(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 11 月 14 日 (14.11.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/214149 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2018/108219

(22) 国际申请日: 2018 年 9 月 28 日 (28.09.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810449138.8 2018年5月11日 (11.05.2018) 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: 李正洋 (LI, Zhengyang); 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)。 李海疆 (LI, Haijiang); 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: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路10号中浩大厦1528-1530室于志光, 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,

(54) **Title:** TEXT KEY INFORMATION IDENTIFICATION METHOD, ELECTRONIC DEVICE, AND READABLE STORAGE MEDIUM

(54) **发明名称:** 文本关键信息的识别方法、电子装置及可读存储介质

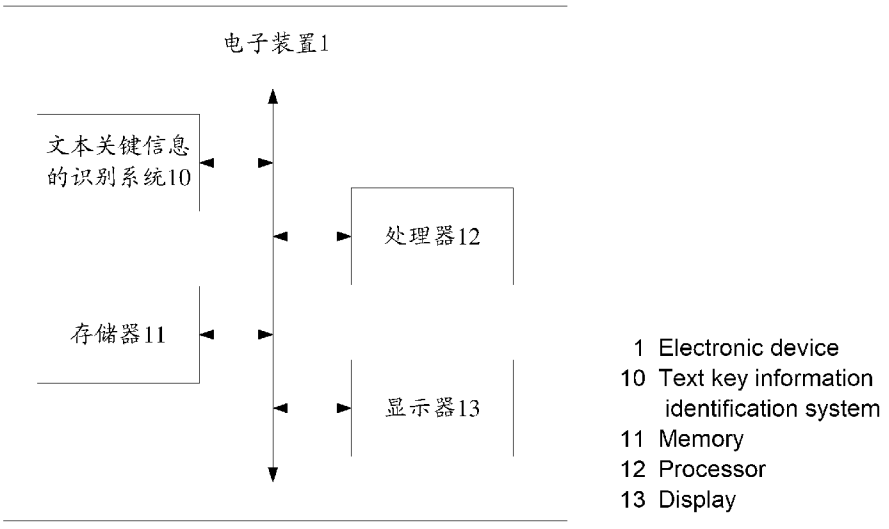


图 1

(57) **Abstract:** A text key information identification method, an electronic device, and a readable storage medium. The method comprises: performing, after receiving a text to be identified, word segmentation on the received text to be identified by using a predetermined word segmentation model to obtain segmented words of the text to be identified (S10), wherein the predetermined word segmentation model is a long short-term memory cycle neural network model obtained by pre-training a preset number of sample statements labeled using a sequence labeling method; and obtaining the score of each segmented word by calculation on the basis of word frequencies, positions, and word spans of the segmented words in the text to be identified and according to a preset scoring formula, ranking the segmented words in the text to be identified in descending order of scores, extracting highly-ranked segmented words as keywords, and obtaining key information of the text to be identified according to the extracted keywords (S20). By means of the method, a user can

WO 2019/214149 A1

[见续页]

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))。

quickly and accurately obtain the key information in a text to be identified.

(57) 摘要：一种文本关键信息的识别方法、电子装置及可读存储介质，该方法包括：在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词（S10）；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息（S20）。该方法能使用户快速准确地获取待识别文本中的关键信息。

文本关键信息的识别方法、电子装置及可读存储介质

优先权申明

本申请基于巴黎公约申明享有 2018 年 05 月 11 日递交的申请号为 CN 2018104491388、名称为“文本关键信息的识别方法、电子装置及可读存储介质”中国专利申请的优先权，该中国专利申请的整体内容以参考的方式结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种文本关键信息的识别方法、电子装置及可读存储介质。

背景技术

用户经常需要从各种文本中快速获取到关键内容如文本的核心观点，然而，一些文本例如证券研究报告内容繁多且主旨不够清晰，特别是对于长达十几页甚至更多的深度报告，需要花费较多的时间去理清分析师的逻辑及相应结论。这既对看报告的用户的专业能力有较高要求，又需要花费用户较长的工作时间，导致用户获取文本中关键内容的效率较低且准确率不高。

发明内容

本申请的目的在于提供一种文本关键信息的识别方法、电子装置及可读存储介质，旨在快速准确地获取文本中的关键信息。

为实现上述目的，本申请第一方面提供一种电子装置，所述电子装置包括存储器、处理器，所述存储器上存储有可在所述处理器上运行的文本关键信息的识别系统，所述文本关键信息的识别系统被所述处理器执行时实现如下步骤：

在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；

基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

此外，为实现上述目的，本申请第二方面还提供一种文本关键信息的识别方法，所述文本关键信息的识别方法包括：

在收到待识别文本后，对收到的待识别文本利用预先确定的分词

模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；

基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

进一步地，为实现上述目的，本申请第三方面还提供一种计算机可读存储介质，所述计算机可读存储介质存储有文本关键信息的识别系统，所述文本关键信息的识别系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述的文本关键信息的识别方法的步骤。

本申请提出的文本关键信息的识别方法、系统及可读存储介质，通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型来对收到的待识别文本进行分词，并基于所述待识别文本中各个分词的词频、位置及词跨度来利用预设评分公式计算得到每个分词的评分，对各个分词按评分高低进行排序，根据排序靠前的分词来获取所述待识别文本的关键信息。由于能利用训练好的长短期记忆循环神经网络模型对待识别文本进行自动分词，且能综合考虑各个分词的词频、位置及词跨度来提取出所述待识别文本中的关键信息，无需用户自身花费较多的时间去理清待识别文本的核心内容，能使用户快速准确地获取待识别文本中的关键信息。

附图说明

图 1 为本申请文本关键信息的识别系统 10 较佳实施例的运行环境示意图；

图 2 为本申请文本关键信息的识别方法一实施例的流程示意图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。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。

需要说明的是，在本申请中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的

技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的保护范围之内。

本申请提供一种文本关键信息的识别系统。请参阅图 1，是本申请文本关键信息的识别系统 10 较佳实施例的运行环境示意图。

在本实施例中，所述的文本关键信息的识别系统 10 安装并运行于电子装置 1 中。该电子装置 1 可包括，但不仅限于，存储器 11、处理器 12 及显示器 13。图 1 仅示出了具有组件 11-13 的电子装置 1，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

所述存储器 11 为至少一种类型的可读计算机存储介质，所述存储器 11 在一些实施例中可以是所述电子装置 1 的内部存储单元，例如该电子装置 1 的硬盘或内存。所述存储器 11 在另一些实施例中也可以是所述电子装置 1 的外部存储设备，例如所述电子装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。进一步地，所述存储器 11 还可以既包括所述电子装置 1 的内部存储单元也包括外部存储设备。所述存储器 11 用于存储安装于所述电子装置 1 的应用软件及各类数据，例如所述文本关键信息的识别系统 10 的程序代码等。所述存储器 11 还可以用于暂时地存储已经输出或者将要输出的数据。

所述处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit, CPU），微处理器或其他数据处理芯片，用于运行所述存储器 11 中存储的程序代码或处理数据，例如执行所述文本关键信息的识别系统 10 等。

所述显示器 13 在一些实施例中可以是 LED 显示器、液晶显示器、触控式液晶显示器以及 OLED（Organic Light-Emitting Diode，有机发光二极管）触摸器等。所述显示器 13 用于显示在所述电子装置 1 中处理的信息以及用于显示可视化的用户界面，例如待识别文本、分词结果、最终识别出的关键词、关键信息等。所述电子装置 1 的部件 11-13 通过系统总线相互通信。

文本关键信息的识别系统 10 包括至少一个存储在所述存储器 11 中的计算机可读指令，该至少一个计算机可读指令可被所述处理器 12 执行，以实现本申请各实施例。

其中，上述文本关键信息的识别系统 10 被所述处理器 12 执行时实现如下步骤：

步骤 S1，在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其

中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型。

本实施例中，文本关键信息的识别系统接收用户发出的包含待识别文本的关键信息识别请求，例如，接收用户通过手机、平板电脑、自助终端设备等终端发送的关键信息识别请求，如接收用户在手机、平板电脑、自助终端设备等终端中预先安装的客户端上发送来的关键信息识别请求，或接收用户在手机、平板电脑、自助终端设备等终端中的浏览器系统上发送来的关键信息识别请求。

在一种可选的实施方式中，本实施例中的待识别文本为国内各券商、机构发布的研究报告，由于中文不像英语等拉丁语系，除了标点符号之外，字之间紧密相连，没有明显的词边界，因此很难将词简单准确的提取出来。在中文中，单字作为最基本的语义单位，虽然也有自己的意义，但表意能力较差，意义较分散，而词的表意能力更强，能更加准确的描述一个事物，在自然语言处理中，通常情况下词（包括单字成词）是最基本的处理单位，因此必须对研究报告进行准确的分词处理。本实施例中采用序列标注法训练一个长短期记忆循环神经网络（Long Short-Term Memory，简称 LSTM）模型来进行分词，该长短期记忆循环神经网络模型的训练过程如下：

获取预设数量（例如，10 万）的样本语句，其中，样本语句中的词语为预设语料库中采用预设序列标注规则标注好的词语。其中，该预设序列标注规则为按照每个字在词中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。例如，在一段文字中，可以将每个字按照在词中的位置进行标注，常用的标记有以下四个 label：B，Begin，表示这个字是一个词的首字；M，Middle，表示这是一个词中间的字；E，End，表示这是一个词的尾字；S，Single，表示这是单字成词。分词的过程就是将一段字符输入模型，然后得到相应的标记序列，再根据标记序列进行分词。举例来说：“达观数据是企业大数据服务商”，经过模型后得到的理想标注序列是：“BMMESBEBMEBME”，最终还原的分词结果是“达观数据/是/企业/大数据/服务商”。该预设语料库使用的是经典的 bakeoff2005 中的微软研究院的切分语料，可将其中的 train 部分拿过来做训练，将 test 作为最终的测试。

将样本语句经过 word2vec 模型词向量化后得到相应的样本向量信息。其中，word2vec 是 2013 年末 google 发布的 word2vec 工具，可以看做是深度学习在 NLP 领域的一个重要应用，虽然 word2vec 只有三层神经网络，但是已经取得非常好的效果。通过 word2vec，可以将一个词表示为词向量，将文字数字化，更好的让计算机理解。使 word2vec 模型，可以方便的找到同义词或联系紧密的词，或者意义

相反的词等。

将预设数量的样本语句按照预设比例（例如，9:1）分为训练集和评估集。在使用训练集对参数进行训练的时候，通常会将一整个训练集分为三个部分。一般分为：训练集（train_set），评估集（valid_set），测试集（test_set）这三个部分。这其实是为了保证训练效果而特意设置的。其中测试集其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。在实际的训练中，训练的结果对于训练集的拟合程度通常还行，但是对于训练集之外的数据的拟合程度通常就不行了。因此通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。本实施例中即采用这种交叉验证（Cross Validation）的方式。K-fold cross-validation 即 K 折交叉验证，初始采样分割成 K 个子样本，一个单独的子样本被保留作为验证模型的数据即本实施例中将预设数量的样本语句按照预设比例划分的评估集，其他 K-1 个样本用来训练即本实施例中将预设数量的样本语句按照预设比例划分的训练集。交叉验证重复 K 次，每个子样本验证一次，平均 K 次的结果或者使用其它结合方式，最终得到一个单一估测。同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10 折交叉验证是最常用的。

具体的，可将所述训练集对应的样本向量信息送入长短期记忆循环神经网络模型进行模型训练，每隔预设时间（例如每进行 1000 次迭代），使用训练得到的模型对所述评估集中样本语句的各个词语的每个字的标注进行识别，并将识别字符与该样本语句中采用预设序列标注规则的标注进行比对，以评估模型识别标注的误差；具体地，在计算误差时，可采用编辑距离作为计算标准。若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数（如 activation 函数、LSTM 层数、输入输出的变量维度等）并重新训练，直至使得训练得到的模型识别标注的误差能够收敛；若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为训练好的模型。

本实施例中采用的是序列标注法来评判模型的误差，目的是最小化这个误差（最大化预测标注的精度）来更新模型的参数，训练模型采用的是反向传播算法，反向传播算法（Backpropagation）是目前用来训练人工神经网络（Artificial Neural Network，ANN）的最常用且最有效的算法。其主要思想是：

（1）将训练集数据输入到 ANN 的输入层，经过隐藏层，最后达到输出层并输出结果，这是 ANN 的前向传播过程；

（2）由于 ANN 的输出结果与实际结果有误差，则计算估计值与实际值之间的误差，并将该误差从输出层向隐藏层反向传播，直至传

播到输入层;

（3）在反向传播的过程中，根据误差调整各种参数的值；不断迭代上述过程，直至收敛。

本实施例中在模型训练过程中关注的是标注的准确性，而在训练结束后实际应用时是根据模型得到的标注来还原得到最终的分词结果。即利用预先确定的分词模型识别出所述待识别文本中各个字的标注得到标注序列，并根据预设序列标注规则将所述待识别文本中各个字按标注序列还原成对应的分词。例如，在对文本中句子‘今天是个好日子’进行词向量化后，再通入训练好的分词模型（长短期记忆循环神经网络模型），得到相应的标注 label“BESSBME”，然后即可根据预先设定好的序列标注规则（例如“BE”、“S”、“BME”形式均作为分词）得到根据预测的 label 得到最终的分词结果“今天”、“是”、“个”、“好日子”。

步骤 S2，基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

本实施例中，在得到待识别文本的各个分词后，可根据无监督的统计类方法对各个分词进行排序打分来提取出关键词或关键内容。具体的，先去除得到的各个分词中的重复词汇及无意义词汇，如将得到的分词文本中的所有词汇提取成一个大的集合（集合中词汇不重复，即每个词在这个集合中只出现一次），得到文本词汇集合后，删除一些无意义的词如“我、你 、的、是”等。将这些代词、介词等无具体意义的词汇预先按经验组成一个 list，然后删除这些无意义的词。接下来在词汇集合中，筛选出的是一些比较专业、有具体意义的词汇。对筛选出的各个分词按照如下公式计算得到 P 值:

$$P=a1* X1+a2* X2+a3* X3$$

其中，X1 为分词的词频评分，a1 为预设的词频权重；X2 为分词出现位置的位置评分，a2 为预设的位置权重；X3 为分词的词跨度评分，a3 为预设的词跨度权重；将各个分词按评分 P 值高低进行排序，排序靠前的预设数量分词作为文本的关键词。可根据词频、位置、词跨度的不同影响或重要性分配不同的权重。

本实施例中考虑词频、位置、词跨度来对不同分词进行评分排序，如可按照词频大小进行排序，然后按照词的位置对每个词进行打分，再进行排序，再用词跨度进行排序，最后在获取其中排序靠前的 TopK（K 的范围是个可自选的范围，根据需求来设定）个词作为关键词，再去提取文本中与关键词有关的句子，即可最终提取出文本的关键信

息，也是整篇文章的主要内容及重要观点。

具体地，对于词频，词频表示一个词在文本中出现的频率。一般认为，如果一个词在文本中出现的越是频繁，那么这个词就越有可能作为文章的核心词。词频简单地统计了词在文本中出现的次数，但是，只依靠词频所得到的关键词有很大的不确定性，对于长度比较长的文本，这个方法会有很大的噪音，因此，还综合考虑位置和词跨度。

对于位置信息，一般情况下，词出现的位置对于词来说有着很大的价值。例如，标题、摘要本身就是作者概括出的文章的中心思想，因此出现在这些地方的词具有一定的代表性，更可能成为关键词。但是，因为每个作者的习惯不同，写作方式不同，关键句子的位置也会有所不同，所以这也是一种很宽泛的得到关键词的方法，一般情况下不会单独使用。本实施例中采用的比例是 5: 5: 1 的方式设定为词位置在 开头、结尾、中间的重要性程度，开头、结尾、中间的划分比例为 10%、10%、80%。例如，如果一个待识别文本一共由一万个词有序组成，位于前百分之十位置的词划分为开头位置，位于最后百分之十位置的词划分为结尾位置，其余作为中间位置，重要程度为 5: 5: 1。例如，假设“试验区”一词一共在待识别文本中出现过 5 次，2 次在开头，1 次在结尾，其余在中间，于是‘试验区’一词关于位置的分数为：5 × 2+5*1+1*2=17。

对于词跨度，词跨度是指一个词或者短语在待识别文本中首次出现和末次出现之间的距离，词跨度越大说明这个词对文本越重要，可以反映文本的主题。一个词的跨度计算公式如下：

$$span_i = \frac{last_i - first_i + 1}{sum}$$

其中，$last_i$表示词在文本中最后出现的位置，$first_i$表示词在文本中第一次出现的位置，sum表示文本中词的总数。在实际应用中，文本中总是有很多噪声（指不是关键词的那些词），使用词跨度来参考可以减少这些噪声。

本实施例通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型来对收到的待识别文本进行分词，并基于所述待识别文本中各个分词的词频、位置及词跨度来利用预设评分公式计算得到每个分词的评分，对各个分词按评分高低进行排序，根据排序靠前的分词来获取所述待识别文本的关键信息。由于能利用训练好的长短期记忆循环神经网络模型对待识别文本进行自动分词，且能综合考虑各个分词的词频、位置及词跨度来提取出所述待识别文本中的关键信息，无需用户自身花费较多的时间去理清待识别文本的核心内容，能使用户快速准确地获取待识别文本中的关

键信息。

如图 2 所示，图 2 为本申请文本关键信息的识别方法一实施例的流程示意图，该文本关键信息的识别方法包括以下步骤：

步骤 S10，在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型。

本实施例中，文本关键信息的识别系统接收用户发出的包含待识别文本的关键信息识别请求，例如，接收用户通过手机、平板电脑、自助终端设备等终端发送的关键信息识别请求，如接收用户在手机、平板电脑、自助终端设备等终端中预先安装的客户端上发送来的关键信息识别请求，或接收用户在手机、平板电脑、自助终端设备等终端中的浏览器系统上发送来的关键信息识别请求。

在一种可选的实施方式中，本实施例中的待识别文本为国内各券商、机构发布的研究报告，由于中文不像英语等拉丁语系，除了标点符号之外，字之间紧密相连，没有明显的词边界，因此很难将词简单准确的提取出来。在中文中，单字作为最基本的语义单位，虽然也有自己的意义，但表意能力较差，意义较分散，而词的表意能力更强，能更加准确的描述一个事物，在自然语言处理中，通常情况下词（包括单字成词）是最基本的处理单位，因此必须对研究报告进行准确的分词处理。本实施例中采用序列标注法训练一个长短期记忆循环神经网络（Long Short-Term Memory，简称 LSTM）模型来进行分词，该长短期记忆循环神经网络模型的训练过程如下：

获取预设数量（例如，10 万）的样本语句，其中，样本语句中的词语为预设语料库中采用预设序列标注规则标注好的词语。其中，该预设序列标注规则为按照每个字在词中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。例如，在一段文字中，可以将每个字按照在词中的位置进行标注，常用的标记有以下四个 label：B，Begin，表示这个字是一个词的首字；M，Middle，表示这是一个词中间的字；E，End，表示这是一个词的尾字；S，Single，表示这是单字成词。分词的过程就是将一段字符输入模型，然后得到相应的标记序列，再根据标记序列进行分词。举例来说：“达观数据是企业大数据服务商”，经过模型后得到的理想标注序列是：“BMMESBEBMEBME”，最终还原的分词结果是“达观数据/是/企业/大数据/服务商”。该预设语料库使用的是经典的 bakeoff2005 中的微软研究院的切分语料，可将其中的 train 部分拿过来做训练，将 test 作为最终的测试。

将样本语句经过 word2vec 模型词向量化后得到相应的样本向量信息。其中，word2vec 是 2013 年末 google 发布的 word2vec 工具，可以看做是深度学习在 NLP 领域的一个重要应用，虽然 word2vec 只有三层神经网络，但是已经取得非常好的效果。通过 word2vec，可以将一个词表示为词向量，将文字数字化，更好的让计算机理解。使 word2vec 模型，可以方便的找到同义词或联系紧密的词，或者意义相反的词等。

将预设数量的样本语句按照预设比例（例如，9:1）分为训练集和评估集。在使用训练集对参数进行训练的时候，通常会将一整个训练集分为三个部分。一般分为：训练集（train_set），评估集（valid_set），测试集（test_set）这三个部分。这其实是为了保证训练效果而特意设置的。其中测试集其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。在实际的训练中，训练的结果对于训练集的拟合程度通常还行，但是对于训练集之外的数据的拟合程度通常就不行了。因此通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。本实施例中即采用这种交叉验证（Cross Validation）的方式。K-fold cross-validation 即 K 折交叉验证，初始采样分割成 K 个子样本，一个单独的子样本被保留作为验证模型的数据即本实施例中将预设数量的样本语句按照预设比例划分的评估集，其他 K-1 个样本用来训练即本实施例中将预设数量的样本语句按照预设比例划分的训练集。交叉验证重复 K 次，每个子样本验证一次，平均 K 次的结果或者使用其它结合方式，最终得到一个单一估测。同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10 折交叉验证是最常用的。

具体的，可将所述训练集对应的样本向量信息送入长短期记忆循环神经网络模型进行模型训练，每隔预设时间（例如每进行 1000 次迭代），使用训练得到的模型对所述评估集中样本语句的各个词语的每个字的标注进行识别，并将识别字符与该样本语句中采用预设序列标注规则的标注进行比对，以评估模型识别标注的误差；具体地，在计算误差时，可采用编辑距离作为计算标准。若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数（如 activation 函数、LSTM 层数、输入输出的变量维度等）并重新训练，直至使得训练得到的模型识别标注的误差能够收敛；若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为训练好的模型。

本实施例中采用的是序列标注法来评判模型的误差，目的是最小化这个误差（最大化预测标注的精度）来更新模型的参数，训练模型采用的是反向传播算法，反向传播算法（Backpropagation）是目前用

来训练人工神经网络（Artificial Neural Network，ANN）的最常用且最有效的算法。其主要思想是:

（1）将训练集数据输入到 ANN 的输入层，经过隐藏层，最后达到输出层并输出结果，这是 ANN 的前向传播过程;

（2）由于 ANN 的输出结果与实际结果有误差，则计算估计值与实际值之间的误差，并将该误差从输出层向隐藏层反向传播，直至传播到输入层;

（3）在反向传播的过程中，根据误差调整各种参数的值；不断迭代上述过程，直至收敛。

本实施例中在模型训练过程中关注的是标注的准确性，而在训练结束后实际应用时是根据模型得到的标注来还原得到最终的分词结果。即利用预先确定的分词模型识别出所述待识别文本中各个字的标注得到标注序列，并根据预设序列标注规则将所述待识别文本中各个字按标注序列还原成对应的分词。例如，在对文本中句子‘今天是个好日子’进行词向量化后，再通入训练好的分词模型（长短期记忆循环神经网络模型），得到相应的标注 label“BESSBME”，然后即可根据预先设定好的序列标注规则（例如“BE”、“S”、“BME”形式均作为分词）得到根据预测的 label 得到最终的分词结果“今天”、“是”、“个”、“好日子”。

步骤 S20，基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

本实施例中，在得到待识别文本的各个分词后，可根据无监督的统计类方法对各个分词进行排序打分来提取出关键词或关键内容。具体的，先去除得到的各个分词中的重复词汇及无意义词汇，如将得到的分词文本中的所有词汇提取成一个大的集合（集合中词汇不重复，即每个词在这个集合中只出现一次），得到文本词汇集合后，删除一些无意义的词如“我、你 、的、是”等。将这些代词、介词等无具体意义的词汇预先按经验组成一个 list，然后删除这些无意义的词。接下来在词汇集合中，筛选出的是一些比较专业、有具体意义的词汇。对筛选出的各个分词按照如下公式计算得到 P 值:

$$P=a1* X1+a2* X2+a3* X3$$

其中，X1 为分词的词频评分，a1 为预设的词频权重；X2 为分词出现位置的位置评分，a2 为预设的位置权重；X3 为分词的词跨度评分，a3 为预设的词跨度权重；将各个分词按评分 P 值高低进行排序，排序靠前的预设数量分词作为文本的关键词。可根据词频、位置、词

跨度的不同影响或重要性分配不同的权重。

本实施例中考虑词频、位置、词跨度来对不同分词进行评分排序，如可按照词频大小进行排序，然后按照词的位置对每个词进行打分，再进行排序，再用词跨度进行排序，最后在获取其中排序靠前的 TopK（K 的范围是个可自选的范围，根据需求来设定）个词作为关键词，再去提取文本中与关键词有关的句子，即可最终提取出文本的关键信息，也是整篇文章的主要内容及重要观点。

具体地，对于词频，词频表示一个词在文本中出现的频率。一般认为，如果一个词在文本中出现的越是频繁，那么这个词就越有可能作为文章的核心词。词频简单地统计了词在文本中出现的次数，但是，只依靠词频所得到的关键词有很大的不确定性，对于长度比较长的文本，这个方法会有很大的噪音，因此，还综合考虑位置和词跨度。

对于位置信息，一般情况下，词出现的位置对于词来说有着很大的价值。例如，标题、摘要本身就是作者概括出的文章的中心思想，因此出现在这些地方的词具有一定的代表性，更可能成为关键词。但是，因为每个作者的习惯不同，写作方式不同，关键句子的位置也会有所不同，所以这也是一种很宽泛的得到关键词的方法，一般情况下不会单独使用。本实施例中采用的比例是 5: 5: 1 的方式设定为词位置在 开头、结尾、中间的重要性程度，开头、结尾、中间的划分比例为 10%、10%、80%。例如，如果一个待识别文本一共由一万个词有序组成，位于前百分之十位置的词划分为开头位置，位于最后百分之十位置的词划分为结尾位置，其余作为中间位置，重要程度为 5: 5: 1。例如，假设“试验区”一词一共在待识别文本中出现过 5 次，2 次在开头，1 次在结尾，其余在中间，于是‘试验区’一词关于位置的分数为：5 × 2+5*1+1*2=17。

对于词跨度，词跨度是指一个词或者短语在待识别文本中首次出现和末次出现之间的距离，词跨度越大说明这个词对文本越重要，可以反映文本的主题。一个词的跨度计算公式如下：

$$span_i = \frac{last_i - first_i + 1}{sum}$$

其中，$last_i$表示词在文本中最后出现的位置，$first_i$表示词在文本中第一次出现的位置，sum表示文本中词的总数。在实际应用中，文本中总是有很多噪声（指不是关键词的那些词），使用词跨度来参考可以减少这些噪声。

本实施例通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型来对收到的待识别文本进行分词，并基于所述待识别文本中各个分词的词频、位置及词跨度来

利用预设评分公式计算得到每个分词的评分，对各个分词按评分高低进行排序，根据排序靠前的分词来获取所述待识别文本的关键信息。由于能利用训练好的长短期记忆循环神经网络模型对待识别文本进行自动分词，且能综合考虑各个分词的词频、位置及词跨度来提取出所述待识别文本中的关键信息，无需用户自身花费较多的时间去理清待识别文本的核心内容，能使用户快速准确地获取待识别文本中的关键信息。

此外，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质存储有文本关键信息的识别系统，所述文本关键信息的识别系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述实施例中的文本关键信息的识别方法的步骤，该文本关键信息的识别方法的步骤S10、S20等具体实施过程如上文所述，在此不再赘述。

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

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件来实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

以上参照附图说明了本申请的优选实施例，并非因此局限本申请的权利范围。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。另外，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

本领域技术人员不脱离本申请的范围和实质，可以有多种变型方案实现本申请，比如作为一个实施例的特征可用于另一实施例而得到又一实施例。凡在运用本申请的技术构思之内所作的任何修改、等同替换和改进，均应在本申请的权利范围之内。

权 利 要 求 书

1. 一种电子装置，其特征在于，所述电子装置包括存储器、处理器，所述存储器上存储有可在所述处理器上运行的文本关键信息的识别系统，所述文本关键信息的识别系统被所述处理器执行时实现如下步骤：

在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；

基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

2. 如权利要求 1 所述的电子装置，其特征在于，所述预先确定的分词模型的训练过程如下：

获取预设数量的样本语句，其中，所述样本语句中的词语为预设语料库中采用预设序列标注规则标注好的词语；

将所述样本语句经过 word2vec 模型词向量化后得到相应的样本向量信息；将预设数量的样本语句按照预设比例分为训练集和评估集；

将所述训练集对应的样本向量信息送入长短期记忆循环神经网络模型进行模型训练，每隔预设时间，使用训练得到的模型对所述评估集的样本语句各个词语中每个字的标注进行识别，并将识别出的标注与该样本语句中采用预设序列标注规则的标注进行比对，以评估模型识别标注的误差；

若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数并重新训练，直至使得训练得到的模型识别标注的误差能够收敛；

若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为训练好的所述预先确定的分词模型。

3. 如权利要求 2 所述的电子装置，其特征在于，所述预设序列标注规则为按照每个字在词语中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。

4. 如权利要求 3 所述的电子装置，其特征在于，所述对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词包括：

利用预先确定的分词模型识别出所述待识别文本中各个字的标注得到标注序列，并根据预设序列标注规则将所述待识别文本中各个字按标注序列还原成对应的分词。

5. 如权利要求 1 所述的电子装置，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

6. 如权利要求 2 所述的电子装置，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

7. 如权利要求 3 所述的电子装置，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

8. 如权利要求 4 所述的电子装置，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨

度评分，a3 为预设的词跨度权重。

9. 一种文本关键信息的识别方法，其特征在于，所述文本关键信息的识别方法包括:

在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；

基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

10. 如权利要求 9 所述的文本关键信息的识别方法，其特征在于，所述预先确定的分词模型的训练过程如下:

获取预设数量的样本语句，其中，所述样本语句中的词语为预设语料库中采用预设序列标注规则标注好的词语；

将所述样本语句经过 word2vec 模型词向量化后得到相应的样本向量信息；将预设数量的样本语句按照预设比例分为训练集和评估集；

将所述训练集对应的样本向量信息送入长短期记忆循环神经网络模型进行模型训练，每隔预设时间，使用训练得到的模型对所述评估集的样本语句各个词语中每个字的标注进行识别，并将识别出的标注与该样本语句中采用预设序列标注规则的标注进行比对，以评估模型识别标注的误差；

若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数并重新训练，直至使得训练得到的模型识别标注的误差能够收敛；

若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为训练好的所述预先确定的分词模型。

11. 如权利要求 10 所述的文本关键信息的识别方法，其特征在于，所述预设序列标注规则为按照每个字在词语中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。

12. 如权利要求 11 所述的文本关键信息的识别方法，其特征在于，所述对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词包括:

利用预先确定的分词模型识别出所述待识别文本中各个字的标注得到标注序列，并根据预设序列标注规则将所述待识别文本中各个字按标注序列还原成对应的分词。

13. 如权利要求 9 所述的文本关键信息的识别方法，其特征在于，

所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

14. 如权利要求 10 所述的文本关键信息的识别方法，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

15. 如权利要求 11 所述的文本关键信息的识别方法，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

16. 如权利要求 12 所述的文本关键信息的识别方法，其特征在于，所述根据预设评分公式计算得到每个分词的评分包括:

对所述待识别文本中的各个分词按照如下公式计算得到每个分词的评分 P 值:

$$P=a1*X1+a2*X2+a3*X3$$

其中，X1 为分词在所述待识别文本中出现频率的词频评分，a1 为预设的词频权重；X2 为分词在所述待识别文本中出现位置的位置评分，a2 为预设的位置权重；X3 为分词在所述待识别文本中的词跨度评分，a3 为预设的词跨度权重。

17. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有文本关键信息的识别系统，所述文本关键信息的识别系统被处理器执行时实现如下步骤:

在收到待识别文本后，对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词；其中，所述预先确定的分词模型为预先通过对采用序列标注法标注的预设数量样本语句进行训练得到的长短期记忆循环神经网络模型；

基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息。

18. 如权利要求 17 所述的计算机可读存储介质，其特征在于，所述预先确定的分词模型的训练过程如下：

获取预设数量的样本语句，其中，所述样本语句中的词语为预设语料库中采用预设序列标注规则标注好的词语；

将所述样本语句经过 word2vec 模型词向量化后得到相应的样本向量信息；将预设数量的样本语句按照预设比例分为训练集和评估集；

将所述训练集对应的样本向量信息送入长短期记忆循环神经网络模型进行模型训练，每隔预设时间，使用训练得到的模型对所述评估集的样本语句各个词语中每个字的标注进行识别，并将识别出的标注与该样本语句中采用预设序列标注规则的标注进行比对，以评估模型识别标注的误差；

若训练得到的模型识别标注的误差出现发散，则调整预设的训练参数并重新训练，直至使得训练得到的模型识别标注的误差能够收敛；

若训练得到的模型识别标注的误差收敛，则结束模型训练，将生成的模型作为训练好的所述预先确定的分词模型。

19. 如权利要求 18 所述的计算机可读存储介质，其特征在于，所述预设序列标注规则为按照每个字在词语中的位置进行标注，标注类型包括：首字标注、中间字标注、尾字标注、单字标注。

20. 如权利要求 19 所述的计算机可读存储介质，其特征在于，所述对收到的待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词包括：

利用预先确定的分词模型识别出所述待识别文本中各个字的标注得到标注序列，并根据预设序列标注规则将所述待识别文本中各个字按标注序列还原成对应的分词。

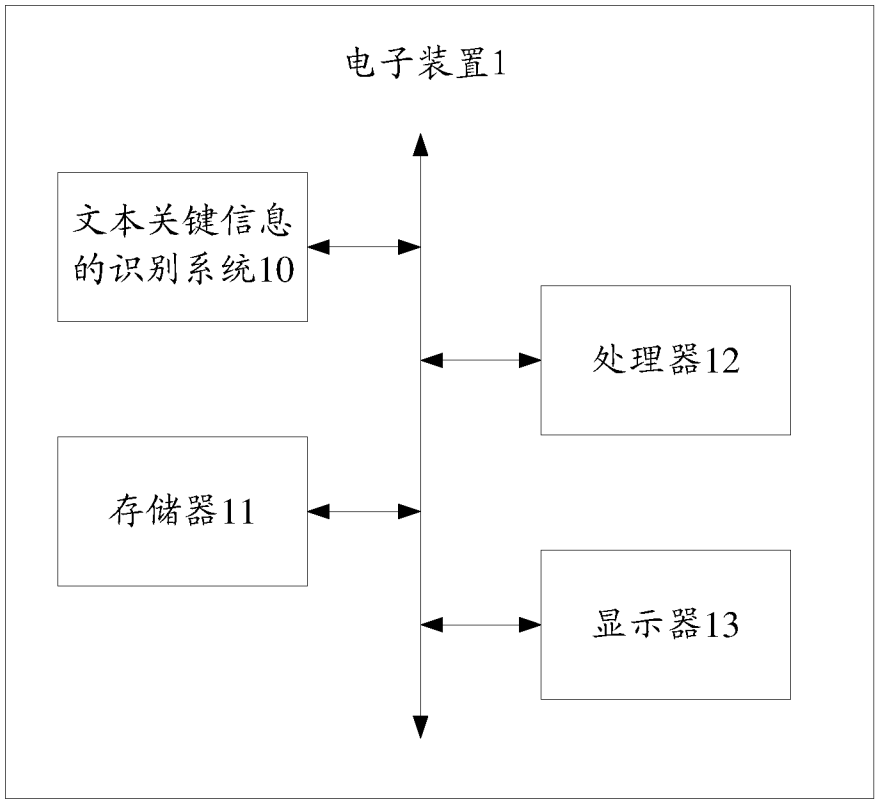

电子装置1
文本关键信息的识别系统10
处理器12
存储器11
显示器13

图 1

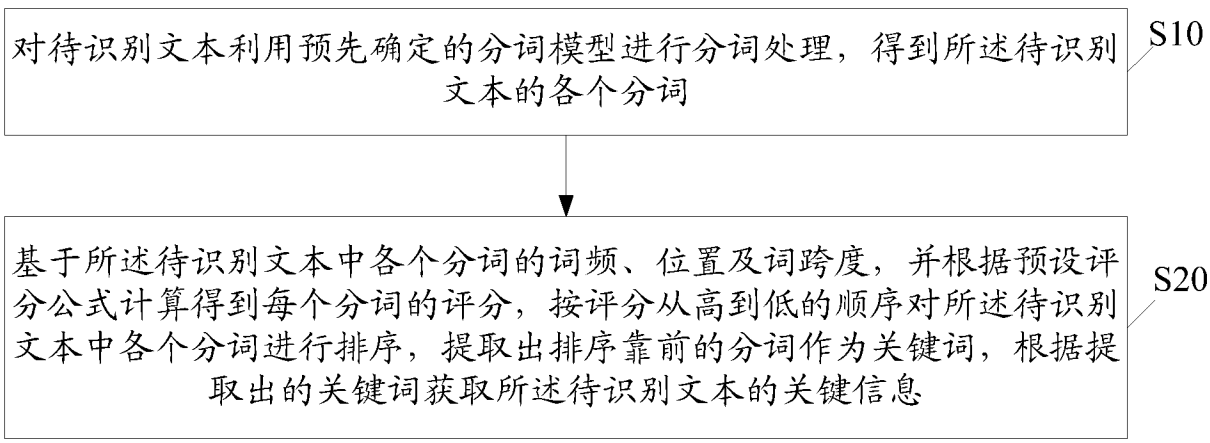

对待识别文本利用预先确定的分词模型进行分词处理，得到所述待识别文本的各个分词
S10
基于所述待识别文本中各个分词的词频、位置及词跨度，并根据预设评分公式计算得到每个分词的评分，按评分从高到低的顺序对所述待识别文本中各个分词进行排序，提取出排序靠前的分词作为关键词，根据提取出的关键词获取所述待识别文本的关键信息
S20

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/108219

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, CNKI, CNPAT, IEEE, GOOGLE: 文本, 识别, 模型, 关键词, 网络, 神经, text, identify, model, keyword, network, neural

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 107544956 A (IFLYTEK CO., LTD.) 05 January 2018 (2018-01-05) description, paragraphs 0149-0165, and figure 4	1-20
Y	CN 106557508 A (ULTRAPOWER SOFTWARE) 05 April 2017 (2017-04-05) description, paragraphs 0046-0073, and figure 1	1-20
A	CN 107562718 A (IFLYTEK CO., LTD.) 09 January 2018 (2018-01-09) entire document	1-20
A	US 2017052945 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 23 February 2017 (2017-02-23) 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
28 January 2019	**11 February 2019**

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

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/108219

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107544956	A	05 January 2018	None			
CN	106557508	A	05 April 2017	None			
CN	107562718	A	09 January 2018	None			
US	2017052945	A1	23 February 2017	US	2016124933	A1	05 May 2016
				JP	2016091078	A	23 May 2016

国际检索报告

国际申请号
PCT/CN2018/108219

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, EPODOC, CNKI, CNPAT, IEEE, GOOGLE: 文本, 识别, 模型, 关键词, 网络, 神经, text, identify, model, keyword, network, neural

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
Y	CN 107544956 A (科大讯飞股份有限公司) 2018年 1月 5日 (2018-01-05) 说明书第0149-0165段，附图4	1-20
Y	CN 106557508 A (北京神州泰岳软件股份有限公司) 2017年 4月 5日 (2017-04-05) 说明书第0046-0073段，附图1	1-20
A	CN 107562718 A (科大讯飞股份有限公司) 2018年 1月 9日 (2018-01-09) 全文	1-20
A	US 2017052945 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2017年 2月 23日 (2017-02-23) 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 1月 28日	2019年 2月 11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 王艳臣 电话号码 86-(10)-53961435

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/108219

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	107544956	A	2018年 1月 5日	无			
CN	106557508	A	2017年 4月 5日	无			
CN	107562718	A	2018年 1月 9日	无			
US	2017052945	A1	2017年 2月 23日	US	2016124933	A1	2016年 5月 5日
				JP	2016091078	A	2016年 5月 23日