(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108647201 A
(43)申请公布日 2018.10.12

(21)申请号 201810300929.4

(22)申请日 2018.04.04

(71)申请人 卓望数码技术(深圳)有限公司
地址 518000 广东省深圳市南山区高新技术产业园南区深港产学研基地大楼西座六楼南翼

(72)发明人 吴岳辉

(74)专利代理机构 广州嘉权专利商标事务所有限公司 44205
代理人 唐致明 洪铭福

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/30(2006.01)

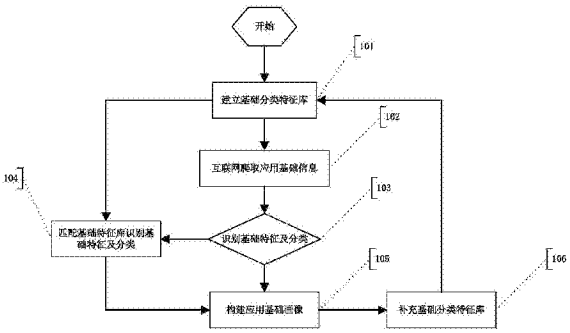
权利要求书1页 说明书5页 附图3页

(54)发明名称

一种基于移动应用的分类识别方法及系统

(57)摘要

本发明公开了一种基于移动应用的分类识别方法，其包括以下步骤：系统建立基础分类特征库；通过互联网爬取移动应用及应用页面中关键信息；系统识别所述关键信息中的分类信息及应用特征关键词；若识别到分类信息，则持续进行分类信息识别；若未识别到分类信息，则进入匹配基础特征库关键词识别；系统对采集到的应用特征关键词进行处理，获取到最优分类，并将新的分类结果补充至基础分类特征库。一种基于移动应用的分类识别系统，其包括：初始化控制模块、识别分类信息模块、匹配关键词模块。其提高了后续爬取到的移动应用分类识别效率和准确性，解决了现有应用分类的低效及无法识别的问题，可广泛应用于互联网应用领域。



CN 108647201 A

1.一种基于移动应用的分类识别方法，其特征在于，其包括以下步骤：

系统建立基础分类特征库；

通过互联网爬取移动应用及应用页面中关键信息；

系统识别所述关键信息中的分类信息及应用特征关键词；

若识别到分类信息，则持续进行分类信息识别；

若未识别到分类信息，则进入匹配基础特征库关键词识别；

系统对采集到的应用特征关键词进行处理，获取到最优分类；

构建应用基础画像，并将新的分类结果补充至基础分类特征库。

2.根据权利要求1所述的基于移动应用的分类识别方法，其特征在于，所述基础分类特征库的基础分类包括：社交类、影音类、游戏类。

3.根据权利要求1或2所述的基于移动应用的分类识别方法，其特征在于，所述步骤系统对采集到的应用特征关键词进行处理，获取到最优分类，其中通过最大公约算法获取最优分类。

4.根据权利要求3所述的基于移动应用的分类识别方法，其特征在于，所述方法还包括：

系统预先设定识别应用分类的默认映射关系；

获取所述爬取的应用描述信息；

通过分词组件获取描述信息中的分词结果，并剔除忽略词库中的内容；

采用最大匹配算法匹配基础分类特征词；

依据匹配阈值判断是否匹配成功，若匹配成功，则直接识别对应基础特征及分类；否则，系统进行映射策略选择。

5.根据权利要求4所述的基于移动应用的分类识别方法，其特征在于，所述映射策略选择包括发送无法识别分类通知至系统以进行手动映射，或自动映射为未识别分类。

6.根据权利要求5所述的基于移动应用的分类识别方法，其特征在于，所述匹配阈值包括设定命中词的个数。

7.根据权利要求6所述的基于移动应用的分类识别方法，其特征在于，当命中情况为非唯一命中或全未命中，则进行手动映射。

8.根据权利要求7所述的基于移动应用的分类识别方法，其特征在于，当自动映射为未识别分类，则系统持续进行分类识别，直至匹配完成。

9.一种基于移动应用的分类识别系统，其特征在于，其包括：

初始化控制模块，用于执行步骤系统建立基础分类特征库；

通过互联网爬取移动应用及应用页面中关键信息；

识别分类信息模块，用于执行步骤系统识别所述关键信息中的分类信息及应用特征关键词；

若识别到分类信息，则持续进行分类信息识别；

匹配关键词模块，用于执行步骤若未识别到分类信息，则进入匹配基础特征库关键词识别；

系统对采集到的应用特征关键词进行处理，获取到最优分类；

构建应用基础画像，并将新的分类结果补充至基础分类特征库。

一种基于移动应用的分类识别方法及系统

技术领域

[0001] 本发明涉及互联网应用领域，具体为基于移动应用的分类识别方法及系统。

背景技术

[0002] 在现有移动终端项目中，通常需要对应用大致分类，以便后续进行统计及识别。

[0003] 一般的分类方法是通过互联网爬取应用商城应用详情页面中的已知类别，该类别通常是对应于该商城本身需要所作的分类，分类标签也是各个商城不一致。对于分类有固定要求的系统，则会设置一些基础分类标签，如未识别，则会通过默认标签识别。

[0004] 然而，使用一般的识别方法，不足之处非常明显，具体在于：

[0005] 1、各应用商城分类不一致，导致后续类别标签高重复率；

[0006] 2、部分应用商城分类不明确，导致无法识别应用分类；

[0007] 3、对于赋予默认标签的应用，后续统计会出现分类偏差；

[0008] 如通过对现有项目中应用分类结果分析得知，正常爬取应用后，普遍出现分类不准确，包括：

[0009] 1、分类重复；

[0010] 2、分类无法识别；

[0011] 3、分类识别错误；

[0012] 4、分类不完整，多类型分类统计不准确。

[0013] 在传统的方法中，固定的认同各个应用商城中分类，导致很多分类不准确，例如百度应用商城中定义“社交通讯”，而在360应用商城中则叫“聊天工具”，且百度应用商城中对于“社交通讯”还有二级分类，即“聊天”、“社交”、“婚恋”、“通讯”，而360应用商城中对应二级分类则叫“社交聊天”、“网络电话”、“视频聊天”、“游戏语音”，因此对于如此多种且意义相近的分类进行只有系统分类处理，将需要一个持续的分析识别过程。为解决当前应用分类不够准确的问题，因此有必要提出一种新的移动应用持续标签识别方法。

发明内容

[0014] 为了解决上述技术问题，本发明的目的是提供一种基于移动应用的分类识别方法及系统。

[0015] 本发明所采用的技术方案是：

[0016] 本发明提供一种基于移动应用的分类识别方法，其包括以下步骤：

[0017] 系统建立基础分类特征库；

[0018] 通过互联网爬取移动应用及应用页面中关键信息；

[0019] 系统识别所述关键信息中的分类信息及应用特征关键词；

[0020] 若识别到分类信息，则持续进行分类信息识别；

[0021] 若未识别到分类信息，则进入匹配基础特征库关键词识别；

[0022] 系统对采集到的应用特征关键词进行处理，获取到最优分类；

[0023] 构建应用基础画像，并将新的分类结果补充至基础分类特征库。

[0024] 作为该技术方案的改进，所述基础分类特征库的基础分类包括：社交类、影音类、游戏类。

[0025] 作为该技术方案的改进，所述步骤系统对采集到的应用特征关键词进行处理，获取到最优分类，其中通过最大公约算法获取最优分类。

[0026] 作为该技术方案的改进，所述方法还包括：

[0027] 系统预先设定识别应用分类的默认映射关系；

[0028] 获取所述爬取的应用描述信息；

[0029] 通过分词组件获取描述信息中的分词结果，并剔除忽略词库中的内容；

[0030] 采用最大匹配算法匹配基础分类特征词；

[0031] 依据匹配阈值判断是否匹配成功，若匹配成功，则直接识别对应基础特征及分类；否则，系统进行映射策略选择。

[0032] 作为该技术方案的改进，所述映射策略选择包括发送无法识别分类通知至系统以进行手动映射，或自动映射为未识别分类。

[0033] 进一步地，所述匹配阈值包括设定命中词的个数。

[0034] 进一步地，当命中情况为非唯一命中或全未命中，则进行手动映射。

[0035] 进一步地，当自动映射为未识别分类，则系统持续进行分类识别，直至匹配完成。

[0036] 另一方面，本发明还提供一种基于移动应用的分类识别系统，其包括：

[0037] 初始化控制模块，用于执行步骤系统建立基础分类特征库；

[0038] 通过互联网爬取移动应用及应用页面中关键信息；

[0039] 识别分类信息模块，用于执行步骤系统识别所述关键信息中的分类信息及应用特征关键词；

[0040] 若识别到分类信息，则持续进行分类信息识别；

[0041] 匹配关键词模块，用于执行步骤若未识别到分类信息，则进入匹配基础特征库关键词识别；

[0042] 系统对采集到的应用特征关键词进行处理，获取到最优分类；

[0043] 构建应用基础画像，并将新的分类结果补充至基础分类特征库。

[0044] 本发明的有益效果是：本发明提供的基于移动应用的分类识别方法及系统，通过改进原有互联网应用分类识别的映射模式，设计了一套依据连续在互联网爬取并积累应用类别分词语义库，后续通过语义匹配和人工映射两种途径来构建应用分类的基础特征映射库，在匹配过程中通过最大匹配分词过程和匹配度阈值等机制来获得精确匹配结果，由此提高后续爬取到的移动应用分类识别效率和准确性，解决了现有应用分类的低效及无法识别的问题；且对于新爬取到的应用，为后续应用统计和报表提供了准确分类内容。

附图说明

[0045] 下面结合附图对本发明的具体实施方式作进一步说明：

[0046] 图1是本发明第一实施例的移动应用持续分类识别方法控制流程示意图；

[0047] 图2是本发明第二实施例的移动应用识别基础分类方法控制流程示意图；

[0048] 图3是本发明第三实施例的模块连接图。

具体实施方式

[0049] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

[0050] 参照图1，本发明提供一种基于移动应用的分类识别方法，其包括以下步骤：

[0051] 系统建立基础分类特征库；

[0052] 通过互联网爬取移动应用及应用页面中关键信息；

[0053] 系统识别所述关键信息中的分类信息及应用特征关键词；

[0054] 若识别到分类信息，则持续进行分类信息识别；

[0055] 若未识别到分类信息，则进入匹配基础特征库关键词识别；

[0056] 系统对采集到的应用特征关键词进行处理，获取到最优分类；

[0057] 构建应用基础画像，并将新的分类结果补充至基础分类特征库。

[0058] 作为该技术方案的改进，所述基础分类特征库的基础分类包括：社交类、影音类、游戏类。

[0059] 作为该技术方案的改进，所述步骤系统对采集到的应用特征关键词进行处理，获取到最优分类，其中通过最大公约算法获取最优分类。

[0060] 作为一具体实施例，其流程说明如下：

[0061] 101)综合互联网商城分类标签，建立分类标签基础库；

[0062] 102)通过后台爬虫服务持续爬取互联网移动应用，及应用详情页面中关键信息；

[0063] 103)通过分类识别模块识别关键信息中的分类信息及描述中应用特征关键词，如果识别到分类信息，则继续后续分类信息识别，否则进入匹配特征库关键词识别；

[0064] 104)采集到的特征关键词会通过最大公约算法获取到最优分类。其中，最大公约算法即最大字频优选法，其将采集的特征关键词拆分成单字，计算各字在基础特征库中的出现频率百分比值，略掉频率等于0的字，最终将结果值排序后取前100的字所对应的关键词作为最后的最优分类。对于匹配到多个分类结果的情况，将通过应用内部分析匹配已分类应用库，进一步筛选应用分类，例如通过应用包名等识别应用库，依据已匹配的历史来推断该应用分类；

[0065] 105)通过以上各项分类识别，大体构建该应用的分类标签，允许适配多个分类，通过后续基础库的不断完善，将不断提高分类结果；

[0066] 106)将新的分类结果补充至分类特征库。

[0067] 实际项目中按照以上流程，首先建立标准分类，如下表1所示：

[0068] 表1

[0069]

标准分类	特征关键词
社交类	社交、聊天、通讯、电话、美容
影音类	视频、语音、电台、铃声、娱乐、特效
游戏类	休闲、益智、养成、射击、模拟、竞速、棋牌

[0070] 后台爬虫服务通过互联网爬取到百度应用商城分类“社交通讯”；和360应用商城分类“社交网络”、“休闲娱乐”分类信息；

[0071] 通过标准库中特征关键词直接匹配到“社交通讯”和“社交网络”，因此建立该匹配关系，百度应用商城分类中的“社交通讯”和360应用商城分类中的“社交网络”下的所有应用将在爬取后属于标准分类中的“社交类”。

[0072] 系统在识别过程中，而未被直接识别出的“休闲娱乐”分类将通过策略配置是否采用人工映射，或者自动映射；如果采用人工映射，则表现在系统会发送提醒或通知管理员登录系统进行设定映射；而采用自动映射，则由系统暂时设定未知标签，等待系统基础特征库丰富后，定期再次进行识别。

[0073] 作为该技术方案的改进，参照图2，所述方法还包括：

[0074] 系统预先设定识别应用分类的默认映射关系；

[0075] 获取所述爬取的应用描述信息；

[0076] 通过分词组件获取描述信息中的分词结果，并剔除忽略词库中的内容；

[0077] 采用最大匹配算法匹配基础分类特征词；

[0078] 依据匹配阈值判断是否匹配成功，若匹配成功，则直接识别对应基础特征及分类；否则，系统进行映射策略选择。

[0079] 作为该技术方案的改进，所述映射策略选择包括发送无法识别分类通知至系统以进行手动映射，或自动映射为未识别分类。

[0080] 进一步地，所述匹配阈值包括设定命中词的个数。

[0081] 进一步地，当命中情况为非唯一命中或全未命中，则进行手动映射。

[0082] 进一步地，当自动映射为未识别分类，则系统持续进行分类识别，直至匹配完成。

[0083] 作为另一具体实施例，107)预先设定好基础的分类特征库及默认映射关系；

[0084] 108)获取由爬虫模块爬取的应用描述信息；

[0085] 109)通过分词组件获取描述信息中文分词结果，并剔除/忽略词库中的内容，包括介词、语气词、连接词等无意义词语；

[0086] 110)采用最大匹配算法匹配分词结果和基础分类特征词库，其中最大匹配算法，即将分词后的词组集合与分类特征词库中的词组集合一一匹配，获取匹配命中的词和命中次数，最后保留命中次数大于限定值的词组，初始设定限定值为1，后续不断积累，可逐步提高限定值，以便后续更精确匹配；

[0087] 111)依据匹配阈值判断是否有匹配结果；如果匹配成功，则直接识别对应基础特征分类映射结果，否则，将无法识别分类通知给系统管理人员以进行手动映射；

[0088] 112)直接通过基础分类特征库的映射结果识别为分类结果；

[0089] 113)系统依据配置的策略选择未识别分类的后续流程；

[0090] 114)人为设定分类结果，该流程由设定分类策略约束，默认为发送提醒和通知策略；

[0091] 115)自动设定为未识别分类，在后续定时任务再次匹配109步骤。

[0092] 以上流程中需要预先对爬取的大段文本分词，设定分词最大词长如设为4，这个长度主要考量了关键特征库中的定义词长，依次计算对应各标准分类特征库的逆向最大匹配算法将360应用商城中的“美图秀秀”中描述内容分词获得有效结果如下表2：

[0093] 表2

[0094]

标准分类	特征关键词	匹配命中词
社交类	社交、聊天、通讯、电话、美容、照片	美容、照片
影音类	视频、语音、电台、铃声、娱乐、特效	特效
游戏类	休闲、益智、养成、射击、模拟、竞速、棋牌	无命中词

[0095] 提前设定的匹配阈值为2，则命中词超过两个，则视为有效命中，如将“美图秀秀”分类为“社交类”，后续设置的关键词越多，命中的结果会增加，则需要调节阈值到合适值，剔除掉命中低的无效分类匹配结果。

[0096] 对于非唯一命中和全未命中的情况，均需要人工干预处理，但有不同的处理策略配置，可以发邮件通知，也可以设置默认分类等，或者设置默认分类后再通知人工干预。

[0097] 参照图3，本发明还提供一种基于移动应用的分类识别系统，其包括：

[0098] 初始化控制模块，用于执行步骤系统建立基础分类特征库；

[0099] 通过互联网爬取移动应用及应用页面中关键信息；

[0100] 识别分类信息模块，用于执行步骤系统识别所述关键信息中的分类信息及应用特征关键词；

[0101] 若识别到分类信息，则持续进行分类信息识别；

[0102] 匹配关键词模块，用于执行步骤若未识别到分类信息，则进入匹配基础特征库关键词识别；

[0103] 系统对采集到的应用特征关键词进行处理，获取到最优分类；

[0104] 构建应用基础画像，并将新的分类结果补充至基础分类特征库。

[0105] 本发明提供的基于移动应用的分类识别方法及系统，通过改进原有互联网应用分类识别的映射模式，设计了一套依据连续在互联网爬取并积累应用类别分词语义库，后续通过语义匹配和人工映射两种途径来构建应用分类的基础特征映射库，在匹配过程中通过最大匹配分词过程和匹配度阈值等机制来获得精确匹配结果，由此提高后续爬取到的移动应用分类识别效率和准确性，解决了现有应用分类的低效及无法识别的问题；且对于新爬取到的应用，为后续应用统计和报表提供了准确分类内容。

[0106] 以上是对本发明的较佳实施进行了具体说明，但本发明创造并不限于所述实施例，熟悉本领域的技术人员在不违背本发明精神的前提下还可做出种种的等同变形或替换，这些等同的变形或替换均包含在本申请权利要求所限定的范围内。

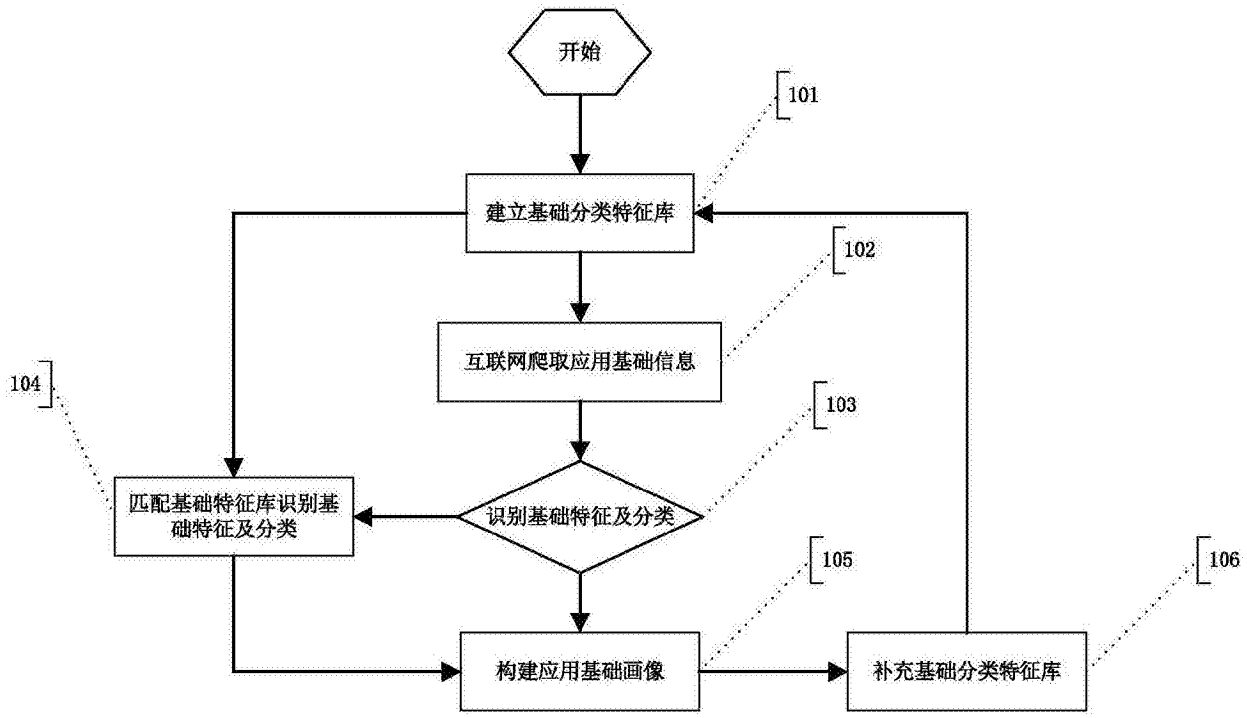

开始
101
建立基础分类特征库
102
互联网爬取应用基础信息
103
识别基础特征及分类
104
匹配基础特征库识别基础特征及分类
105
构建应用基础画像
106
补充基础分类特征库

图1

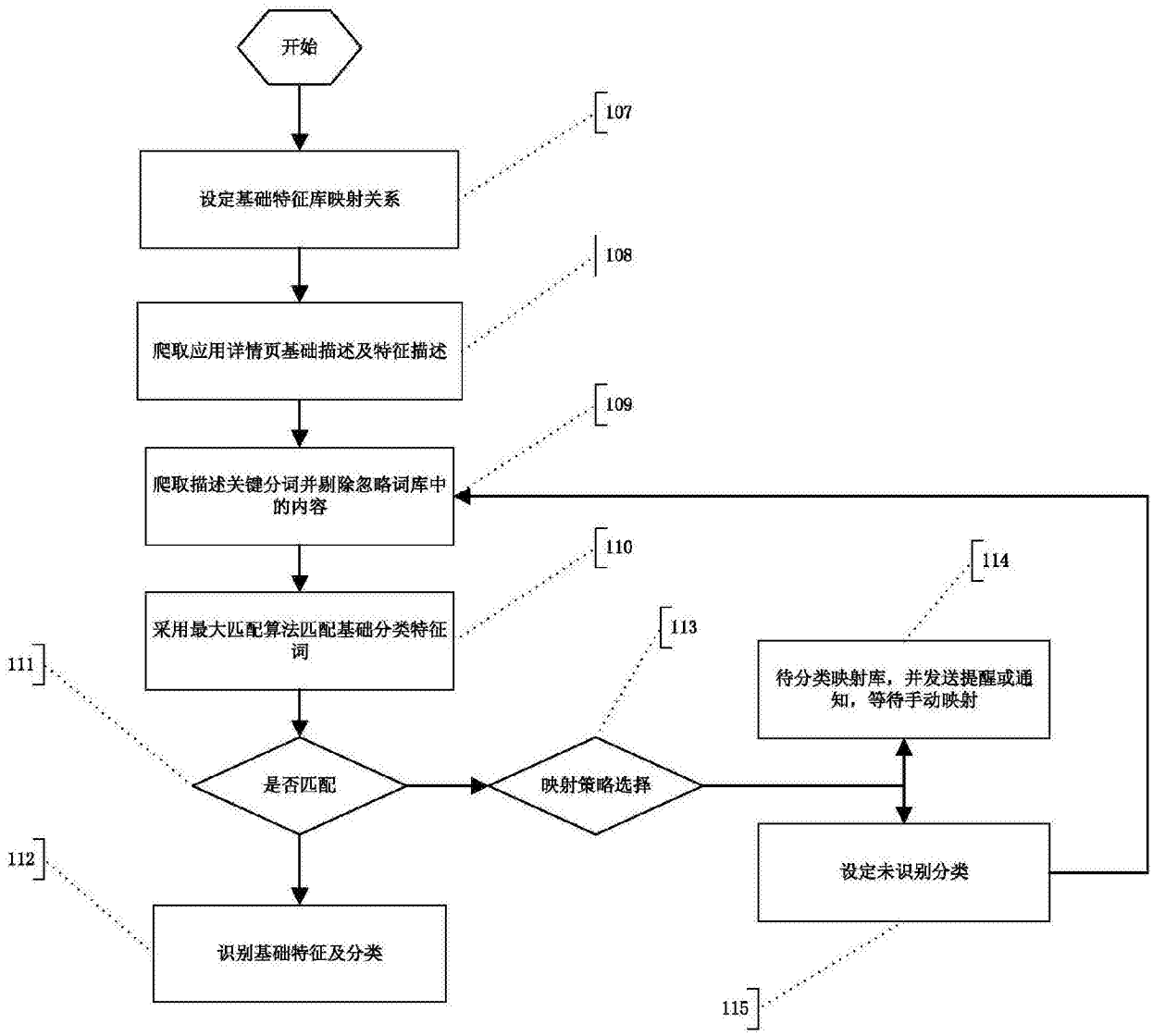

开始
107
设定基础特征库映射关系
108
爬取应用详情页基础描述及特征描述
109
爬取描述关键分词并剔除忽略词库中的内容
110
采用最大匹配算法匹配基础分类特征词
111
是否匹配
112
识别基础特征及分类
113
映射策略选择
114
待分类映射库，并发送提醒或通知，等待手动映射
设定未识别分类
115

图2

图3