(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 7 月 4 日 (04.07.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/128124 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2018/090878

(22) 国际申请日: 2018 年 6 月 12 日 (12.06.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201711484243.7 2017年12月29日 (29.12.2017) CN

(71) 申请人: 中国银联股份有限公司 (CHINA UNIONPAY CO., LTD.) [CN/CN]; 中国上海市浦东新区含笑路 36 号银联大厦, Shanghai 200135 (CN)。

(72) 发明人: 王琪 (WANG, Qi); 中国上海市浦东新区含笑路36号银联大厦, Shanghai 200135 (CN)。何东杰 (HE, Dongjie); 中国上海市浦东新区含笑路36号银联大厦, Shanghai 200135 (CN)。杨洁 (YANG, Jie); 中国上海市浦东新区含笑路36号银联大厦, Shanghai 200135 (CN)。

(74) 代理人: 北京同达信恒知识产权代理有限公司 (TDIP & PARTNERS); 中国北京市海淀区宝盛南路1号院20号楼8层101-01, Beijing 100192 (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, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) **Title:** TEXT QUALITY INDEX OBTAINING METHOD AND APPARATUS

(54) **发明名称:** 一种文本质量指标获取方法及装置

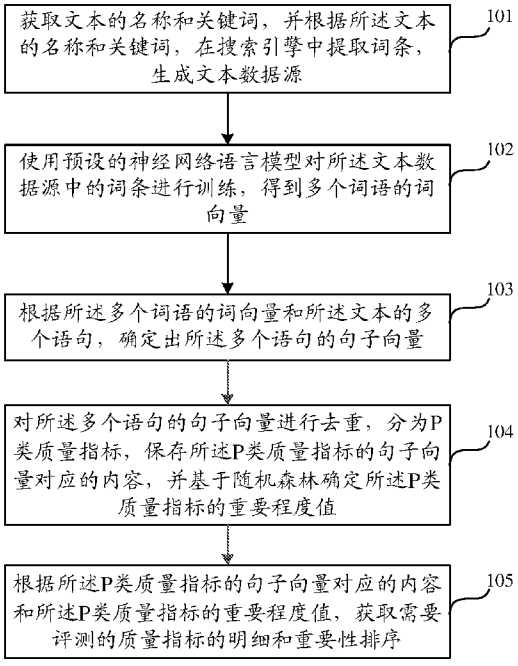

图 1

(57) **Abstract:** Disclosed in the present invention are a text quality index obtaining method and apparatus. The method comprises: obtaining a name and keywords of a text, and generating a text data source; training entries in the text data source by using a preset neural network language model to obtain word vectors of multiple words; determining sentence vectors of multiple sentences; de-replicating the sentence vectors of the multiple sentences, a P-type quality index being involved, saving content corresponding to the sentence vectors of the P-type quality index, and on the basis of a random forest, determining an important degree value of the P-type quality

WO 2019/128124 A1

[见续页]

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))。

index; and according to the content corresponding to the sentence vectors of the P-type quality index and the important degree value of the P-type quality index, obtaining details and importance ranks of quality indexes needing to be evaluated. By quantifying sentences of open source software into vectors, a quality index set is obtained, the subsequent ranking accuracy is improved, and the important degree values of quality indexes are obtained on the basis of a random forest, so that the obtained quality index result is more accurate and detailed.

(57) 摘要：本发明公开了一种文本质量指标获取方法及装置，该方法包括获取文本的名称和关键词，生成文本数据源，使用预设的神经网络语言模型对文本数据源中的词条进行训练，得到多个词语的词向量，确定出多个语句的句子向量，对多个语句的句子向量进行去重，分为P类质量指标，保存P类质量指标的句子向量对应的内容，并基于随机森林确定P类质量指标的重要程度值，根据P类质量指标的句子向量对应的内容和P类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。通过将开源软件的语句量化为向量，得到质量指标集合，提高了后续排序的准确率，基于随机森林得到质量指标的重要程度值，使得获取的质量指标结果更加准确和细化。

一种文本质量指标获取方法及装置

本申请要求在 2017 年 12 月 29 日提交中国专利局、申请号为 201711484243.7、发明名称为“一种文本质量指标获取方法及装置”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

本发明实施例涉及语句分析技术领域，尤其涉及一种文本质量指标获取方法及装置。

背景技术

开源软件的广泛应用已经成为一种趋势。金融行业出于技术成熟度和安全合规方面的考虑，对开源软件的应用保持审慎的态度。所以在使用一个开源软件之前应对软件进行完备科学的评估，通常通过建立评测模型对开源软件进行评测，基于模型评测诸如 kakfa，rabbitmq，rootwrap 等开源软件，在此过程中，我们发现了如下问题：首先，由于缺乏自动化的过程和工具，部分步骤通过人工抓取，每个评测指标及相应内容选取非常耗时并相对主观。其次，开源软件评测指标数量大，不同软件对于不同指标评测的敏感度不尽相同，有效地选取评测指标才能有效地评估软件。

现有的软件自动分类方法通常利用包含网页，日志等内容的文本来表征对象，通过数据挖掘技术对软件文本进行自动分类，将软件文本集合按照主题进行聚类，聚类的结果是每个文本自动归属于某个主题，从而间接实现对词条等对象的自动分类。现有方案下的数据源只是简单利用关键词进行聚类，不包含语义以及和上下文的关联，这样孤立的分类对更加抽象或者是表征含义更丰富的对象进行分类效果很差，同时很难对更长的量如句子进行识别分类。

发明内容

本发明实施例提供一种文本质量指标获取方法及装置，用以实现自动化获取文本的质量指标，提高了准确性。

第一方面，本发明实施例提供的一种文本质量指标获取方法，包括：

获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源；

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；

根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量；

对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数；

根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

可选的，所述使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量，包括：

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量；

对每个词条进行遍历，得到多个词语的词向量。

可选的，所述根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量，包括：

将所述文本的多个语句进行分词；

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

可选的，所述对多个语句的句子向量进行去重，分为 P 类质量指标，包括：

将所述多个语句的句子向量进行补齐;

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离;

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

可选的，所述基于随机森林确定所述 P 类质量指标的重要程度值，包括:

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数;

根据所述样本个数和构建森林的树的棵数构建决策树;

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数;

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

第二方面,，本发明实施例还提高了一种文本质量指标获取装置，包括:

生成单元，用于获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源;

确定单元，用于使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；以及根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量;

去重单元，用于对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数;

处理单元，用于根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

可选的，所述确定单元具体用于:

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量;

对每个词条进行遍历，得到多个词语的词向量。

可选的，所述确定单元具体用于：

将所述文本的多个语句进行分词；

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

可选的，所述去重单元具体用于：

将所述多个语句的句子向量进行补齐；

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离；

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

可选的，所述去重单元具体用于：

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数；

根据所述样本个数和构建森林的树的棵数构建决策树；

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数；

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

相应的，本发明实施例还提供了一种计算设备，包括：

存储器，用于存储程序指令；

处理器，用于调用所述存储器中存储的程序指令，按照获得的程序执行上述文本质量指标获取方法。

相应的，本发明实施例还提供了一种计算机存储介质，所述计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令用于使计算机执行上述文本质量指标获取方法。

第三方面，本发明实施例提供一种电子设备，包括：处理器、存储器、

收发机、总线接口，其中处理器、存储器与收发机之间通过总线接口连接；

所述收发机，用于获取文本的名称和关键词；

所述处理器，用于根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源；

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；

根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量；

对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数；

根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序；

所述存储器，用于存储一个或多个可执行程序，可以存储所述处理器在执行操作时所使用的数据；

所述总线接口，用于提供接口。

第四方面，本发明实施例提供一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令用于使所述计算机执行上述第一方面中任一实施例所述文本质量指标获取方法。

第五方面，本发明实施例提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行上述第一方面中任一实施例所述文本质量指标获取方法。

本发明实施例表明，通过获取文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源，使用预设的神经网络语言模型对文本数据源中的词条进行训练，得到多个词语的词向量，根据多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量，对多个语句的句子向量进行去

重，分为 P 类质量指标，保存 P 类质量指标的句子向量对应的内容，并基于随机森林确定 P 类质量指标的重要程度值，根据 P 类质量指标的句子向量对应的内容和 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。通过将文本的语句量化为向量，得到质量指标集合，提高了后续排序的准确率，基于随机森林得到质量指标的重要程度值，使得获取的质量指标结果更加准确和细化。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域的普通技术人员来讲，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 为本发明实施例提供的一种文本质量指标获取方法的流程示意图；

图 2 为本发明实施例提供的一种生成词向量的示意图；

图 3 为本发明实施例提供的一种文本质量指标获取装置的结构示意图；

图 4 为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作进一步地详细描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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它实施例，都属于本发明保护的范围。

图 1 示例性的示出了本发明实施例提供的一种文本质量指标获取方法的流程，该流程可以由文本质量指标获取装置执行。在本发明实施例中，为了更好的解释本发明实施例所提供的文本质量指标获取方法，下面将以分布式开源软件 kaska 为例，来描述该文本质量指标获取的流程。

如图 1 所示，该流程具体包括:

步骤 101，获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源。

在本发明实施例中，文本可以包括各种软件等使用文本来表达内容的事物，以开源软件为例，通过在搜索引擎中输入需要评测的开源软件的名称和关键词，提取词条，可以形成文本数据源。例如，通过搜索引擎搜索开源软件“Kafka”“kafka 功能”等关键词组合，得到搜索结果。通过通用的爬虫技术对搜索结果的前 N 个词条（假设为 1000 条，词条越多，指标越全面），获取结果信息的 HTML（HyperText Markup Language，超文本标记语言）标签如 title（标题）、text（文本）等，将结果存为一个文本文件。

可选的，文本数据源的获取方式不限于通过搜索引擎获取词条的 Title 标签，也可以通过解析网页，进行聚类分析等更多复杂预处理方式得到。

步骤 102，使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量。

具体的，可以使用预设的神经网络语言模型对文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测当前词语的词向量。然后对每个词条进行遍历，就可以得到多个词语的词向量。该预设的神经网络语言模型（如 CBOW（Continuous Bagof-Words，连续词袋）模型）可以是预设了一些参数的神经网络语言模型。

举例来说，使用基于神经网络语言模型对文本数据源中的词条进行训练，得到每个词语的词向量，通过词条语句中前后文单词如 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$ 来预测当前单词 w_t 的向量表示。例如，其中一个单词为“发布”，则通过其前后文的单词如“版本”、“发布”、“时间”、“周期”、“产品”、“活跃度”等前后文，具体的可以如图 2 所示的预测词向量的流程。

可选的，上述 CBOW 模型也可以替换为改进 CBOW 模型或其他类似功能的模型。

步骤 103，根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量。

在得到多个词语的词向量之后，就可以先将开源软件的多个语句进行分词，然后使用该多个词语的词向量对分词后的语句进行遍历，将多个语句中的词语转换为词向量，确定出多个语句的句子向量。

针对开源软件中的每一个语句进行分词，并使用步骤 102 中得到的词向量对分词后的结果进行遍历，得到每一个语句的句子向量（共 N 个句子向量，N 为正整数）。例如，其中一个语句的内容为“软件的贡献者人数”，则提取“软件”、“贡献者”、“人数”三个词的对应向量为V_1,V_2,V_3，那么对应的句子向量就可以得到 V=(V_1,V_2,V_3)。

步骤 104，对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数。

将步骤 103 中得到的多个语句的句子向量进行补齐，针对该多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离，可以将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。该第一阈值和第二阈值可以依据经验设置，其中， 第一阈值大于第二阈值。例如，第一阈值可以设置为 1，第二阈值可以设置为 0.1。

对得到的 N 个句子向量进行补齐（以最长的向量长度为准）。对每一个句子向量，遍历其他句子向量，计算向量之间的欧式距离，如果距离小于阈值（假设取值为 1），那么两个句子向量可以认为是同一类。如果两个向量之间的距离小于 0.1，说明两个句子几乎相同，保留其中之一即可，完成去重。最终，所有的语句在去掉相同句子向量的基础上被分为 P 类，也就是 P 类质量指标。完成分类后，保存每一类的句子向量对应的内容。

可选的，上述句子向量的分类、去重、确定质量指标除了本发明实施例

所示提供的方法得到外，也能通过改进算法，分类聚类过程得到近似处理结果。

得到该 P 类质量指标之后，可以根据该 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数，根据样本个数和构建森林的树的棵数构建决策树，然后遍历所有的决策树中质量指标的特征，在一次循环中，特征出现一次计数值加 1，得到特征在森林中出现的次数，最后再根据每个特征在森林中出现的次数，就可以得到各类质量指标的重要程度值。

经过去重后的 P 类质量指标集合，经过补齐后的向量深度相同为 n，则所有的特征数为 P*n。通过随机森林生成决策树训练集的策略，从 P 类句子向量中通过重采样来获得训练样本。重复 S 次，产生 S 棵树。然后采用下述的流程对结果进行统计:

Begin

$\{F_1, F_2, \ldots, F_n\} = BuildRandomForest(F, N, f_s)$ //随机森林训练

for $1 \le n \le S$ do //遍历所有的树

for $1 \le i \le M$ do //遍历所有的特征数

if $f_i \in F_n$ then

$\theta_i = \theta_i + 1$;

for $1 \le p \le P$ do //遍历所有的指标

$$Q_n = \frac{1}{n}\sum_{q=1}^{n}\theta_n(P-1)+q$$

其中，需要说明的是，Q_n为质量指标的重要程度值; S 是森林中树的个数; P 为质量指标的个数; n 为每个质量指标对应的句子向量的深度; f_s为选取的特征; M 为选取的特征数（P*n）。

首先，确定每次形成决策树利用的样本个数以及构建森林的树的棵树 S（随机选取），根据确定的每次随机选取的样本个数和树的棵树构建决策树。然后，遍历所有的决策树中的特征，在一次循环中，特征数出现过一次就在计数值上加 1，$\theta_i = \theta_i + 1$。特征遍历结束后，得到每一个特征在森林中出现的

次数。最后，对 P 类质量指标进行排序计算。根据每个特征在整个森林中出现的次数，得到针对某一类指标的重要程度值。即各个评测的质量指标对应的 Q 的值，值越大说明评测的质量指标越重要。如表 1 所示，表 1 中各个评测的质量指标对应的 Q 的值即为统计结果，值越大说明质量指标越重要，如果一些值远远小于其他值，那么这个评测的质量指标可以忽略不计。

表 1

	N_1	N_2	N_3	……	N_n	重要程度值
p_1	c_1	c_2	c_3	……	c_n	$Q_1=\frac{c_1+c_2+\cdots+c_n}{n}$
p_2	⋮	⋮	⋮	……	⋮	Q_2
p_3	⋮	⋮	⋮	……	⋮	Q_3
p_4	⋮	⋮	⋮	……	⋮	Q_4
p_5	⋮	⋮	⋮	……	⋮	Q_5
⋮	⋮	⋮	⋮	……	⋮	⋮
p_n	⋮	⋮	⋮	……	⋮	Q_n

步骤 105，根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

具体的，根据步骤 104 中保存 P 类质量质量的句子向量对应的内容，找出各类质量指标 p_n 的每个特征对应的名称，最终根据 p_n 筛选和排序得到需要评测的质量指标的明细，以及重要性排序。该质量指标的明细也就是该质量指标的句子向量对应的内容。

上述实施例表明，通过获取文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源，使用预设的神经网络语言模型对文本数据源中的词条进行训练，得到多个词语的词向量，根据多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量，对多个语句的句子向量进行去重，分为 P 类质量指标，保存 P 类质量指标的句子向量对应的内容，并基于随机森林确定 P 类质量指标的重要程度值，根据 P 类质量指标的句子向量对应的

内容和 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。通过将文本的语句量化为向量，得到质量指标集合，提高了后续排序的准确率，基于随机森林得到质量指标的重要程度值，使得获取的质量指标结果更加准确和细化。

基于相同的技术构思，图 3 示例性的示出了本发明实施例提高的一种文本质量指标获取装置，该装置可以执行文本质量指标获取的流程。

如图 3 所示，该装置包括:

生成单元 301，用于获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源;

确定单元 302，用于使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；以及根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量;

去重单元 303，用于对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数;

处理单元 304，用于根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

可选的，所述确定单元 302 具体用于:

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量;

对每个词条进行遍历，得到多个词语的词向量。

可选的，所述确定单元 302 具体用于:

将所述文本的多个语句进行分词;

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

可选的，所述去重单元 303 具体用于:

将所述多个语句的句子向量进行补齐;

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离；

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

可选的，所述去重单元 303 具体用于：

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数；

根据所述样本个数和构建森林的树的棵数构建决策树；

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数；

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

基于相同的构思，本发明还提供一种电子设备，如图 4 所示，包括处理器 401、存储器 402、收发机 403、总线接口 404，其中处理器 401、存储器 402 与收发机 403 之间通过总线接口 404 连接；

所述收发机 403，用于获取文本的名称和关键词；

所述处理器 401，用于根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源；

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；

根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量；

对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数；

根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序；

所述存储器 402，用于存储一个或多个可执行程序，可以存储所述处理器在执行操作时所使用的数据；

所述总线接口 404，用于提供接口。

所述处理器 401 具体用于，使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量；

对每个词条进行遍历，得到多个词语的词向量。

所述处理器 401 具体用于：

将所述文本的多个语句进行分词；

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

所述处理器 401 具体用于：

将所述多个语句的句子向量进行补齐；

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离；

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

可选的，所述处理器 401 具体用于：

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数；

根据所述样本个数和构建森林的树的棵数构建决策树；

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数；

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

本发明实施例提供一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令用于使所述计算机执行上述

第一方面中任一实施例所述文本质量指标获取方法。

本发明实施例提供一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行上述第一方面中任一实施例所述文本质量指标获取方法。

本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和／或方框图来描述的。应理解可由计算机程序指令实现流程图和／或方框图中的每一流程和／或方框、以及流程图和／或方框图中的流程和／或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要

求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

权 利 要 求

1、一种文本质量指标获取方法，其特征在于，包括:

获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源;

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量;

根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量;

对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数;

根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

2、如权利要求 1 所述的方法，其特征在于，所述使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量，包括:

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量;

对每个词条进行遍历，得到多个词语的词向量。

3、如权利要求 1 所述的方法，其特征在于，所述根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量，包括:

将所述文本的多个语句进行分词;

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

4、如权利要求 1 所述的方法，其特征在于，所述对多个语句的句子向量进行去重，分为 P 类质量指标，包括:

将所述多个语句的句子向量进行补齐;

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离；

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

5、如权利要求 1 所述的方法，其特征在于，所述基于随机森林确定所述 P 类质量指标的重要程度值，包括:

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数；

根据所述样本个数和构建森林的树的棵数构建决策树；

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数；

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

6、一种文本质量指标获取装置，其特征在于，包括:

生成单元，用于获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源；

确定单元，用于使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；以及根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量；

去重单元，用于对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数；

处理单元，用于根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序。

7、如权利要求 6 所述的装置，其特征在于，所述确定单元具体用于:

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量；

对每个词条进行遍历，得到多个词语的词向量。

8、如权利要求 6 所述的装置，其特征在于，所述确定单元具体用于：

将所述文本的多个语句进行分词；

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

9、如权利要求 6 所述的装置，其特征在于，所述去重单元具体用于：

将所述多个语句的句子向量进行补齐；

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离；

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

10、如权利要求 6 所述的装置，其特征在于，所述去重单元具体用于：

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数；

根据所述样本个数和构建森林的树的棵数构建决策树；

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数；

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

11、一种电子设备，其特征在于，包括处理器、存储器、收发机、总线接口，其中处理器、存储器与收发机之间通过总线接口连接；

所述收发机，用于获取文本的名称和关键词；

所述处理器，用于根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源；

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量；

根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语

句的句子向量;

对所述多个语句的句子向量进行去重，分为 P 类质量指标，保存所述 P 类质量指标的句子向量对应的内容，并基于随机森林确定所述 P 类质量指标的重要程度值，P 为正整数;

根据所述 P 类质量指标的句子向量对应的内容和所述 P 类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序;

所述存储器，用于存储一个或多个可执行程序，可以存储所述处理器在执行操作时所使用的数据;

所述总线接口，用于提供接口。

12、如权利要求 11 所述的设备，其特征在于，所述处理器具体用于:

使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，通过词条语句中当前词语的前后文词语预测所述当前词语的词向量;

对每个词条进行遍历，得到多个词语的词向量。

13、如权利要求 11 所述的设备，其特征在于，所述处理器具体用于:

将所述文本的多个语句进行分词;

使用所述多个词语的词向量对分词后的语句进行遍历，将所述多个语句中的词语转换为词向量，确定出多个语句的句子向量。

14、如权利要求 11 所述的设备，其特征在于，所述处理器具体用于:

将所述多个语句的句子向量进行补齐;

针对所述多个句子向量中任意一个句子向量，遍历其他的句子向量，计算向量之间的欧式距离;

将欧式距离小于第一阈值的两个句子向量确定为同一类质量指标，将欧式距离小于第二阈值的两个句子向量确定为相同的句子向量，进行去重，得到 P 类质量指标。

15、如权利要求 11 所述的设备，其特征在于，所述处理器具体用于:

根据所述 P 类质量指标，确定每次形成决策树利用的样本个数和构建森林的树的棵数;

根据所述样本个数和构建森林的树的棵数构建决策树;

遍历所有的决策树中质量指标的特征，在一次循环中，所述特征出现一次计数值加 1，得到所述特征在森林中出现的次数;

根据每个特征在森林中出现的次数，得到各类质量指标的重要程度值。

16、一种计算机存储介质，其特征在于，所述计算机可读存储介质存储有计算机可执行指令，所述计算机可执行指令用于使计算机执行权利要求 1 至 5 任一项所述的方法。

17、一种计算机程序产品，其特征在于，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行权利要求 1~5 任一所述方法。

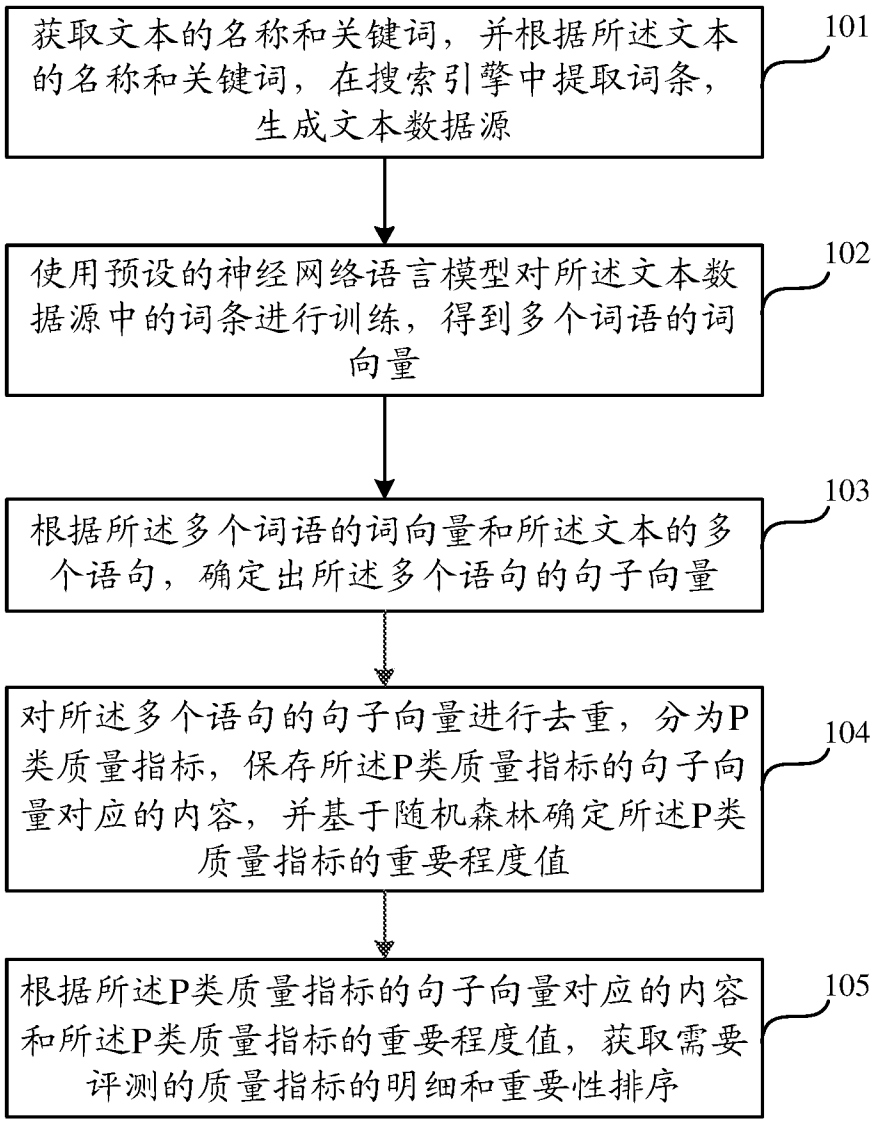
获取文本的名称和关键词，并根据所述文本的名称和关键词，在搜索引擎中提取词条，生成文本数据源
101
使用预设的神经网络语言模型对所述文本数据源中的词条进行训练，得到多个词语的词向量
102
根据所述多个词语的词向量和所述文本的多个语句，确定出所述多个语句的句子向量
103
对所述多个语句的句子向量进行去重，分为P类质量指标，保存所述P类质量指标的句子向量对应的内容，并基于随机森林确定所述P类质量指标的重要程度值
104
根据所述P类质量指标的句子向量对应的内容和所述P类质量指标的重要程度值，获取需要评测的质量指标的明细和重要性排序
105

图 1

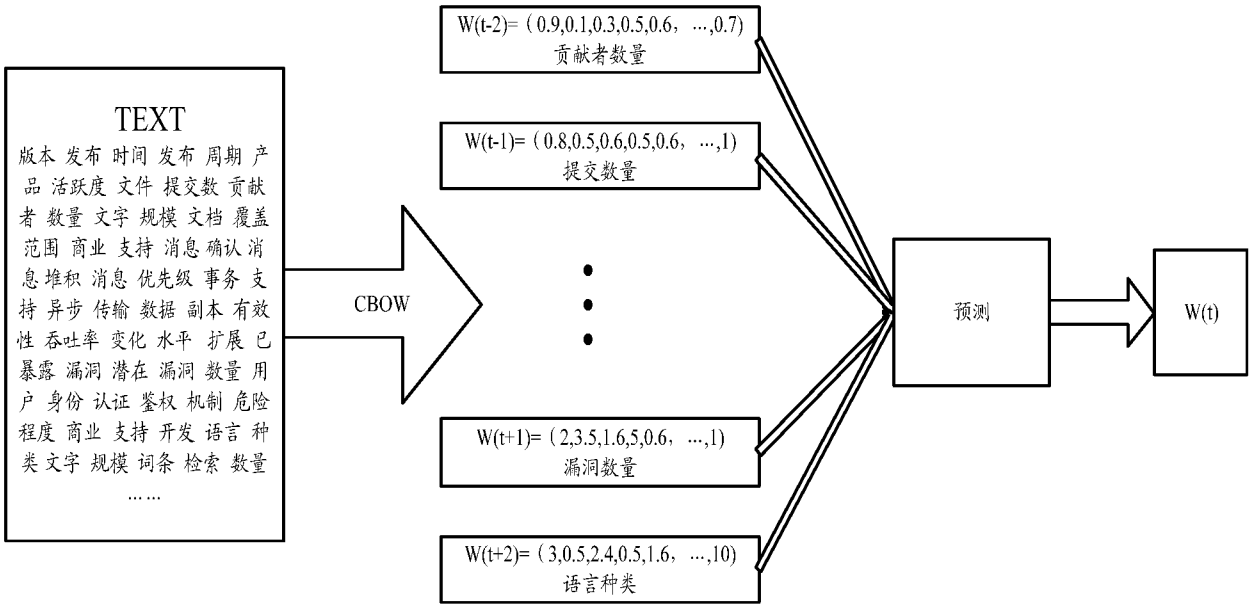
TEXT
版本 发布 时间 发布 周期 产品 活跃度 文件 提交数 贡献者 数量 文字 规模 文档 覆盖范围 商业 支持 消息 确认 消息 堆积 消息 优先级 事务 支持 异步 传输 数据 副本 有效性 吞吐率 变化 水平 扩展 已暴露 漏洞 潜在 漏洞 数量 用户 身份 认证 鉴权 机制 危险程度 商业 支持 开发 语言 种类 文字 规模 词条 检索 数量
……
CBOW
W(t-2)=（0.9,0.1,0.3,0.5,0.6，…,0.7)
贡献者数量
W(t-1)=（0.8,0.5,0.6,0.5,0.6，…,1)
提交数量
W(t+1)=（2,3.5,1.6,5,0.6，…,1)
漏洞数量
W(t+2)=（3,0.5,2.4,0.5,1.6，…,10)
语言种类
预测
W(t)

图 2

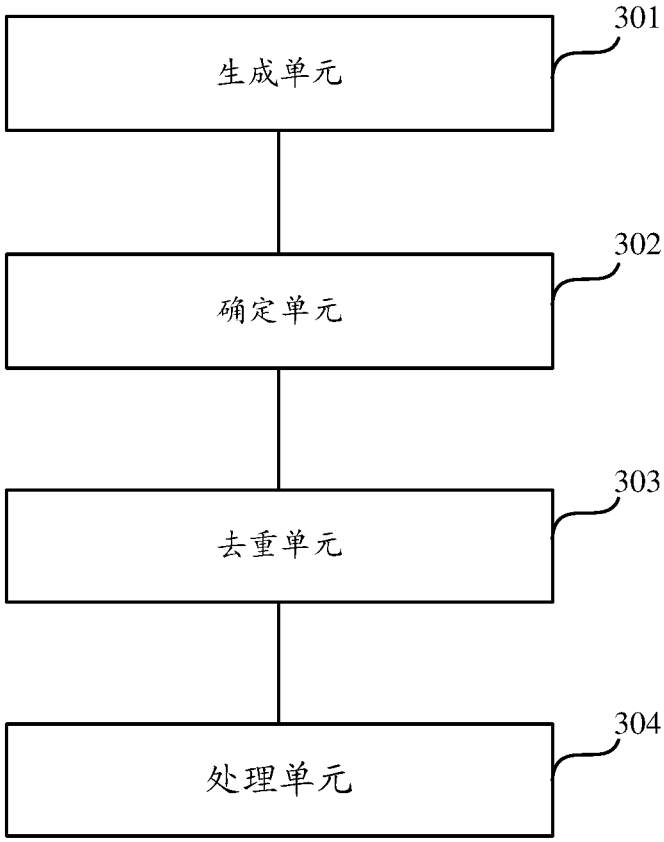
生成单元
301
确定单元
302
去重单元
303
处理单元
304

图 3

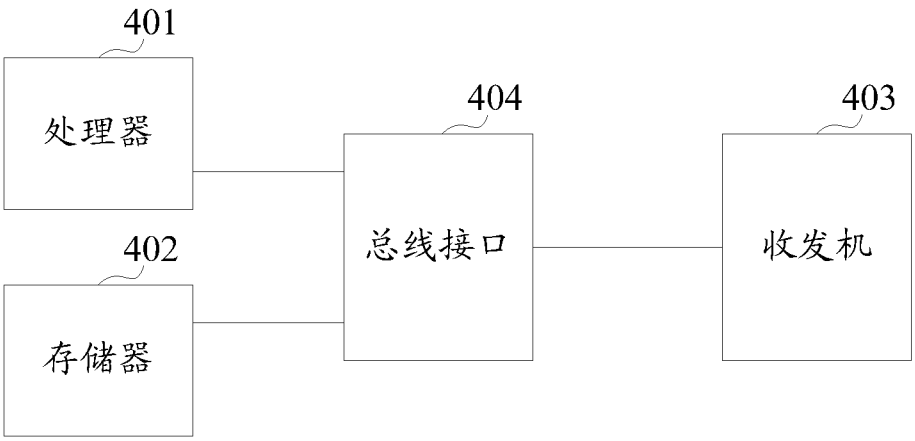

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/090878

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)

G06F17/-

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)

CNABS, CNTXT, CNKI, VEN, USTXT, EPTXT, WOTXT: 文本, 软件, 源码, 代码, 关键词, 质量, 随机森林, 神经网络, 训练, 向量, text, software, source code, code, keyword, quality, random forests, neural network, train, vector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106021410 A (INSTITUTE OF SOFTWARE, CHINESE ACADEMY OF SCIENCES) 12 October 2016 (2016-10-12) entire document	1-17
A	CN 106326346 A (SHANGHAI GAOXIN COMPUTER SYSTEMS CO., LTD.) 11 January 2017 (2017-01-11) entire document	1-17
A	US 2011173191 A1 (MICROSOFT CORP.) 14 July 2011 (2011-07-14) entire document	1-17

☐ 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
26 August 2018	**25 September 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China (ISA/CN) **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/CN2018/090878

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 106021410 A	12 October 2016	None	
CN 106326346 A	11 January 2017	None	
US 2011173191 A1	14 July 2011	US 8990124 B2	24 March 2015

国际检索报告

国际申请号
PCT/CN2018/090878

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)

G06F17/-

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）)

CNABS,CNTXT,CNKI,VEN,USTXT,EPTXT,WOTXT:文本，软件，源码，代码，关键词，质量，随机森林，神经网络，训练，向量，text,software,source code,code,keyword,quality,random forests,neural network,train,vector

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
A	CN 106021410 A (中国科学院软件研究所) 2016年 10月 12日 (2016-10-12) 全文	1-17
A	CN 106326346 A (上海高欣计算机系统有限公司) 2017年 1月 11日 (2017-01-11) 全文	1-17
A	US 2011173191 A1 (MICROSOFT CORP) 2011年 7月 14日 (2011-07-14) 全文	1-17

☐ 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
2018年 8月 26日	2018年 9月 25日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 王国海 电话号码 86-(20)-28958137

表 PCT/ISA/210 (第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/090878

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 106021410 A	2016年 10月 12日	无	
CN 106326346 A	2017年 1月 11日	无	
US 2011173191 A1	2011年 7月 14日	US 8990124 B2	2015年 3月 24日