

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 3 月 19 日 (19.03.2020)



(10) 国际公布号
WO 2020/052547 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01) **G06F 16/332** (2019.01)

(21) 国际申请号: PCT/CN2019/105123

(22) 国际申请日: 2019 年 9 月 10 日 (10.09.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811076259.9 2018 年 9 月 14 日 (14.09.2018) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: 高喆 (GAO, Zhe); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。康杨杨 (KANG, Yangyang); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。周笑添 (ZHOU, Xiaotian); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。孙常龙 (SUN, Changlong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。刘晓钟 (LIU, Xiaozhong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。司罗 (SI, Luo); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING NEW WORDS IN SPAM MESSAGE, AND ELECTRONIC DEVICE

(54) 发明名称: 短信垃圾新词识别方法、装置及电子设备

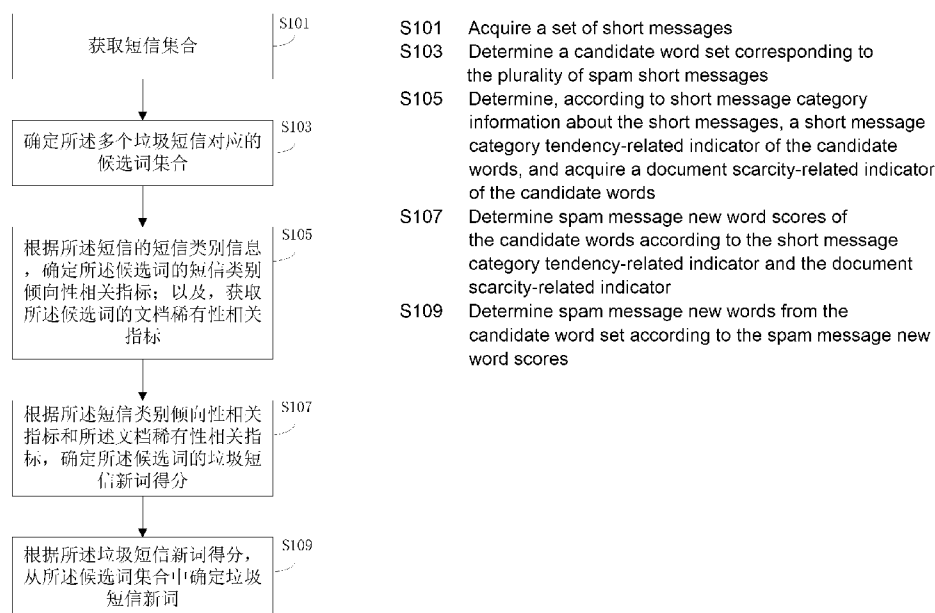


图 1

(57) Abstract: Disclosed in the present application are a method and an apparatus for identifying new words in a spam message, and an electronic device. Said method comprises: acquiring a set of short messages; determining a candidate word set corresponding to a plurality of spam messages; determining, according to short message category information, a short message category tendency-related indicator of candidate words, and acquiring a document scarcity-related indicator of the candidate words; determining spam message new word scores of the candidate words according to the short message category tendency-related indicator and the document scarcity-re-

(74) 代理人：北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, 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))。

lated indicator; determining spam message new words among the candidate word set according to the spam message new word scores. This processing method makes most of the new words recalled according to the short message category tendency-related indicator all have a representative meaning for the spam messages, so as to avoid recalling of many normal words. Therefore, the accuracy rate of recalling can be effectively improved. In addition, such a processing method enables low-frequency new words in the spam messages to be recalled according to the document scarcity-related indicator. Therefore, the accuracy rate of recalling can be effectively improved.

(57) 摘要: 本申请公开了短信垃圾新词识别方法、装置及电子设备。其中, 所述方法包括: 获取短信集合; 确定多个垃圾短信对应的候选词集合; 根据短信类别信息, 确定候选词的短信类别倾向性相关指标; 以及, 获取候选词的文档稀有性相关指标; 根据短信类别倾向性相关指标和文档稀有性相关指标, 确定候选词的垃圾短信新词得分; 根据垃圾短信新词得分, 从候选词集合中确定垃圾短信新词。采用这种处理方式, 使得根据短信类别倾向性相关指标召回的大多数新词均对垃圾短信具有代表意义, 避免召回很多正常的词汇; 因此, 可以有效提升召回率。同时, 这种处理方式, 使得根据文档稀有性的相关指标可召回低频的垃圾短信新词; 因此, 可以有效提升召回率。

短信垃圾新词识别方法、装置及电子设备

本申请要求 2018 年 09 月 14 日递交的申请号为 201811076259.9、发明名称为“短信垃圾新词识别方法、装置及电子设备”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。

5

技术领域

本申请涉及文本挖掘技术领域，具体涉及短信垃圾新词识别方法和装置，以及电子设备。

10 背景技术

一种典型的短信发送场景是，商家通过网络平台向消费者发送短信，以便于将商品促销等信息及时送至消费者处，从而确保商家销售计划的有效实施，并提升用户体验。然而，伴随着这些有益效果的同时，也出现了大量垃圾短信。垃圾短信泛滥，已经严重影响消费者正常生活、网络平台形象乃至社会稳定。

15 随着互联网技术的不断发展，越来越多的网络平台利用短信内容安全系统对商对客（Business-to-Customer，B2C）的短信进行内容分析，并进行智能短信拦截和通道优化。其中，垃圾短信新词发现是短信内容安全系统的一个重要功能，有效地挖掘垃圾短信新词可提高垃圾短信识别模型准确率，可以及时地应对线上的垃圾短信变种，并可以为短信审核人员提供每天线上新出现的垃圾变种词。目前，垃圾短信新词识别方法主要分为
20 两类：基于有监督的新词发现的垃圾短信新词识别方法、和基于无监督的新词发现的垃圾短信新词识别方法。其中，有监督的新词发现技术要依赖序列标注结果实现，该方法成本高，不容易获取语料；无监督的新词发现技术，通过某些指标来计算未包括在垃圾短信词典中的候选短语的成词概率，根据成词概率判断候选短语是否为垃圾短信新词，该方法成本低且效率高。由于无监督的新词发现技术具有成本低且效率高的优点，因此
25 成为一种最为常用的垃圾短信新词识别方案。

然而，在实现本发明过程中，发明人发现基于无监督的新词发现技术的垃圾短信新词识别技术方案至少存在如下问题：一方面，由于简单地依赖简单的频率特征（如词的出现频率等）进行识别，而没有考虑类别的分布，因此容易召回很多正常的词汇，由此导致后续审核成本较高；另一方面，由于要依赖简单的频率特征计算成词概率，而垃圾

短信中变异的新词出现频率低，因此这类词的成词概率就会低，这样就无法识别出这类垃圾短信新词，出现低频垃圾短信新词无法被召回的问题。

综上所述，现有技术存在垃圾短信新词召回率且召回率均较低的问题。

5 发明内容

本申请提供短信垃圾新词识别方法，以解决现有技术存在的垃圾短信新词召回率且召回率均较低的问题。本申请另外提供短信垃圾新词识别装置，以及电子设备。

本申请提供一种短信垃圾新词识别方法，包括：

获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

10 确定所述多个垃圾短信对应的候选词集合；

根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，
获取所述候选词的文档稀有性相关指标；

根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的
垃圾短信新词得分；

15 根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

可选的，所述确定所述多个垃圾短信对应的候选词集合，包括：

通过分词算法，获取所述垃圾短信包括的词，作为短信词；

若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候
选词。

20 可选的，所述候选词规则包括候选词的字数小于预设字数。

可选的，所述短信类别倾向性相关指标包括以下指标的至少一项：交叉熵，优势率，
互信息；

所述文档稀有性相关指标包括：倒文档频率 IDF。

可选的，所述根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定
25 所述候选词的垃圾短信新词得分，包括：

将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述
垃圾短信新词得分。

可选的，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，
包括：

30 将所述垃圾短信新词得分大于得分阈值的所述候选词作为垃圾短信新词。

可选的，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取并展示所述垃圾短信新词得分大于得分阈值的新词；

接收针对目标新词的确定指令；

5 将所述目标新词为所述垃圾短信新词。

可选的，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词；

确定所述候选的垃圾短信新词的成词概率；

10 根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

可选的，所述根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

将所述成词概率大于成词概率阈值的候选的垃圾短信新词作为所述垃圾短信新词。

可选的，所述根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

获取并展示所述成词概率大于成词概率阈值的候选的垃圾短信新词；

接收用户输入的针对目标新词的确定指令；

将所述目标新词为所述垃圾短信新词。

可选的，所述确定所述候选的垃圾短信新词的成词概率，包括：

20 确定所述候选的垃圾短信新词的内部结合度；以及，确定所述候选的垃圾短信新词的边界自由度；

根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短信新词的成词概率。

可选的，所述根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短信新词的成词概率，包括：

25 将所述内部凝聚度和所述外部自由度的平均值作为所述成词概率。

可选的，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

30 确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度；

根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

可选的，所述根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

将所述垃圾短信词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词作为垃圾短信新词。

可选的，所述根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

获取并展示所述垃圾短信词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词；

接收针对目标候选新词的确定指令；

将所述目标候选新词为所述垃圾短信新词。

可选的，所述根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度，包括：

将所述语义相似度的最大值作为所述垃圾短信词汇相似度。

可选的，所述确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度，包括：

确定所述候选新词的词向量；

根据所述预设垃圾短信词汇的词向量和所述候选新词的词向量，确定所述预设垃圾词汇与所述候选新词之间的语义相似度。

本申请还提供一种垃圾短信新词识别方法，包括：

获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

确定所述多个垃圾短信对应的候选词集合；

根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；

根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

可选的，所述根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述短信类别倾向性得分大于得分阈值的所述候选词，作为候选的垃圾短信新

词；

确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

本申请还提供一种垃圾短信新词识别方法，包括：

5 获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

确定所述多个垃圾短信对应的候选词集合；

确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；

根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

10 本申请还提供一种垃圾短信新词识别装置，包括：

短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

15 指标确定单元，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；

得分确定单元，用于根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；

新词确定单元，用于根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

20 可选的，所述候选词集确定单元包括：

通过分词算法，获取所述垃圾短信包括的词，作为短信词；

若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候选词。

25 可选的，所述得分确定单元，具体用于将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述垃圾短信新词得分。

可选的，所述新词确定单元包括：

候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词；

成词概率确定子单元，用于确定所述候选的垃圾短信新词的成词概率；

30 第一新词确定子单元，用于根据所述成词概率，从所述候选的垃圾短信新词中确定

所述垃圾短信新词。

可选的，所述新词确定单元包括：

候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

5 第一相似度确定子单元，用于确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度；

第二相似度确定子单元，用于根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

10 第二新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

本申请还提供一种电子设备，包括：

处理器；以及

15 存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

20 本申请还提供一种垃圾短信新词识别装置，包括：

短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

25 得分确定单元，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；

新词确定单元，用于根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

可选的，所述新词确定单元包括：

30 候选词选取子单元，用于获取所述短信类别倾向性得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

相似度确定子单元，用于确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

本申请还提供一种电子设备，包括：

5 处理器；以及

存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

本申请还提供一种垃圾短信新词识别装置，包括：

短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

15 第一语义相似度确定单元，用于确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；

第二语义相似度确定单元，用于根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；

20 新词确定单元，用于根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

本申请还提供一种电子设备，包括：

处理器；以及

25 存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

30 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中存储有指令，当其在计算机上运行时，使得计算机执行上述各种方法。

本申请还提供一种包括指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述各种方法。

与现有技术相比，本申请具有以下优点：

本申请实施例提供的一种短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词；这种处理方式，使得根据短信类别倾向性相关指标召回的大多数新词均对垃圾短信具有代表意义，避免召回很多正常的词汇；因此，可以有效提升召回率。同时，这种处理方式，使得根据文档稀有性的相关指标可召回低频的垃圾短信新词；因此，可以有效提升召回率。综上所述，这种处理方式，使得可筛选出既在垃圾短信中具有一定的稀有性，且在类别上比较倾向于垃圾短信的新词；因此，可以有效提升召回率和召回率，从而提升新词发现准确度。

本申请实施例提供的另一种短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；所述短信类别包括垃圾短信类或正常短信类；根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词；这种处理方式，使得根据短信类别倾向性相关指标召回的大多数新词均对垃圾短信具有代表意义，避免召回很多正常的词汇；因此，可以有效提升召回率。

本申请实施例提供的又一种短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词；这种处理方式，使得通过词嵌入的方式对候选词进行了语义扩展，并根据候选词与垃圾短信词汇之间的语义相似度，确定所述垃圾短信新词，因此不会忽略一些稀有的新词；因此，可以有效提升垃圾短信新词的召回率。

附图说明

图 1 是本申请提供的一种短信垃圾新词识别方法的实施例的流程图；

图 2 是本申请提供一种短信垃圾新词识别方法的实施例的具体流程图；

图 3 是本申请提供一种短信垃圾新词识别方法的实施例的具体流程图；

图 4 是本申请提供一种短信垃圾新词识别方法的实施例的具体流程图；

图 5 是本申请提供一种短信垃圾新词识别装置的实施例的示意图；

5 图 6 是本申请提供的电子设备的实施例的示意图；

图 7 是本申请提供一种短信垃圾新词识别方法的实施例的流程图；

图 8 是本申请提供一种短信垃圾新词识别装置的实施例的示意图；

图 9 是本申请提供的电子设备的实施例的示意图；

图 10 是本申请提供一种短信垃圾新词识别方法的实施例的流程图；

10 图 11 是本申请提供一种短信垃圾新词识别装置的实施例的示意图；

图 12 是本申请提供的电子设备的实施例的示意图。

具体实施方式

15 在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施的限制。

在本申请中，提供了垃圾短信新词识别方法和装置，以及电子设备。在下面的实施例中逐一对各种方案进行详细说明。

第一实施例

20 请参考图 1，其为本申请提供一种垃圾短信词汇识别方法实施例的流程图，该方法的执行主体包括垃圾短信词汇识别装置。本申请提供一种垃圾短信词汇识别方法包括：

步骤 S101：获取短信集合。

25 所述短信，又称为短消息或短信息，包括但不限于手机短信，也可以是即时消息等其它形式的短信。

所述短信集合包括多个垃圾短信和多个正常短信。其中，将垃圾短信的短信类别标注为垃圾短信，将正常短信的短信类别标注为正常短信。

步骤 S103：确定所述多个垃圾短信对应的候选词集合。

30 所述候选词，又称为候选新词，包括所述多个垃圾短信中出现的垃圾短信词典以外的词，但由于垃圾短信的分词结果并非新词，因此不包括垃圾短信的分词结果。所述垃

圾短信词典包括多个已经确定的垃圾短信词汇。

在一个示例中，步骤 S103 可包括如下子步骤：1) 通过分词算法，获取所述垃圾短信包括的词，作为短信词；2) 若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候选词。

5 1) 通过分词算法，获取所述垃圾短信包括的词，作为短信词。

具体实施时，可采用现有的分词算法对垃圾短信进行分词处理。现有的分词算法，可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。分词算法属于较为成熟的现有技术，此处不再赘述，可根据实际需求选取
10 任意一种现有分词算法。

例如，垃圾短信内容包括“新农合保险”，分词结果包括如下短信词：“新”、“农”、“合”和“保险”。

2) 若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候选词。

15 所述候选词规则，包括但不限于：候选词的字数小于预设字数。所述预设字数可以根据业务需求设置，例如，设置为 4 等等。

例如，对于上述垃圾短信内容“新农合保险”，分词结果：“新”、“农”、“合”和“保险”，所述至少两个相邻短信词构成的组合词包括：“新农”、“农合”、“合保险”。

20 需要说明的是，步骤 S103 也可以采用其它具体实施方式，只要能够确定所述多个垃圾短信对应的候选词集合即可。

步骤 S105：根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标。

本申请实施例提供的方法，引入短信类别信息，根据所述候选词在垃圾短信和正常
25 短信中的分布情况，获取所述候选词的短信类别倾向性相关指标。其中，短信类别可以是垃圾短信类或正常短信类。

所述短信类别倾向性，可以是正常短信倾向性，也可以是垃圾短信倾向性，例如，候选词 A 在正常短信中出现 10 次，在垃圾短信中出现 2 次，则候选词 A 的正常短信倾向性得分可以是 $10/2=5$ ，垃圾短信倾向性得分可以是 $2/10=0.2$ 。

30 所述短信类别倾向性相关指标，是指能够反映所述候选词的短信类别倾向性的指

标，也就是说，根据所述短信类别倾向性相关指标，可确定所述候选词的短信类别倾向性。

所述短信类别倾向性相关指标，包括但不限于以下指标的至少一项：交叉熵、优势率、互信息。

5 交叉熵 (Cross Entropy) 是 Shannon 信息论中一个重要概念，在本申请提供的方法中，主要用于度量两个概率（所述候选词的垃圾短信概率和正常短信概率）分布间的差异性信息。

优势率 (Odds Ratio) 只适用于二元分类的情况，其特点是只关心文本特征对于目标类的分值。Pos 表示目标类（如垃圾短信），neg 表示非目标类（如正常短信）。

10 互信息 (Mutual Information) 是信息论里一种有用的信息度量，它可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性。

本申请实施例提供的方法，还引入所述候选词的文档稀有性相关指标，以便于挖掘出来出现频率较低的垃圾短信中变异的新词等等。

15 所述文档稀有性相关指标，是指反映所述候选词的文档稀有性的指标，包括但不限于：逆文档频率 (inverse document frequency, IDF)。IDF 又称反文档频率，是文档频率的倒数，主要用于概念 TF-IDF (term frequency - inverse document frequency) 中。

在本实施例中，所述文档稀有性相关指标包括所述候选词在所述短信集内的文档稀有性相关指标。

20 步骤 S107：根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分。

通过上述步骤获得每一个所述候选词的所述短信类别倾向性相关指标和所述文档稀有性相关指标之后，就可以综合这两类指标，确定所述候选词的垃圾短信新词得分。

25 在一个示例中，步骤 S107 可采用如下方式实现：将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述垃圾短信新词得分。

其中，各个指标的权重，可根据业务需求确定。当所述文档稀有性相关指标的权重较大时，会提升低频候选词的得分，有利于筛选出较多低频的候选词，但可能包括较多的正常词汇；当所述短信类别倾向性相关指标的权重较大时，会提升垃圾短信倾向性高的候选词的得分，有利于筛选出对垃圾文本具有代表性的高频的候选词，但可能忽略部

分低频的候选词。

例如，所述候选词 A 在垃圾短信中出现 10 次，在正常短信中出现 1 次，简化的垃圾短信倾向性得分为 $10/1=10$ ；所述候选词 B 在垃圾短信中出现 2 次，在正常短信中出现 1 次，简化的垃圾短信倾向性得分为 $2/1=2$ ，如果仅按照垃圾短信倾向性得分，高频词 A 更容易被筛选，但是低频词 B 可能是个很关键的垃圾词汇，为了让 B 这样的词汇也同
5 样容易被采集到，就要考虑加入文档稀有性指标，对低频词的权重增大。譬如高频词 A 的 idf 是 2，低频词 B 的 idf 是 6，那么按照交叉熵、优势率和互信息这三个指标占 0.5，idf 占 0.5 的权重，此时最终得分为高频词 $A=0.5*10+0.5*2=6$ ，低频词 $B=0.5*2+0.5*6=4$ ，这样不仅可以筛选出高频词 A，也能够筛选出低频词 B。

10 步骤 S109：根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

本步骤根据每一个所述候选词的垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

在一个示例中，将所述垃圾短信新词得分大于得分阈值的所述候选词作为垃圾短信新词。

15 在另一个示例中，步骤 S109 包括如下子步骤：1) 获取并展示所述垃圾短信新词得分大于得分阈值的新词；2) 接收用户输入的针对目标新词的确定指令；3) 将所述目标新词为所述垃圾短信新词。

所述得分阈值，可根据业务需求确定。所述得分阈值越高，则垃圾短信新词的噪声越小，但垃圾短信新词的召全率越低，可能丢掉一些垃圾短信新词；所述得分阈值越
20 低，则垃圾短信新词的召全率越高，但垃圾短信新词的噪声越大，人工审核量越大。

请参看图 2，其为本申请的垃圾短信新词识别方法的实施例的具体流程图。在另一个示例中，步骤 S109 可包括如下子步骤：

步骤 S201：获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词。

25 步骤 S203：确定所述候选的垃圾短信新词的成词概率。

在一个示例中，所述确定所述候选的垃圾短信新词的成词概率的步骤，可包括如下子步骤：1) 确定所述候选的垃圾短信新词的内部结合度；以及，确定所述候选的垃圾短信新词的边界自由度；2) 根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短信新词的成词概率。

30 具体实施时，所述根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短

信新词的成词概率的步骤，可采用如下方式实现：将所述内部凝聚度和所述外部自由度的平均值作为所述成词概率。

步骤 S205：根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

- 5 具体实施时，可将所述成词概率大于成词概率阈值的候选的垃圾短信新词作为所述垃圾短信新词；也可以采用如下子步骤确定所述垃圾短信新词：1) 获取并展示所述成词概率大于成词概率阈值的候选的垃圾短信新词；2) 接收用户输入的针对目标新词的确定指令；3) 将所述目标新词为所述垃圾短信新词。

所述成词概率阈值，可根据业务需求确定。在本实施例中，为避免将根据所述垃圾
10 短信新词得分筛选出的低频的候选词过滤掉，可将所述成词概率阈值设置大些。

本申请实施例提供的方法，通过获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词；确定所述候选的垃圾短信新词的成词概率；根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词；这种处理方式，使得确定出的垃圾短信新词还是更加具有实际意义的词，如“80后”，避免筛选出不具有实际意义的
15 的垃圾短信新词，如“新合”；因此，可以有效提升垃圾短信新词的有效性。

请参看图 3，其为本申请的垃圾短信新词识别方法的实施例的具体流程图。在另一个示例中，步骤 S109 可包括如下子步骤：

步骤 S301：获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词。

20 步骤 S303：确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度。

所述至少一个预设垃圾短信词汇，包括但不限于：垃圾短信词典中的词汇。

在一个示例中，步骤 S303 可包括如下子步骤：1) 确定所述候选的垃圾短信新词的词向量；2) 根据所述预设垃圾短信词汇的词向量和所述候选的垃圾短信新词的词向量，
25 确定所述预设垃圾词汇与所述候选的垃圾短信新词之间的语义相似度。

1) 确定所述候选的垃圾短信新词的词向量。

要确定所述候选的垃圾短信新词的词向量，可采用如下方式实现：离线或在线计算所述短信集中所有短信的基于字的语言模型的 embedding(词嵌入，词向量)，如 N-Gram 或 Skip-Gram 语言模型，或采用 cbow、glove 等方式，以此确定所述候选的垃圾短信新
30 词的词向量。采用这种处理方式，可以有效提升词向量的准确度，例如，短信 A “买卖

发票，加我微信”，其中“发票”是常见词；短信 B 是“买卖发曝，加我微信”，“发曝”是新词，此时“发曝”本身词频比较低，但是 embedding 刻画了该词经常出现的上下文，因此“发曝”和“发票”在 embedding 是比较相似的。

2) 根据所述预设垃圾短信词汇的词向量和所述候选的垃圾短信新词的词向量，确定
5 所述预设垃圾词汇与所述候选的垃圾短信新词之间的语义相似度。

获取到所述预设垃圾短信词汇的词向量和所述候选的垃圾短信新词的词向量之后，就可以通过计算两个词向量之间的余弦距离等方式，确定所述预设垃圾词汇与所述候选的垃圾短信新词之间的语义相似度。

步骤 S305: 根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相
10 似度。

在获得每一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度后，就可以根据这些语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度。

所述垃圾短信词汇相似度，包括一个词与已有垃圾短信词汇之间的语义相似度。
15 在一个示例中，步骤 S305 采用如下方式实现：将每一个预设垃圾短信词汇分别与
所述候选的垃圾短信新词之间的语义相似度中，最大的语义相似度作为所述垃圾短信词
汇相似度。采用这种处理方式，可以将所述候选的垃圾短信新词的垃圾短信词汇相似度
设定为该词和语义最相近的垃圾短信词汇之间的语义相似度。

步骤 S307: 根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述
20 垃圾短信新词。

在一个示例中，可将所述垃圾短信词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词作为垃圾短信新词。所述第一相似度阈值和所述第二相似度阈值，可根据业务需求设定。

在另一个示例中，步骤 S307 也可以采用如下子步骤：1) 获取并展示所述垃圾短信
25 词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词；2)
接收用户输入的针对目标新词的确定指令；3) 将所述目标新词为所述垃圾短信新词。

具体实施时，可通过一个分段概率密度函数，根据所述垃圾短信词汇相似度得到所述候选的垃圾短信新词的采样概率，根据采样概率确定所述垃圾短信新词。在本实施例中，采用如下分段概率密度函数：

$$p(x) = \begin{cases} 1.2 & (0.7 < x \leq 1) \\ 0.7 & (0.3 < x \leq 0.7) \\ 1.2 & (0 \leq x \leq 0.3) \end{cases}$$

其中， x 表示所述垃圾短信词汇相似度， $p(x)$ 表示采样概率。由该函数可见，所述第一相似度阈值可设置为 0.7，所述第二相似度阈值可设置为 0.3。

本申请实施例提供的方法，通过对候选词进行词嵌入的处理，确定候选词的词向量，并根据已有垃圾短信词汇的词向量和候选词的词向量，确定已有垃圾词汇与候选词之间的语义相似度，进而确定候选词的垃圾短信词汇相似度，并通过分段概率密度函数确定候选词的采样概率，将采用概率大的候选词展示给审核人员审核；这种处理方式，使得对候选词进行了语义扩展，新词识别不仅仅从频率角度出发，在考虑到候选词的语义后，一方面不会忽略稀有的新词，由此可突出那些与已有垃圾词汇语义比较相近的候选词，另一方面可以筛选出与已有垃圾词汇语义完全不相近的候选词，由此可找到一些新的垃圾短信模式，这种短信的上下文可能并未出现过，但是既然该词通过了短信类别倾向性的审核和成词概率的审核，那么对现有垃圾词汇可能是个很好的补充；因此，可以有效提升垃圾短信新词识别的准确度。

例如，“买卖发票，加我微信”，其中发票是常见词，某条短信是“买卖发票，加我微信”，发票是新词，此时“发票”本身词频比较低，但是词向量(embedding)刻画了该词经常出现的上下文，因此“发票”和“发票”在词向量是比较相似的，“发票”作为与已有垃圾词汇语义比较相近的候选词被筛选出来，供审核人员审核。

请参看图 4，其为本申请的垃圾短信新词识别方法的实施例的示意图。由图 4 可见，本实施例首先通过步骤 S103 确定候选词集合，该集合包括多个候选词；然后，通过步骤 S105，根据标签（即所述短信类别信息），确定候选词的诸如交叉熵、优势率和互信息等短信类别倾向性相关指标，以及确定 IDF 等文档稀有性相关指标，并通过步骤 S107 确定候选词的垃圾短信新词得分，再根据该得分对候选词集合中的候选词进行第一层筛选，将得分低于或等于得分阈值的候选词从该集合中去除；接下来，再对该集合中保留下来的候选词进行第二层筛选，即根据候选词的内部凝聚度和外部自由度，获取候选词的成词概率，根据成词概率进行第二层筛选，将成词概率低于或等于成词概率阈值的候选词从该集合中去除；接下来，再对该集合中保留下来的候选词进行第三层筛选，即通过词嵌入方式，计算候选词与已有垃圾短信词汇的余弦相似度，并通过分段概率密度函数获取候选词的采样概率，根据采样概率最终确定出展示给人工审核的垃圾短信新

词。采用这种处理方式，通过基于三层筛选模式的无监督新词发现技术来挖掘短信中的垃圾词汇，能够有效提升垃圾短信新词识别的准确度和召回率，从而解决人工全量审核垃圾词汇效率低下的痛点，该方案可为审核人员及时提供线上新出现的垃圾词汇，来进行风险管控；另外，挖掘出的垃圾词汇可以构建黑库，还可以提高垃圾文本的识别率。

- 5 从上述实施例可见，本申请实施例提供的短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词；
- 10 这种处理方式，使得根据短信类别倾向性相关指标召回的大多数新词均对垃圾短信具有代表意义，避免召回很多正常的词汇；因此，可以有效提升召准率。同时，这种处理方式，使得根据文档稀有性的相关指标可召回低频的垃圾短信新词；因此，可以有效提升召全率。综上所述，这种处理方式，使得可筛选出既在垃圾短信中具有一定的稀有性，且在类别上比较倾向于垃圾短信的新词；因此，可以有效提升召准率和召全率，从而提
- 15 升新词发现准确度。

在上述的实施例中，提供了一种短信垃圾新词识别方法，与之相对应的，本申请还提供一种短信垃圾新词识别装置。该装置是与上述方法的实施例相对应。

第二实施例

- 请参看图 5，其为本申请的短信垃圾新词识别装置的实施例的示意图。由于装置实
- 20 施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的短信垃圾新词识别装置实施例仅仅是示意性的。

本申请另外提供一种短信垃圾新词识别装置，包括：

短信集获取单元 501，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

- 25 候选词集确定单元 503，用于确定所述多个垃圾短信对应的候选词集合；

指标确定单元 505，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；

得分确定单元 507，用于根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；

- 30 新词确定单元 509，用于根据所述垃圾短信新词得分，从所述候选词集合中确定垃

圾短信新词。

可选的，所述候选词集确定单元 503 包括：

通过分词算法，获取所述垃圾短信包括的词，作为短信词；

若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候

5 选词。

可选的，所述得分确定单元 507，具体用于将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述垃圾短信新词得分。

可选的，所述新词确定单元 509 包括：

候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的新词，作为候

10 选的垃圾短信新词；

成词概率确定子单元，用于确定所述候选的垃圾短信新词的成词概率；

第一新词确定子单元，用于根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

可选的，所述新词确定单元 509 包括：

15 候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

第一相似度确定子单元，用于确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度；

20 第二相似度确定子单元，用于根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

第二新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

第三实施例

25 请参考图 6，其为本申请的电子设备实施例的示意图。由于设备实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的设备实施例仅仅是示意性的。

本实施例的一种电子设备，该电子设备包括：处理器 601 和存储器 602；所述存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多
30 个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信

的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

5 第四实施例

在上述的实施例中，提供了一种短信垃圾新词识别方法，与之相对应的，本申请还提供一种短信垃圾新词识别方法。该方法是与上述方法具有相同的技术构思。

请参看图 7，其为本申请的短信垃圾新词识别方法的实施例的流程图。由于该方法实施例与上述方法实施例相对应，所以描述得比较简单，相关之处参见上述方法实施例的部分说明即可。

本申请另外提供一种短信垃圾新词识别方法，包括：

步骤 S701：获取短信集合。

步骤 S703：确定多个垃圾短信对应的候选词集合。

步骤 S705：根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分。

在一个示例中，所述短信类别倾向性得分采用如下方式计算：将短信类别倾向性相关指标的加权平均值作为所述短信类别倾向性得分。

步骤 S707：根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

在一个示例中，将所述短信类别倾向性得分大于得分阈值的所述候选词作为垃圾短信新词。所述得分阈值，可根据业务需求设置。在短信类别倾向性得分为垃圾短信倾向性得分的情况下，得分阈值设置的越小，通过的候选词越多，但可能将所有倾向于垃圾短信的词都选取出来，其中会存在很多非垃圾短信新词；得分阈值设置的越大，通过的候选词越少，有利于筛选出对垃圾文本具有代表性的高频的候选词，但可能忽略部分低频的候选词。

在又一个示例中，可对所述短信类别倾向性得分大于得分阈值的所述候选词进行如下后续的处理：1) 根据所述候选词的成词概率，对其进行第二层的筛选；2) 根据所述候选词的垃圾短信词汇相似度，通过一个分段概率密度函数，根据所述垃圾短信词汇相似度得到所述候选词的采样概率，根据采样概率确定所述垃圾短信新词。采用这种处理方式，使得确定出的垃圾短信新词还是更加具有实际意义的词，并且还可以突出那些与

已有垃圾词汇语义比较相近的候选词，以及，与已有垃圾词汇语义完全不相近的候选词。

从上述实施例可见，本申请实施例提供的短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；所述短信类别包括垃圾短信类或正常短信类；根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词；这种处理方式，使得根据短信类别倾向性相关指标召回的大多数新词均对垃圾短信具有代表意义，避免召回很多正常的词汇；因此，可以有效提升召回率。

第五实施例

在上述的实施例中，提供了一种短信垃圾新词识别方法，与之相对应的，本申请还提供一种短信垃圾新词识别装置。该装置是与上述方法的实施例相对应。

请参看图 8，其为本申请的短信垃圾新词识别装置的实施例的示意图。由于装置实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的短信垃圾新词识别装置实施例仅仅是示意性的。

本申请另外提供一种短信垃圾新词识别装置，包括：

短信集获取单元 801，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元 803，用于确定所述多个垃圾短信对应的候选词集合；

得分确定单元 805，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；

新词确定单元 807，用于根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

可选的，所述新词确定单元 807 包括：

候选词选取子单元，用于获取所述短信类别倾向性得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

相似度确定子单元，用于确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

第六实施例

请参考图 9，其为本申请的电子设备实施例的示意图。由于设备实施例基本相似于

方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的设备实施例仅仅是示意性的。

本实施例的一种电子设备，该电子设备包括：处理器 901 和存储器 902；所述存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该
5 短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

第七实施例

10 在上述的实施例中，提供了一种短信垃圾新词识别方法，与之相对应的，本申请还提供一种短信垃圾新词识别方法。该方法是与上述方法具有相同的技术构思。

请参看图 10，其为本申请的短信垃圾新词识别方法的实施例的流程图。由于该方法实施例与上述方法实施例相对应，所以描述得比较简单，相关之处参见上述方法实施例的部分说明即可。

15 本申请另外提供一种短信垃圾新词识别方法，包括：

步骤 S1001：获取短信集合。

步骤 S1003：确定多个垃圾短信对应的候选词集合。

步骤 S1005：确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度。

20 步骤 S1007：根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度。

步骤 S1009：根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

在一个示例中，步骤 S1009 可采用如下方式实现：根据所述候选词的垃圾短信词汇相似度，通过一个分段概率密度函数，根据所述垃圾短信词汇相似度得到所述候选词的
25 采样概率，根据采样概率确定所述垃圾短信新词。采用这种处理方式，可以突出那些与已有垃圾词汇语义比较相近的候选词，以及，与已有垃圾词汇语义完全不相近的候选词。

从上述实施例可见，本申请实施例提供的短信垃圾新词识别方法，通过获取短信集合；确定所述多个垃圾短信对应的候选词集合；确定至少一个预设垃圾短信词汇分别与
30 所述候选词之间的语义相似度；根据所述语义相似度，确定所述候选新词的垃圾短信词

汇相似度；根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词；这种处理方式，使得通过词嵌入的方式对候选词进行了语义扩展，并根据候选词与垃圾短信词汇之间的语义相似度，确定所述垃圾短信新词，因此不会忽略一些稀有的新词；因此，可以有效提升垃圾短信新词的召回率。

5 第八实施例

在上述的实施例中，提供了一种短信垃圾新词识别方法，与之相对应的，本申请还提供一种短信垃圾新词识别装置。该装置是与上述方法的实施例相对应。

请参看图 11，其为本申请的短信垃圾新词识别装置的实施例的示意图。由于装置实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的短信垃圾新词识别装置实施例仅仅是示意性的。

本申请另外提供一种短信垃圾新词识别装置，包括：

短信集获取单元 1101，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元 1103，用于确定所述多个垃圾短信对应的候选词集合；

15 第一语义相似度确定单元 1105，用于确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；

第二语义相似度确定单元 1107，用于根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；

20 新词确定单元 1109，用于根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

第九实施例

请参考图 12，其为本申请的电子设备实施例的示意图。由于设备实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的设备实施例仅仅是示意性的。

25 本实施例的一种电子设备，该电子设备包括：处理器 1201 和存储器 1202；所述存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；根据所述语义相似度，确定
30 所述候选新词的垃圾短信词汇相似度；根据所述垃圾短信词汇相似度，从所述候选词集

合中确定垃圾短信新词。

本申请虽然以较佳实施例公开如上，但其并不是用来限定本申请，任何本领域技术人员在不脱离本申请的精神和范围内，都可以做出可能的变动和修改，因此本申请的保护范围应当以本申请权利要求所界定的范围为准。

- 5 在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

- 10 1、计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存（PRAM）、静态随机存取存储器（SRAM）、动态随机存取存储器（DRAM）、其他类型的随机存取存储器（RAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、快闪记忆体或其他内存技术、只读光盘只读存储器（CD-ROM）、数字多功能光盘（DVD）或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体（transitory media），如调制的数据信号和载波。
- 15

- 20 2、本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

权 利 要 求 书

1.一种垃圾短信新词识别方法，其特征在于，包括：

获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

确定所述多个垃圾短信对应的候选词集合；

- 5 根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，
获取所述候选词的文档稀有性相关指标；

根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；

根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

- 10 2.根据权利要求 1 所述的方法，其特征在于，所述确定所述多个垃圾短信对应的候选词集合，包括：

通过分词算法，获取所述垃圾短信包括的词，作为短信词；

若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候选词。

- 15 3.根据权利要求 2 所述的方法，其特征在于，所述候选词规则包括候选词的字数小于预设字数。

4.根据权利要求 1 所述的方法，其特征在于，

所述短信类别倾向性相关指标包括以下指标的至少一项：交叉熵，优势率，互信息；

所述文档稀有性相关指标包括：倒文档频率 IDF。

- 20 5.根据权利要求 1 所述的方法，其特征在于，所述根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分，包括：

将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述垃圾短信新词得分。

- 25 6.根据权利要求 1 所述的方法，其特征在于，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

将所述垃圾短信新词得分大于得分阈值的所述候选词作为垃圾短信新词。

7.根据权利要求 1 所述的方法，其特征在于，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取并展示所述垃圾短信新词得分大于得分阈值的新词；

- 30 接收针对目标新词的确定指令；

将所述目标新词为所述垃圾短信新词。

8.根据权利要求 1 所述的方法，其特征在于，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词；

5 确定所述候选的垃圾短信新词的成词概率；

根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

9.根据权利要求 8 所述的方法，其特征在于，所述根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

将所述成词概率大于成词概率阈值的候选的垃圾短信新词作为所述垃圾短信新词。

10 10.根据权利要求 8 所述的方法，其特征在于，所述根据所述成词概率，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

获取并展示所述成词概率大于成词概率阈值的候选的垃圾短信新词；

接收用户输入的针对目标新词的确定指令；

将所述目标新词为所述垃圾短信新词。

15 11.根据权利要求 8 所述的方法，其特征在于，所述确定所述候选的垃圾短信新词的成词概率，包括：

确定所述候选的垃圾短信新词的内部凝聚度；以及，确定所述候选的垃圾短信新词的外部自由度；

20 根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短信新词的成词概率。

12.根据权利要求 11 所述的方法，其特征在于，所述根据所述内部凝聚度和所述外部自由度，获取所述候选的垃圾短信新词的成词概率，包括：

将所述内部凝聚度和所述外部自由度的平均值作为所述成词概率。

25 13.根据权利要求 1 所述的方法，其特征在于，所述根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度；

30 根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

14.根据权利要求 13 所述的方法，其特征在于，所述根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

5 将所述垃圾短信词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词作为垃圾短信新词。

15.根据权利要求 13 所述的方法，其特征在于，所述根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词，包括：

10 获取并展示所述垃圾短信词汇相似度大于第一相似度阈值或小于第二相似度阈值的所述候选的垃圾短信新词；

接收针对目标候选新词的确定指令；

将所述目标候选新词为所述垃圾短信新词。

16.根据权利要求 13 所述的方法，其特征在于，所述根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度，包括：

15 将所述语义相似度的最大值作为所述垃圾短信词汇相似度。

17.根据权利要求 13 所述的方法，其特征在于，所述确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度，包括：

确定所述候选新词的词向量；

20 根据所述预设垃圾短信词汇的词向量和所述候选新词的词向量，确定所述预设垃圾词汇与所述候选新词之间的语义相似度。

18.一种垃圾短信新词识别方法，其特征在于，包括：

获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

确定所述多个垃圾短信对应的候选词集合；

根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；

25 根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

19.根据权利要求 18 所述的方法，其特征在于，所述根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词，包括：

获取所述短信类别倾向性得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

30 确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

20.一种垃圾短信新词识别方法，其特征在于，包括：

获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

5 确定所述多个垃圾短信对应的候选词集合；

确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；

根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；

根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

21.一种短信垃圾新词识别装置，其特征在于，包括：

10 短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

指标确定单元，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；

15 得分确定单元，用于根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；

新词确定单元，用于根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

22.根据权利要求 21 所述的装置，其特征在于，所述候选词集确定单元包括：

20 通过分词算法，获取所述垃圾短信包括的词，作为短信词；

若至少两个相邻短信词构成的组合词符合候选词规则，则将所述组合词作为所述候选词。

23.根据权利要求 21 所述的装置，其特征在于，

25 所述得分确定单元，具体用于将所述短信类别倾向性相关指标和所述文档稀有性相关指标的加权平均值作为所述垃圾短信新词得分。

24.根据权利要求 21 所述的装置，其特征在于，所述新词确定单元包括：

候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的新词，作为候选的垃圾短信新词；

成词概率确定子单元，用于确定所述候选的垃圾短信新词的成词概率；

30 第一新词确定子单元，用于根据所述成词概率，从所述候选的垃圾短信新词中确定

所述垃圾短信新词。

25.根据权利要求 21 所述的装置，其特征在于，所述新词确定单元包括：

候选词选取子单元，用于获取所述垃圾短信新词得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

5 第一相似度确定子单元，用于确定至少一个预设垃圾短信词汇分别与所述候选的垃圾短信新词之间的语义相似度；

第二相似度确定子单元，用于根据所述语义相似度，确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

10 第二新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

26.一种电子设备，其特征在于，包括：

处理器；以及

15 存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性相关指标；以及，获取所述候选词的文档稀有性相关指标；根据所述短信类别倾向性相关指标和所述文档稀有性相关指标，确定所述候选词的垃圾短信新词得分；根据所述垃圾短信新词得分，从所述候选词集合中确定垃圾短信新词。

20 27.一种短信垃圾新词识别装置，其特征在于，包括：

短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

25 得分确定单元，用于根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；

新词确定单元，用于根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

28.根据权利要求 27 所述的装置，其特征在于，所述新词确定单元包括：

30 候选词选取子单元，用于获取所述短信类别倾向性得分大于得分阈值的所述候选词，作为候选的垃圾短信新词；

相似度确定子单元，用于确定所述候选的垃圾短信新词的垃圾短信词汇相似度；

新词确定子单元，用于根据所述垃圾短信词汇相似度，从所述候选的垃圾短信新词中确定所述垃圾短信新词。

29.一种电子设备，其特征在于，包括：

5 处理器；以及

存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；根据所述短信的短信类别信息，确定所述候选词的短信类别倾向性得分；根据所述短信类别倾向性得分，从所述候选词集合中确定垃圾短信新词。

30.一种短信垃圾新词识别装置，其特征在于，包括：

短信集获取单元，用于获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；

候选词集确定单元，用于确定所述多个垃圾短信对应的候选词集合；

15 第一语义相似度确定单元，用于确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；

第二语义相似度确定单元，用于根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；

20 新词确定单元，用于根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

31.一种电子设备，其特征在于，包括：

处理器；以及

25 存储器，用于存储实现短信垃圾新词识别方法的程序，该设备通电并通过所述处理器运行该短信垃圾新词识别方法的程序后，执行下述步骤：获取短信集合；所述短信集合包括多个垃圾短信和多个正常短信；确定所述多个垃圾短信对应的候选词集合；确定至少一个预设垃圾短信词汇分别与所述候选词之间的语义相似度；根据所述语义相似度，确定所述候选新词的垃圾短信词汇相似度；根据所述垃圾短信词汇相似度，从所述候选词集合中确定垃圾短信新词。

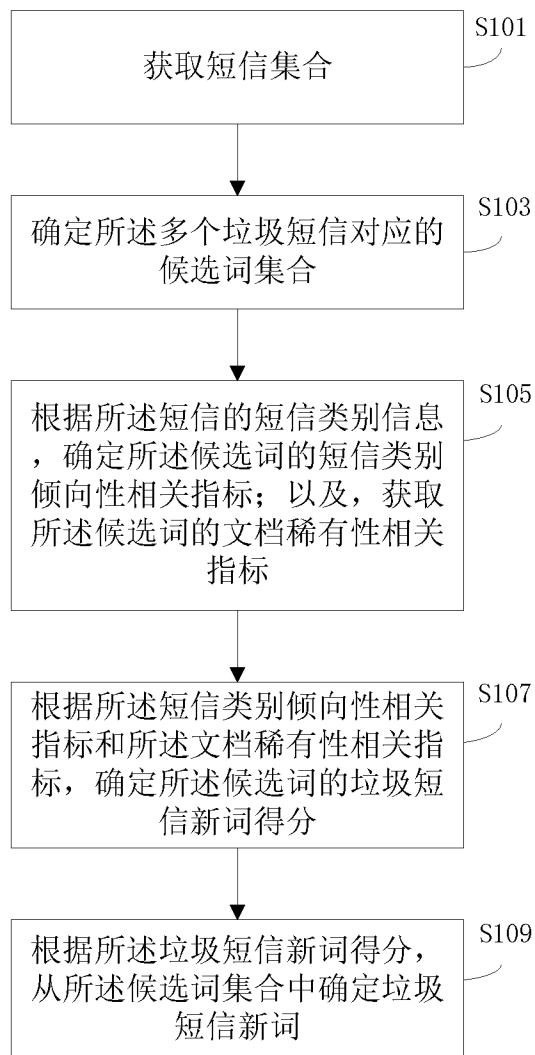


图 1

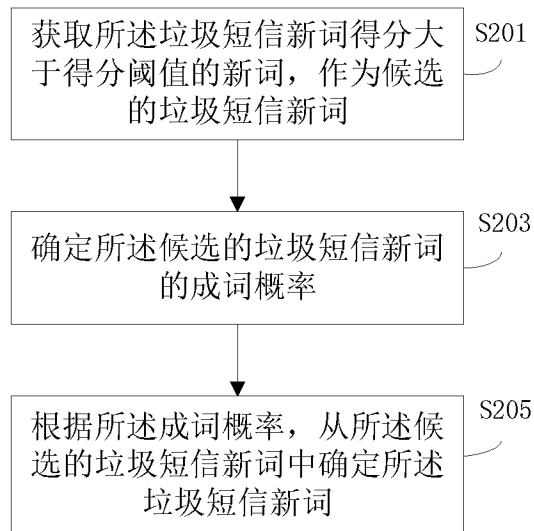


图 2

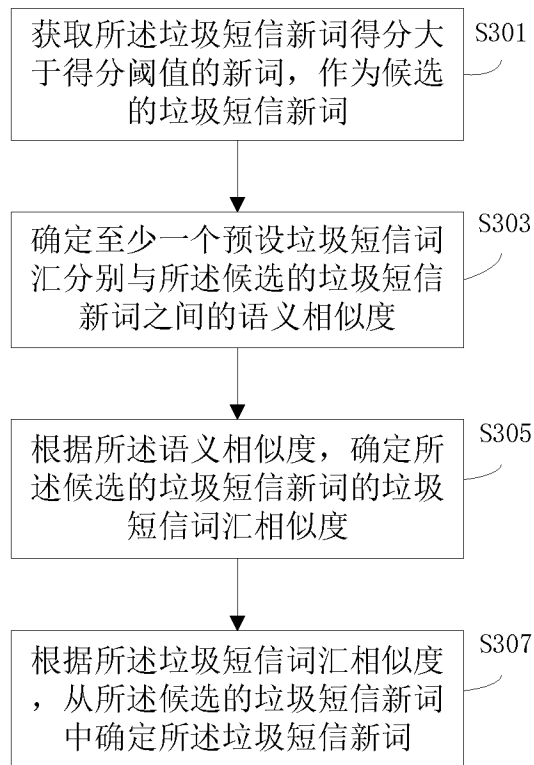


图 3

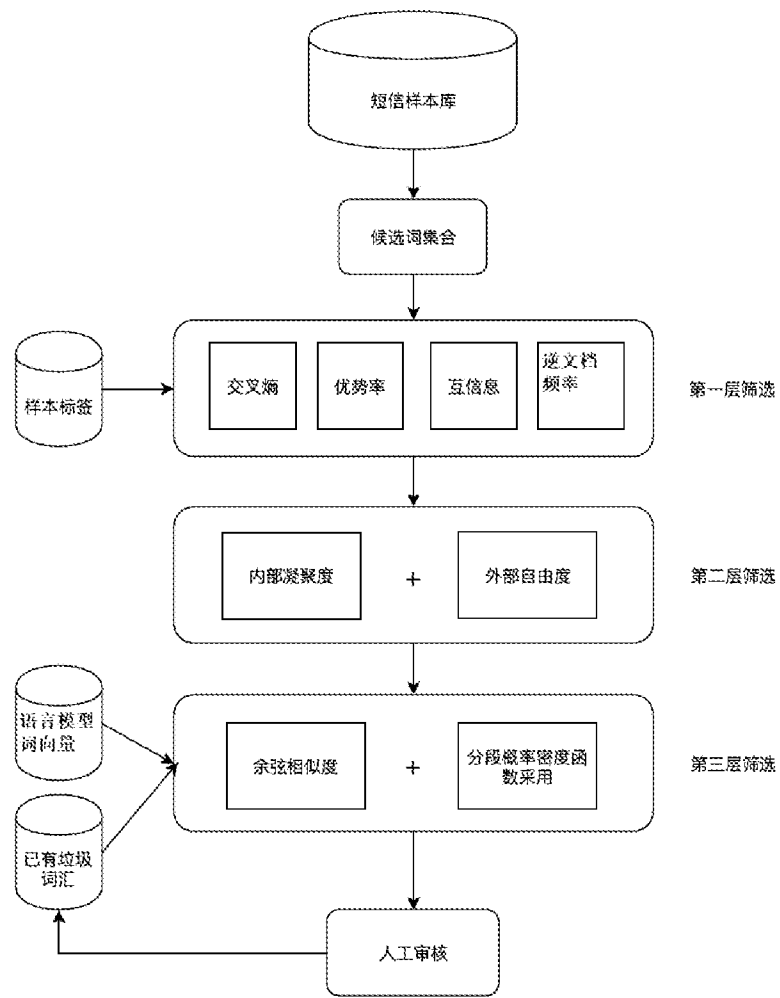


图 4

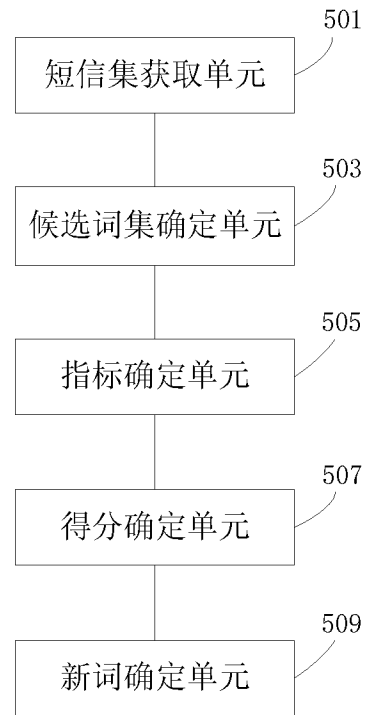


图 5

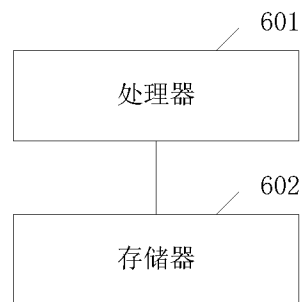


图 6

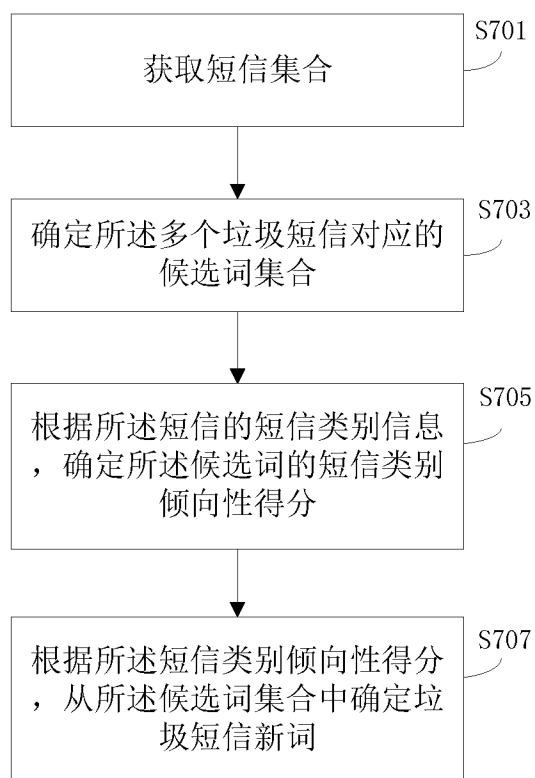


图 7

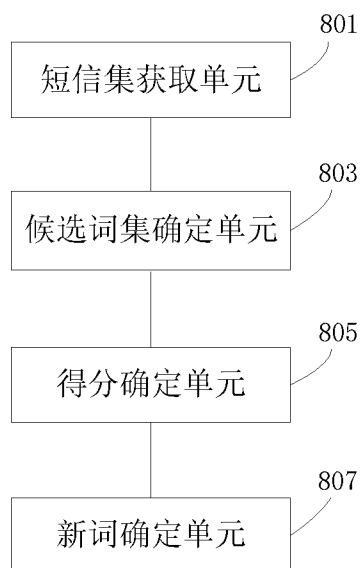


图 8

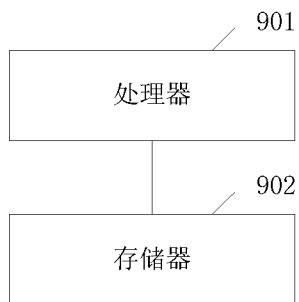


图 9

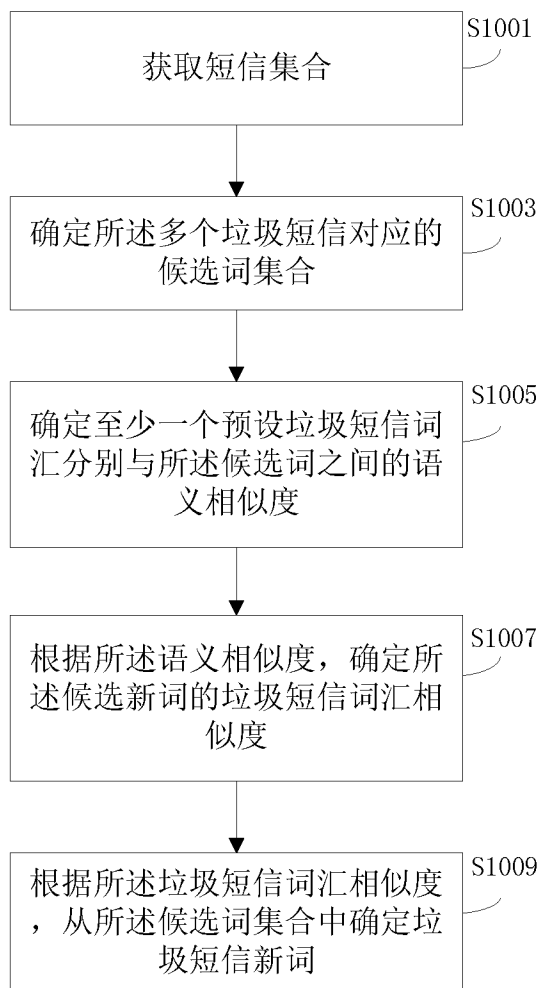


图 10

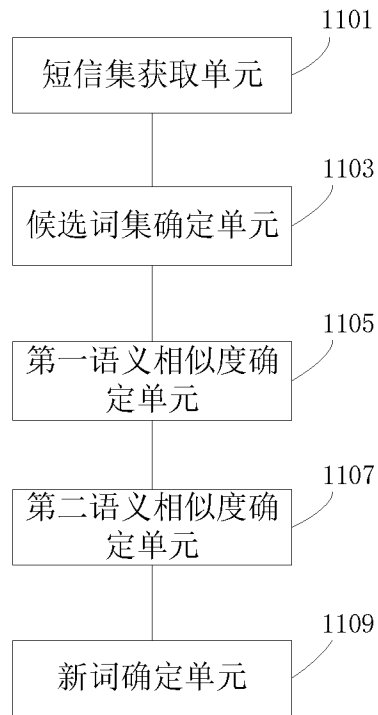


图 11

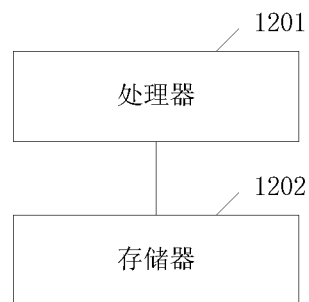


图 12

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/105123

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(2006.01)i; G06F 16/332(2019.01)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; H04M; H04W; H04L

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, ISI: 垃圾, 短信, 识别, IDF, 语义, 相似度, 匹配, 词向量, spam, short message, recognition, semantic, similarity, match, word vector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107193804 A (HOHAI UNIVERSITY) 22 September 2017 (2017-09-22) description, paragraphs [0058]-[0116]	1-19, 21-29
A	CN 107193804 A (HOHAI UNIVERSITY) 22 September 2017 (2017-09-22) description, paragraphs [0058]-[0116]	20, 30, 31
X	CN 105516499 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.) 20 April 2016 (2016-04-20) description, paragraphs [0049]-[0069]	20, 30, 31
A	CN 105516499 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.) 20 April 2016 (2016-04-20) description, paragraphs [0049]-[0069]	1-19, 21-29
X	CN 106878347 A (BEIJING QIHOO TECHNOLOGY CO., LTD.) 20 June 2017 (2017-06-20) description, paragraphs [0074]-[0076]	20, 30, 31
A	CN 106878347 A (BEIJING QIHOO TECHNOLOGY CO., LTD.) 20 June 2017 (2017-06-20) description, paragraphs [0074]-[0076]	1-19, 21-29



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

“D” document cited by the applicant in the international application

“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

06 November 2019

Date of mailing of the international search report

29 November 2019

Name and mailing address of the ISA/CN

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

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/105123

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2017084267 A1 (LE HOLDINGS (BEIJING) CO., LTD. et al.) 26 May 2017 (2017-05-26) entire document	1-31

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/105123

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107193804	A	22 September 2017	None			
CN	105516499	A	20 April 2016	None			
CN	106878347	A	20 June 2017	None			
WO	2017084267	A1	26 May 2017	CN	105893410	A	24 August 2016
				US	2017139899	A1	18 May 2017

国际检索报告

国际申请号

PCT/CN2019/105123

A. 主题的分类 G06F 17/27(2006.01)i; G06F 16/332(2019.01)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; H04M; H04W; H04L 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, EPDOC, CNKI, ISI: 垃圾, 短信, 识别, IDF, 语义, 相似度, 匹配, 词向量, spam, short message, recognition, semantic, similarity, match, word vector																										
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段</td> <td>1-19, 21-29</td> </tr> <tr> <td>A</td> <td>CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段</td> <td>20, 30-31</td> </tr> <tr> <td>X</td> <td>CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段</td> <td>20, 30-31</td> </tr> <tr> <td>A</td> <td>CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段</td> <td>1-19, 21-29</td> </tr> <tr> <td>X</td> <td>CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段</td> <td>20, 30-31</td> </tr> <tr> <td>A</td> <td>CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段</td> <td>1-19, 21-29</td> </tr> <tr> <td>A</td> <td>WO 2017084267 A1 (乐视控股北京有限公司等) 2017年 5月 26日 (2017 - 05 - 26) 全文</td> <td>1-31</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段	1-19, 21-29	A	CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段	20, 30-31	X	CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段	20, 30-31	A	CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段	1-19, 21-29	X	CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段	20, 30-31	A	CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段	1-19, 21-29	A	WO 2017084267 A1 (乐视控股北京有限公司等) 2017年 5月 26日 (2017 - 05 - 26) 全文	1-31
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																								
X	CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段	1-19, 21-29																								
A	CN 107193804 A (河海大学) 2017年 9月 22日 (2017 - 09 - 22) 说明书第58-116段	20, 30-31																								
X	CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段	20, 30-31																								
A	CN 105516499 A (北京奇虎科技有限公司等) 2016年 4月 20日 (2016 - 04 - 20) 说明书第49-69段	1-19, 21-29																								
X	CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段	20, 30-31																								
A	CN 106878347 A (北京奇虎科技有限公司) 2017年 6月 20日 (2017 - 06 - 20) 说明书第74-76段	1-19, 21-29																								
A	WO 2017084267 A1 (乐视控股北京有限公司等) 2017年 5月 26日 (2017 - 05 - 26) 全文	1-31																								
☐ 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 2019年 11月 6日		国际检索报告邮寄日期 2019年 11月 29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		授权官员 余永佼 电话号码 86-(010)-53961745																								

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/105123

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	107193804	A	2017年 9月 22日	无	
CN	105516499	A	2016年 4月 20日	无	
CN	106878347	A	2017年 6月 20日	无	
WO	2017084267	A1	2017年 5月 26日	CN 105893410 A	2016年 8月 24日
				US 2017139899 A1	2017年 5月 18日