(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 2 月 7 日 (07.02.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/024838 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/097748

(22) 国际申请日: 2018 年 7 月 30 日 (30.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710642152.5 2017年7月31日 (31.07.2017) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518000 (CN)。

(72) 发明人: 谢润泉 (XIE, Runquan); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518000 (CN)。 连凤宗 (LIAN, Fengzong); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518000 (CN)。 黄泽谦 (HUANG, Zeqian); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518000 (CN)。

(74) 代理人: 深 圳 市 隆 天 联 鼎 知 识 产 权 代 理 有 限 公 司 (SHENZHEN LUNGTIN LIANDING INTELLECTUAL PROPERTY AGENT LTD.); 中国广东省深圳市福田区南园路上田大厦4A刘抗美, 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,

(54) **Title:** SEARCH ITEM GENERATION METHOD AND RELEVANT APPARATUS

(54) **发明名称**: 检索项的生成方法及相关装置

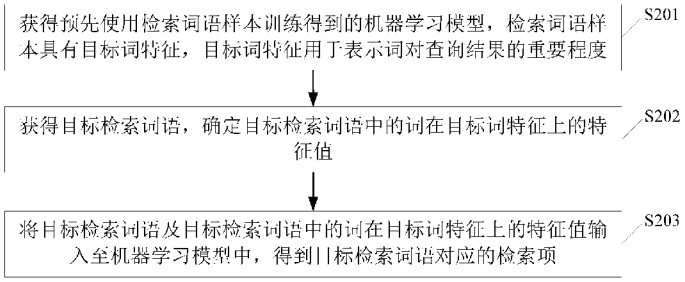

图 3

S201 ACQUIRE A MACHINE LEARNING MODEL OBTAINED IN ADVANCE BY TRAINING USING SEARCH TERM SAMPLES, THE SEARCH TERM SAMPLES HAVE A TARGET WORD FEATURE, AND THE TARGET WORD FEATURE IS USED FOR INDICATING THE IMPORTANCE OF THE WORD TO A QUERY RESULT

S202 OBTAIN TARGET SEARCH TERMS, AND DETERMINE A CHARACTERISTIC VALUE, ON THE TARGET WORD FEATURE, OF A WORD IN THE TARGET SEARCH TERMS

S203 INPUT THE TARGET SEARCH TERMS AND THE CHARACTERISTIC VALUE, ON THE TARGET WORD FEATURE, OF THE WORD IN THE TARGET SEARCH TERMS INTO THE MACHINE LEARNING MODEL, TO OBTAIN A SEARCH ITEM CORRESPONDING TO THE TARGET SEARCH TERMS

(57) **Abstract:** Provided is a search item generation method. In the method, a machine learning model is used to determine the leave or stay condition of each word in search terms; the machine learning model is trained by search term samples having a target word feature; compared to an existing search item generation method, the target word feature can reflect the importance of a word to a query result; using the machine learning model trained by the target word feature is more precise in terms of word selection result in the search term, thereby generating a more accurate search item, thus avoiding the recalling of irrelevant query results. Furthermore, also provided is a relevant apparatus for generating a search item used for ensuring that the method is realised and applied in practice.

WO 2019/024838 A1

[见续页]

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, 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))。

(57) 摘要：本申请提供了一种检索项的生成方法，该方法使用机器学习模型来确定检索词语中各个词的去留情况，机器学习模型是由具有目标词特征的检索词语样本训练而成的，相比现有的检索项的生成方法而言，目标词特征能反映词对查询结果重要程度，使用该种目标词特征训练而成的机器学习模型，对检索词语中词选择结果更加精确，从而生成更加准确的检索项，进而避免召回不相关的查询结果。另外，本申请还提供了用于生成检索项的相关装置，用以保证上述方法在实际中的实现及应用。

检索项的生成方法及相关装置

技术领域

本申请涉及检索技术领域，更具体地，涉及检索项的生成方法及相关装置、电子设备、计算机可读存储介质。

背景技术

搜索，是指获得检索词语（query）后，按照一定的查询策略从网络中搜集与检索词语相匹配的信息，将信息组织及处理后进行展示等操作。

具体地，在搜索时会对检索词语进行分词操作，得到若干独立的词（term），并检索与各个词相关的信息列表，再求取每个词所对应的信息列表的交集，从而得到召回文档。通常地，使用所有词的信息列表求交得到的召回文档数量过少且优质文档不能召回，因此需要对检索词语获得的词进行筛选，选择出用于检索的词，这些词的集合可以称为检索项，然后仅使用所选择的词的信息列表获得召回文档即可。

技术问题

目前在生成检索项时，选择的词准确度不高，容易造成召回文档与检索词语相关性较差的问题。

技术解决方案

有鉴于此，本申请提供了一种检索项的生成方法，使用该方法生成的检索项得到的查询结果准确度更高。

第一方面，本申请提供了一种检索项的生成方法，包括：

获得预先由检索词语样本训练得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；

获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；

将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

在一种示例性实施例中，所述机器学习模型的训练方法包括：

获得检索词语样本后，确定所述检索词语样本中的词的去留情况，所述去留情况用于表示所述词是否用于生成检索项；

确定所述词在所述目标词特征上的特征值；

对所述检索词语样本、所述词的去留情况及所述词的特征值进行训练后，得到所述机器学习模型。

在一种示例性实施例中，所述确定所述词在所述目标词特征上的特征值，包括：

获得依据语料生成的目标词特征值确定工具；

使用所述目标词特征值确定工具，确定所述词在所述目标词特征上的特征值。

在一种示例性实施例中，所述确定所述检索词样本中的词的去留情况，包括：

获得所述检索词样本对应的召回文档；

确定所述检索词样本中的词在所述召回文档中的特征；

根据所述特征，确定所述检索词样本中的词的去留情况。

在一种示例性实施例中，所述根据所述特征，确定所述检索词样本中的词的去留情况，包括：

根据所述特征计算所述检索词样本中的词的分数，并根据所述分数确定所述词的去留情况。

在一种示例性实施例中，所述将所述目标检索词语及所述特征值输入至所述机器学习

模型中，得到所述目标检索词语对应的检索项，包括：

将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语中的词的保留概率；

根据所述保留概率，确定所述目标检索词语中的词的去留情况，其中必留的词的集合为检索项。

在一种示例性实施例中，本申请提供的一种检索项的生成方法还包括：

获得非机器学习模型为所述目标检索词语生成的修正检索项；

使用所述修正检索项，对所述机器学习模型生成的检索项进行修正。

在一种示例性实施例中，所述使用所述修正检索项，对所述机器学习模型生成的检索项进行修正，包括：

判断所述机器学习模型确定为非必留的词是否被所述非机器学习模型确定为符合必留条件的词；

若是，则将所述机器学习模型确定为非必留的词修正为必留。

在一种示例性实施例中，所述目标词特征包括以下几项中的任意一项或多项的组合：词性特征、句法依存特征、反文档频率、点互信息、词向量类特征、重要性特征、紧密度、词维度特征。

在一种示例性实施例中，本申请提供的一种检索项的生成方法还包括：

结合所述目标词特征对所述机器学习模型生成的检索项进行修正。

第二方面，本申请提供了一种检索项的生成装置，包括：

机器学习模型获得器，配置为获得预先由检索词语样本训练得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；

目标词特征值获得器，配置为获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；

检索项生成器，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

在一种示例性实施例中，本申请提供的一种检索项的生成装置还包括：机器学习模型训练器，配置为训练检索词语样本训练，以得到机器学习模型；

其中，所述机器学习模型训练器包括：

词去留情况确定器，配置为获得检索词语样本后，确定所述检索词语样本中的词的去留情况，所述去留情况用于表示所述词是否用于生成检索项；

目标词特征值确定器，配置为确定所述词在所述目标词特征上的特征值；

机器学习模型生成器，配置为对所述检索词语样本、所述词的去留情况及所述词的特征值进行训练后，得到所述机器学习模型。

在一种示例性实施例中，所述目标词特征值确定器在执行确定所述词在所述目标词特征上的特征值的步骤时，用于：

获得依据语料生成的目标词特征值确定工具；

使用所述目标词特征值确定工具，确定所述词在所述目标词特征上的特征值。

在一种示例性实施例中，所述词去留情况确定器在执行确定所述检索词样本中的词的去留情况的步骤时，用于：获得所述检索词样本对应的召回文档；确定所述检索词样本中的词在所述召回文档中的特征；根据所述特征，确定所述检索词样本中的词的去留情况。

在一种示例性实施例中，所述词去留情况确定器在执行确定所述检索词语样本中的词的去留情况的步骤时，用于：根据所述特征计算所述检索词样本中的词的分数，并根据所述分数确定所述词的去留情况。

在一种示例性实施例中，所述检索项生成器包括：

保留概率确定器，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语中的词的保留概率；

检索项确定器，配置为根据所述保留概率，确定所述目标检索词语中的词的去留情况，其中必留的词的集合为检索项。

在一种示例性实施例中，本申请提供的一种检索项的生成装置还包括：

检索项修正器，配置为获得非机器学习模型为所述目标检索词语生成的修正检索项；以及使用所述修正检索项，对所述机器学习模型生成的检索项进行修正。

在一种示例性实施例中，所述检索项修正器，包括：

非必留词修正器，配置为判断所述机器学习模型确定为非必留的词是否被所述非机器学习模型确定为符合必留条件的词；若是，则将所述机器学习模型确定为非必留的词修正为必留。

在一个示例中，上述目标词特征包括以下几项中的任意一项或多项的组合：

词性特征、句法依存特征、反文档频率、点互信息、词向量类特征、重要性特征、紧密度、词维度特征。

在一种示例性实施例中，本申请提供的一种检索项的生成装置还包括：

第二修正器，配置为结合所述目标词特征对所述机器学习模型生成的检索项进行修正。

第三方面，本申请提供了一种检索项的生成设备，包括：

输入器，配置为输入检索词语样本；

处理器，配置为训练所述检索词语样本训练，以得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；以及将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

第四方面，本申请提供了一种电子设备，所述电子设备包括：

处理器；

配置为存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行上述检索项的生成方法。

第五方面，本申请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，所述计算机程序可由处理器执行完成上述检索项的生成方法。

有益效果

由以上技术方案可知，本申请提供了一种检索项的生成方法，该方法使用机器学习模型来确定检索词语中各个词的去留情况，机器学习模型是使用具有目标词特征的检索词语样本训练而成的，相比现有的检索项的生成方法而言，目标词特征能反映词对查询结果重要程度的特征，使用该种目标词特征训练而成的机器学习模型，能够更加准确地确定检索词语中词的去留情况，从而生成更加准确的检索项，进而避免召回不相关的查询结果。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术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，还可以根据提供的附图获得其他的附图。

图 1 及 2 为使用不同的检索项得到的查询结果示意图；
图 3 为本申请提供的检索项的生成方法的一种流程图；
图 4 为本申请提供的检索词语的搜索量曲线的一种示意图；
图 5 为本申请提供的训练集的生成方式的一种流程图；
图 6 为本申请提供的检索词语的句法依存树的一种示意图；
图 7 为本申请提供的词与其他词的紧密度特征的一种示意图;
图 8 为本申请提供的检索项的生成装置的一种结构示意图；
图 9 为本申请提供的检索项的生成装置的另一种结构示意图；
图 10 是图 9 对应实施例中机器学习模型训练器的细节结构示意图；
图 11 是图 8 对应检索项生成器的细节结构示意图；
图 12 为本申请提供的检索项的生成装置的又一种结构示意图；
图 13 为本申请提供的检索项的生成设备的一种硬件结构示意图。

本发明的最佳实施方式

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，所描述的实施例仅仅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。

搜索，是指获得检索词语（query）后，按照一定的查询策略从网络中搜集与检索词语相匹配的信息，将信息组织及处理后进行展示等操作。其中，检索词语也可以称为搜索语句，搜集与检索词语相匹配的信息这一过程也可以称为召回文档。

在搜索时，会对检索词语进行分词操作。若使用所有词召回文档，则所得到的文档数量过少且优质文档不能召回。例如，检索词语为“儿歌简谱大全 100 首”，分词结果包括“儿歌”“简谱”“大全”及“100 首”。若使用所有的词检索，则召回文档如图 1 所示，检索结果与检索词语相关性不高。在用户检索的场景中，用户检索体验较差。

因此，为了保证召回文档的准确性，需要对检索词语获得的词进行筛选，选择出用于检索的词及可以丢弃的词。其中，用于检索的词可以称为必留词，丢弃的词可以称为非必留词；选择的过程可以称为查非必留词或查必留词；选择出的用于检索的词的集合可以称为查询项或检索项。在实际应用中，使用舍弃部分词后的检索项进行检索，会比使用完整的检索词语效果更好。

可以理解的是，确定必留词及非必留词是检索中的一个重要步骤，确定结果准确与否能够决定召回文档是否优质。目前的检索项的生成方法，对哪些词属于必留词哪些词属于非必留词的判断准确度不高，生成的检索项不够准确，从而造成召回的文档也不够准确。

究其原因，目前的检索项的生成方法，是根据词赋权得到每个词的权重，按照权重对词进行排序，然后结合模板及后验规则来判断哪些词可以保留，哪些词可以丢弃。例如，检索词语为“民国时期人物”，将该检索词语进行分词操作后得到的词包括“民国”“时期”及“人物”，其中词赋权方法为词“民国”设置的权重为 0.8357，为词“时期”设置的权重为 0.8045，为词“人物”设置的权重为 0.3494，根据权重大小的排序，将词“民国”及词“时期”确定为必留词，将词“人物”确定为非必留词。

在使用词赋权设置词权重时，该方法仅仅使用词的独立特征如词出现次数等对每个词进行判断，并未考虑词与词之间的联系性特征，因此导致词的去留结果不准确。

在实际应用中，词的去留结果不准确的一种表现是，有意义的片段词语中的一部分被丢弃，这种情况会导致语义漂移，召回不相关文档。例如，检索词语为“怎样查看陌生人的朋友圈”，其分词结果包括“怎样”“查看”“陌生”“人”“的”“朋友”及“圈”。其中“朋友”和“圈”是两个紧密度比较高的词，目前的选择方法可能会出现丢弃“朋友”保

留“圈”，或丢弃“圈”保留“朋友”的情况，从而出现语义漂移，检索到不相关文档。

词的去留结果不准确的又一种表现是，一些对检索词语较为关键的词被丢弃，相反一些对检索词语并不关键的词被保留。以上述检索词语“民国时期人物”为例，从权重值上可知词“时期”比词“人物”被保留的可能性更大，但是依据人工经验可知“人物”这个词更应该被保留。

对此，本申请提供了一种检索项的生成方法，该方法根据词对查询结果的重要程度来判断词的去留情况，在较大程度上提高了所生成的检索项的准确度，进而提高了检索结果的准确度。仍以图 1 中的检索词语“儿歌简谱大全 100 首”为例，本申请针对该检索词语生成的检索项包括“儿歌”及“简谱”，使用该检索项检索到的文档如图 2 所示，对比图 1 及图 2 可知，本申请得到的检索项召回的文档明显比图 1 中召回的文档更加准确。

如图 3 所示，本申请提供的检索项的生成方法具体包括以下步骤 S201~S203。

S201：获得预先使用检索词语样本训练得到的机器学习模型，其中检索词语样本具有目标词特征，目标词特征用于表示词对查询结果的重要程度；

其中，机器学习是一种人工智能的科学，其使用已有经验数据自动学习出模型，用于对未知数据进行结果预测。在搜索领域中的生成检索项这一具体应用场景下，已有经验数据是检索词语（query），机器学习模型是使用机器学习算法对包含多条检索词语样本的训练集进行训练得到的，例如机器学习模型的一种训练方式可以为但不局限于 GBDT（Gradient Boosting Decision Tree，梯度提升决策树）。

机器学习模型在训练时所使用的检索词语样本具有目标词特征，需要说明的是，目标词特征可以表示出词对于查询结果（即召回文档）的重要程度。目标词特征可以是一项或多项，例如其可以包括以下几项中的任意一项或多项：词性特征、句法依存特征、反文档频率特征、点互信息特征、词向量类特征、重要性（Importance，IMP）特征、紧密度特征、词维度特征。以下将介绍这几项目标词特征的含义及计算方式，此处并不赘述。需要说明的是，在实际应用中目标词特征并不局限于以上几种，还可以是其他。

具体来讲，目标词特征对于查询结果的重要程度可以体现在，目标词特征不仅包含词的独立特征，还包含词与其他词的联系性特征，该联系性特征也可以称为词与词之间的相对特征。以上述几种目标词特征为例，联系性特征可以具体包括句法依存特征、点互信息特征、词向量类特征及紧密度特征等。

词与其他词的联系性特征可以避免有意义的词片段中的一部分丢弃，从而避免语义漂移导致的查询到不相关的结果，例如防止将“朋友圈”中的词“朋友”或词“圈”丢弃，从而避免查询到与“朋友”相关的文档或查询到与“圈”相关的文档。

另外，检索词语样本具有对应的查询结果，目标词特征的值可以是根据查询结果的命中情况计算得到的，因此，词对于查询结果的重要程度还可以体现在，目标词特征可以反映查询结果是否准确。这种能够反映查询结果命中情况的目标词特征，体现了查询结果的准确度，因此在一定程度上可以提高机器学习模型预测结果的准确度。

例如， 用户在搜索引擎中输入某条检索词语，搜索引擎向用户展示查询到的相关文档，用户点击某条展示文档则说明该文档命中。因此便可以使用该条检索词语、该条检索词语的召回文档及各个文档的命中与否等情况，来计算该检索词语中各个词的一项目标词特征，该项目标词特征用于训练机器学习模型。以上述几项目标词特征为例，计算的 IMP 特征便可以达到该效果。

S202：获得目标检索词语，确定目标检索词语中的词在目标词特征上的特征值。

其中，目标检索词语即待生成检索项的检索词语，也就是待确定词去留情况的检索词语。在用户搜索的应用场景中，目标检索词语可以是用户输入的检索词语如“插口灯泡怎么换”。在获得目标检索词语后，可以首先对目标检索词语进行分词操作，如分词结果为

“插口”“灯泡”“怎么”“换”，然后计算每个词在目标词特征上的特征值。

需要说明的是，该目标词特征为上述训练机器学习模型所使用的目标词特征。若目标词特征为多项，则确定每个词在每项目标词特征上的特征值，也就是说这种情况下，每个词具有多个目标词特征值。

假设，目标词特征包括紧密度，某目标检索词语为“一代宗师宋慧乔结局”，该目标检索词语分词后的结果为“一代”“宗师”“宋”“慧”“乔”“结局”，计算各个词的紧密度值为：一代 0.678 宗师 0.057 宋 0.568 慧 0.670 乔 0.061 结局，因此可知，词“一代”和词“宗师”之间的紧密度为 0.678，词“宗师”和词“宋”之间的紧密度为“0.057”，词“宋”和词“慧”之间的紧密度为“0.568”等等。可见，词“一代”和词“宗师”的紧密度 0.678 远大于词“宗师”和词“宋”的紧密度 0.057。

目标词特征可能并非仅仅包括参数，因此目标词特征的特征值也不一定仅仅包括数值。例如目标词特征包括词性特征的情况下，词性特征的特征值就是词性标记。以词集合为“插口”“灯泡”“怎么”“换”为例，各个词的词性特征的特征值分别为名词、名词、代词及动词。

S203：将目标检索词语及目标检索词语中词的目标词特征值输入至机器学习模型中，得到目标检索词语对应的检索项。

其中，词在目标词特征上的特征值可以简称为目标词特征值，在得到目标检索词语中词的目标词特征值后，便可以将目标检索词语及目标词特征值输入至机器学习模型中。

机器学习模型具有自身的计算方式，可以计算得到每个词的保留概率，进而可以根据保留概率确定词的去留情况。例如，目标检索词语为“上班甲醛超标怎么办”的词去留情况为：上班（非必留）甲醛（保留）超标（保留）怎么（非必留）办（非留）。

在一个示例中，可以按照由大到小的保留概率对词进行排序，从中选择保留概率排序在前的预设数量的词作为保留的词，其余词则作为可以舍弃的词。在另一个示例中，可以从中选择保留概率超过预设阈值的词作为保留的词，其余词则作为可以舍弃的词。当然，还可以是其他的去留选择方式，并不局限于以上两种。

确定为保留的词的集合可以称为检索项，使用检索项去进行查询，得到该检索词语对应的查询结果。或者，在使用机器学习模型确定词的去留情况时，机器学习模型可以仅仅输出词的保留概率，然后由其他模块根据保留概率生成检索项。

由以上的技术方案可知，本申请提供了一种检索项的生成方法，该方法使用机器学习模型来确定检索词语中各个词的去留情况，机器学习模型是使用具有目标词特征的检索词语样本训练而成的，相比现有的检索项的生成方法而言，目标词特征能反映词对查询结果重要程度的特征，使用该种目标词特征训练而成的机器学习模型，对检索词语中词选择结果更加精确，避免召回不相关的查询结果。

现有的检索项的生成方法的另一个问题是，需要依据用户的点击行为日志，对检索词语中的词的去留进行判断选择，搜索量不同的检索词语，其在点击行为日志中的数据量也不同，搜索量较少的检索词语中的词的去留判断准确性较低。

具体地，可以根据点击行为日志得到各个检索词语的搜索量曲线，如图 4 所示，搜索量曲线包括头部及尾部，头部对应的是搜索量较大的检索词语，尾部对应的是搜索量较小的检索词语。例如“微信”是一个处于头部的检索词语，“微信通讯录”是一个处于尾部的检索词语。处于头部的检索词语的搜索量较大，其在点击行为日志中的相关数据也较多，相反，处于尾部的检索词语的搜索量较小，其在点击行为日志中的相关数据也较少。在使用点击行为日志选择处于尾部检索词语的词时，相关数据较少，从而选择准确度不高。

然而，本申请提供的检索项的生成方法，并不依赖于点击行为日志进行判断，对于头部检索词语和尾部检索词语均能有较好的确定结果。

本发明的实施方式

以下具体说明机器学习模型的训练过程。

如图 5 所示，机器学习模型的训练过程包括如下步骤 S401~S404。

S401：获得检索词语样本，并对检索词语样本进行分词操作得到若干词。

在实施前，可以预先收集一些基础数据，如用户的查询日志、查询日志中的检索词语召回的文档。检索词语样本可以从基础数据中的查询日志中得到，例如收集大量用户的查询日志，将这些用户的查询日志中的检索词语作为检索词语样本。对每条检索词语样本进行分词操作，得到每条检索词语样本的词结果。

S402：确定词的去留情况。

其中，针对每条检索词语样本的每个词，需要确定该词的去留情况，去留情况用于表示该词是否用于生成检索项。一种确定方式是由人工进行标注，但该种方式需要耗费人工成本且主观性较强容易出现标准不准确的情况，因此，可以使用如下基于统计的确定的方式。

获得检索词语样本对应的召回文档，例如用户在搜索引擎中输入某条检索词语，搜索引擎会展示一定数量的文档，若该条检索词语作为用于机器学习训练的样本，则将搜索引擎展示的全部文档、或排序在前的部分文档、或用户点击的文档作为召回文档。

得到检索词语样本对应的召回文档后，统计检索词语样本的词在召回文档中的特征如出现次数，根据所述特征，确定所述检索词样本中的词的去留情况。

在一个示例中，根据统计的特征计算词的分数，并根据分数的大小确定词的去留情况。

计算词分数的一种具体方式为，统计词在召回文档的标题、摘要、标红信息、正文等位置出现的次数，根据词出现的位置、该位置对应的权重、及在该位置出现的次数，来计算词的分数。其中，不同的位置具有不同的权重，这样因为词在不同位置出现，其对查询结果的重要程度不同，例如词在某召回文档的标题、摘要及标红信息中出现，更能代表该召回的文档准确，因此，这些位置的权重更高一些。

根据分数的大小确定词去留情况的一种实现方式为，根据分数的大小进行排序，将排序在前的预设数量的词确定为必留，其余词确定为非必留。另一种实现方式是，将分数达到预设分数阈值的词确定为必留，其余词确定为非必留。

如上基于统计方式自动构建的训练集可能含有噪音，因此可以使用人工标注的方式对上述自动标注的结果进行修正，例如将某必留词修正为非必留词，或者将某非必留词修正为必留词。在人工修正时，由于人工主观标准不同，可能存在多种修正结果，其中非必留词较多的修正结果可以称为激进结果，非必留词较少的修正结果可以称为保守结果。

例如，检索词语样本为“上班甲醛超标怎么办”，保守结果和激进结果如下表 1 所示，其中 1 表示必留，0 表示非必留。

表 1

	保守结果	激进结果
上班	1	0
甲醛	1	1
超标	1	1
怎么	0	0
办	0	0

S403：确定词在目标词特征上的特征值。

需要说明的是，目标词特征指的是步骤 S201 中的目标词特征，这种目标词特征可以表征词对于查询结果的重要程度。需要说明的是，以上步骤 S402 与步骤 S403 并无执行顺序上的限定，可以是先执行步骤 S402，也可以是先执行步骤 S403，也可以是同时执行。

其中词在目标词特征上的特征值可以称为目标词特征值，目标词特征值可以是根据预先生成的目标词特征词典或目标词特征树进行计算。以下介绍几种目标词特征值的具体计算方式。

1、对于词性特征，即确定词的词性。

例如，检索词语“插口灯泡怎么换”的分词结果为“插口”“灯泡”“怎么”“换”，词性分别为名词、名词、代词及动词。

需要说明的是，名词、实体词、人名、地名等词性的重要程度大于形容词及动词等，形容词及动词等词性大于连词、介词、助词及语句词等无意义的词。在训练机器学习模型及使用机器学习模型确定词的去留时，重要性高的词性的权重大于重要性低的词性，从而被保留的概率更大。

2、对于句法依存特征。获得预先依据语料生成的句法依存树，从句法依存树中提取词的句法依存特征。

具体地，可以预先搜集语料如各种新闻、论文等，在这些语料中提取语句的句法依存树。其中语料可以包含检索词语样本召回的文档，也可以包含其他文档，也可以两者都包含。

例如，检索词语“插口灯泡怎么换”的句法依存树如图 6 所示，其中 HED（head）指的是核心关系，ATT（attribute）指的是定中关系，SBV（subject-verb）指的是主谓关系，ADV（adverbial）状中关系。

从句法依存树中可以词的句法依存特征，例如词“灯泡”有两个句法依存特征，分别为外指向的 ATT 关系及内指向的 SBV 关系，分别可以记为：灯泡_to_ATT = 1 及灯泡_from_SBV = 1。

需要说明的是，某些关系如主谓宾关系、核心关系指向及被指向的词重要性更高一些，因此，在训练机器学习模型及使用机器学习模型确定词的去留时，重要性高的句法依存特征的权重大于重要性低的句法依存特征，从而被保留的概率更大。

3、关于反文档频率（inverse document frequency，IDF）。其是文档频率（DF document frequency）的倒数，与文档的召回数量相关，反文档频率越大则召回的文档越少，反文档频率越小则召回的文档越多。

在计算词的反文档频率时，获得预先依据语料生成的反文档频率词典（IDF 词典），反文档频率词典中包含多个词的反文档频率，因此可以从反文档频率词典中查询词的反文档频率。

反文档频率可以包括单词的反文档频率（uni_idf）、两词的反文档频率（bi_idf）及基于两词的反文档频率得到的综合两词反文档频率特征。在生成反文档频率词典时，可以通过以下方式来计算词的各种反文档频率。

其中，uni_idf 可以由总文件数目除以包含该词的文件的数目，再将得到的商取对数得到。bi_idf 的计算方式与 uni_idf 的计算方式相同，只不过在计算 bi_idf 时，只有两个词均出现在文件中且出现的间隔距离小于预设的距离阈值如 5 才作为两个词出现在某个文件中。

得到词与其他词的 bi_idf 后，基于 bi_idf 计算该词的综合两词反文档频率，如最大两词反文档频率（max_bi_idf）、最小两词反文档频率（min_bi_idf）、平均两词反文档频率（avg_bi_idf）、变异两词反文档频率（cov_bi_idf）。其中，变异两词反文档频率的计算方式可以是该词的 avg_bi_idf 除以该词的各个 bi_idf 的标准差。

如图 7 所示，词“小说”分别与词“网游”、词“长枪”、词“主角”、词“的”及词“用”的 bi_idf 值为 3.33003，5.40573，2.4141，1.54426，3.27171，通过计算可知 max_bi_idf 为 5.40573，min_bi_idf 为 1.54426，avg_bi_idf 为 3.193166，cov_bi_idf 为 0.402179。

需要说明的是，两词的反文档频率 bi_idf 能够反映两个词的联系，这个目标词特征能

够提高词去留的判断准确性。

4、对于点互信息（Pointwise Mutual Information ，PMI）特征。其统计两个词在文档中同时出现的概率，可以用来衡量两个词的语义相关性。

确定词的点互信息特征时，可以获得预先依据语料生成的点互信息特征词典（PMI 词典），点互信息特征词典中包含多个词的点互信息特征，因此可以从点互信息特征词典中查询词的点互信息特征。

与反文档频率类似，点互信息特征可以包括两词的点互信息特征及基于两词的点互信息特征得到的综合两词点互信息特征。在生成点互信息特征词典时，可以通过以下方式来计算词的各种点互信息特征。

两词点互信息特征 PMI(term1，term2)的计算方式为：

PMI(term1,term2)=p (term1,term2)/p(term1)*p(term2) ；其中，p(term1,term2)是词 1 和词 2 在文本中同时出现的概率，p(term1)是词 1 在文本中出现的概率，p(term2)是 term2 在文本中出现的概率。其中，词在文本中出现的概率的计算方式是，词出现的文本数除以计算总文本数。

得到词与多个其他词的 PMI 后，基于各个 PMI 计算该词的综合 PMI，如最大 PMI、最小 PMI、平均 PMI、变异 PMI。计算方式可以参见上述反文档频率，此处并不赘述。

需要说明的是，两词 PMI 能够反映两个词的语义相关性，这个目标词特征能够提高词去留的判断准确性。

5、对于词向量类特征。词向量类特征可以反映词在语义上的特征，具体体现为可以反映词与检索词语在语义（如主题意图）是否相似。

在计算词的词向量类特征时，可以获得预先根据语料训练得到的词向量提取工具，使用词向量提取工具训练词的词向量类特征。例如词向量提取工具为 word2vec。

词向量提取工具提取词的词向量类特征的具体方式可以有多种，例如一种方式可以是，计算词与检索词语中其他词的向量余弦相似性，然后类似 bi_idf 可以求出词向量类特征；又如一种方式可以是将检索词语中所有词的平均词向量作为检索词语的词向量，然后计算词的词向量与检索词语的词向量的余弦相似性。

6、对于重要性（Importance，IMP）特征。基于有明确点击意图的检索词语，迭代的计算每个词的 IMP 值和词在每个检索词语中的重要性占比，多轮迭代后，每个词的 IMP 值逐渐收敛，直至前后两轮 IMP 的值变化不大，从而得到 IMP 词典。其中明确点击意图可以从用户的点击日志中确定出来的。

相比于反向文档频率等目标词特征，这个目标词特征能表示词对于查询结果的重要程度，因此在训练机器学习模型及使用机器学习模型确定词的去留情况时，该目标词特征的权重可能会比反向文档频率的权重更大。

计算词的 IMP 特征时，可以获得检索词语样本、检索词语样本的召回文档及召回文档的命中情况，通过多轮迭代计算 IMP。

7、对于紧密度特征。该目标词特征可以体现词与左右词的紧密程度。例如，检索词语“一代宗师宋慧乔结局”的紧密度结果为：一代 0.678 宗师 0.057 宋 0.568 慧 0.670 乔 0.061 结局。其中，词“一代”和词“宗师”的紧密度为 0.678，词“宗师”和词“宋”的紧密度为 0.057 等等。

在计算词的紧密度特征时，一种计算方式是，获得预先依据语料生成的检索词语关系树，使用检索词语关系树计算词的紧密度特征。具体地，根据各个检索词语内的词的包含关系构建检索词语关系树，基于检索词语关系树自下而上逐层计算并多次递归计算相邻词的紧密度。具体地，可以计算词所在节点的权重及连接相邻词的边的权重，使用两个权重计算相邻词的紧密度。

另一种计算方式是，根据词与其他词在同一文档中出现的概率计算。需要说明的是，

出现的位置间隔要小于预设间隔阈值。又一种计算方式是，计算词的多种紧密度特征，然后计算多种紧密度特征的综合值。又一种计算方式是，在通过以上任意方式得到词的紧密度特征后，可以使用词间的依存关系对紧密度特征进行修正。

该目标词特征可以体现词与其他词的联系程度，在构建机器学习模型时，使用该目标词特征可以提高机器学习模型的准确度，从而提高机器学习模型对输入的检索词语中词的去留情况的判断准确度。

8、关于词维度特征。该特征针对检索词语，可以具体为检索词语中包含的词个数、检索词语的长度等。

步骤 S401 得到检索词语样本的词后，可以计算词在以上 8 种目标词特征上的特征值。当然，在实际应用中，也可以是其他目标词特征值，只要目标词特征值能够表示目标词特征对查询结果的重要程度即可。

需要说明的是，以上各种词典及各种树结构都是由语料生成的，用于确定词在目标词特征值上的特征值，因此其可以被称为目标词特征值确定工具。

S404：将检索词语样本、检索词语样本中词的去留情况及词的目标词特征组成训练集。

其中，步骤 S401 得到了检索词语样本，步骤 S402 得到了检索词语样本中词的去留情况，本步骤即步骤 S403 得到词的目标词特征，此三项便可以组成训练集，训练集进而用于训练机器学习模型。

需要说明的是，在得到机器学习模型后，使用机器训练模型对某条检索词语中词的去留情况进行预测时，需要输入该条检索词语中词的目标词特征值。该条检索词语中词的目标词特征值也可以按照以上方法进行计算。

在实际应用中，还可以获得非机器学习模型为目标检索词语生成的修正检索项，使用修正检索项，对机器学习模型生成的检索项进行修正。将机器学习模型的确定结果（即目标检索词语对应的检索项）与其他方式（也就是“非机器学习模型”）如词赋权方式确定的结果进行结合，使用其他方式确定的结果对机器学习模型的确定结果进行修正。其中其他方式确定的结果可以称为修正检索项。

具体他，判断机器学习模型对应的确定结果中，被确定为非必留的词是否被其他方式确定为保留概率较大的词，若是，则将该非必留词修改为必留词。其中保留概率较大可以体现为该词被其他方式确定为排序在前预设数量的词。在进行此种修正后，若必留词的个数超过检索词语中词个数对应的阈值，则将所有必留词中保留概率较小的词修改为非必留词。

例如，针对检索词语“Swisse 多种维生素哺乳期能吃吗”，使用机器学习模型确定的必留词包括“维生”“素”“哺乳”及“期”，确定的非必留词包括“Swisse”“多种”“能”“吃”及“吗”。但假设使用词赋权方式计算的权重最大词为“Swisse”，权重最小词为“期”，则将词“Swisse”由非必留词修改为必留词。修正后的必留词包括 5 个，分别为“Swisse”“维生”“素”“哺乳”及“期”。假设预设阈值为 4 个，此时必留词的个数大于该预设阈值，则需要将所有必留词中权重最小的词“期”修改为非必留词，因此，最终的必留词包括“Swisse”“维生”“素”及“哺乳”。

或者，还可以结合目标词特征如紧密度特征、句法依存特征对机器学习模型得到的确定结果进行修正。例如，机器学习模型将词“一代”和“宗师”中的某个词确定为必留，某个词确定为非必留，而由紧密度特征确定出两个词之间的紧密度较高，因此将其中的非必留词修改为必留词。

可见，结合多种确定方式所得到的词去留情况结果更加准确。

以下通过实验数据说明本申请提供的词确定方法的准确度。

在一个实验中，收集到 105829 条检索词语，按照 8 与 2 的比例划分为训练集及测试集。使用 GBDT 方法对训练集进行训练后得到机器学习模型，对机器学习模型进行调优后，使用测试集对机器学习模型的确定结果进行测试，得到该机器学习模型确定的词去留情况准确度为 0.80702。在另一项实验中，本申请生成的检索项相较于现有技术而言胜出率为 67.5%。

见图 8，其示出了本申请提供的检索项的生成装置的一种结构，具体包括：机器学习模型获得器 701、目标词特征值获得器 702 及检索项生成器 703。

机器学习模型获得器 701，配置为获得预先由检索词语样本训练得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；

目标词特征值获得器 702，配置为获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；

检索项生成器 703，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

见图 9，其示出了本申请提供的检索项的生成装置的另一种结构，具体包括：机器学习模型获得器 701、目标词特征值获得器 702、检索项生成器 703 及机器学习模型训练器 704。

机器学习模型训练器 704，配置为训练检索词语样本训练，以得到机器学习模型；

其中，如图 10 所示，所述机器学习模型训练器 704 具体包括：词去留情况确定器 7041、目标词特征值确定器 7042、及机器学习模型生成器 7043。

词去留情况确定器 7041，配置为获得检索词语样本后，确定所述检索词语样本中的词的去留情况，所述去留情况用于表示所述词是否用于生成检索项；

目标词特征值确定器 7042，配置为确定所述词在所述目标词特征上的特征值；

机器学习模型生成器 7043，用配置为对所述检索词语样本、所述词的去留情况及所述词的特征值进行训练后，得到所述机器学习模型。

在一个示例中，所述目标词特征值确定器 7042 在执行确定所述词在所述目标词特征上的特征值的步骤时，具体用于：获得依据语料生成的目标词特征值确定工具；以及使用所述目标词特征值确定工具，确定所述词在所述目标词特征上的特征值。

在一个示例中，所述词去留情况确定器 7041 在执行确定所述检索词样本中的词的去留情况的步骤时，具体用于：获得所述检索词样本对应的召回文档；确定所述检索词样本中的词在所述召回文档中的特征；根据所述特征，确定所述检索词样本中的词的去留情况。

在一个示例中，所述词去留情况确定器 7041 在执行确定所述检索词语样本中的词的去留情况的步骤时，具体用于：根据所述特征计算所述检索词样本中的词的分数，并根据所述分数确定所述词的去留情况。

在一个示例中，如图 11 所示，所述检索项生成器 703 包括：保留概率确定器 7031 及检索项确定器 7032。

保留概率确定器 7031，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语中的词的保留概率；

检索项确定器 7032，配置为根据所述保留概率，确定所述目标检索词语中的词的去留情况，其中必留的词的集合为检索项。

见图 12，其示出了本申请提供的检索项的生成装置的又一种结构，具体包括：机器学习模型获得器 701、目标词特征值获得器 702、检索项生成器 703、机器学习模型训练器 704 及检索项修正器 705。

检索项修正器 705，配置为获得非机器学习模型为所述目标检索词语生成的修正检索项；以及使用所述修正检索项，对所述机器学习模型生成的检索项进行修正。

在一个示例中，所述检索项修正器 705 包括：非必留词修正器。

非必留词修正器，配置为判断所述机器学习模型确定为非必留的词是否被所述非机器学习模型确定为符合必留条件的词；若是，则将所述机器学习模型确定为非必留的词修正为必留。

在一个示例中，所述目标词特征包括以下几项中的任意一项或多项的组合：

词性特征、句法依存特征、反文档频率、点互信息、词向量类特征、重要性特征、紧密度、词维度特征。

在一种示例性实施例中，本申请提供的一种检索项的生成装置还包括：

第二修正器，配置为结合所述目标词特征对所述机器学习模型生成的检索项进行修正。

下面对本申请实施例提供的检索项的生成设备的硬件结构进行描述。图 13 为本申请实施例提供的检索项的生成设备的硬件结构示意图，参照图 13，该设备可以包括：输入器 1、处理器 2、存储器 3 及通信总线 4。

其中输入器 1、处理器 2、存储器 3 通过通信总线 4 完成相互间的通信；

可选地，输入器 1 用于输入数据；处理器 2 用于执行程序；存储器 3 用于存放程序。

程序可以包括程序代码，所述程序代码包括处理器的操作指令。

处理器 2 可能是一个中央处理器 CPU，或者是特定集成电路 ASIC（Application Specific Integrated Circuit），或者是被配置成实施本申请实施例的一个或多个集成电路。

存储器 3 可能包含高速 RAM 存储器，也可能还包括非易失性存储器（non-volatile memory），例如至少一个磁盘存储器。

输入器 1，用于输入检索词语样本。

处理器 2，用于训练所述检索词语样本训练，以得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；以及将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

在一种示例性实施例中，本申请还提供一种电子设备，该电子设备包括：

处理器；

配置为存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行上述示例性实施例所述的检索项的生成方法。

该实施例中电子设备的处理器执行操作的具体方式已经在有关该检索项的生成方法的实施例中执行了详细描述，此处将不做详细阐述说明。

在一种示例性实施例中，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计算机程序，所述计算机程序可由处理器执行完成上述示例性实施例所述的检索项的生成方法。

需要说明的是，本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、物品或者设备不仅包括那些要素，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、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括上述要素的过程、方法、物品或者设备中还存在另外的相同要素。

对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权利要求

1、一种检索项的生成方法，其中，包括：

获得预先由检索词语样本训练得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；

获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；

将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

2、根据权利要求1所述的检索项的生成方法，其中，所述机器学习模型的训练方法包括：

获得检索词语样本后，确定所述检索词语样本中的词的去留情况，所述去留情况用于表示所述词是否用于生成检索项；

确定所述词在所述目标词特征上的特征值；

对所述检索词语样本、所述词的去留情况及所述词的特征值进行训练后，得到所述机器学习模型。

3、根据权利要求2所述的检索项的生成方法，其中，所述确定所述词在所述目标词特征上的特征值，包括：

获得依据语料生成的目标词特征值确定工具；

使用所述目标词特征值确定工具，确定所述词在所述目标词特征上的特征值。

4、根据权利要求2所述的检索项的生成方法，其中，所述确定所述检索词样本中的词的去留情况，包括：

获得所述检索词样本对应的召回文档；

确定所述检索词样本中的词在所述召回文档中的特征；

根据所述特征，确定所述检索词样本中的词的去留情况。

5、根据权利要求4所述的检索项的生成方法，其中，所述根据所述特征，确定所述检索词样本中的词的去留情况，包括：

根据所述特征计算所述检索词样本中的词的分数，并根据所述分数确定所述词的去留情况。

6、根据权利要求1所述的检索项的生成方法，其中，所述将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项，包括：

将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语中的词的保留概率；

根据所述保留概率，确定所述目标检索词语中的词的去留情况，其中必留的词的集合为检索项。

7、根据权利要求1至6任意一项所述的检索项的生成方法，其中，还包括：

获得非机器学习模型为所述目标检索词语生成的修正检索项；

使用所述修正检索项，对所述机器学习模型生成的检索项进行修正。

8、根据权利要求7所述的检索项的生成方法，其中，所述使用所述修正检索项，对所述机器学习模型生成的检索项进行修正，包括：

判断所述机器学习模型确定为非必留的词是否被所述非机器学习模型确定为符合必留条件的词；

若是，则将所述机器学习模型确定为非必留的词修正为必留。

9、根据权利要求1所述的检索项的生成方法，其中，所述目标词特征包括以下几项中的任意一项或多项的组合：

词性特征、句法依存特征、反文档频率、点互信息、词向量类特征、重要性特征、紧密度、词维度特征。

10、根据权利要求1所述的检索项的生成方法，其中，还包括：

结合所述目标词特征对所述机器学习模型生成的检索项进行修正。

11、一种检索项的生成装置，其中，包括：

机器学习模型获得器，配置为获得预先由检索词语样本训练得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；

目标词特征值获得器，配置为获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；

检索项生成器，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

12、根据权利要求11所述的检索项的生成装置，其中，还包括：机器学习模型训练器，配置为训练检索词语样本训练，以得到机器学习模型；

其中，所述机器学习模型训练器包括：

词去留情况确定器，配置为获得检索词语样本后，确定所述检索词语样本中的词的去留情况，所述去留情况用于表示所述词是否用于生成检索项；

目标词特征值确定器，配置为确定所述词在所述目标词特征上的特征值；

机器学习模型生成器，配置为对所述检索词语样本、所述词的去留情况及所述词的特征值进行训练后，得到所述机器学习模型。

13、根据权利要求11所述的检索项的生成装置，其中，所述检索项生成器包括：

保留概率确定器，配置为将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语中的词的保留概率；

检索项确定器，配置为根据所述保留概率，确定所述目标检索词语中的词的去留情况，其中必留的词的集合为检索项。

14、根据权利要求11至13任意一项所述的检索项的生成装置，其中，还包括：

检索项修正器，配置为获得非机器学习模型为所述目标检索词语生成的修正检索项；以及使用所述修正检索项，对所述机器学习模型生成的检索项进行修正。

15、根据权利要求14所述的检索项的生成装置，其中，所述检索项修正器，包括：

非必留词修正器，配置为判断所述机器学习模型确定为非必留的词是否被所述非机器学习模型确定为符合必留条件的词；若是，则将所述机器学习模型确定为非必留的词修正为必留。

16、一种检索项的生成设备，其中，包括：

输入器，配置为输入检索词语样本；

处理器，配置为训练所述检索词语样本训练，以得到的机器学习模型，其中所述检索词语样本包含词，所述词具有目标词特征，所述目标词特征用于表示所述词对查询结果的重要程度；获得目标检索词语，并确定所述目标检索词语中的词在所述目标词特征上的特征值；以及将所述目标检索词语及所述特征值输入至所述机器学习模型中，得到所述目标检索词语对应的检索项。

17、一种电子设备，其中，所述电子设备包括：

处理器；

配置为存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行权利要求1-10任意一项所述的检索项的生成方法。

18、一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序可由处理器执行完成权利要求1-10任意一项所述的检索项的生成方法。

说明书附图

[1] 以歌做戈：乐界开幕第一人
摘要: 作为第一个把简谱从日本引进我国的音乐教育家...入选为"二十世纪华人音乐经典"的第一首歌
[2] 这里有好听的英文歌...
摘要: 今天老爱介绍的是一首来自 rihanna 的 we found love.这首歌曲大家应该听过很多次，熟悉的节奏，具有...
[3] 【岁月如歌】：《八路军军歌》
摘要: 简谱 为电友带来快乐； 为客户创造价值； 为自己创造形象； 为各种年龄段全方位学会快速的脱颖）...
[4] 一路走来一路歌，音乐才女备战之路
摘要: 乐理、简谱视唱、五线谱视唱一齐上，寒假的时间...原文标题《2016 中考，我们一起走过》 文/长沙码...
[5] 76. 中华文化精髓及智慧语录[同一首歌]
摘要: 人皆可以为尧舜。 (孟子) 100. 得道多助，失道寡助。 (孟子) 101. 天时不如地利，地利不如人和。 ...

图 1

[1] 幼儿园最喜欢的儿歌简谱大全（幼师必备）
摘要: 申报人：张卫宁、邓玉平 联系方式：13488600105、13366888525 学前教育高校教师支持 1 群（已满，只
[2] 幼儿园小班儿歌（简谱）大全
摘要: 儿歌：大气球 你有红气球，我有绿气球， 碰碰球，好像点点头， 你有红气球，我...
[3] 游戏律动 | 幼儿园小班上学期儿歌（简谱）大全
摘要: 儿歌：幼儿园像我家 幼儿园，像我家， 老师爱我我爱她，老师叫我好性急， 爱说老师像妈妈， 儿歌...
[4] 幼儿园儿歌弹唱简谱 13 首，幼师收藏好了！
摘要: 《幼儿园儿歌弹唱简谱 13 首》图文源于网络，我们致力于推广视频、专业知识，如涉及原作者权益 ..
[5] 教师篇：幼儿园儿歌弹唱简谱 18 首
摘要: 分享幼儿园家长育儿知识等内容 ..

图 2

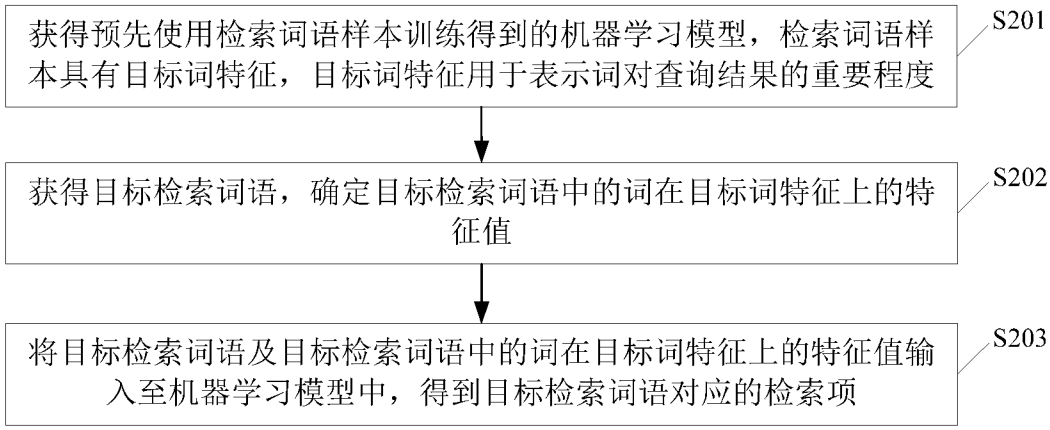
获得预先使用检索词语样本训练得到的机器学习模型，检索词语样本具有目标词特征，目标词特征用于表示词对查询结果的重要程度
S201
获得目标检索词语，确定目标检索词语中的词在目标词特征上的特征值
S202
将目标检索词语及目标检索词语中的词在目标词特征上的特征值输入至机器学习模型中，得到目标检索词语对应的检索项
S203

图 3

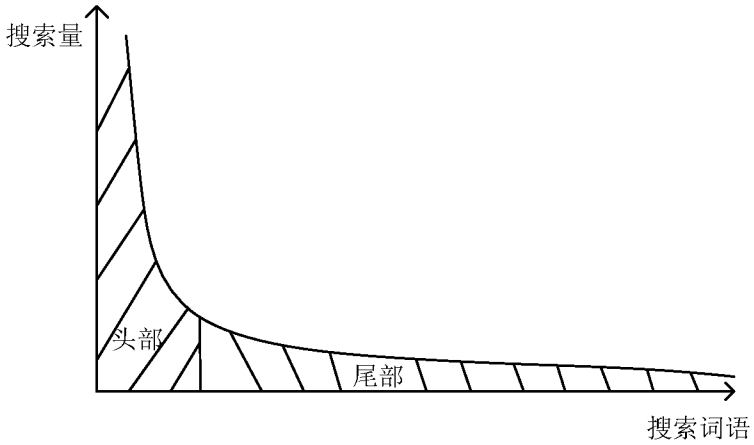
搜索量
头部
尾部
搜索词语

图 4

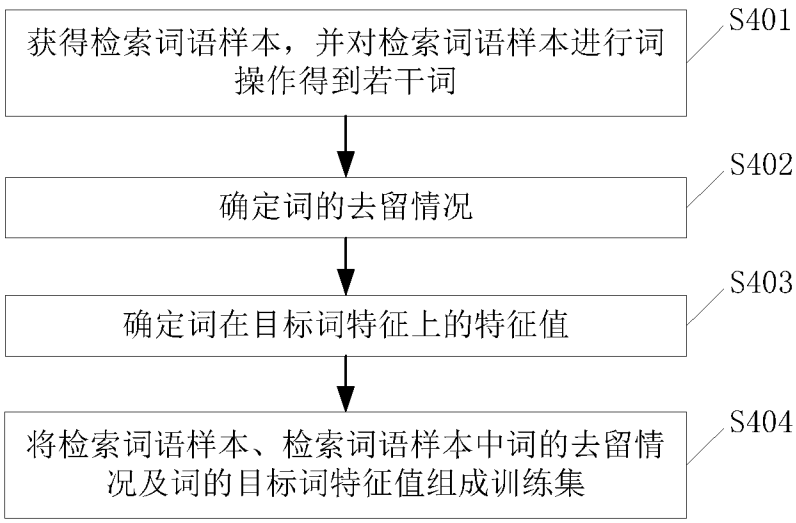
获得检索词语样本，并对检索词语样本进行词操作得到若干词
S401
确定词的去留情况
S402
确定词在目标词特征上的特征值
S403
将检索词语样本、检索词语样本中词的去留情况及词的目标词特征值组成训练集
S404

图 5

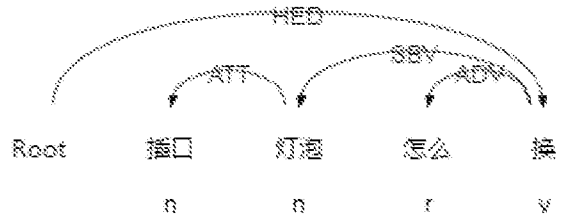
HED
ATT
SBV
ADV
Root
插口
灯泡
怎么
换
n
n
r
v

图 6

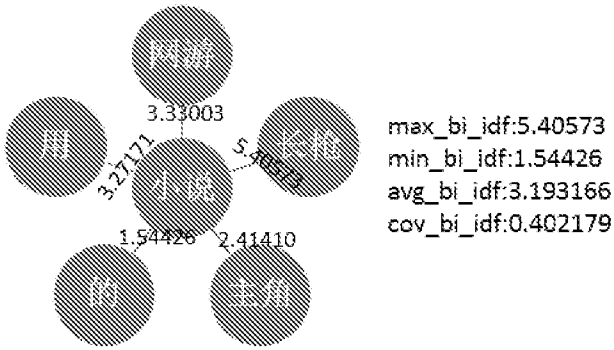
网游
3.33003
周
3.27171
长跑
5.40573
小说
1.54426
2.41410
跑
主角
max_bi_idf:5.40573
min_bi_idf:1.54426
avg_bi_idf:3.193166
cov_bi_idf:0.402179

图 7

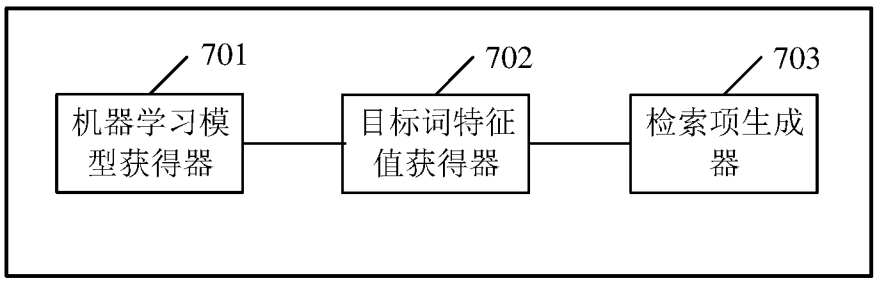
701
机器学习模型获得器
702
目标词特征值获得器
703
检索项生成器

图 8

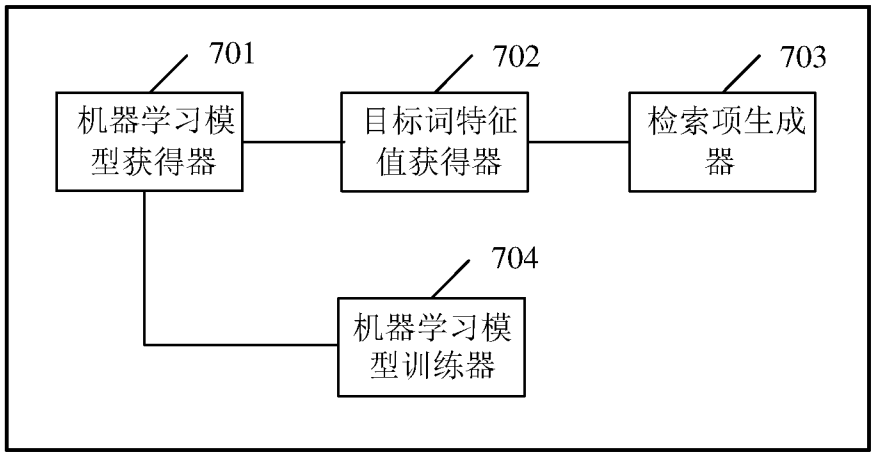
701
机器学习模型获得器
702
目标词特征值获得器
703
检索项生成器
704
机器学习模型训练器

图 9

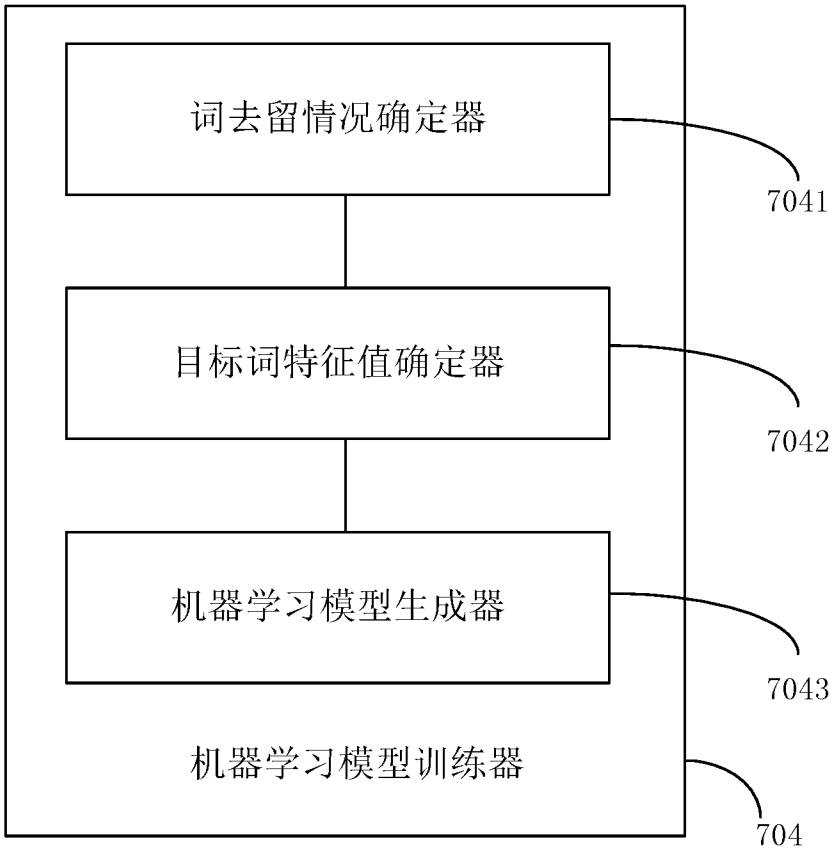
词去留情况确定器
7041
目标词特征值确定器
7042
机器学习模型生成器
7043
机器学习模型训练器
704

图 10

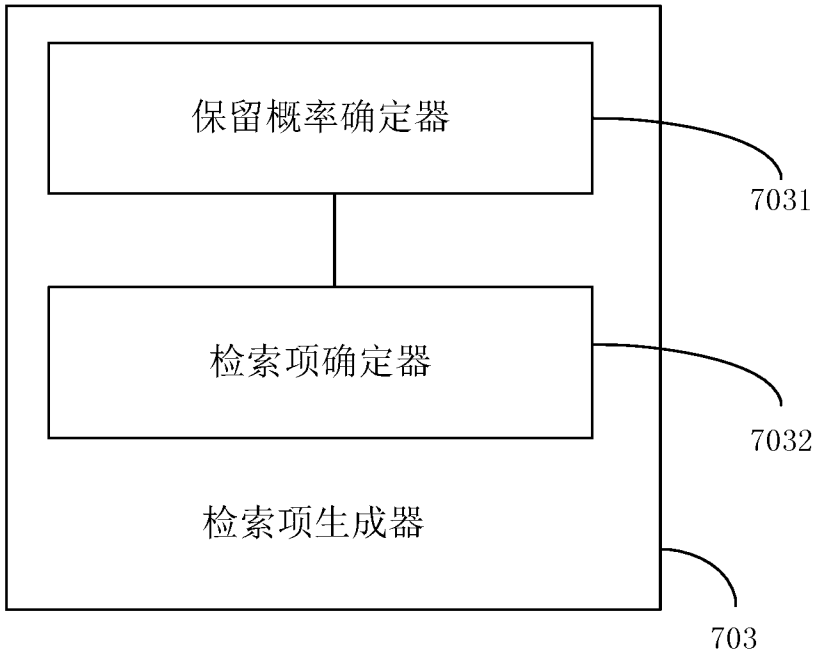
保留概率确定器
7031
检索项确定器
7032
检索项生成器
703

图 11

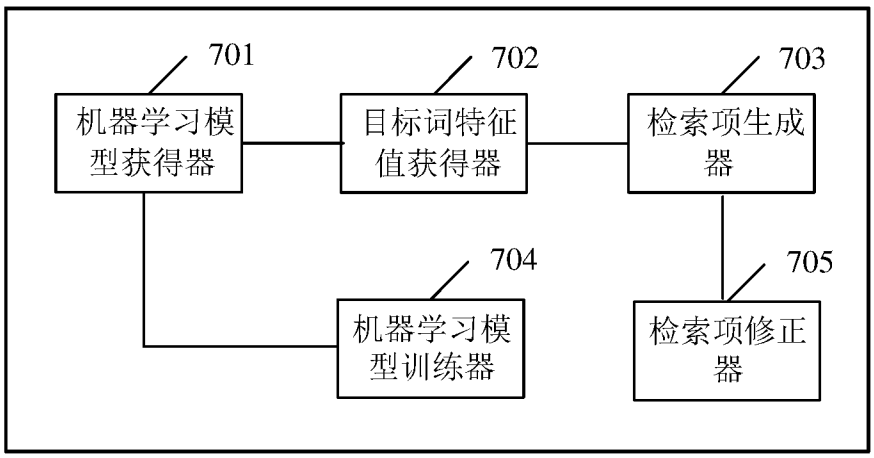
701
机器学习模型获得器
702
目标词特征值获得器
703
检索项生成器
704
机器学习模型训练器
705
检索项修正器

图 12

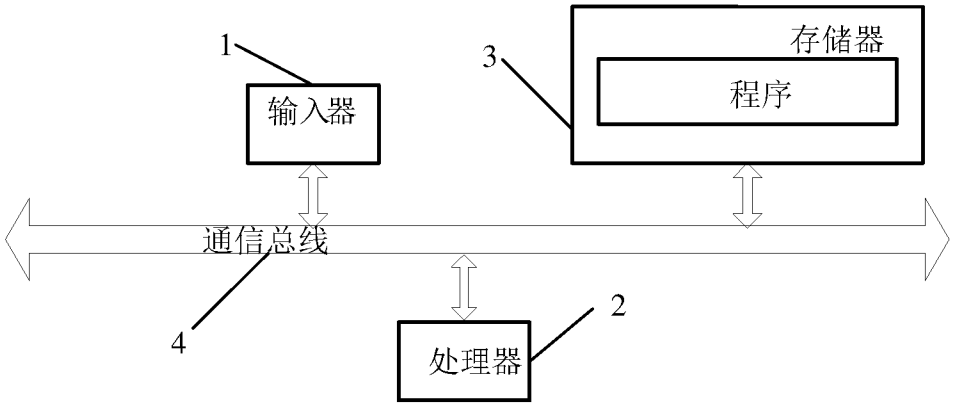

图 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/097748

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(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)

USTXT; EPTXT; CNTXT; CNABS; WOTXT; VEN; CNKI: 检索项, 样本, 检索, 模型, 检索词, 训练, 程度, 特征值, search term, sample, search, model, search word, train, degree, eigenvalues

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104376065 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 25 February 2015 (2015-02-25) description, paragraphs 7-16, 62 and 108-134, and figures 1-3	1-18
A	CN 106934063 A (SHENYANG AEROSPACE UNIVERSITY) 07 July 2017 (2017-07-07) entire document	1-18
A	CN 101650731 A (ZHEJIANG UNIVERSITY) 17 February 2010 (2010-02-17) entire document	1-18
A	CN 103838793 A (DALIAN LINGDONG TECHNOLOGY DEVELOPMENT CO., LTD.) 04 June 2014 (2014-06-04) entire document	1-18
A	US 7788276 B2 (YAHOO INC.) 31 August 2010 (2010-08-31) entire document	1-18

☐ 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
11 October 2018	**29 October 2018**

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/CN2018/097748

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 104376065 A	25 February 2015	None	
CN 106934063 A	07 July 2017	None	
CN 101650731 A	17 February 2010	None	
CN 103838793 A	04 June 2014	None	
US 7788276 B2	31 August 2010	US 2009055380 A1	26 February 2009

国际检索报告

国际申请号 PCT/CN2018/097748

A. 主题的分类

G06F 17/30(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）)

USTXT;EPTXT;CNTXT;CNABS;WOTXT;VEN;CNKI:检索项，样本，检索，模型，检索词，训练，程度，特征值，search term, sample, search, model, search word, train, degree, eigenvalues

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 104376065 A (百度在线网络技术北京有限公司) 2015年 2月 25日 (2015-02-25) 说明书第7-16、62、108-134及图1-3	1-18
A	CN 106934063 A (沈阳航空航天大学) 2017年 7月 7日 (2017-07-07) 全文	1-18
A	CN 101650731 A (浙江大学) 2010年 2月 17日 (2010-02-17) 全文	1-18
A	CN 103838793 A (大连灵动科技发展有限公司) 2014年 6月 4日 (2014-06-04) 全文	1-18
A	US 7788276 B2 (YAHOO INC) 2010年 8月 31日 (2010-08-31) 全文	1-18

☐ 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 10月 11日	2018年 10月 29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 夏雪 电话号码 86-(20)-28950718

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/097748

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 104376065 A	2015年 2月 25日	无	
CN 106934063 A	2017年 7月 7日	无	
CN 101650731 A	2010年 2月 17日	无	
CN 103838793 A	2014年 6月 4日	无	
US 7788276 B2	2010年 8月 31日	US 2009055380 A1	2009年 2月 26日