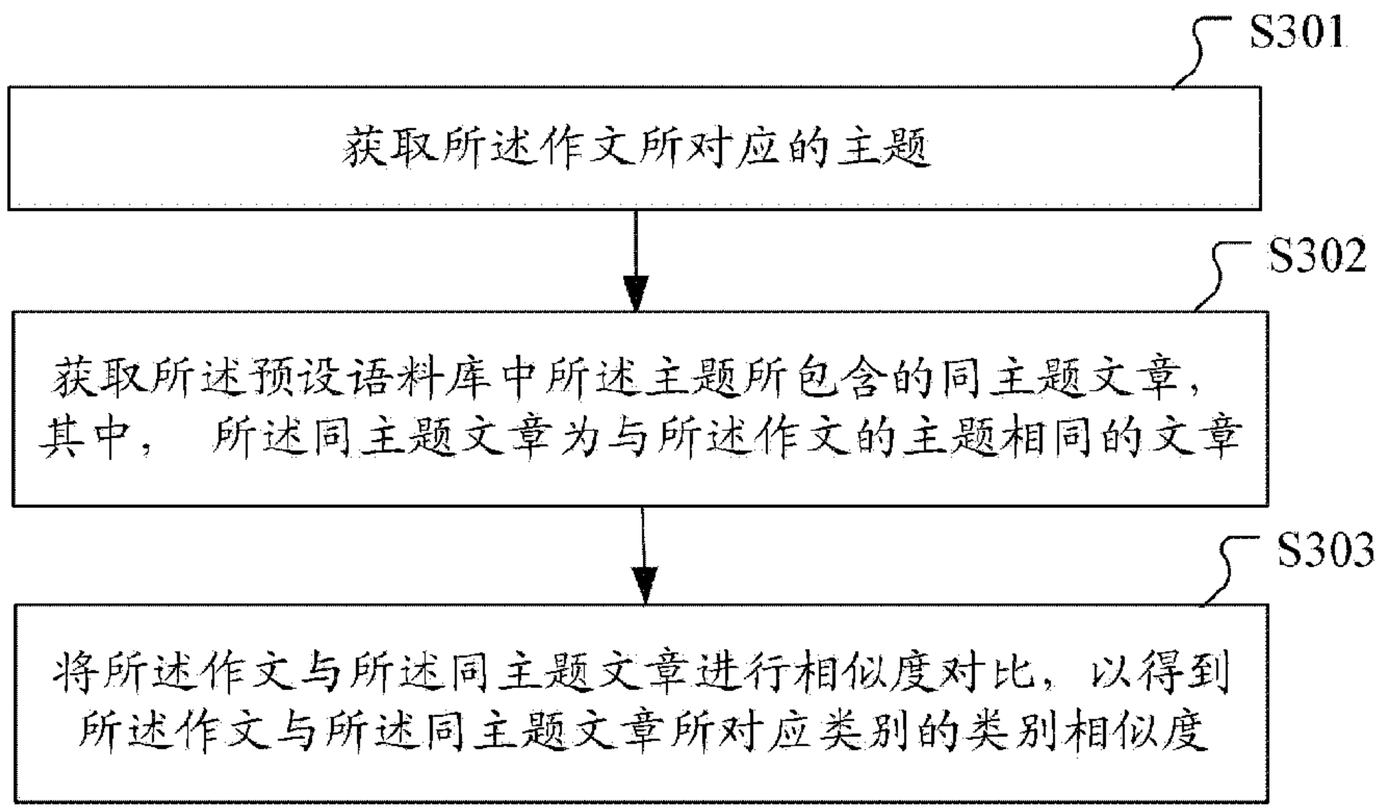

S301
获取所述作文所对应的主题
S302
获取所述预设语料库中所述主题所包含的同主题文章，其中， 所述同主题文章为与所述作文的主题相同的文章
S303
将所述作文与所述同主题文章进行相似度对比，以得到所述作文与所述同主题文章所对应类别的类别相似度

图 3

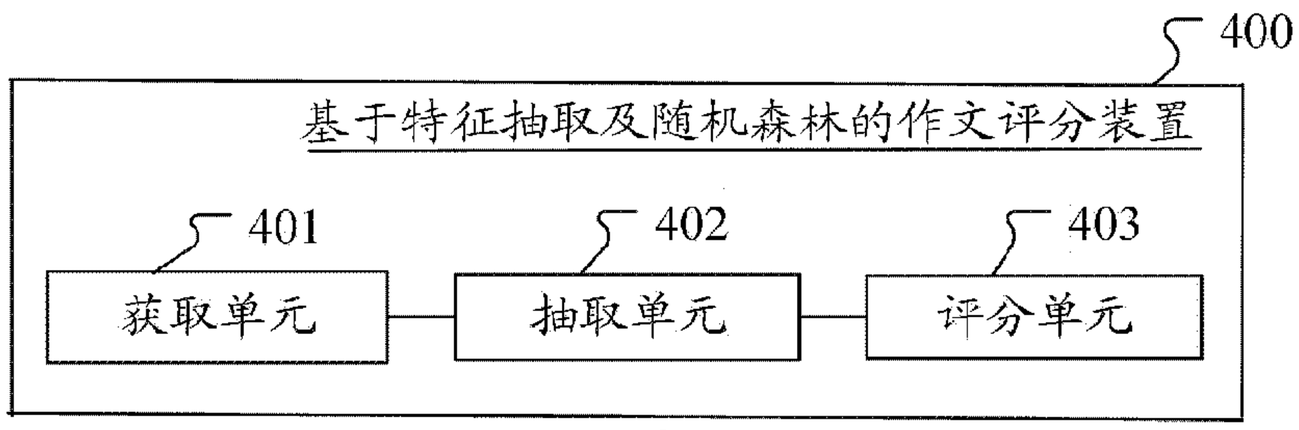

400
基于特征抽取及随机森林的作文评分装置
401
获取单元
402
抽取单元
403
评分单元

图 4

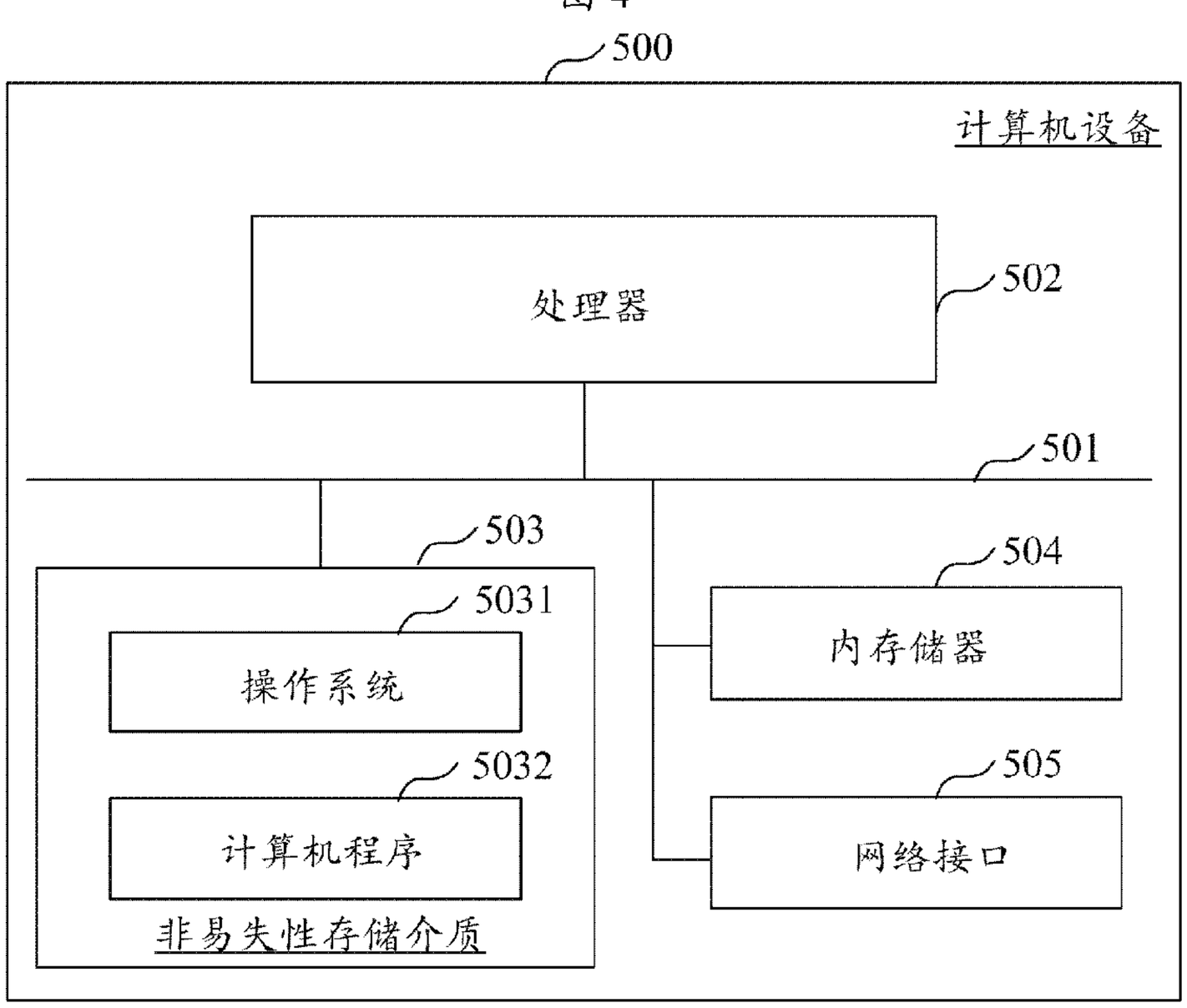

500
计算机设备
处理器
502
501
503
5031
操作系统
5032
计算机程序
非易失性存储介质
504
内存储器
505
网络接口

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/118737

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/205(2020.01)i; G06F 40/194(2020.01)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; G09B

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)

CNPAT; WPI; EPODOC; CNKI; IEEE: 文章, 文本, 作文, 评分, 自动, 模型, article, text, essay, scoring, auto+, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107506360 A (IFLYTEK CO., LTD.) 22 December 2017 (2017-12-22) description, paragraphs [0085]-[0163]	1-20
X	CN 110851599 A (SUN YAT-SEN UNIVERSITY) 28 February 2020 (2020-02-28) description, paragraphs [0062]-[0095]	1-20
X	CN 110069768 A (GLOBALWISDOM INC. et al.) 30 July 2019 (2019-07-30) abstract, description, paragraphs [0009]-[0013]	1-20
X	CN 111104789 A (CENTRAL CHINA NORMAL UNIVERSITY) 05 May 2020 (2020-05-05) description, paragraphs [0038]-[0126]	1-20
A	CN 110795538 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 14 February 2020 (2020-02-14) entire document	1-20
A	WO 2005045786 A1 (EDUCATIONAL TESTING SERVICE) 19 May 2005 (2005-05-19) 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	Date of mailing of the international search report
02 April 2021	**16 April 2021**
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/CN2020/118737

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107506360	A	22 December 2017	None			
CN	110851599	A	28 February 2020	None			
CN	110069768	A	30 July 2019	None			
CN	111104789	A	05 May 2020	None			
CN	110795538	A	14 February 2020	None			
WO	2005045786	A1	19 May 2005	US	2011027769	A1	03 February 2011
				US	2005142529	A1	30 June 2005

国际检索报告

国际申请号
PCT/CN2020/118737

A. 主题的分类

G06F 40/205(2020.01)i; G06F 40/194(2020.01)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; G09B

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））

CNPAT;WPI;EPODOC;CNKI;IEEE: 文章，文本，作文，评分，自动，模型，article，text，essay，scoring，auto+，model

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 107506360 A (科大讯飞股份有限公司) 2017年 12月 22日 (2017-12-22) 说明书第[0085]-[0163]段	1-20
X	CN 110851599 A (中山大学) 2020年 2月 28日 (2020-02-28) 说明书第[0062]-[0095]段	1-20
X	CN 110069768 A (北京博智天下信息技术有限公司 等) 2019年 7月 30日 (2019-07-30) 摘要，说明书第[0009]-[0013]段	1-20
X	CN 111104789 A (华中师范大学) 2020年 5月 5日 (2020-05-05) 说明书第[0038]-[0126]段	1-20
A	CN 110795538 A (腾讯科技深圳有限公司) 2020年 2月 14日 (2020-02-14) 全文	1-20
A	WO 2005045786 A1 (EDUCATIONAL TESTING SERVICE) 2005年 5月 19日 (2005-05-19) 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
2021年 4月 2日	2021年 4月 16日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 金霞 电话号码 86-(10)-53961524

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/118737

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	107506360	A	2017年 12月 22日	无			
CN	110851599	A	2020年 2月 28日	无			
CN	110069768	A	2019年 7月 30日	无			
CN	111104789	A	2020年 5月 5日	无			
CN	110795538	A	2020年 2月 14日	无			
WO	2005045786	A1	2005年 5月 19日	US	2011027769	A1	2011年 2月 3日
				US	2005142529	A1	2005年 6月 30日