(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
2017 年 1 月 19 日 (19.01.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/008440 A1

(51) 国际专利分类号:
H04M 1/663 (2006.01)

(21) 国际申请号: PCT/CN2015/097746

(22) 国际申请日: 2015 年 12 月 17 日 (17.12.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510408711.7 2015 年 7 月 13 日 (13.07.2015) CN

(71) 申请人: **小米科技有限责任公司 (XIAOMI INC.)** [CN/CN]; 中国北京市海淀区清河中街 68 号华润五彩城购物中心二期 13 层, Beijing 100085 (CN)。

(72) 发明人: **秦秋平 (QIN, Qiuping)**; 中国北京市海淀区清河中街 68 号华润五彩城购物中心二期 13 层, Beijing 100085 (CN)。 **陈志军 (CHEN, ZhiJun)**; 中国北京市海淀区清河中街 68 号华润五彩城购物中心二期 13 层, Beijing 100085 (CN)。 **龙飞 (LONG, Fei)**; 中国北京市海淀区清河中街 68 号华润五彩城购物中心二期 13 层, Beijing 100085 (CN)。

(74) 代理人: **北京派特恩知识产权代理有限公司 (CHINA PAT INTELLECTUAL PROPERTY OFFICE)**; 中国北京市海淀区海淀南路 21 号中关村知识产权大厦 B 座 2 层, Beijing 100080 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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))。

(54) **Title**: METHOD, APPARATUS AND SYSTEM FOR DETERMINING DISTURBING TELEPHONE NUMBER

(54) **发明名称** : 骚扰电话号码确定方法、装置和系统

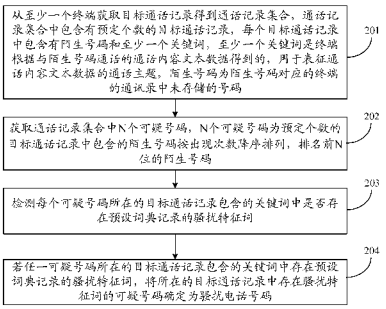


图 2

201 Acquiring target call records from at least one terminal to obtain a call record set, wherein the call record set contains a predetermined number of target call records, each of the target call records contains an unfamiliar number and at least one key word, the at least one key word is obtained by the terminal according to call content text data of calls made with the unfamiliar number for characterizing a call theme of the call content text data, and the unfamiliar number is a number which is not stored in an address book of the terminal corresponding to the unfamiliar number

202 Acquiring N suspicious numbers from the call record set, wherein the N suspicious numbers are top N unfamiliar numbers in the unfamiliar numbers, contained in the predetermined number of target call records, that are ranked in a descending order according to the frequency of occurrence

203 Detecting whether a disturbance feature word recorded in a pre-set dictionary exists in key words contained in a target call record where each of the suspicious numbers is located

204 If the disturbance feature word recorded in the pre-set dictionary exists in the key words contained in the target call record of any suspicious number, determining the suspicious number, if the target call record where the suspicious number is located has the disturbance feature word

(57) **Abstract**: A method, apparatus and system for determining a disturbing telephone number, which falls within the technical field of security. The method comprises: acquiring target call records from a terminal to obtain a call record set, wherein the call record set contains a predetermined number of target call records; acquiring N suspicious numbers from the call record set; detecting whether a disturbance feature word recorded in a pre-set dictionary exists in key words contained in a target call record where each of the suspicious numbers is located; and determining a suspicious number, if the target call record where the suspicious number is located has the disturbance feature word, as a disturbing telephone number.

(57) **摘要**: 一种骚扰电话号码确定方法、装置和系统，属于安全技术领域。所述方法包括：从终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录；获取通话记录集合中 N 个可疑号码；检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词；将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

WO 2017/008440 A1

骚扰电话号码确定方法、装置和系统

本申请基于申请号为201510408711.7、申请日为2015年07月13日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

本公开实施例涉及安全技术领域，特别涉及一种骚扰电话号码确定方法、装置和系统。

背景技术

骚扰电话号码通常可以包括推销保险的电话号码、房产中介的电话号码和进行诈骗的电话号码，这些骚扰电话号码长期干扰着人们的正常生活。

相关技术中有一种骚扰电话号码确定方法，该方法首先由大量用户通过终端来确认接听到的号码是否为骚扰电话号码，并将确认的骚扰电话号码上传至服务器，服务器将大量用户确认的骚扰电话号码进行汇总生成骚扰电话号码表。

发明内容

为了解决相关技术问题，本公开实施例提供了一种骚扰电话号码确定方法、装置和系统。所述技术方案如下:

第一方面，提供一种骚扰电话号码确定方法，该方法包括:

从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，陌生号码为

陌生号码对应的终端的通讯录中未存储的号码;

获取通话记录集合中 N 个可疑号码，N 个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码;

检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词;

若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

本公开实施例的第一方面，能够方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，提高了骚扰电话号码的获取效率。

结合第一方面，在第一种可实现方式中，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码，包括:

获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码;

根据预设的骚扰特征词与骚扰类型的对应关系，确定第一通话记录包含的骚扰特征词的骚扰类型，骚扰类型表征第一可疑电话号码的骚扰目的，第一通话记录为目标通话记录集合中的任一通话记录;

将第一通话记录包含的骚扰特征词的骚扰类型确定为第一通话记录的骚扰类型;

统计目标通话记录集合中每种骚扰类型对应的通话记录的数量;

根据每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例;

若目标通话记录集合中存在比例大于比例阈值的骚扰类型，则确定第一可疑电话号码为骚扰电话号码。

本公开实施例第一方面的第一种可实现方式，提高了确定的骚扰电话号码的准确性。

结合第一种可实现方式，在第二种可实现方式中，该方法还包括:

将目标通话记录集合中比例最大的骚扰类型作为骚扰电话号码的骚扰类型。

本公开实施例第一方面的第二种可实现方式，确定了骚扰电话号码的骚扰类型。

结合第一方面，或第一种可实现方式，或第二种可实现方式，在第三种可实现方式中，每个目标通话记录中包含有两个电话号码，两个电话号码中包含有陌生号码，获取通话记录集合中 N 个可疑号码，包括:

对通话记录集合进行去重复处理，去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，同类通话记录为包含相同的两个电话号码的目标通话记录;

获取处理后的通话记录集合中 N 个可疑号码。

本公开实施例第一方面的第三种可实现方式，增加了最终确定的骚扰电话号码的准确性。

第二方面，提供一种骚扰电话号码确定方法，该方法包括:

获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码;

将通话语音数据转化为通话内容文本数据;

获取通话内容文本数据的至少一个关键词，至少一个关键词为表征通话内容文本数据的通话主题的词;

根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词;

将目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌

生号码和至少一个关键词确定陌生号码是否为骚扰电话号码。

本公开实施例的第二方面，使服务器能够方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，提高了骚扰电话号码的获取效率。

结合第二方面，在第一种可实现方式中，获取通话内容文本数据的至少一个关键词，包括:

对内容文本数据进行分词处理得到包含有至少一个词的分词结果;

获取分词结果中每个词的关键词因子，关键词因子正相关于每个词的主题相关度，每个词的主题相关度用于指示每个词与通话内容文本数据的通话主题相关的程度;

按照关键词因子的大小将分词结果中的词降序排列;

将排名前 M 位的词作为内容文本数据的关键词。

本公开实施例第二方面的第一种可实现方式，提高了获取的关键词与主题的相关程度。

结合第一种可实现方式，在第二种可实现方式中，获取分词结果中每个词的关键词因子，包括:

根据预设类型词典获取分词结果中每个词的类型，类型用于表征每个词的特征;

根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值;

获取分词结果中每个词的词频;

将每个词的权值与词频的乘积作为每个词的关键词因子。

本公开实施例第二种可实现方式，获取了可信度较高的关键词因子，提高了关键因子的可靠性。

结合第二种可实现方式，在第三种可实现方式中，获取分词结果中每

个词的关键词因子，包括:

去除分词结果中类型为语气词、人称代词和修饰词的词得到处理后的分词结果;

获取处理后的分词结果中每个词的关键词因子。

本公开实施例第二方面第三种可实现方式，减少了终端的计算量，提高了关键词因子的可信度。

第三方面，提供一种骚扰电话号码确定装置，该装置包括:

集合获取模块，被配置为从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，陌生号码为陌生号码对应的终端的通讯录中未存储的号码;

号码获取模块，被配置为获取通话记录集合中 N 个可疑号码，N 个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码;

检测模块，被配置为检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词;

确定模块，被配置为在任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词时，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

结合第三方面，在第一种可实现方式中，确定模块，包括:

集合获取子模块，被配置为获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码;

特征词确定子模块，被配置为根据预设的骚扰特征词与骚扰类型的对

应关系，确定第一通话记录包含的骚扰特征词的骚扰类型，骚扰类型表征第一可疑电话号码的骚扰目的，第一通话记录为目标通话记录集合中的任一通话记录；

记录确定子模块，被配置为将第一通话记录包含的骚扰特征词的骚扰类型确定为第一通话记录的骚扰类型；

统计子模块，被配置为统计目标通话记录集合中每种骚扰类型对应的通话记录的数量；

比例确定子模块，被配置为根据每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例；

号码确定子模块，被配置为在目标通话记录集合中存在比例大于比例阈值的骚扰类型时，确定所述第一可疑电话号码为骚扰电话号码。

结合第一种可实现方式，在第二种可实现方式中，该装置还包括：

类型确定模块，被配置为将目标通话记录集合中比例最大的骚扰类型作为骚扰电话号码的骚扰类型。

结合第三方面，或第一种可实现方式，或第二种可实现方式，在第三种可实现方式中，每个目标通话记录中包含有两个电话号码，两个电话号码中包含有陌生号码，

号码获取模块，被配置为对通话记录集合进行去重复处理，去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，同类通话记录为包含相同的两个电话号码的目标通话记录；获取处理后的通话记录集合中 N 个可疑号码。

第四方面，提供一种骚扰电话号码确定装置，该装置包括：

终端获取模块，被配置为获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码；

转化模块，被配置为将通话语音数据转化为通话内容文本数据；

关键词获取模块，被配置为获取通话内容文本数据的至少一个关键词，至少一个关键词为表征通话内容文本数据的通话主题的词；

目标生成模块，被配置为根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词；

记录提供模块，被配置为将目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌生号码和至少一个关键词确定陌生号码是否为骚扰电话号码。

结合第四方面，在第一种可实现方式中，关键词获取模块，包括：

分词子模块，被配置为对内容文本数据进行分词处理得到包含有至少一个词的分词结果；

因子获取子模块，被配置为获取分词结果中每个词的关键词因子，关键词因子正相关于每个词的主题相关度，每个词的主题相关度用于指示每个词与通话内容文本数据的通话主题相关的程度；

因子排序子模块，被配置为按照关键词因子的大小将分词结果中的词降序排列；

关键词确定子模块，被配置为将排名前 M 位的词作为内容文本数据的关键词。

结合第一种可实现方式，在第二种可实现方式中，因子获取子模块，被配置为根据预设类型词典获取分词结果中每个词的类型，类型用于表征每个词的特征；根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值；获取分词结果中每个词的词频；将每个词的权值与词频的乘积作为每个词的关键词因子。

结合第二种可实现方式，在第三种可实现方式中，因子获取子模块，被配置为去除分词结果中类型为语气词、人称代词和修饰词的词得到处理

后的分词结果；获取处理后的分词结果中每个词的关键词因子。

第五方面，提供一种骚扰电话号码确定装置，该装置包括：

处理器；

配置为存储处理器的可执行指令的存储器；

其中，处理器被配置为：

从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，陌生号码为陌生号码对应的终端的通讯录中未存储的号码；

获取通话记录集合中 N 个可疑号码，N 个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码；

检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词；

若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

第六方面，提供一种骚扰电话号码确定装置，该装置包括：

处理器；

配置为存储处理器的可执行指令的存储器；

其中，处理器被配置为：

获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码；

将通话语音数据转化为通话内容文本数据；

获取通话内容文本数据的至少一个关键词，至少一个关键词为表征通

话内容文本数据的通话主题的词;

根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词;

将目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌生号码和至少一个关键词确定陌生号码是否为骚扰电话号码。

第七方面，提供一种骚扰电话号码确定系统，该系统包括：服务器和终端，

服务器包括第三方面提供的骚扰电话号码确定装置;

终端包括第四方面提供的骚扰电话号码确定装置。

第八方面，提供一种骚扰电话号码确定系统，该系统包括：服务器和终端，

服务器包括第五方面提供的骚扰电话号码确定装置;

终端包括第六方面提供的骚扰电话号码确定装置。

本公开实施例的提供的技术方案可以包括以下有益效果:

本公开实施例提供的骚扰电话号码确定方法、装置和系统，通过将通话记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性的，并不能限制本公开。

附图说明

为了更清楚地说明本公开的实施例，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本公开的

一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 是本公开各个实施例提供的骚扰电话号码确定方法所涉及的实施环境的示意图；

图 2 是根据一示例性实施例示出的一种骚扰电话号码确定方法的流程图；

图 3 是根据一示例性实施例示出的另一种骚扰电话号码确定方法的流程图；

图 4-1 是根据一示例性实施例示出的另一种骚扰电话号码确定方法的流程图；

图 4-2 是图 4-1 示出的骚扰电话号码确定方法中获取关键词的流程图；

图 4-3 是图 4-1 示出的骚扰电话号码确定方法中获取关键词因子的流程图；

图 4-4 是图 4-1 示出的骚扰电话号码确定方法中确定骚扰电话号码的流程图；

图 5-1 是根据一示例性实施例示出的一种骚扰电话号码确定装置的框图；

图 5-2 是图 5-1 示出的骚扰电话号码确定装置中确定模块的框图；

图 5-3 是本公开实施例提供的另一种骚扰电话号码确定装置的框图；

图 6-1 是根据一示例性实施例示出的另一种骚扰电话号码确定装置的框图；

图 6-2 是图 6-1 示出的骚扰电话号码确定装置中的关键词获取模块的框图；

图 7 是根据一示例性实施例示出的一种骚扰电话号码确定装置的框图；

图 8 是根据一示例性实施例示出的一种骚扰电话号码确定装置的框图；

图 9 是根据一示例性实施例示出的一种骚扰电话号码确定系统的框图；

图 10 是根据一示例性实施例示出的一种骚扰电话号码确定系统的框图。

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开实施例的原理。

具体实施方式

为了使本公开实施例的目的、技术方案和优点更加清楚，下面将结合附图对本公开实施例作进一步地详细描述，显然，所描述的实施例仅仅是本公开一部份实施例，而不是全部的实施例。基于本公开中的实施例，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，都属于本公开保护的范围。

图 1 是本公开各个实施例提供的骚扰电话号码确定方法所涉及的实施环境的示意图，该实施环境可以包括：终端 11 和服务器 12。

终端 11 可以是手机、平板电脑、膝上型便携计算机和台式计算机等等。

服务器 12 可以是一台服务器，或者由若干台服务器组成的服务器集群，或者是一个云计算服务中心。

终端 11 和服务器 12 之间可以建立有有线连接或无线连接。此外，本实施环境不对终端 11 的数量作出限制。

图 2 是根据一示例性实施例示出的一种骚扰电话号码确定方法的流程图，本实施例以该骚扰电话号码确定方法应用于图 1 所示实施环境中的服务器 12 来举例说明，该骚扰电话号码确定方法可以包括如下几个步骤：

在步骤 201 中，从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，

陌生号码为陌生号码对应的终端的通讯录中未存储的号码。

在步骤 202 中，获取通话记录集合中 N 个可疑号码，N 个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码。

在步骤 203 中，检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词。

在步骤 204 中，若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

综上所述，本公开实施例提供的骚扰电话号码确定方法，通过将通话记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

图 3 是根据一示例性实施例示出的另一种骚扰电话号码确定方法的流程图，本实施例以该骚扰电话号码确定方法应用于图 1 所示实施环境中的终端 11 来举例说明，该骚扰电话号码确定方法可以包括如下几个步骤：

在步骤 301 中，获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码。

在步骤 302 中，将通话语音数据转化为通话内容文本数据。

在步骤 303 中，获取通话内容文本数据的至少一个关键词，至少一个关键词为表征通话内容文本数据的通话主题的词。

在步骤 304 中，根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词。

在步骤 305 中，将目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌生号码和至少一个关键词确定陌生号码是否为骚扰电话号码。

综上所述，本公开实施例提供的骚扰电话号码确定方法，通过获取目标通话记录，该目标通话记录中包含有陌生号码和至少一个关键词，并将该目标通话记录提供给服务器，服务器用于根据该目标通话记录中的陌生号码和至少一个关键词来确定该陌生号码是否为骚扰电话号码，达到了提高骚扰电话号码获取效率的效果。

图 4-1 是根据一示例性实施例示出的另一种骚扰电话号码确定方法的流程图，本实施例以该骚扰电话号码确定方法应用于图 1 所示实施环境中来举例说明，该骚扰电话号码确定方法可以包括如下几个步骤：

在步骤 401 中，终端获取与陌生号码通话的通话语音数据和通话记录，该通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码。

每当终端开始通话时，可以根据通讯录来检测通话对端的电话号码是否为陌生号码，在通话对端的电话号码为陌生号码时，终端可以同步获取通话语音数据，并生成一个通话记录，该通话记录中可以记录有通话对端的陌生号码和该终端自身的电话号码，或者该通话记录中可以仅记录通话对端的陌生号码。

在步骤 402 中，终端将通话语音数据转化为通话内容文本数据。

终端在获取通话语音数据后，可以将通话语音数据转化为文本格式的通话内容文本数据并存储，该文本格式可以为 txt 或 ASCII（American Standard Code for Information Interchange，美国标准信息交换代码）等。

示例性的，可以通过自动语音识别技术（Automatic Speech Recognition）将获取的通话语音数据转化为通话内容文本数据，该自动语音识别技术可以参考现有技术，本公开实施例不再赘述。

可选的，本步骤可以和步骤 401 同时执行，即终端在与陌生号码通话时，获取与陌生号码的通话语音数据，并实时的将语音数据转化为通话内容文本数据，这样在通话结束时终端就可以获取完整的通话内容文本数据。

在步骤 403 中，终端获取通话内容文本数据的至少一个关键词，该至少一个关键词为表征该通话内容文本数据的通话主题的词。

在终端获取了通话内容文本数据之后，终端可以对该通话内容文本数据进行分析，并获取其中可以表征通话主题的词。示例性的，获取的通话内容文本数据是讨论某一个明星的一段话，则终端获取的关键词可能是该明星的姓名。

如图 4-2 所示，本步骤分可以为下面 4 个子步骤:

在子步骤 4031 中，终端对内容文本数据进行分词处理得到包含有至少一个词的分词结果。

终端获取的通话内容文本数据通常包含多个完整的语句，终端想要从中获取关键词，首先需要对该通话内容文本数据进行分词处理，将其中包含的完整的语句都拆分成一个个词。

其中，进行分词处理的方法可以包括基于字典、词库匹配的分词方法、基于词频度统计的分词方法和基于知识理解的分词方法这三大类，每一个大类又可以包括多种分词算法，比如基于字典、词库匹配的分词方法可以包括最大正向匹配法、最大反向匹配法和最短路径分词法等等，又比如基于词频度统计的分词方法中可以包括建立基于统计的分词模型，并根据该分词模型来进行分词处理的算法。可选的，本公开实施例可以从该三大类分词处理的方法中任选至少一种分词算法来对获取的通话内容文本数据进行分词处理。

在子步骤 4032 中，终端获取分词结果中每个词的关键词因子。

终端在获取了通话内容文本数据的分词结果之后，可以获取该分词结

果中每个词的关键词因子，该关键词因子可以正相关于每个词的主题相关度，该主题相关度用于指示一个词与通话内容文本数据的通话主题相关的程度，即一个词的关键词因子越大，则该词与通话内容文本数据的通话主题的相关程度越高。

如图 4-3 所示，本步骤可以包括下面 5 个子步骤:

在子步骤 4032a 中，终端根据预设类型词典获取分词结果中每个词的类型，该类型用于表征每个词的特征。

终端在获取了通话内容文本数据的分词结果之后，首先可以根据预设的类型词典来获取该分词结果中每个词的类型，该类型可以用于表征每个词的特征。示例性的，该类型可以包括骚扰特征词，专有名词，人物姓名，日期，语气词，人称代词，修饰词和地名等等，预设类型词典中可以记录有大部分常用词汇的类型。

在子步骤 4032b 中，终端去除分词结果中类型为语气词、人称代词和修饰词的词得到处理后的分词结果。

终端在获取了分词结果中每个词的类型之后，可以去除其中类型为语气词、人称代词和修饰词的词，得到处理后的分词结果。这样可以减少后续步骤中终端的计算量，提高终端在实施本公开实施例提供的骚扰电话号码确定方法时的运行速度。

其中，类型为语气词、人称代词和修饰词的词通常都和通话主题无关，且出现频率可能较高，示例性的，语气词可以包括了、么、呢、吧、啊等；人称代词可以包括你、我、他等；修饰词可以包括大、高、认真、生动、美丽等，去除这些词不但不会影响后续获取关键词，还可以增加获取的关键词的准确性。此外，终端还可以去除其它类型的与主题相关度较小的词，本公开实施例不作出限制。

需要说明的是，本步骤为可选步骤，即终端也可以不进行本步骤，在

子步骤 4032a 之后直接执行子步骤 4032c。

在子步骤 4032c 中，终端根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值。

在终端去除了与通话主题无关的类型的词之后，终端可以继续获取处理后的分词结果中每个词的关键词因子。终端可以根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值。其中，预设的类型与权值的对应关系，可以预先根据类型与通话主题可能的相关度来进行设置，示例性的，可以给类型为骚扰特征词的词设置最高的权值，给类型为专有名词和人物姓名的词设置较高的权值，给类型为日期和地名的词设置较低的权值，比如骚扰特征词对应的权值为 2，专有名词和人物姓名对应的权值为 1.5，日期和地名对应的权值为 1 等等。

在子步骤 4032d 中，终端获取分词结果中每个词的词频。

终端在为分词结果中每个词分配权值后，可以获取每个词在分词结果中的词频，其中词频可以表示某个词在分词结果中出现的次数，比如分词结果中共有 100 个词，“楼盘”这个词出现 12 次，则“楼盘”在该分词结果中的词频可以为 12；或者，词频可以表示某个词在分词结果中出现的次数除以该分词结果中的总词数所得的数值，比如分词结果中共有 100 个词，“楼盘”这个词出现 12 次，则“楼盘”在该分词结果中的词频可以为 12/100=0.12。需要说明的是，在使用多种分词算法进行分词处理，并获取多种分词结果时，可以在每种分词结果中分别统计每个词的词频。

需要说明的是，本步骤也可以在子步骤 4032c 之前执行，或者，本步骤和子步骤 4032c 同时执行，本公开实施例不作出限制。

在子步骤 4032e 中，终端将每个词的权值与词频的乘积作为每个词的关键词因子。

终端在获取了每个词的权值和乘积后，可以将权值与词频的乘积作为

关键词因子。

在子步骤 4033 中，终端按照关键词因子的大小将分词结果中的词降序排列。

终端在获取了分词结果中每个词的关键词因子后，可以按照关键词因子的大小将分词结果中的词降序排列，即关键词因子越大的词排名越靠前。

需要说明的是，在终端通过多种分词算法进行分词得到多种分词结果时，可以对每种分词结果中的词分别进行排名，或者将多种分词结果中的词合并进行排名，本公开实施例不作出限制。

在子步骤4034中，终端将排名前M位的词作为内容文本数据的关键词。

终端可以将排名前 M 位的词作为内容文本数据的关键词。

需要说明的是，在终端通过多种分词算法进行分词处理得到多种分词结果的情况下，该前 M 位的词可以为每种分词结果中的前 M 位的词，或者该前 M 位的词为合并排列中前 M 位的词。其中 M 为预设的值。

示例性的，M 为 1，终端根据两种分词算法得到了两种分词结果 F1 和 F2。F1 包含词保险、看房和恭喜，且保险的关键词因子为 10，看房的关键词因子为 12，恭喜的关键词因子为 8。F2 包含词奖品、张三和李四，且奖品的关键词因子为 13，张三的关键词因子为 7，李四的关键词因子为 3。

F1 中的词按照关键词因子降序排列为：1、看房；2、保险；3、恭喜。F2 中的词按照关键词因子降序排列为：1、奖品；2、张三；3、李四。终端可以将这两个排列中排名首位的词看房和奖品作为关键词。

而 F1 和 F2 中包含的词按照关键词因子的大小降序合并排列为：1、奖品；2、看房；3、保险；4、恭喜；5、张三；6、李四。终端也可以将该合并排列的首位奖品作为关键词。

需要说明的是，终端还可以设置有关键词因子阈值，终端可以利用该关键词因子阈值对排名前 M 位的词作进一步的筛选，去除排名前 M 位的词

中关键词因子小于关键词因子阈值的词，并将剩余的词作为关键词，这样可以提高最终确定的关键词的准确性，或者终端还可以直接将分词结果中关键词因子大于该关键词因子阈值的词作为关键词。

至此，已经获取了内容文本数据的关键词。需要说明的是，步骤 402 和步骤 403 为终端获取关键词的步骤，这两个步骤或这两个步骤中的部分步骤也可以由服务器来完成，本公开实施例不作出限制。示例性的，终端可以将通话语音数据或通话内容文本数据上传至服务器，由服务器来获取其中的关键词。

在步骤 404 中，终端根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词。

终端在获取了关键词之后，可以将关键词以及该关键词对应的通话记录合并得到目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词。示例性的，可以将关键词插入通话记录中对应的字段以生成目标通话记录。

在步骤 405 中，终端将预定时间内所有通话记录提供给服务器。

终端在获取了目标通话记录之后，可以将预定时间内所有的通话记录都提供给服务器，该所有通话记录可以包含有目标通话记录和普通通话记录，其中普通通话记录为不包含关键词的普通的通话记录。示例性的，终端可以每过一个月将该月内所有的通话记录上传至服务器。

此外，终端还可以仅将目标通话记录提供给服务器，本公开实施例不作出限制。

在步骤 406 中，服务器获取至少一个终端在预定时间内所有通话记录。

服务器可以获取至少一个终端在预定时间内所有通话记录，示例性的，服务器可以获取 1000 个终端在一个月内所有的通话记录，且服务器可以每隔一个月从这 1000 个终端获取一次。

在步骤 407 中，服务器从至少一个终端在预定时间内所有通话记录中获取目标通话记录得到通话记录集合。

在终端向服务器提供的是所有通话记录时，服务器可以从中筛选出目标通话记录，并将目标通话记录组成通话记录集合。筛选方法根据不同情况可以分为：

1）在终端向服务器提供了通讯录时，服务器可以根据该通讯录来筛选其中包含有陌生号码的通话记录，并将这些包含有陌生号码的电话号码作为目标通话记录。

2）在终端未向服务器提供通讯录，仅向服务器提供通话记录时，服务器可以根据通话记录中是否记录有姓名，并将未记录有姓名的通话记录作为目标通话记录，需要说明的是，终端在生成通话记录时，若通话对端的号码为通讯录中记录的联系人的号码，通常会在通话记录中附上该联系人的姓名。

此外，在终端仅向服务器提供了目标通话记录时，服务器可以直接将获取的目标通话记录组成通话记录集合。

在步骤 408 中，服务器对通话记录集合进行去重复处理。

目标通话记录中除了记录有通话对端的陌生号码，还可以记录有生成目标通话记录的终端的电话号码。其中，去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，即存在多个包含的两个电话号码一致的目标通话记录时，可以按照包含的关键词个数将这些目标通话记录进行降序排列，并仅保留首位的目标通话记录。同类通话记录为包含相同的两个电话号码的目标通话记录，该两个号码为通话双方的电话号码。

示例性的，当目标通话记录 A、目标通话记录 B 和目标通话记录 C 为同类通话记录时，目标通话记录 A 包含有关键词保险、看房、恭喜和奖品，

目标通话记录 B 包含有关键词张三，目标通话记录 C 包含有关键词李四和投资，则按照包含的关键词个数将这些目标通话记录进行降序排列：

1、目标通话记录 A（4 个关键词）;

2、目标通话记录 C（2 个关键词）;

3、目标通话记录 B（1 个关键词）。

服务器可以丢弃目标通话记录 C 和目标通话记录 B，并保留包含关键词最多的目标通话记录 A，这样可以增加本公开实施例提供的骚扰电话号码确定方法确定的骚扰电话号码的准确性。

在步骤 409 中，服务器获取处理后的通话记录集合中 N 个可疑号码。

其中，N 个可疑号码可以为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码，N 为预设值。

由于骚扰电话的拨打方通常会向大量的陌生号码拨打电话，因而服务器在对通话记录集合进行去重复处理之后，可以首先统计通话记录集合包含的目标通话记录中每个陌生号码的出现次数，并按照该出现次数将所有陌生号码进行降序排列，将排名前 N 位的陌生号码确定为可疑号码。由于骚扰电话的拨打方通常会向大量的陌生号码拨打骚扰电话，因而骚扰电话号码的拨打方的号码通常会大量的出现与目标通话记录中。

示例性的，服务器统计得到的陌生号码出现次数为：陌生号码 1 出现 1000 次，陌生号码 2 出现 900 次，陌生号码 3 出现 800 次，陌生号码 4 出现 1200 次，陌生号码 5 出现 2000 次，将这些陌生号码按照出现次数降序排列：

1、陌生号码 5;

2、陌生号码 4;

3、陌生号码 1;

4、陌生号码 2;

5、陌生号码 3。

在 N 预设为 3 时，可以得到可疑号码为陌生号码 5、陌生号码 4 和陌生号码 1。

此外，服务器还可以设置出现次数阈值，过滤掉排名前 N 位的陌生号码中出现次数小于该出现次数阈值的陌生号码，并将剩下的陌生号码作为可疑号码，以提高可疑号码的准确性，或者服务器可以直接将通话记录集合中出现次数大于该出现次数阈值的陌生号码作为可疑号码。

本步骤能够进一步的提高本公开实施例提供的骚扰电话号码确定方法确定的骚扰电话号码的准确性。

需要说明的是，服务器也可以将可疑号码确定为骚扰电话号码，本公开实施例不作出限制。

在步骤 410 中，服务器检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词。

服务器在获取了可疑号码之后，可以根据预设词典来检测每个可疑号码所在的目标通话记录包含的关键词中是否存在骚扰特征词，其中预设词典中可以记录有骚扰特征词。示例性的，预设词典中记录的骚扰特征词可以包括:

保险、保险公司、保险产品、风险、投资;

二手房、看房、学区房、业主、新房，换房;

恭喜、手机号码、电话号码、抽奖、中奖、奖品。

服务器可以检测每个可疑号码所在的目标通话记录包含的关键词中是否存在这些词。

在步骤 411 中，若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，服务器将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

服务器在根据预设词典检测了每个可疑号码是否存在骚扰特征词之后，可以将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。本步骤可以进一步提高确定的骚扰电话号码的准确性，减小对正常号码（非骚扰电话号码）误判的可能性。

如图 4-4 所示，本步骤可以包括下面 7 个子步骤:

在子步骤 4111 中，服务器获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码。

服务器在检测了所有可疑号码所在的目标通话记录是否包含有骚扰特征词之后，可以获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码，第一可疑号码可以出现于多个目标通话记录中，服务器可以通过这多个目标通话记录组成目标通话记录集合。

在子步骤 4112 中，服务器根据预设的骚扰特征词与骚扰类型的对应关系，确定第一通话记录包含的骚扰特征词的骚扰类型。

其中，骚扰类型表征第一可疑电话号码的骚扰目的，第一通话记录为目标通话记录集合中的任一通话记录。即服务器在获取了目标通话记录集合之后，可以根据预设的骚扰特征词与骚扰类型的对应关系，确定目标通话记录集合中每个目标通话记录包含的骚扰特征词的骚扰类型。示例性的，预设的骚扰特征词与骚扰类型的对应关系可以如表 1 所示:

表 1

骚扰类型	骚扰特征词
推销	保险、保险公司、保险产品、风险、投资
中介	二手房、看房、学区房、业主、新房，换房

诈骗	恭喜、手机号码、电话号码、抽奖、中奖、奖品

在表 1 中，第一行为表头，用于对表 1 中的两列进行说明，左边一列代表骚扰类型，右边一列代表每种骚扰类型包含的骚扰特征词，服务器可以根据该表确定第一通话记录包含的骚扰特征词的骚扰类型，比如第一通话记录中包含的骚扰特征词为“恭喜”，则根据表 1 可以得知“恭喜”的骚扰类型为“诈骗”。

在子步骤 4113 中，服务器将第一通话记录包含的骚扰特征词的骚扰类型确定为第一通话记录的骚扰类型。

服务器在获取了目标通话记录集合中每个通话记录包含的骚扰特征词的骚扰类型之后，可以将该骚扰特征词的骚扰类型确定为该通话记录的骚扰类型。示例性的，目标通话记录集合中的一个通话记录包含有骚扰类型为“诈骗”的骚扰特征词，则服务器可以将该通话记录的骚扰类型确定为“诈骗”。

需要说明的是，在目标通话记录集合中的通话记录包含有多个骚扰类型的骚扰特征词时，可以将这多个骚扰类型均确定为该通话记录的骚扰类型，示例性的，通话记录 A 包含有“诈骗”和“中介”两种骚扰类型的骚扰特征词，则通话记录 A 的骚扰类型为“诈骗”和“中介”。

在子步骤 4114 中，服务器统计目标通话记录集合中每种骚扰类型对应的通话记录的数量。

服务器在获取了目标通话记录中每个通话记录的骚扰类型之后，可以统计目标通话记录集合中每种骚扰类型对应的通话记录的数量。

示例性的，通话记录 A 的骚扰类型为“诈骗”和“中介”，通话记录 B 的骚扰类型为“推销”，通话记录 C 的骚扰类型为“中介”。则服务器的统计结果为:

骚扰类型“中介”对应的通话记录为通话记录 A 和通话记录 C 两个，

骚扰类型“诈骗”对应的通话记录为通话记录 A 一个，骚扰类型“推销”对应的通话记录为通话记录 B 一个。

在子步骤 4115 中，服务器根据每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例。

服务器在统计了目标通话记录集合中每种骚扰类型对应的通话记录的数量之后，可以根据每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例。

示例性的，目标通话记录集合中骚扰类型“推销”对应的通话记录有 60 个，骚扰类型“中介”对应的通话记录有 20 个，该目标通话记录集合的通话记录共有 100 个，则骚扰类型“推销”在该目标通话记录集合中的比例为 0.6，骚扰类型“中介”在该目标通话记录集合中的比例为 0.2。

在子步骤 4116 中，若目标通话记录集合中存在比例大于比例阈值的骚扰类型，服务器则确定第一可疑电话号码为骚扰电话号码。

服务器在获取了每种骚扰类型在目标通话记录集合中的比例之后，可以检测这些比例中是否存在大于比例阈值的值，若存在，则确定该第一可疑电话号码为骚扰电话号码。示例性的，比例阈值为 0.5，骚扰类型“推销”在目标通话记录集合中的比例为 0.6，大于该比例阈值，则可以确定该目标通话记录集合对应的第一可疑电话号码为骚扰电话号码。

需要说明的是，由于骚扰电话中骚扰方说的内容大都相差不多，因而一个骚扰电话号码所在的多个通话记录中的骚扰特征词的骚扰类型的相似度会较高，即根据比例阈值对可疑号码进行筛选能够进一步的提高确定的骚扰电话号码的准确性。

在子步骤 4117 中，服务器将目标通话记录集合中比例最大的骚扰类型

作为骚扰电话号码的骚扰类型。

服务器在确定了骚扰电话号码之后，可以将该骚扰电话号码对应的目标通话记录集合中比例最大的骚扰类型作为该骚扰电话号码的骚扰类型，该骚扰类型可以提供给用户一个对骚扰电话号码清晰的认识，提高用户体验。示例性的，目标通话记录集合中包含“推销”和“中介”这两种骚扰类型的通话记录，而骚扰类型“推销”在该目标通话记录集合中的比例为0.6，骚扰类型“中介”在该目标通话记录集合中的比例为0.2，则可以确定该骚扰电话号码的骚扰类型为“推销”。

此外，子步骤4117为对骚扰电话号码分类的步骤，具体地，服务器将目标通话记录集合中比例最大的骚扰类型作为骚扰电话号码的骚扰类型，是可选的步骤。

需要说明的是，目标通话记录集合中的通话记录均为目标通话记录。

在步骤412中，服务器根据确定的骚扰电话号码生成即时骚扰号码表。

服务器在获取了骚扰电话号码之后，可以根据骚扰电话号码生成即时骚扰号码表，示例性的，生成的即时骚扰电话号码表可以如表2所示。

表2

骚扰电话号码
111***1111
111***1112
111***1113
……

在表2中，第一行为表头，用于对表2中的数据进行说明，表2中的数字代表骚扰电话号码。

需要说明的是，在服务器确定了骚扰电话号码的骚扰类型时，服务器还可以根据骚扰电话号码和骚扰类型共同生成即时骚扰号码表，也就是说即时骚扰号码表中不但记录有骚扰电话号码，还记录有骚扰电话号码的骚

扰类型。示例性的，如表 3 所示：

表 3

骚扰电话号码	骚扰类型
111***1111	推销
111***1112	中介
111***1113	诈骗
……	……

在表 3 中，第一行为表头，用于对表 3 中的两列数据进行说明，左边一列代表骚扰电话号码，右边一列代表每个骚扰电话号码的骚扰类型。比如第二行的骚扰电话号码“111***1111”的骚扰类型为“推销”。

在步骤 413 中，服务器根据即时骚扰号码表更新历史骚扰号码表。

服务器在获取了即时骚扰号码表之后，可以以该即时骚扰号码表来更新历史骚扰号码表，其中历史骚扰号码表可以为服务器根据本次之前获取的即时骚扰号码表生成的，此外，在本次即为第一次的到即时骚扰号码表时，可以将该即时骚扰号码表作为历史骚扰号码表。在更新历史骚扰号码表时，可以给每条骚扰电话号码附上时间戳（Timestamp），以表示该条骚扰电话号码记录进历史骚扰号码表的时间。服务器可以每过一段时间就生成一次即使骚扰号码表，并根据最新获取的即时骚扰号码表更新一次历史骚扰号码表。

需要说明的是，时间戳，或称 Unix 时间（Unix time），是一种时间表示方式，定义为从格林威治时间 1970 年 01 月 01 日 00 时 00 分 00 秒（北京时间 1970 年 01 月 01 日 08 时 00 分 00 秒）起至现在的总秒数。

本步骤在更新历史骚扰号码表时，可以遵循下面三条原则：

1）历史骚扰号码表中没有而在即时骚扰号码表中有的骚扰电话号码，更新到历史骚扰号码表中。

2）历史骚扰号码表中和即时骚扰号码表中都有的骚扰电话号码，更新

历史骚扰号码表中该骚扰电话号码的时间戳。

3）历史骚扰号码表中有而即时骚扰号码表中没有的骚扰电话号码，比较当前时间与该骚扰电话号码的时间戳，如果时间差大于阈值（可以预先设定），则删除历史骚扰号码表中的该骚扰电话号码。这样可以删除掉历史骚扰号码表中长时间未使用的骚扰电话号码，避免骚扰电话号码更换使用者而产生的误判。

本步骤结束后，服务器获取的历史骚扰号码表中记录的骚扰电话号码会具有相当高的准确性。

在步骤 414 中，服务器向终端提供更新后的历史骚扰号码表。

服务器在更新历史骚扰号码表之后，可以向终端提供该更新后的历史骚扰号码表，示例性的服务器可以主动向终端发送该历史骚扰号码表，或者由终端向服务器下载该历史骚扰号码表。

终端在获取了更新后的历史骚扰号码表之后，终端的用户可以根据其中记录的骚扰电话号码以及骚扰电话号码的骚扰类型来对骚扰电话号码的来电做出进一步的处理。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定方法，通过关键词因子的大小来判断分词结果中的词是否为关键词，达到了提高关键词的准确性的效果。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定方法，通过将目标通话记录中出现次数最多的 N 个号码确定为可疑号码，达到了增加最终确定的骚扰电话号码的准确性的效果。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定方法，通过目标通话记录集合中各骚扰类型对应的目标通话记录在目标通话记录集合中的比例来筛选出骚扰电话号码，达到了进一步增加确定的骚扰电话号码的准确性的效果。

综上所述，本公开实施例提供的骚扰电话号码确定方法，通过将通话记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

下述为本公开装置实施例，可以配置为执行本公开方法实施例。对于本公开装置实施例中未披露的细节，请参照本公开方法实施例。

图 5-1 是根据一示例性实施例示出的一种骚扰电话号码确定装置的框图，本实施例以该骚扰电话号码确定装置应用于图 1 所示实施环境中的服务器 12 中来举例说明，该骚扰电话号码确定装置可以包括:

集合获取模块 510，被配置为从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，陌生号码为陌生号码对应的终端的通讯录中未存储的号码。

号码获取模块 520，被配置为从集合获取模块 510 获取的通话记录集合中获取 N 个可疑号码，N 个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码。

检测模块 530，被配置为检测号码获取模块 520 获取的每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词。

确定模块 540，被配置为在检测模块 530 检测出的任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词时，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

综上所述，本公开实施例提供的骚扰电话号码确定装置，通过将通话

记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

图 5-2 是图 5-1 示出的骚扰电话号码确定装置中确定模块 540 的框图。

可选的，确定模块 540，包括:

集合获取子模块 541，被配置为获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码。

特征词确定子模块 542，被配置为根据预设的骚扰特征词与骚扰类型的对应关系，确定集合获取子模块 541 获取的目标通话记录集合中第一通话记录包含的骚扰特征词的骚扰类型，骚扰类型表征第一可疑电话号码的骚扰目的，第一通话记录为目标通话记录集合中的任一通话记录。

记录确定子模块 543，被配置为将特征词确定子模块 542 确定的第一通话记录包含的骚扰特征词的骚扰类型确定为第一通话记录的骚扰类型。

统计子模块 544，被配置为统计记录确定子模块 543 确定的目标通话记录集合中每种骚扰类型对应的通话记录的数量。

比例确定子模块 545，被配置为根据统计子模块 544 统计的每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例。

号码确定子模块 546，被配置为在比例确定子模块 545 确定的目标通话记录集合中存在比例大于比例阈值的骚扰类型时，确定所述第一可疑电话号码为骚扰电话号码。

如图 5-3 所示，其为本公开实施例示提供的另一种骚扰电话号码确定装

置的框图。

可选的，该装置还包括:

类型确定模块 550，被配置为将比例确定子模块确定的目标通话记录集合中比例最大的骚扰类型作为骚扰电话号码的骚扰类型。

可选的，每个目标通话记录中包含有两个电话号码，两个电话号码中包含有陌生号码。

号码获取模块 520，被配置为对集合获取模块 510 获取的通话记录集合进行去重复处理，去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，同类通话记录为包含相同的两个电话号码的目标通话记录；获取处理后的通话记录集合中 N 个可疑号码。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定装置，通过将目标通话记录中出现次数最多的 N 个号码确定为可疑号码，达到了增加最终确定的骚扰电话号码的准确性的效果。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定装置，通过目标通话记录集合中各骚扰类型对应的目标通话记录在目标通话记录集合中的比例来筛选出骚扰电话号码，达到了进一步增加确定的骚扰电话号码的准确性的效果。

综上所述，本公开实施例提供的骚扰电话号码确定装置，通过将通话记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

如图 5-1 和图 5-3 所示的骚扰电话号码确定装置，配置为执行如图 2 和图 4-1 所示的方法流程。

如图 5-2 所示的确定模块，配置为执行如图 4-4 所示的方法流程。

图 6-1 是根据一示例性实施例示出的另一种骚扰电话号码确定装置的框图，本实施例以该骚扰电话号码确定装置应用于图 1 所示实施环境中的终端 11 中来举例说明，该骚扰电话号码确定装置可以包括:

终端获取模块 610，被配置为获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码。

转化模块 620，被配置为将终端获取模块 610 获取的通话语音数据转化为通话内容文本数据。

关键词获取模块 630，被配置为获取转化模块 620 获取的通话内容文本数据的至少一个关键词，至少一个关键词为表征通话内容文本数据的通话主题的词。

目标生成模块 640，被配置为根据关键词获取模块 630 获取的至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词。

记录提供模块 650，被配置为将目标生成模块 640 生成的目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌生号码和至少一个关键词确定陌生号码是否为骚扰电话号码。

综上所述，本公开实施例提供的骚扰电话号码确定装置，通过获取目标通话记录，该目标通话记录中包含有陌生号码和至少一个关键词，并将该目标通话记录提供给服务器，服务器用于根据该目标通话记录中的陌生号码和至少一个关键词来确定该陌生号码是否为骚扰电话号码，达到了提高骚扰电话号码获取效率的效果。

图 6-2 是图 6-1 示出的骚扰电话号码确定装置中的关键词获取模块 630 的框图，该关键词获取模块 630，包括:

分词子模块 631，被配置为对内容文本数据进行分词处理得到包含有至

少一个词的分词结果。

因子获取子模块 632，被配置为获取分词子模块 631 得到的分词结果中每个词的关键词因子，关键词因子正相关于每个词的主题相关度，每个词的主题相关度用于指示每个词与通话内容文本数据的通话主题相关的程度。

因子排序子模块 633，被配置为按照因子获取子模块 632 获取的关键词因子的大小将分词结果中的词降序排列。

关键词确定子模块 634，被配置为将因子排序子模块 633 的排列中排名前 M 位的词作为内容文本数据的关键词。

可选的，因子获取子模块 632，被配置为根据预设类型词典获取分词子模块 631 得到的分词结果中每个词的类型，类型用于表征每个词的特征；根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值；获取分词结果中每个词的词频；将每个词的权值与词频的乘积作为每个词的关键词因子。

可选的，因子获取子模块 632，被配置为去除分词子模块 631 得到的分词结果中类型为语气词、人称代词和修饰词的词得到处理后的分词结果；获取处理后的分词结果中每个词的关键词因子。

需要补充说明的是，本公开实施例提供的骚扰电话号码确定装置，通过关键词因子的大小来判断分词结果中的词是否为关键词，达到了提高关键词的准确性的效果。

综上所述，本公开实施例提供的骚扰电话号码确定装置，通过获取目标通话记录，该目标通话记录中包含有陌生号码和至少一个关键词，并将该目标通话记录提供给服务器，服务器再根据该目标通话记录中的陌生号码和至少一个关键词来确定该陌生号码是否为骚扰电话号码，达到了提高骚扰电话号码获取效率的效果。

如图 6-1 所示的骚扰电话号码确定装置，配置为执行如图 3 所示的方法流程。

如图 6-2 所示的关键词获取模块，配置为执行如图 4-2 所示的方法流程。

图 7 是根据一示例性实施例示出的一种用于确定骚扰电话号码的装置 700 的框图。例如，装置 700 可以是移动电话，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理等。

参照图 7，装置 700 可以包括以下一个或多个组件：处理组件 702，存储器 704，电源组件 706，多媒体组件 708，音频组件 710，输入/输出（I/O）接口 712，传感器组件 714，以及通信组件 716。

处理组件 702 通常控制装置 700 的整体操作，诸如与显示，电话呼叫，数据通信，相机操作和记录操作相关联的操作。处理组件 702 可以包括一个或多个处理器 720 来执行指令，以完成上述的方法的全部或部分步骤。此外，处理组件 702 可以包括一个或多个模块，便于处理组件 702 和其他组件之间的交互。例如，处理组件 702 可以包括多媒体模块，以方便多媒体组件 708 和处理组件 702 之间的交互。

存储器 704 被配置为存储各种类型的数据以支持在装置 700 的操作。这些数据的示例包括用于在装置 700 上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。存储器 704 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

电源组件 706 为装置 700 的各种组件提供电力。电源组件 706 可以包括电源管理系统，一个或多个电源，及其他与为装置 700 生成、管理和分

配电力相关联的组件。

多媒体组件 708 包括在所述装置 700 和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中，多媒体组件 708 包括一个前置摄像头和/或后置摄像头。当装置 700 处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

音频组件 710 被配置为输出和/或输入音频信号。例如，音频组件 710 包括一个麦克风（MIC），当装置 700 处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器 704 或经由通信组件 716 发送。在一些实施例中，音频组件 710 还包括一个扬声器，配置为输出音频信号。

I/O 接口 712 为处理组件 702 和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

传感器组件 714 包括一个或多个传感器，配置为为装置 700 提供各个方面的状态评估。例如，传感器组件 714 可以检测到装置 700 的打开/关闭状态，组件的相对定位，例如所述组件为装置 700 的显示器和小键盘，传感器组件 714 还可以检测装置 700 或装置 700 一个组件的位置改变，用户与装置 700 接触的存在或不存在，装置 700 方位或加速/减速和装置 700 的温度变化。传感器组件 714 可以包括接近传感器，被配置用来在没有任何

的物理接触时检测附近物体的存在。传感器组件 714 还可以包括光传感器，如 CMOS 或恭喜 D 图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件 714 还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

通信组件 716 被配置为便于装置 700 和其他设备之间有线或无线方式的通信。装置 700 可以接入基于通信标准的无线网络，如 WiFi，2G 或 3G，或它们的组合。在一个示例性实施例中，通信组件 716 经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件 716 还包括近场通信（NFC）模块，以促进短程通信。例如，在 NFC 模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

在示例性实施例中，装置 700 可以被一个或多个应用专用集成电路（ASIC）、数字信号处理器（DSP）、数字信号处理设备（DSPD）、可编程逻辑器件（PLD）、现场可编程门阵列（FPGA）、控制器、微控制器、微处理器或其他电子元件实现，配置为执行上述方法。

在示例性实施例中，还提供了一种包括指令的非临时性计算机可读存储介质，例如包括指令的存储器 704，上述指令可由装置 700 的处理器 720 执行以完成上述方法。例如，所述非临时性计算机可读存储介质可以是 ROM、随机存取存储器（RAM）、CD-ROM、磁带、软盘和光数据存储设备等。

一种非临时性计算机可读存储介质，当所述存储介质中的指令由装置 700 的处理器执行时，使得装置 700 能够执行上述各个实施例提供的骚扰电话号码确定方法。

图 8 是根据一示例性实施例示出的一种用于确定骚扰电话号码的装置 800 的框图。例如，装置 800 可以被提供为一服务器。参照图 8，装置 800

包括处理组件 822，其进一步包括一个或多个处理器，以及由存储器 832 所代表的存储器资源，配置为存储可由处理组件 822 执行的指令，例如应用程序。存储器 832 中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件 822 被配置为执行指令，以执行上述骚扰电话号码确定方法。

装置 800 还可以包括一个电源组件 826 被配置为执行装置 800 的电源管理，一个有线或无线网络接口 850 被配置为将装置 800 连接到网络，和一个输入输出（I/O）接口 858。装置 800 可以操作基于存储在存储器 832 的操作系统，例如 Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM 或类似。

图 9 是根据一示例性实施例示出的一种骚扰电话号码确定系统，该系统包括：服务器 910 和终端 920。

服务器 910 可以包括图 5-1 所示的骚扰电话号码确定装置或图 5-3 所示的骚扰电话号码确定装置。

终端 920 可以包括图 6-1 所示的骚扰电话号码确定装置。

图 10 是根据一示例性实施例示出的一种骚扰电话号码确定系统，该系统包括：服务器 1010 和终端 1020。

服务器 1010 可以包括图 8 所示的骚扰电话号码确定装置。

终端 1020 可以包括图 7 所示的骚扰电话号码确定装置。

本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开实施例的其它实施方案。本申请旨在涵盖本公开实施例的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开实施例的一般性原理并包括本公开实施例未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开实施例的真正范围和精神由下面的权利要求指出。

应当理解的是，本公开实施例并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本公开实施例的范围仅由所附的权利要求来限制。

工业实用性

本公开实施例通过将通话记录集合中出现次数最多的 N 个陌生号码确定为可疑号码，再将可疑号码所在的目标通话记录包含的关键词中存在骚扰特征词的可疑号码确定为骚扰电话号码，其中关键词为表征通话主题的词，方便快捷的确定出通话记录集合包含的大量电话号码中的骚扰电话号码，达到了提高骚扰电话号码的获取效率的效果。

权利要求书

1、一种骚扰电话号码确定方法，包括:

从至少一个终端获取目标通话记录得到通话记录集合，所述通话记录集合中包含有预定个数的目标通话记录，每个所述目标通话记录中包含有陌生号码和至少一个关键词，所述至少一个关键词是所述终端根据与陌生号码通话的通话内容文本数据得到的，用于表征所述通话内容文本数据的通话主题，所述陌生号码为所述陌生号码对应的终端的通讯录中未存储的号码;

获取所述通话记录集合中 N 个可疑号码，所述 N 个可疑号码为所述预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码;

检测每个所述可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词;

若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

2、根据权利要求 1 所述的方法，其中，所述将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码，包括:

获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，所述第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码;

根据预设的骚扰特征词与骚扰类型的对应关系，确定第一通话记录包含的骚扰特征词的骚扰类型，所述骚扰类型表征所述第一可疑电话号码的骚扰目的，所述第一通话记录为所述目标通话记录集合中的任一通话记录;

将所述第一通话记录包含的骚扰特征词的骚扰类型确定为所述第一通话记录的骚扰类型;

统计所述目标通话记录集合中每种骚扰类型对应的通话记录的数量;

根据所述每种骚扰类型对应的通话记录的数量与所述目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在所述目标通话记录集合中的比例;

若所述目标通话记录集合中存在比例大于比例阈值的骚扰类型，则确定所述第一可疑电话号码为骚扰电话号码。

3、根据权利要求 2 所述的方法，其中，所述方法还包括:

将所述目标通话记录集合中比例最大的骚扰类型作为所述骚扰电话号码的骚扰类型。

4、根据权利要求 1 至 3 任一所述的方法，其中，每个所述目标通话记录中包含有两个电话号码，所述两个电话号码中包含有陌生号码，所述获取所述通话记录集合中 N 个可疑号码，包括:

对所述通话记录集合进行去重复处理，所述去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，所述同类通话记录为包含相同的两个电话号码的目标通话记录;

获取处理后的通话记录集合中 N 个可疑号码。

5、一种骚扰电话号码确定方法，其中，所述方法包括:

获取与陌生号码通话的通话语音数据和通话记录，所述通话记录中包含有所述陌生号码，所述陌生号码为通讯录中未存储的号码;

将所述通话语音数据转化为通话内容文本数据;

获取所述通话内容文本数据的至少一个关键词，所述至少一个关键词为表征所述通话内容文本数据的通话主题的词;

根据所述至少一个关键词和所述通话记录生成目标通话记录，所述目标通话记录中包含有所述陌生号码和所述至少一个关键词；

将所述目标通话记录提供给服务器，所述服务器用于根据所述目标通话记录中的所述陌生号码和所述至少一个关键词确定所述陌生号码是否为骚扰电话号码。

6、根据权利要求 5 所述的方法，其中，所述获取所述通话内容文本数据的至少一个关键词，包括：

对所述内容文本数据进行分词处理得到包含有至少一个词的分词结果；

获取所述分词结果中每个词的关键词因子，所述关键词因子正相关于所述每个词的主题相关度，所述每个词的主题相关度用于指示所述每个词与所述通话内容文本数据的通话主题相关的程度；

按照所述关键词因子的大小将所述分词结果中的词降序排列；

将排名前 M 位的词作为所述内容文本数据的关键词。

7、根据权利要求 6 所述的方法，其中，所述获取所述分词结果中每个词的关键词因子，包括：

根据预设类型词典获取所述分词结果中每个词的类型，所述类型用于表征所述每个词的特征；

根据预设的类型与权值的对应关系，按照所述分词结果中每个词的类型为所述每个词分配权值；

获取所述分词结果中每个词的词频；

将所述每个词的权值与词频的乘积作为所述每个词的关键词因子。

8、根据权利要求 6 所述的方法，其中，所述获取所述分词结果中每个词的关键词因子，包括：

去除所述分词结果中类型为语气词、人称代词和修饰词的词得到处

理后的分词结果;

获取所述处理后的分词结果中每个词的关键词因子。

9、一种骚扰电话号码确定装置，包括:

集合获取模块，被配置为从至少一个终端获取目标通话记录得到通话记录集合，所述通话记录集合中包含有预定个数的目标通话记录，每个所述目标通话记录中包含有陌生号码和至少一个关键词，所述至少一个关键词是所述终端根据与陌生号码通话的通话内容文本数据得到的，用于表征所述通话内容文本数据的通话主题，所述陌生号码为所述陌生号码对应的终端的通讯录中未存储的号码;

号码获取模块，被配置为获取所述通话记录集合中 N 个可疑号码，所述 N 个可疑号码为所述预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码;

检测模块，被配置为检测每个所述可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词;

确定模块，被配置为在任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词时，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

10、根据权利要求 9 所述的装置，其中，所述确定模块，包括:

集合获取子模块，被配置为获取包含所述第一可疑电话号码的目标通话记录组成的目标通话记录集合，所述第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码;

特征词确定子模块，被配置为根据预设的骚扰特征词与骚扰类型的对应关系，确定第一通话记录包含的骚扰特征词的骚扰类型，所述骚扰类型表征所述第一可疑电话号码的骚扰目的，所述第一通话记录为所述目标通话记录集合中的任一通话记录;

记录确定子模块，被配置为将所述第一通话记录包含的骚扰特征词的骚扰类型确定为所述第一通话记录的骚扰类型；

统计子模块，被配置为统计所述目标通话记录集合中每种骚扰类型对应的通话记录的数量；

比例确定子模块，被配置为根据所述每种骚扰类型对应的通话记录的数量与所述目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在所述目标通话记录集合中的比例；

号码确定子模块，被配置为在所述目标通话记录集合中存在比例大于比例阈值的骚扰类型时，确定所述第一可疑电话号码为骚扰电话号码。

11、根据权利要求 9 所述的装置，其中，所述装置还包括：

类型确定模块，被配置为将所述目标通话记录集合中比例最大的骚扰类型作为所述骚扰电话号码的骚扰类型。

12、根据权利要求 9 至 11 任一所述的装置，其中，每个所述目标通话记录中包含有两个电话号码，所述两个电话号码中包含有陌生号码，

所述号码获取模块，被配置为对所述通话记录集合进行去重复处理，所述去重复处理用于丢弃至少两个同类通话记录中除包含关键词最多的目标通话记录之外的通话记录，所述同类通话记录为包含相同的两个电话号码的目标通话记录；获取处理后的通话记录集合中 N 个可疑号码。

13、一种骚扰电话号码确定装置，包括：

终端获取模块，被配置为获取与陌生号码通话的通话语音数据和通话记录，所述通话记录中包含有所述陌生号码，所述陌生号码为通讯录中未存储的号码；

转化模块，被配置为将所述通话语音数据转化为通话内容文本数据；

关键词获取模块，被配置为获取所述通话内容文本数据的至少一个关键词，所述至少一个关键词为表征所述通话内容文本数据的通话主题

的词；

目标生成模块，被配置为根据所述至少一个关键词和所述通话记录生成目标通话记录，所述目标通话记录中包含有所述陌生号码和所述至少一个关键词；

记录提供模块，被配置为将所述目标通话记录提供给服务器，所述服务器用于根据所述目标通话记录中的所述陌生号码和所述至少一个关键词确定所述陌生号码是否为骚扰电话号码。

14、根据权利要求 13 所述的装置，其中，所述关键词获取模块，包括：

分词子模块，被配置为对所述内容文本数据进行分词处理得到包含有至少一个词的分词结果；

因子获取子模块，被配置为获取所述分词结果中每个词的关键词因子，所述关键词因子正相关于所述每个词的主题相关度，所述每个词的主题相关度用于指示所述每个词与所述通话内容文本数据的通话主题相关的程度；

因子排序子模块，被配置为按照所述关键词因子的大小将所述分词结果中的词降序排列；

关键词确定子模块，被配置为将排名前 M 位的词作为所述内容文本数据的关键词。

15、根据权利要求 14 所述的装置，其中，所述因子获取子模块，被配置为根据预设类型词典获取所述分词结果中每个词的类型，所述类型用于表征所述每个词的特征；根据预设的类型与权值的对应关系，按照所述分词结果中每个词的类型为所述每个词分配权值；获取所述分词结果中每个词的词频；将所述每个词的权值与词频的乘积作为所述每个词的关键词因子。

16、根据权利要求 13 所述的装置，其中，

所述因子获取子模块，被配置为去除所述分词结果中类型为语气词、人称代词和修饰词的词得到处理后的分词结果；获取所述处理后的分词结果中每个词的关键词因子。

17、一种骚扰电话号码确定装置，包括：

处理器；

配置为存储所述处理器的可执行指令的存储器；

其中，所述处理器被配置为：

从至少一个终端获取目标通话记录得到通话记录集合，所述通话记录集合中包含有预定个数的目标通话记录，每个所述目标通话记录中包含有陌生号码和至少一个关键词，所述至少一个关键词是所述终端根据与陌生号码通话的通话内容文本数据得到的，用于表征所述通话内容文本数据的通话主题，所述陌生号码为所述陌生号码对应的终端的通讯录中未存储的号码；

获取所述通话记录集合中 N 个可疑号码，所述 N 个可疑号码为所述预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前 N 位的陌生号码；

检测每个所述可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词；

若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码。

18、一种骚扰电话号码确定装置，包括：

处理器；

配置为存储所述处理器的可执行指令的存储器；

其中，所述处理器被配置为：

获取与陌生号码通话的通话语音数据和通话记录，所述通话记录中包含有所述陌生号码，所述陌生号码为通讯录中未存储的号码；

将所述通话语音数据转化为通话内容文本数据；

获取所述通话内容文本数据的至少一个关键词，所述至少一个关键词为表征所述通话内容文本数据的通话主题的词；

根据所述至少一个关键词和所述通话记录生成目标通话记录，所述目标通话记录中包含有所述陌生号码和所述至少一个关键词；

将所述目标通话记录提供给服务器，所述服务器用于根据所述目标通话记录中的所述陌生号码和所述至少一个关键词确定所述陌生号码是否为骚扰电话号码。

19、一种骚扰电话号码确定系统，包括：服务器和终端，

所述服务器包括权利要求 9 至 12 任一所述的骚扰电话号码确定装置；

所述终端包括权利要求 13 至 16 任一所述的骚扰电话号码确定装置。

20、一种骚扰电话号码确定系统，包括：服务器和终端，

所述服务器包括权利要求 17 所述的骚扰电话号码确定装置；

所述终端包括权利要求 18 所述的骚扰电话号码确定装置。

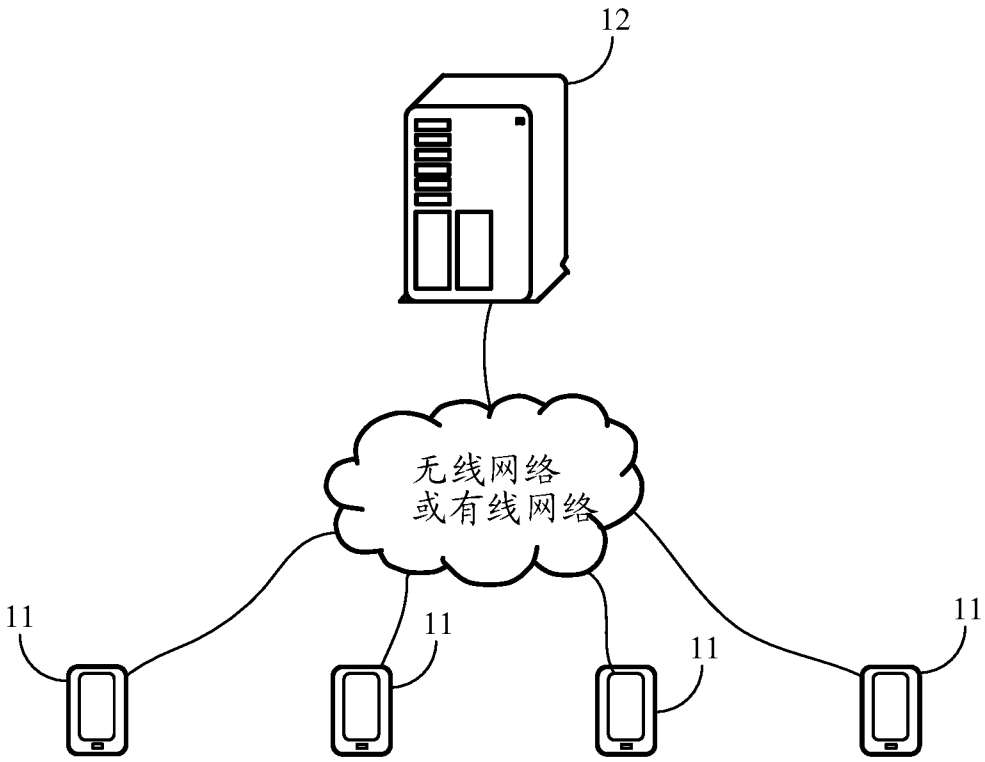


图 1

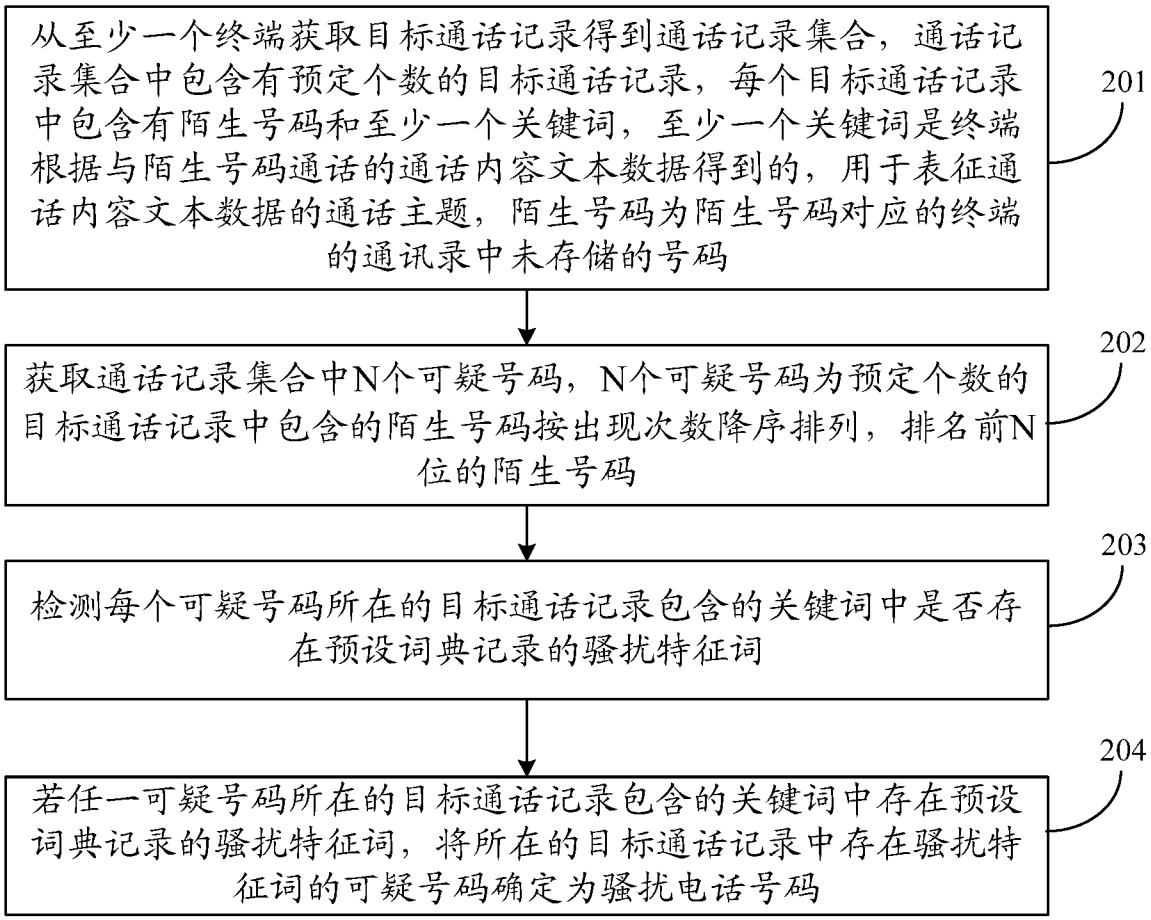

从至少一个终端获取目标通话记录得到通话记录集合，通话记录集合中包含有预定个数的目标通话记录，每个目标通话记录中包含有陌生号码和至少一个关键词，至少一个关键词是终端根据与陌生号码通话的通话内容文本数据得到的，用于表征通话内容文本数据的通话主题，陌生号码为陌生号码对应的终端的通讯录中未存储的号码
201
获取通话记录集合中N个可疑号码，N个可疑号码为预定个数的目标通话记录中包含的陌生号码按出现次数降序排列，排名前N位的陌生号码
202
检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词
203
若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码
204

图 2

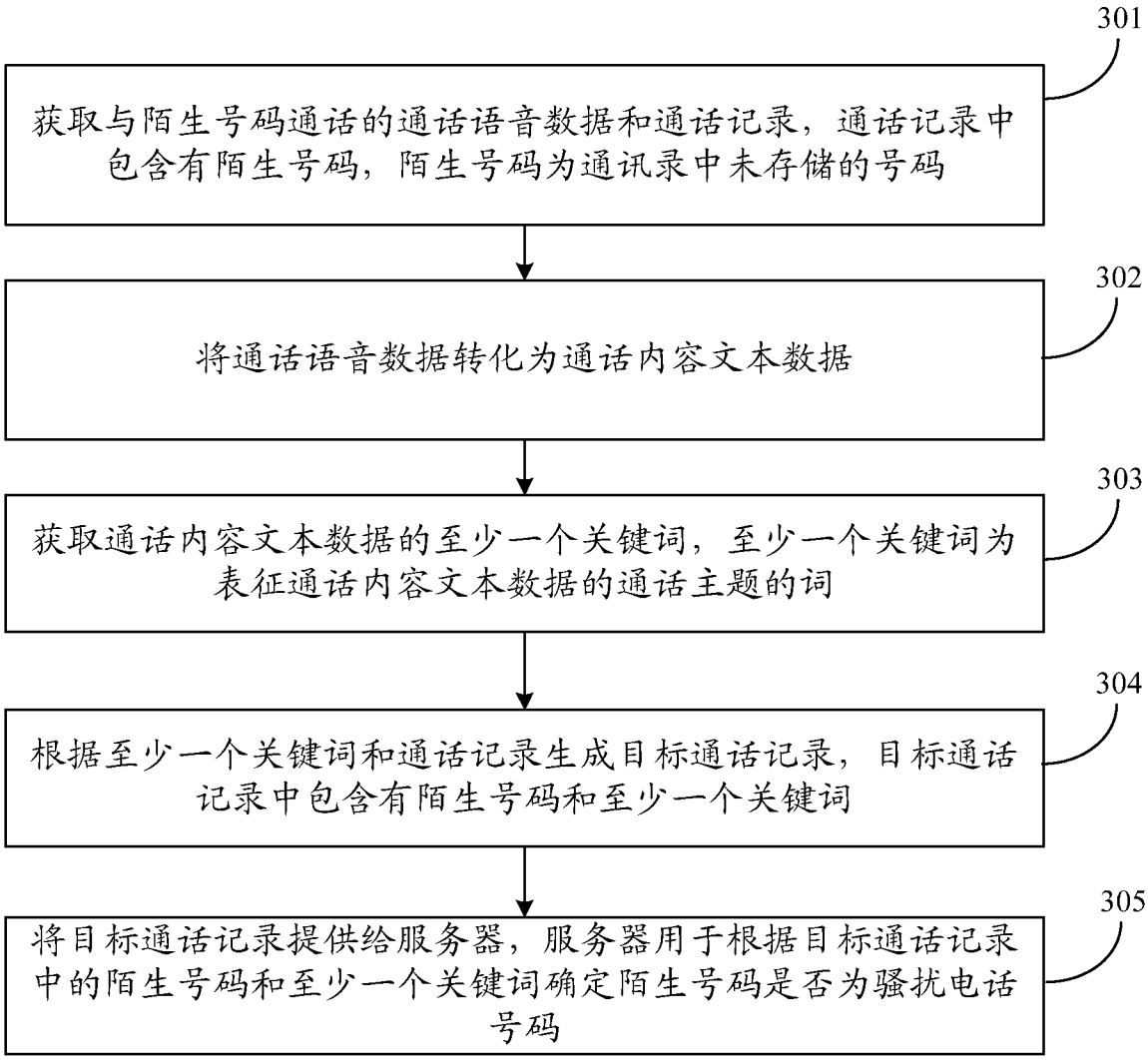

301
获取与陌生号码通话的通话语音数据和通话记录，通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码
302
将通话语音数据转化为通话内容文本数据
303
获取通话内容文本数据的至少一个关键词，至少一个关键词为表征通话内容文本数据的通话主题的词
304
根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词
305
将目标通话记录提供给服务器，服务器用于根据目标通话记录中的陌生号码和至少一个关键词确定陌生号码是否为骚扰电话号码

图 3

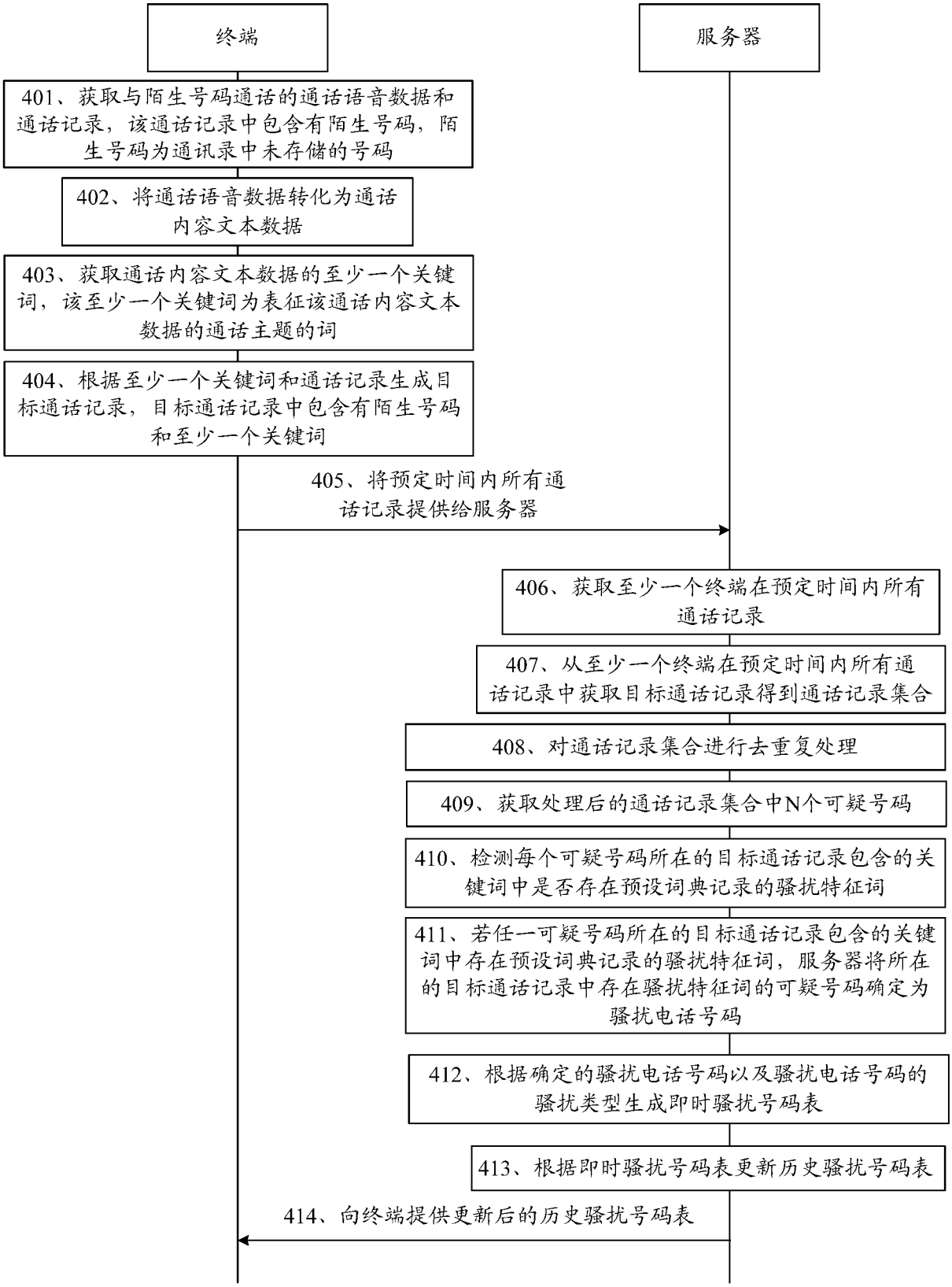

终端
服务器
401、获取与陌生号码通话的通话语音数据和通话记录，该通话记录中包含有陌生号码，陌生号码为通讯录中未存储的号码
402、将通话语音数据转化为通话内容文本数据
403、获取通话内容文本数据的至少一个关键词，该至少一个关键词为表征该通话内容文本数据的通话主题的词
404、根据至少一个关键词和通话记录生成目标通话记录，目标通话记录中包含有陌生号码和至少一个关键词
405、将预定时间内所有通话记录提供给服务器
406、获取至少一个终端在预定时间内所有通话记录
407、从至少一个终端在预定时间内所有通话记录中获取目标通话记录得到通话记录集合
408、对通话记录集合进行去重复处理
409、获取处理后的通话记录集合中N个可疑号码
410、检测每个可疑号码所在的目标通话记录包含的关键词中是否存在预设词典记录的骚扰特征词
411、若任一可疑号码所在的目标通话记录包含的关键词中存在预设词典记录的骚扰特征词，服务器将所在的目标通话记录中存在骚扰特征词的可疑号码确定为骚扰电话号码
412、根据确定的骚扰电话号码以及骚扰电话号码的骚扰类型生成即时骚扰号码表
413、根据即时骚扰号码表更新历史骚扰号码表
414、向终端提供更新后的历史骚扰号码表

图 4-1

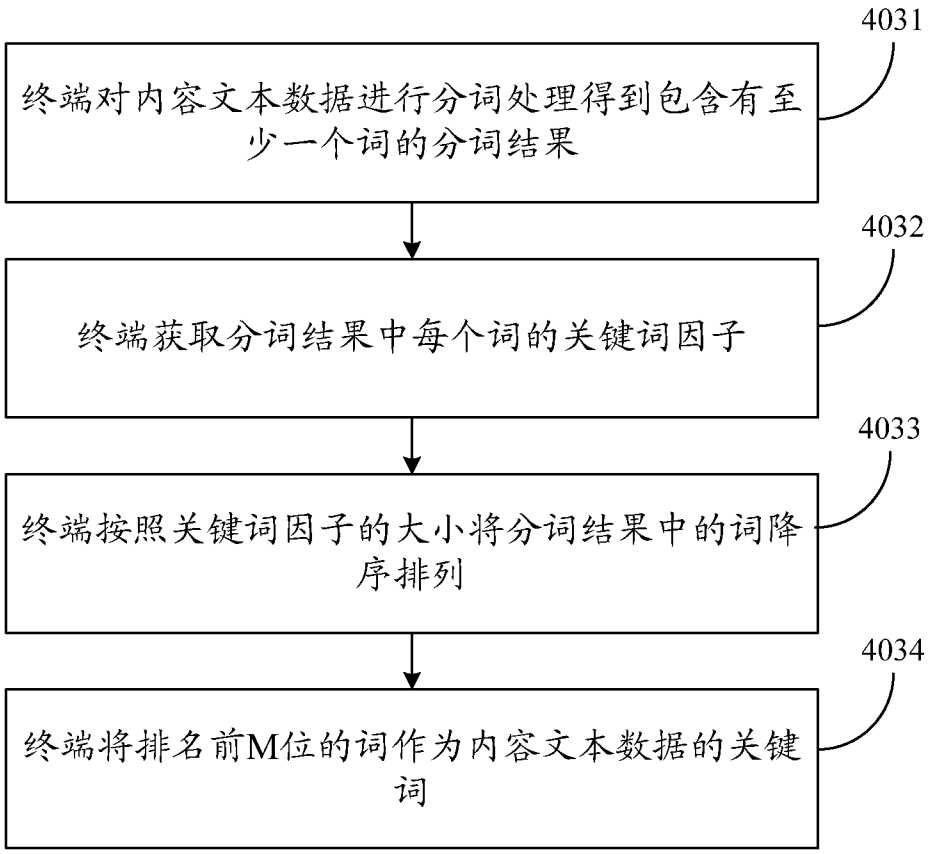

4031
终端对内容文本数据进行分词处理得到包含有至少一个词的分词结果
4032
终端获取分词结果中每个词的关键词因子
4033
终端按照关键词因子的大小将分词结果中的词降序排列
4034
终端将排名前M位的词作为内容文本数据的关键词

图 4-2

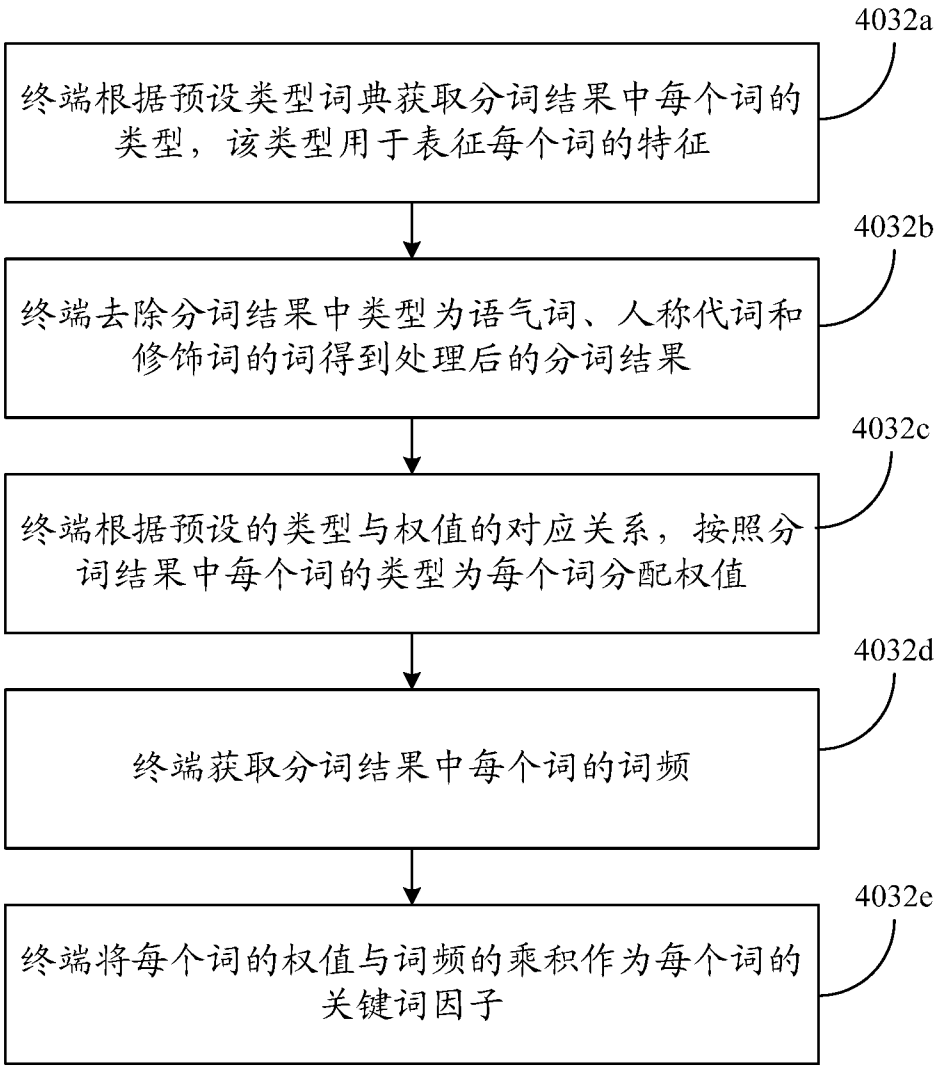

4032a
终端根据预设类型词典获取分词结果中每个词的类型，该类型用于表征每个词的特征
4032b
终端去除分词结果中类型为语气词、人称代词和修饰词的词得到处理后的分词结果
4032c
终端根据预设的类型与权值的对应关系，按照分词结果中每个词的类型为每个词分配权值
4032d
终端获取分词结果中每个词的词频
4032e
终端将每个词的权值与词频的乘积作为每个词的关键词因子

图 4-3

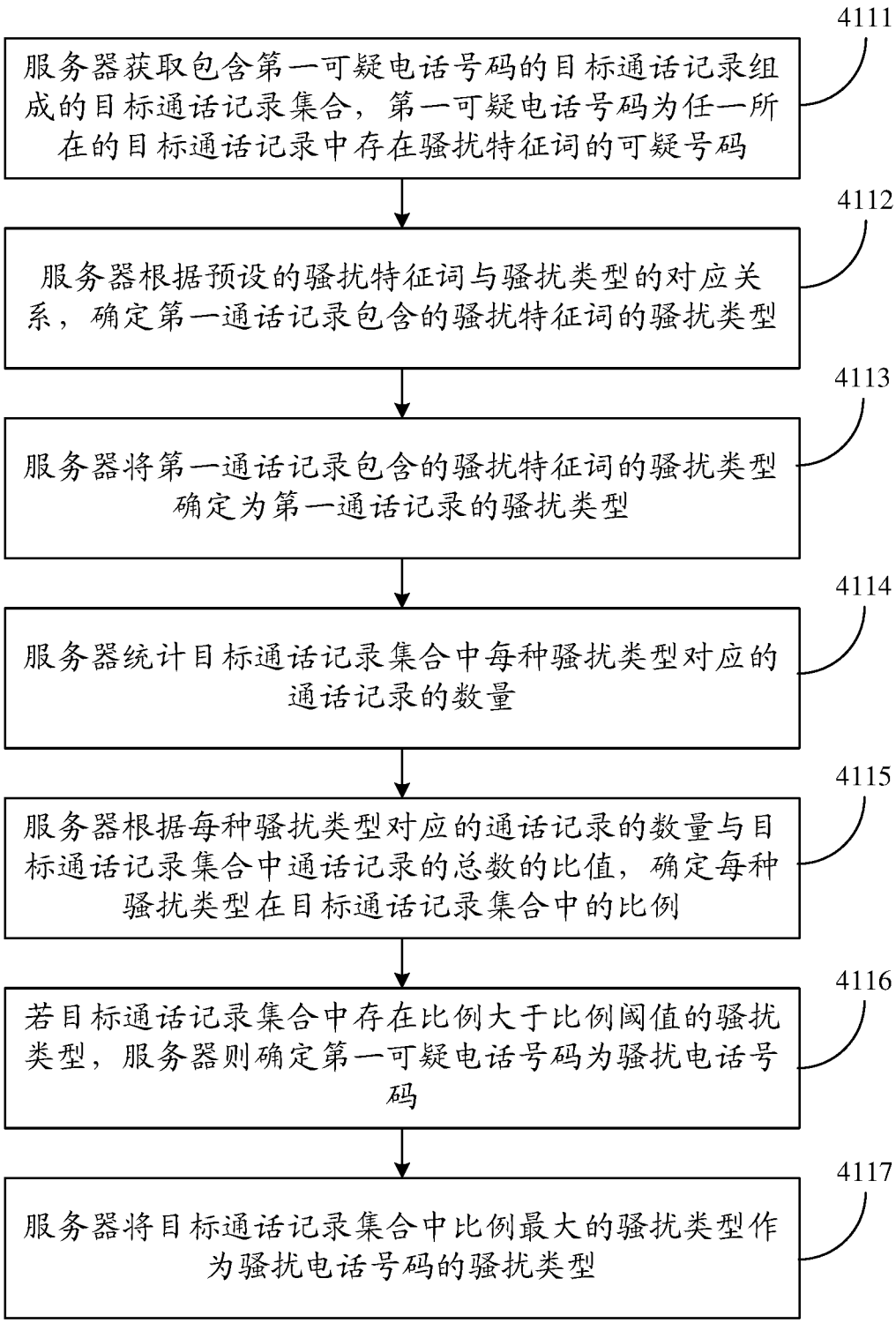

4111
服务器获取包含第一可疑电话号码的目标通话记录组成的目标通话记录集合，第一可疑电话号码为任一所在的目标通话记录中存在骚扰特征词的可疑号码
4112
服务器根据预设的骚扰特征词与骚扰类型的对应关系，确定第一通话记录包含的骚扰特征词的骚扰类型
4113
服务器将第一通话记录包含的骚扰特征词的骚扰类型确定为第一通话记录的骚扰类型
4114
服务器统计目标通话记录集合中每种骚扰类型对应的通话记录的数量
4115
服务器根据每种骚扰类型对应的通话记录的数量与目标通话记录集合中通话记录的总数的比值，确定每种骚扰类型在目标通话记录集合中的比例
4116
若目标通话记录集合中存在比例大于比例阈值的骚扰类型，服务器则确定第一可疑电话号码为骚扰电话号码
4117
服务器将目标通话记录集合中比例最大的骚扰类型作为骚扰电话号码的骚扰类型

图 4-4

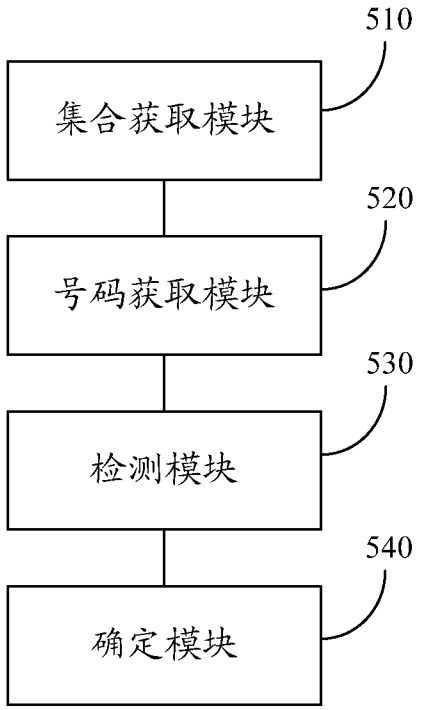

510
集合获取模块
520
号码获取模块
530
检测模块
540
确定模块

图 5-1

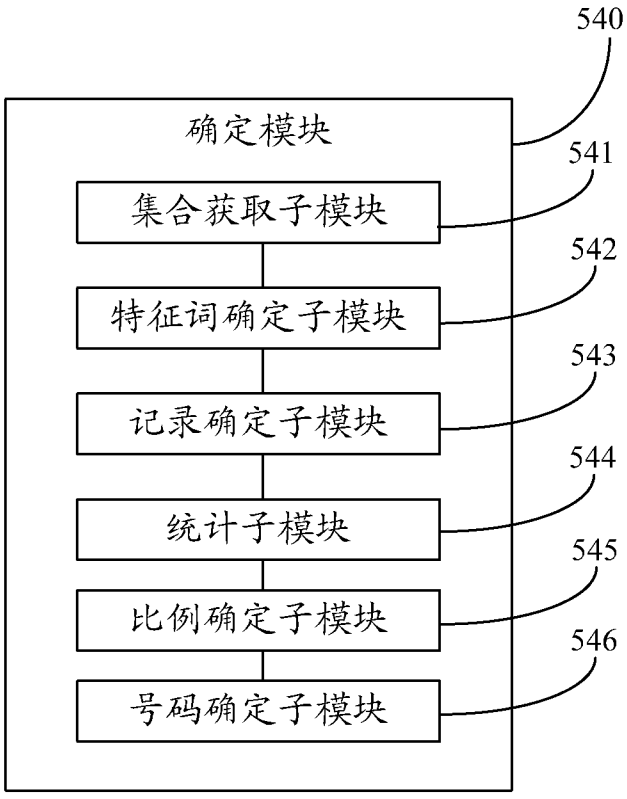

540
确定模块
541
集合获取子模块
542
特征词确定子模块
543
记录确定子模块
544
统计子模块
545
比例确定子模块
546
号码确定子模块

图 5-2

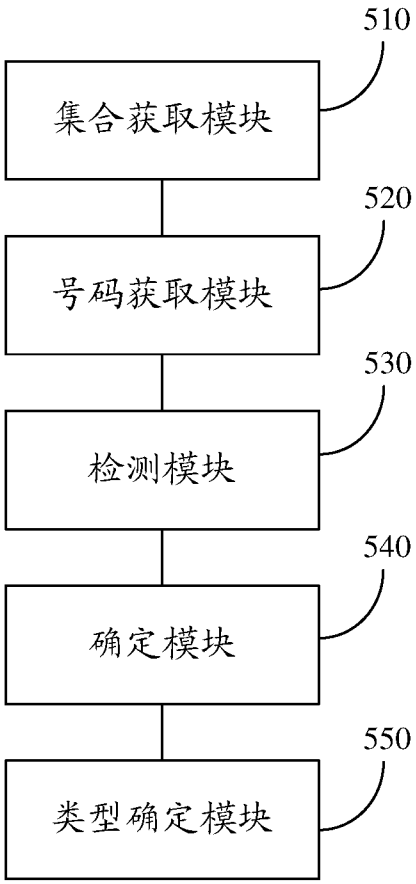


图 5-3

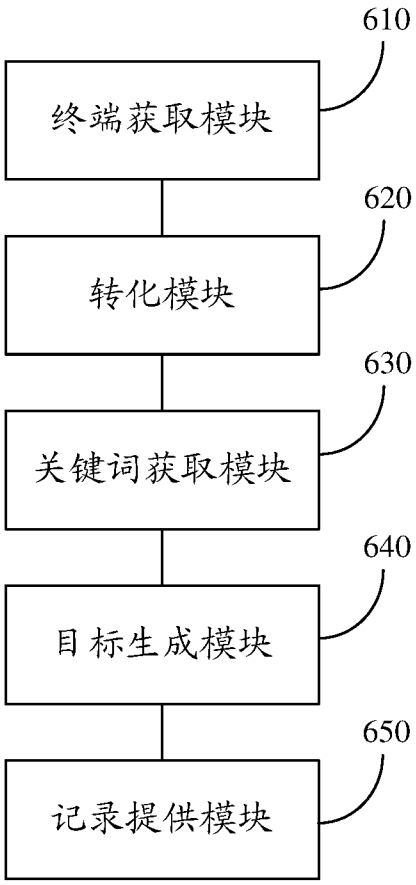


图 6-1

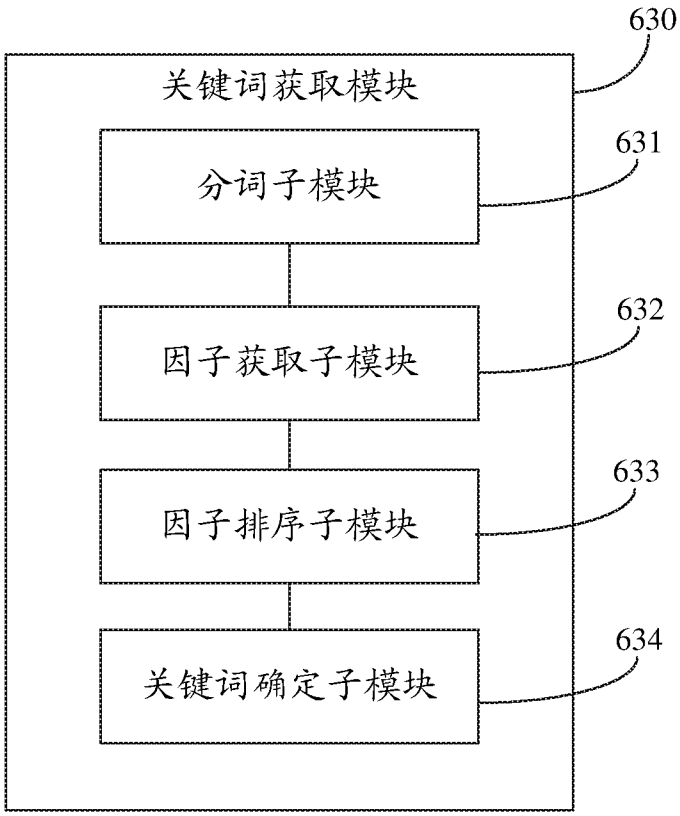


图 6-2

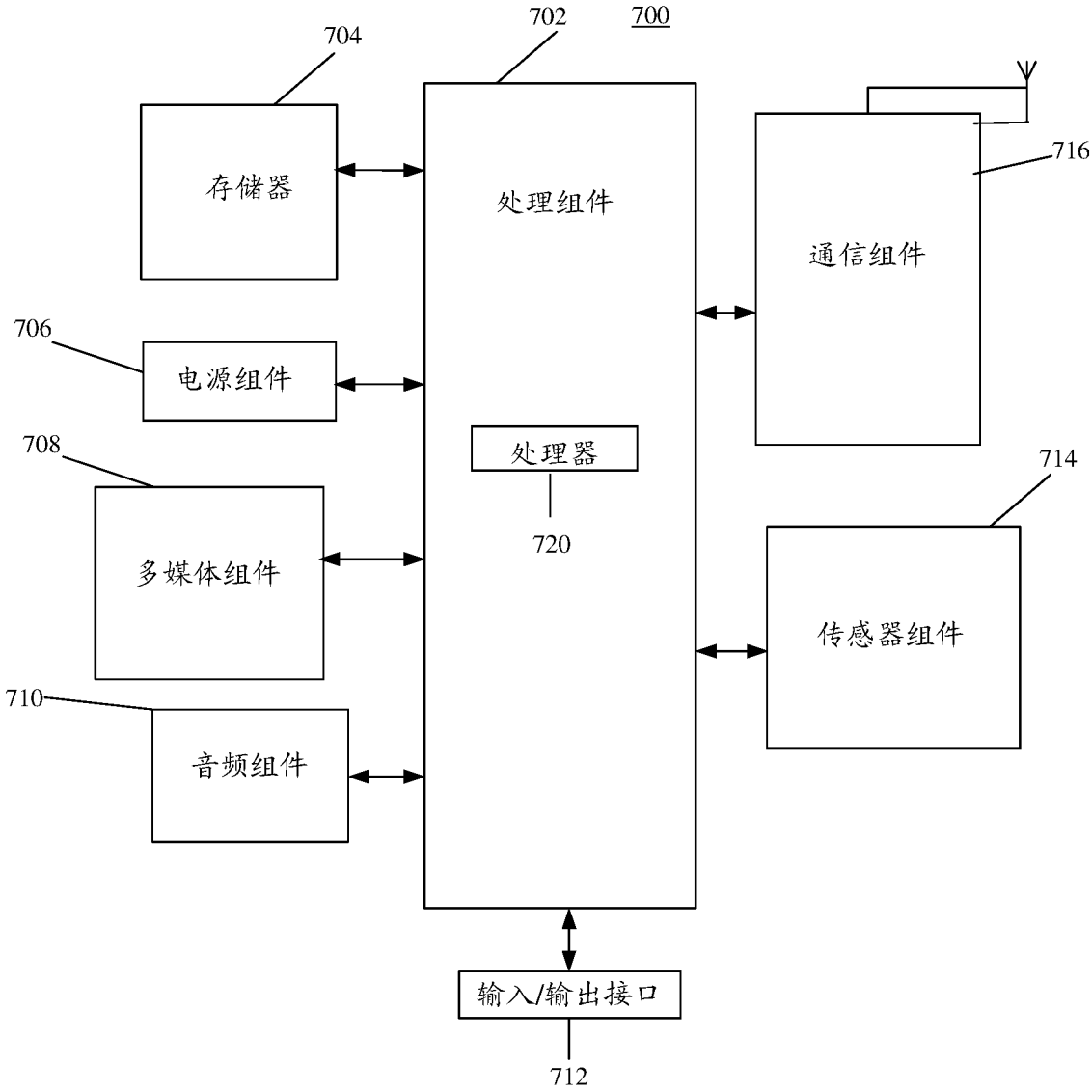

704
702
700
存储器
处理组件
716
通信组件
706
电源组件
处理器
720
708
多媒体组件
714
传感器组件
710
音频组件
输入/输出接口
712

图 7

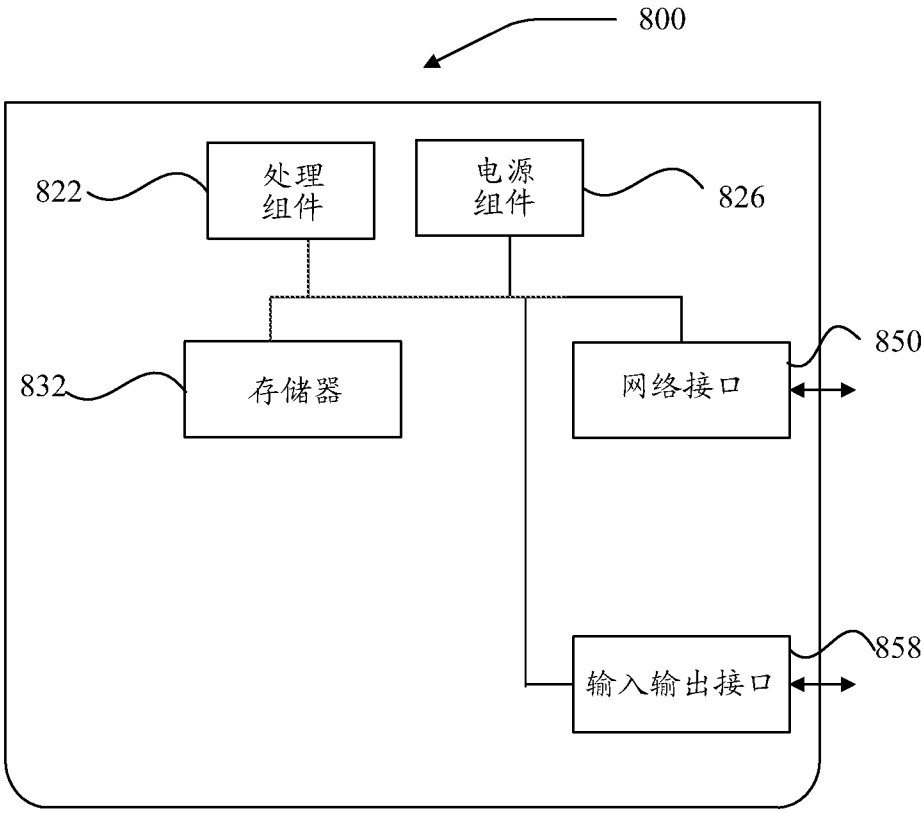

800
822
处理组件
电源组件
826
850
832
存储器
网络接口
858
输入输出接口

图 8

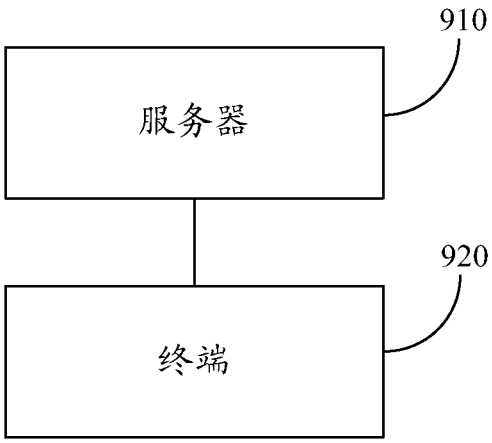

910
服务器
920
终端

图 9

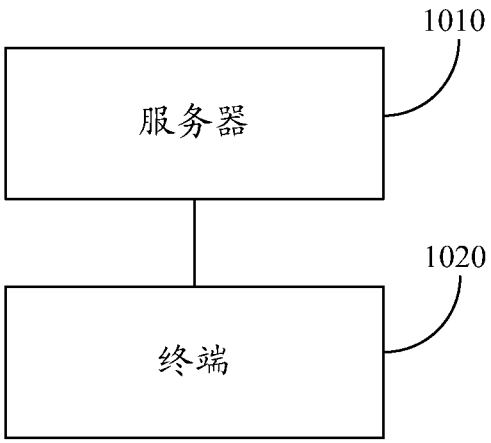


图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/097746

A. CLASSIFICATION OF SUBJECT MATTER

H04M 1/663 (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)

H04M, H04W

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)

CNKI, CNPAT, WPI, EPODOC: promote, intermediary, defraud, content, disturb+, harassment, telephone, number, blacklist, identify, detect, determin+, keyword?, text, context, voice, message, converse

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 105100366 A (XIAOMI TECHNOLOGY CO., LTD.), 25 November 2015 (25.11.2015), claims 1-20	1-20
A	CN 104580650 A (GUANG DONG OPPO MOBILE TELECOMMUNICATIONS CO., LTD.), 29 April 2015 (29.04.2015), description, paragraphs [0005]-[0010]	1-20
A	CN 104580725 A (GUANG DONG OPPO MOBILE TELECOMMUNICATIONS CO., LTD.), 29 April 2015 (29.04.2015), the whole document	1-20
A	CN 101472007 A (CHINA MOBILE COMMUNICATIONS CORPORATION et al.), 01 July 2009 (01.07.2009), the whole document	1-20

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 March 2016 (17.03.2016)	**24 March 2016 (24.03.2016)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **ZHAO, Xiaomin** Telephone No.: (86-10) **62414041**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/097746

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105100366 A	25 November 2015	None	
CN 104580650 A	29 April 2015	None	
CN 104580725 A	29 April 2015	None	
CN 101472007 A	01 July 2009	None	

国际检索报告

国际申请号
PCT/CN2015/097746

A. 主题的分类

H04M 1/663(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)

H04M, H04W

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）)

CNKI, CNPAT, WPI, EPODOC：骚扰、干扰、推销、中介、诈骗、电话、号码、黑名单、识别、检测、确定、关键词、文本、内容、语音、通话、disturb+、harassment、telephone、number、blacklist、identify、detect、determin+、keword?、text、context、voice、message、converse

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
PX	CN 105100366 A (小米科技有限责任公司) 2015年 11月 25日 (2015-11-25) 权利要求1-20	1-20
A	CN 104580650 A (广东欧珀移动通信有限公司) 2015年 4月 29日 (2015-04-29) 说明书第[0005]-[0010]段	1-20
A	CN 104580725 A (广东欧珀移动通信有限公司) 2015年 4月 29日 (2015-04-29) 全文	1-20
A	CN 101472007 A (中国移动通信集团公司 等) 2009年 7月 1日 (2009-07-01) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2016年 3月 17日	2016年 3月 24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 赵晓敏 电话号码 (86-10)62414041

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/097746

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	105100366	A	2015年 11月 25日	无	
CN	104580650	A	2015年 4月 29日	无	
CN	104580725	A	2015年 4月 29日	无	
CN	101472007	A	2009年 7月 1日	无	