手机应用网络流量分析与恶意行为检测研究

苏欣　刘绪崇　张悦　童宇　赵薇　著

湖南大学出版社

Android
手机应用网络流量分析与恶意行为检测研究

苏欣　刘绪崇　张悦　童宇　赵薇　著

湖南大学出版社

内 容 简 介

本著作在总结过去研究工作的基础上阐述 Android 手机应用网络流量分析与恶意行为检测的基础理论、技术方法及应用实践。本著作从“基础理论—技术方法—应用实践”的角度来构建 Android 手机应用网络流量分析与恶意行为检测技术框架。本著作适合具有一定 Android 应用开发、网络安全知识背景的专业技术人员阅读。

图书在版编目（CIP）数据

Android 手机应用网络流量分析与恶意行为检测研究/苏欣，刘绪崇，张悦，童宇，赵薇著．—长沙：湖南大学出版社，2016.10

ISBN 978-7-5667-1217-2

Ⅰ.①A… Ⅱ.①苏… Ⅲ.①移动电话—计算机网络—流量—研究 ②移动电话机—计算机网络—网络安全—研究

Ⅳ.①TN929.53

中国版本图书馆 CIP 数据核字（2016）第 237975 号

Android 手机应用网络流量分析与恶意行为检测研究

Android SHOUJI YINGYONG WANGLUO LIULIANG FENXI YU EYIXINGWEIJIANCEYANJIU

作　　者：苏　欣　刘绪崇　张　悦　童　宇　赵　薇　**著**
责任编辑：刘　旺　**责任校对**：全　健
印　　装：虎彩印艺股份有限公司
开　　本：710×1000　16 开　**印张**：13.75　**字数**：212 千
版　　次：2016 年 10 月第 1 版　**印次**：2016 年 10 月第 1 次印刷
书　　号：ISBN 978-7-5667-1217-2
定　　价：38.00 元

出 版 人：雷　鸣
出版发行：湖南大学出版社
社　　址：湖南·长沙·岳麓山　　**邮　　编**：410082
电　　话：0731-88822559(发行部),88821174(编辑室),88821006(出版部)
传　　真：0731-88649312(发行部),88822264(总编室)
网　　址：http://www.hnupress.com
电子邮箱：liuwangfriend66@126.com

序 言

近几年，Android 在国内的发展极其迅猛，这除了相关产品具备强大的功能与丰富的应用外，更是因为它优良的性能表现吸引着用户。2015 年，Android 设备的出货量已经达到 13.71 亿台，研究机构预计在 2016 年底，Android 设备的数量将超过 Windows 电脑。与此同时，Android 应用的发展也十分迅猛，在 2016 年第一季度，Android 应用市场的应用下载数量达到了 172 亿个，同比增长 8.2%。正如我们从生活中强烈感受到的那样，移动计算显然是信息技术领域近年来得到巨大发展的一个热点方向。现在，各种移动设备已经无时无刻不伴随在我们的左右，我们花在移动设备上的时间要比花在电脑上的时间多得多：办公用的电脑在下班后就会被我们遗弃在办公桌上，而家里的电脑在我们早上急匆匆去上班时甚至没有打开的机会，这种变化是前所未有的。与电脑不同的是，我们的移动设备始终是保持开机的，而且连接着工作与家庭这两个世界，因此也成为了坏人们眼中更具价值的攻击目标。

目前市面上有许多关于 Android 应用安全攻防的书籍。但是这些书籍主要关注于逆向工程、代码分析，几乎没有书籍关注如何从网络流量和大数据的角度来分析 Android 应用及 Android 恶意应用具有的恶意行为，并如何去检测。而本著作写作的主要目的正是从网络流量和大数据的角度出发来说明 Android 应用在网络流量方面的行为特征，并根据这些行为特征来区分 Android 应用中存在的恶意行为。

本文从专业的研究人员角度出发，分三部分展示了 Android 手机应用网络流量分析与恶意行为检测的方法。在第一部分，作者首先介绍了

Android 手机应用研究的背景、意义以及相关技术。在第二部分，作者系统地介绍了 Android 手机应用网络流量生成方式，如何通过网络流量识别 Android 应用以及在网络流量数据中挖掘有价值的数据进行 Android 应用推荐。在第三部分，作者介绍了两种检测 Android 恶意行为的方法，分别是基于分类算法和基于深度学习的检测方法，并对 Android 平台上的恶意应用、不安全广告库和移动僵尸网络进行检测。

本文面向有志于从事 Android 应用研究的专业人员。读者应该具备计算机操作系统、计算机网络、数据挖掘、机器学习等预备知识。作者在文后都给出了详细的参考文献，读者可以找到必要的背景知识。

本文涉及的内容非常广泛、专业，而且融入了作者多年的 Android 应用研究经验。由于作者的知识和经验有限，著作中不当甚至错误之处在所难免，诚恳期待广大读者提出宝贵意见。

苏　欣

2016 年 5 月 30 日

前 言

Android 应用是在 Android 操作系统上进行设计的应用程序，适用于不同类型的 Android 移动设备，比如，智能手机、平板电脑等。它为移动用户在日常生活等方面提供丰富多样的功能。随着安卓（Android）手机的快速发展，Android 手机应用的数量呈指数增加。Android 手机产生的网络流量增长速度迅猛，Android 手机及其相应的应用所生成的网络流量所带来的新情况逐渐成为国内外的研究热点。与此同时，由于 Android 系统和应用具有开源等特点，恶意应用制造者通过嵌入恶意代码来执行恶意行为以达到他们非法的目的。如何让用户了解 Android 应用的网络流量和检测 Android 应用的恶意行为也成为国内外的研究热点。因此本文围绕 Android 应用网络流量分析和恶意行为检测这两大问题展开深入研究，主要工作及创新点如下：

（1）Android 应用网络流量自动生成技术。如果要对 Android 应用网络流量进行分析，网络流量数据是必不可少的。但是运营商的网络流量数据包含大量的用户隐私数据导致难以获取。手动执行 Android 应用产生网络流量的方式不适合于大量 Android 应用的情况。为了解决前述问题，该工作设计并实现一种 Android 应用网络流量自动生成系统 AndroGenerator。该系统首先设计一种自动执行的方法来自动执行大量的 Android 应用，并采集应用产生的网络流量；随后提取网络流量的属性，比如数据包个数、字节数等；再根据采集到的流量数据的特征和模式来模拟生成 Android 应用的网络流量。实验结果表明，第一，该系统的自动执行应用算法可以触发 Android 应用的大部分的网络行为；第二，该系统模

拟生成的 Android 应用网络流量与实际环境中的 Android 应用网络流量具有较高的相似度，可以为 Android 应用网络流量分析研究工作提供数据。

(2) 融入网络流量开销因素的 Android 应用推荐方法。大部分的 Android 应用完成其功能需要访问网络从而产生网络流量，并带来手机流量套餐资费的消耗。目前主流的 Android 应用市场在推荐应用的时候主要考虑 Android 应用的流行度(比如应用评分等)，而忽略了该应用产生的网络流量开销。因此，如何使用户所下载的应用同时具备高流行度和低网络流量开销是本工作要解决的问题。本工作首先从 Android 官方市场中的 22 个主流的应用类别分别下载排名前 100 的应用。其次，采集这 2 200 个应用运行时产生的网络流量，并从每秒流量开销、不同类型的流量开销、不同类型流量所占的比例等几个方面来测量每个应用的流量开销情况。在完成测量后，根据得到的测量结果提出了一个基于网络流量开销的应用推荐技术。该技术可以和现有的 Android 市场的推荐算法相结合来为用户推荐不仅具有较高的流行度，同时节省网络流量开销的 Android 应用。

(3) HTTP 流的 Android 应用识别方法。为了解决已有的方法在识别 Android 应用上存在的识别准确率低、识别不够全面的问题，本工作提出一种从 Android 应用产生 HTTP 流中提取特征签名的识别方法。不同于传统的网络特征签名，该特征签名不仅包括具有特殊字符串的 HTTP 流，也包括具有普通字符串的 HTTP 流。该方法首先识别具有特殊字符串的 HTTP 流，再通过时间窗口和恢复 HTTP 响应流中的压缩内容的方法来关联具有普通字符串的 HTTP 流。随后从这些 HTTP 流中提取的 HTTP 特征签名来识别网络中运行的 Android 应用，并统计这些 Android 应用产生的网络流量大小。实验结果表明该方法的 Android 应用识别率相比于其他已有的方法的识别率提高了 35%～81%。

(4) 基于 HTTP 流挖掘技术的 Android 恶意应用和不安全广告库检测方法。根据已有研究工作，发现大部分的 Android 应用及其嵌入的广告库与服务器之间的通信都是基于 HTTP 协议。根据这个现象，本工作

提出针对 Android 应用和广告库产生的 HTTP 流进行属性挖掘并检测 Android 恶意应用和不安全广告库。该方法首先分析 Android 正常应用和 Android 恶意应用、安全广告库和不安全广告库之间的在 HTTP 流量属性上的区别，并分析 HTTP 流量属性和恶意行为之间的关联；其次，根据比较后得到正常应用和恶意应用，安全广告库和不安全广告库的 HTTP 流量属性，使用分类算法建立分类模型来对 Android 恶意应用和不安全广告库进行分类检测；最后，从检测到的 Android 恶意应用和不安全广告库中提取 HTTP 指纹特征来进一步归类 Android 恶意应用和不安全广告库。实验结果表明，该方法在检测 Android 恶意应用和不安全广告库分别达到 97.67%和 95.86%的准确率，并可以对检测到的恶意应用和不安全广告库进行归类。

(5) 基于深度学习的 Android 应用行为特征选择方法。深度学习算法是一种特征学习算法，属于机器学习算法的一个分支。它采用深信度(DBN)网络架构，从原理上看是特征的“质变”，DBN 的输出是输入特征的另一种表达。通过合理设计 DBN 网络结构可以使输出特征维度远小于输入维度，在输出集合上运用普通的机器学习分类算法便可以实现恶意应用检测。在实验中，采用上述同样数据集，最优准确率达到 98.3%，并且在开放测试中得到 99.4%的准确率和召回率。

(6) 移动僵尸网络检测方法。移动僵尸网络 (Mobile Botnet) 是一种从传统僵尸网络(Botnet)进化而来的新型网络攻击方式，为黑客提供了隐匿、灵活且高效的一对多命令与控制信道 (Command and Control channel, C&C) 机制，可以控制大量僵尸主机实现信息窃取、分布式拒绝服务攻击和垃圾邮件发送等攻击目的。该工作提出一种与移动僵尸网络结构和 C&C 协议无关，不需要分析数据包的特征负载的移动僵尸网络检测方法。该方法首先使用预过滤规则对捕获的流量进行过滤，去掉与移动僵尸网络无关的流量；其次对过滤后的流量属性进行统计并提取网络流量属性；接着使用基于 umerged X－means 聚类算法的两步聚类方法对包含了 C&C 信道的流量和其他正常程序的网络流量的混合数据集进行分析与聚类，从而达到对移动僵尸网络检测的目的。在实验阶段，该

方法在检测移动僵尸网络的准确率可以达到 98.34%。

综上所述，本文首先通过设计 AndroGenerator 来生成 Android 应用的网络流量；然后在网络流量数据的基础上展开一系列的关于 Android 应用网络流量的分析和恶意行为的检测，比如 Android 应用识别、Android 应用网络流量测量、Android 恶意应用和不安全广告库的检测等，从而让用户对使用的 Android 应用的网络行为有更深刻的认识，同时保证了 Android 用户在使用 Android 应用的安全性。本文得到网络侦查技术湖南省重点实验室开放研究基金(2016WLFZZC008)，网络犯罪侦查湖南省普通高等学校重点实验室，湖南省教育厅优秀青年项目(16B085)，湖南警察学院院局合作项目(2015YJHZ06)的资助与支持。

目 次

第 1 章 绪 论

1.1 研究背景 …… 1
1.2 研究意义 …… 4
1.3 本文的主要研究内容 …… 5
1.4 组织结构 …… 9

第 2 章 Android 系统及其应用的相关技术介绍

2.1 Android 综述 …… 11
2.2 Android 应用流量分析关键技术 …… 23
2.3 Android 应用恶意行为检测关键技术 …… 27
2.4 小结 …… 32

第 3 章 Android 应用网络流量自动生成方法研究

3.1 引言 …… 33
3.2 问题描述 …… 35
3.3 AndroGenerator 概述 …… 36
3.4 Android 应用自动执行组件设计 …… 37
3.5 流量解析组件设计 …… 44
3.6 Android 网络流量生成器组件设计 …… 47
3.7 实验 …… 49
3.8 小结 …… 60

第 4 章 Android 应用流量开销测量及推荐算法研究

4.1 引言 …… 62
4.2 Android 应用网络流量开销测量分析 …… 64
4.3 基于网络流量开销考虑的 Android 应用推荐方法 …… 72

4.4 实验 …… 82
4.5 小结 …… 93
第 5 章 基于 HTTP 特征签名的 Android 应用识别方法
5.1 引言 …… 95
5.2 基于网络流量的 Android 应用识别方法存在的问题 …… 97
5.3 HTTP 特征签名提取 …… 100
5.4 实验 …… 107
5.5 小结 …… 115
第 6 章 HTTP 流挖掘技术的 Android 恶意行为检测
6.1 引言 …… 117
6.2 Android 恶意行为载体定义与分析 …… 119
6.3 HTTP 流量与恶意行为关联分析 …… 120
6.4 Android 恶意应用与不安全广告库检测 …… 138
6.5 基于 HTTP 指纹相似度的归类算法 …… 140
6.6 实验 …… 142
6.7 小结 …… 147
第 7 章 基于深度学习的 Android 应用行为特征学习
7.1 深度学习介绍 …… 148
7.2 基于深度信念网络的特征学习方法 …… 162
7.3 方法实现 …… 164
7.4 实验评估 …… 166
7.5 小结 …… 172
第 8 章 基于命令与控制通信信道的移动僵尸网络检测
8.1 引言 …… 174
8.2 相关工作 …… 176
8.3 检测方法 …… 178
8.4 实验 …… 183
8.5 小结 …… 187

结 论 …… 188
参考文献 …… 193

第1章 绪 论

本章首先阐述了本文的研究背景和意义，其次归纳和概括了本文的主要工作，最后介绍本文的组织结构。

1.1 研究背景

随着宽带无线接入技术与移动终端技术的快速发展和融合，移动互联网呈爆炸式发展，改变了人们的生活方式及思维方式。人们迫切地希望能够通过手机及时快捷地使用移动应用，如即时通信、手机游戏、视频应用、位置服务、铃声下载、邮件收发、移动音乐等。在这样的情况下，移动互联网迎来机遇，并飞速发展。

目前流行的手机操作系统主要包括 Android、iOS 和 Windows。在 2014 年第二季度，我国的智能设备的普及数量达到了 4 亿台。其中，基于 Android 系统的智能设备的市场占有率从 2011 年第一季度的 14%增长到 2014 年第二季度的 79%，高居市场第一位[1]，如图 1.1 所示。Android 操作系统是谷歌设计的开源平台，面世以来受到了手机界史无前例的欢迎，众多手机生产厂商都以 Android 系统为平台进行智能手机的开发与生成，这使得 Android 操作系统占据智能手机操作系统市场的第一份额[2]。该系统基于开源的 Linux 操作系统进行开发，底层使用了 C 或 C++语言提高硬件访问速度，应用层采用了简单强大的 Java 语言。

与 Android 智能设备快速发展相对应的是 Android 应用的层出不

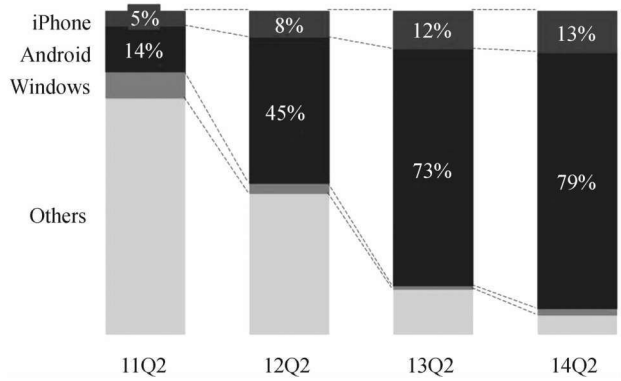

图 1.1　智能手机用户分布

穷。这些应用功能丰富，涵盖了游戏、音乐、即时通信、邮件、视频、办公等多个和人们日常生活、娱乐、工作相关的方面。图 1.2 中列出了 Google Play[3]、Apple app store[4]等几个流行的移动应用市场的应用数量。其中，在 2014 年 Google Play 的应用数目已经超过了 100 万，成为提供移动应用数量最多的市场。这些数据体现了 Android 智能设备及其应用发展的迅速。

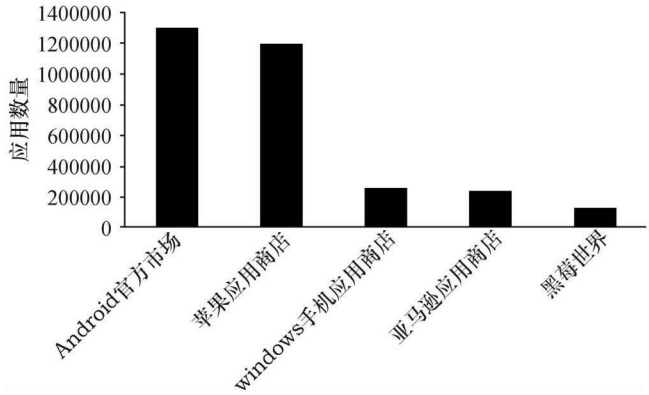

图 1.2　各种应用商店应用数目分布图

然而 Android 手机及其相应的应用在飞速发展的同时，许多的新的问题也相应出现：

(1)Android 应用流量猛增

大多数 Android 应用需要访问网络来实现应用功能，在使用应用时会产生一定的流量数据。用户在移动智能设备上观看视频、下载歌曲、玩在线游戏、收发邮件、视频语音聊天等，都会产生一定的网络流量数据。随着 Android 手机的普及和应用的流行，Android 应用所生成的网络流量逐年增加。图 1.3 展示了从 2008—2013 年全球移动应用产生的流量数据。

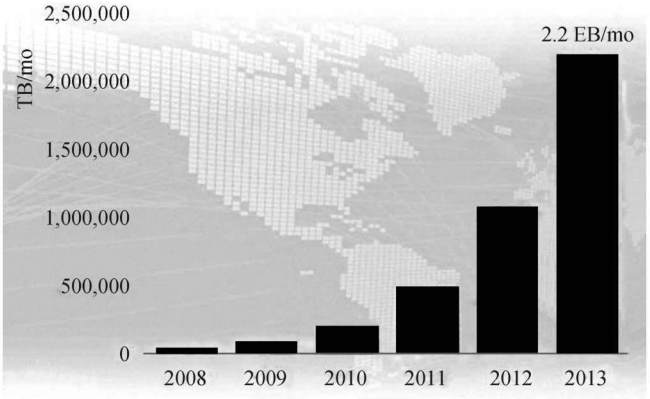

图 1.3 全球移动应用流量数据(2008—2013)

相比于 2008 年，2013 年的移动应用产生的流量增长了 66 倍，预计在 2015 年将达到 75EB 的数据量[5]。与此同时，为了抢占移动互联网这个市场，越来越多的公司也推出自己的移动应用来适应飞速发展的移动互联网。移动应用产生的网络流量猛增带来了新的问题：第一，随着移动智能设备的普及，越来越多的移动智能设备接入到企业、机关单位等大型的局域网里面，如何识别运行在网内的移动应用、统计其产生的流量大小是网络管理员急需解决的问题；第二，移动应用产生的网络流量的快速增长导致移动用户手机流量套餐消耗的增长，如何让移动用户了解自己使用的移动应用的流量开销情况，并保证移动用户可以从移动应用商店下载到满足功能的并消耗较少的流量套餐的应用，这对于提升移动应用的用户体验，防止流量套餐被恶意扣费有着重要的意义；第三，移动应用的网络流量中也包含恶意流量，如何检测这些恶意流量从而达到检测移动

恶意应用，保护移动用户的安全的目的。

(2)安全隐患

随着 Android 手机及其相应应用的流行，人们把生活工作各个方面交给了 Android 手机及其各种应用。这使得 Android 手机及其应用成为了用户隐私信息最为丰富的平台。越来越多的恶意应用制造者都瞄准了这个平台。

恶意应用是拥有恶意行为的应用的统称，包括僵尸程序、间谍应用、特洛伊木马等。2004 年 6 月，在塞班手机上出现的 Cabir 病毒是第一种手机病毒[6]。2011 年 3 月，Android 市场中出现一类似窃取手机 IMEI、手机的 Android 版本等信息为目的的 Android 恶意应用，在被 Android 市场下架之前，有超过 200 万的 Android 手机下载该类应用[7]。根据百度安全实验室报告[8]，2014 第二季度新增的恶意应用数量超过去年同期 41%，Android 恶意应用占移动恶意应用的比例超过了 90%。同时该报告还指出恶意扣费、泄露敏感信息、远程控制等恶意行为为 Android 恶意应用的主要表现形式。甚至有些恶意应用使用多种恶意手段结合来实现恶意应用制造者的不法目的，使得 Android 恶意应用的检测变得更加困难。

1.2 研究意义

不同于传统的个人电脑上的应用，移动应用具有两个特点：第一，应用种类多样化。比如 Google Play 提供了 31 种不同种类的应用供用户下载；第二，大部分 Android 应用利用 HTTP 和外部服务器建立连接[9]。因此，传统的流量识别算法不适用于识别移动应用，比如协议识别[10]、特征载荷匹配[11]、数据挖掘算法[12]等。同时，这些流量数据都需要用户进行支付，但是用户却对在使用移动应用时到底生成了什么样的流量，哪些流量是必要的，哪些是由第三方库所产生的并不清楚。一些移动恶意应用通过网络来与远程服务器进行通信，接收来自服务器的命令来攻击局

域网内的其他未被感染的移动设备。所以准确地识别这些应用对于网络管理、流量计费、维护网络安全等工作必不可少。

Android智能手机作为目前全球智能手机市场占有率排名第一的智能手机，且Android应用本身具有可反编译、添加代码、再重新编译的特点，遭到了越来越多的恶意应用制造者的青睐，其恶意应用制造的过程相比于iOS更加简单快捷。Android恶意应用会对手机和用户造成各种各样的严重危害，主要包括：

①恶意扣费：伪装成免费应用，在用户不知情的情况下误导用户付费；或者是故意颠倒“同意付费”和“不同意付费”的按键选项，恶意误导用户付费。

②窃取用户隐私和机密：恶意软件通过网络或短信方式窃取用户的通讯录，通讯记录，手机号码，邮件，地理位置，手机中已安装的软件，各种账号、密码等隐私资料。

③后台下载应用：在用户不知情的情况下载应用；或者以版本需要更新的提示或者其他提示诱骗用户下载恶意应用。

④远程控制：黑客利用恶意应用远程控制用户手机，同时在手机后台隐秘地进行恶意行为，比如安装未经许可的应用、卸载杀毒应用、发送短信、拨打电话等。

因此，如何有效准确地分析Android应用所产生的流量以及对其恶意行为进行检测成为了急需解决的问题。本文的研究受到国家重点基础研究发展计划“973未来互联网业务流量与网络拓扑模型”(2012CB315805)子课题，国家自然科学基金“基于多核处理器的高性能深度数据包检测技术研究”(61173167)等项目支持。在这些项目的资助下，研究取得良好进展，相关研究成果已经在这些项目中得到应用并发挥着积极作用。

1.3 本文的主要研究内容

与传统的PC应用相比，Android应用具有便捷、灵活等特点。如何

有效准确地分析 Android 应用产生的网络流量，并检测恶意行为是本文的研究目标。针对用户更好更放心使用 Android 应用的需求，在对研究现状充分调研的基础上，本文拟在 Android 应用流量分析和恶意行为检测的关键技术和方法上开展研究。

(1)Android 应用网络流量生成方法研究

网络流量数据是进行流量分析与研究的数据基础，由于运营商的流量数据包含很多用户隐私信息，比如手机号码、IMEI 等，一般情况下不会将其公开供其他研究人员使用。同时，手动执行应用生成流量的方式不适合处理数量巨大的 Android 应用。因此，如何为 Android 应用流量分析工作提供必要的网络流量数据是急需解决的问题。本文设计一种自动执行应用的 Android 应用网络流量生成的方法来生成 Android 应用的网络流量。该方法首先调用两个开源工具，Monkey Runner 和 Hierarchy Viewer 的 API 来设计一种自动执行 Android 应用的方法来执行大量 Android 应用，并采集所生成的网络流量。随后提取网络流量属性并生成配置文件。最后根据配置文件来模拟生成与现实流量相似的 Android 应用网络流量。实验结果表明：第一，设计的自动执行应用算法能达到的应用 Activity 覆盖率比传统的方法提高了 30%～67%；第二，模拟生成的 Android 应用网络流量与真实网络环境中的 Android 应用网络流量具有较高的相似度。

(2)基于网络流量开销的 Android 市场应用推荐方法研究

很多 Android 应用需要网络的支持才能完成自身的功能，比如 QQ、微信等，并且不同于传统的有线网络，移动用户在使用 3G/4G 网络是需要支付一定费用的。因此，让用户了解在使用 Android 应用所带来的网络流量开销等问题可以帮助用户更好地规划自己的手机流量套餐并选择更合适的 Android 应用。该研究首先从 Android 官方市场中的 22 个主流的应用类别分别下载排名前 100 的应用。然后，对这 2200 个应用自动执行并采集网络流量，并根据网络流量开销、每秒网络流量开销、不同类型的网络流量开销、不同类型网络流量所占的比例以及功能类似的应用的网络流量开销比较等几个方面来检测每个应用的网络流量开销情况。