清 华 大 学

**综 合 论 文 训 练**

题目：构建WiFi环境下Web用户体验与网络性能的关系模型

系 别：计算机科学与技术系

专 业：计算机科学与技术专业

姓 名：凌肯

指导教师：裴 丹 副教授

2015年6月12日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

暂无

**关键词：** 关键词；关键词；

ABSTRACT

Wi-Fi

**Keywords:** Key; Key;

目 录

[第1章 引言 1](#_Toc422122255)

[1.1 研究背景 1](#_Toc422122256)

[1.2 相关工作 2](#_Toc422122257)

[1.3 主要工作 3](#_Toc422122258)

[1.4 论文结构 4](#_Toc422122259)

[第2章 无线干扰的计算 5](#_Toc422122260)

[2.1 基于双向吞吐率的干扰计算方法 5](#_Toc422122261)

[2.2 被动测量方式的干扰计算方法 6](#_Toc422122262)

[2.2.1 基于丢包率的干扰率计算方法 6](#_Toc422122263)

[2.2.2 基于数据包个数统计的计算方法 6](#_Toc422122264)

[2.3 基于单向吞吐率的干扰计算方法 7](#_Toc422122265)

[2.4 基于时间延迟的干扰计算方法 8](#_Toc422122266)

[2.4.1 包级别的时间流程 8](#_Toc422122267)

[2.4.2 干扰的分类 9](#_Toc422122268)

[2.4.3 时间延迟的代入 10](#_Toc422122269)

[2.5 本章小结 11](#_Toc422122270)

[第3章 测量系统流程 12](#_Toc422122271)

[3.1 系统流程图 12](#_Toc422122272)

[3.2 数据包抓取 13](#_Toc422122273)

[3.2.1 单一路由器上的数据抓取 13](#_Toc422122274)

[3.2.2 有线端数据包的输出格式 13](#_Toc422122275)

[3.2.3 无线端数据包的输出格式 15](#_Toc422122276)

[3.3 数据包匹配 16](#_Toc422122277)

[3.3.1 数据包摘要 16](#_Toc422122278)

[3.3.2 摘要的选取 17](#_Toc422122279)

[3.3.3 算法流程 17](#_Toc422122280)

[3.4 无线发送过程还原 18](#_Toc422122281)

[3.4.1 无线协议的简化自动机模型 18](#_Toc422122282)

[3.4.2 异常处理 19](#_Toc422122283)

[3.4.3 算法流程 20](#_Toc422122284)

[3.5 数据分析 22](#_Toc422122285)

[3.5.1 发送总时间 22](#_Toc422122286)

[3.5.2 等待延迟 22](#_Toc422122287)

[3.5.3 重传延迟 22](#_Toc422122288)

[3.6 干扰值计算 23](#_Toc422122289)

[3.6.1 数据处理 23](#_Toc422122290)

[3.6.2 结果输出 23](#_Toc422122291)

[3.7 本章小结 24](#_Toc422122292)

[第4章 实验测试 26](#_Toc422122293)

[4.1 实验条件 26](#_Toc422122294)

[4.1.1 实验环境 26](#_Toc422122295)

[4.1.2 实验设备 26](#_Toc422122296)

[4.2 实验流程 26](#_Toc422122297)

[4.2.1 基准测试 26](#_Toc422122298)

[4.2.2 载波侦听情形 27](#_Toc422122299)

[4.2.3 隐藏终端情形 27](#_Toc422122300)

[4.2.4 多干扰者混合情形 29](#_Toc422122301)

[4.3 实验结果 29](#_Toc422122302)

[4.3.1 数据处理 29](#_Toc422122303)

[4.3.2 隐藏终端情形 31](#_Toc422122304)

[4.3.3 载波侦听情形 33](#_Toc422122305)

[4.3.4 多干扰者混合情形 33](#_Toc422122306)

[4.4 系统开销 35](#_Toc422122307)

[4.5 实际部署的干扰输出 37](#_Toc422122308)

[4.6 本章小结 38](#_Toc422122309)

[第5章 总结与展望 39](#_Toc422122310)

[5.1 主要结论 39](#_Toc422122311)

[5.2 工作展望 39](#_Toc422122312)

[插图索引 41](#_Toc422122313)

[表格索引 43](#_Toc422122314)

[参考文献 44](#_Toc422122315)

[致 谢 46](#_Toc422122316)

[声 明 47](#_Toc422122317)

[附录A 外文资料的调研阅读报告 48](#_Toc422122318)

# 引言

## 研究背景

随着智能手机和平板电脑等智能设备的普及，越来越多的设备只能通过无线保真技术（Wireless Fidelity，以下简称Wi-Fi）或蜂窝网络技术等无线方式接入互联网。同时，因为无线上网的便捷性，越来越多的用户也选择采用无线方式而非传统的有线方式上网。据统计，在2013年整个互联网流量的55%是通过Wi-Fi接入网络的[3]。不仅如此，ABI Research预测从2012年开始的5年内支持Wi-Fi的设备将从50亿台增加至200亿台[1]。这些都表明Wi-Fi已经成为当今互联网最主要的网络接入方式。

相对于传统的有线网络，Wi-Fi网络的环境更加复杂，其服务受到更多因素的影响，服务的稳定性也更加难以保证。在Wi-Fi普及的潮流下，为了给人们提供更方便的上网服务，很多商家都在公共场所布置了大量的无线接入点（Access Point，简称为AP），也就是我们常说的路由器。在教室、商场、餐厅甚至家庭中都会有一个甚至多只多个无线接入点。在给用户提供了方便的上网接入同时，这些无线接入点的相互干扰，可能会影响Wi-Fi的服务性能，给用户带来难以预测的影响。一方面，因为缺乏足够足够的用户信息，网络管理者也难以及时发现用户遇到的问题。另外，由于影响无线网络的因素众多，网络管理者也难以找到影响用户网络使用的主要原因，不能有针对性地调整网络。

在互联网众多的服务中，基于HTTP的Web网页浏览服务着数以十亿计的用户，是互联网中最重要也是最基础的应用[7]。网页浏览对于网络性能比较敏感，当网络性能不好时，经常出现网页加载缓慢甚至网页无法加载的问题。这些问题会严重影响Web用户体验。然而，用户体验至关重要，它与用户对一个服务提供商的忠诚度密切相关，直接影响着服务提供商可能获得的利润。例如一个用户进行网购时他发现网站无法及时显示他查询的物品或者他进入购物车后迟迟没有反应，他很可能更换到另一个网购网站完成购物，这对于最初的购物网站就是一个损失。统计显示平均增加500毫秒的延迟会导致Bing搜索的利润减少1.2%[6]，平均增加400毫秒的延迟会导致Google搜索的利润减少0.74%[4]。由此可见，Web用户体验关系重大。

Web用户体验非常重要，但是想要简单地获得反映Web用户体验的评价标准却非常困难。与用户相关的信息最主要记录在客户端上，用户与设备之间的交互也是在客户端才能直接获得。如果希望在客户端上抓取并记录用户的信息用于分析，必须让用户额外安装相应软件。这钟不便给大规模部署实验带来了困难，也使得网络运营商和服务提供商难以有针对性地提升用户体验。

综上所述，Wi-Fi环境下的Web浏览是当今互联网的主要服务，研究Web用户体验有着巨大的价值和意义。然而，目前缺乏一种有效的在非客户端上测量Wi-Fi环境下Web用户体验的方式。另外，Wi-Fi的多个因素均可能影响用户体验，不同因素对Web用户体验的影响程度也有待研究。本文拟采用决策树基于在路由器上采集的信息构建测量参数与用户体验的关系模型，利用该模型预测不同情况下的Web用户体验，并且在这个模型的基础上指出对用户体验影响关系较大的Wi-Fi参数，给提升用户体验提出指导意见。

## 相关工作

路由器在家庭网络中扮演着网络入口这样一个重要的角色，在上面有可能获得更加精确的关于用户设备及用户网页浏览情况的信息，而这些信息是在NAT之后就难以分辨的。基于此，BISmark[5]团队在OpenWRT路由器上部署了Mirage，用于更加精确地监测在NAT之后的家庭网络的性能变化。BISmark团队在全球超过30个国家部署了测量路由器。他们的研究结果显示接入延迟（Last-mile latency）在整个互联网的延迟中占有很大比例，而接入延迟增加十几毫秒就可能导致网页加载时间增加几百毫秒。他们的实验结果同时显示通过对DNS和TCP连接进行缓存可以有效介绍网页加载时间。

传统研究Web用户体验的方法主要依赖于在客户端安装软件记录用户的信息。WProf[7]是一个植入于浏览器中的插件，目标是分析网页加载时间的不同组成部分。通过监测浏览器的行为，WProf绘制了用户访问一个网页时浏览器不同事件的依赖关系图，并且绘制了相应时间的关键路径。研究结果显示网页中的计算部分占据了关键路径时间的大约35%，同时对JavaScript的异步调用以为会阻塞HTML的解析也会显著增加网页的加载时间。

Athula Balachandran[2]首次引入机器学习的方法分析网络性能参数和观看视频的用户体验的关系。因为网络参数之间的关联性很强，传统的分析单个网络参数与用户体验的关系难以找到根本的影响因素。他们考虑了视频类型、视频缓冲比、视频缓冲次数、视频平均传输编码率等因素，与用户体验指标——用户观看视频的持续时间进行关联，利用决策树构建了关系模型，并且提出了普遍的用于预测用户体验的框架。

AT&T[3]团队构建了蜂窝网络中测量参数与Web用户体验的模型。他们在蜂窝网络基站上分别采集网络和用户浏览两方面的信息。前者包括基站信号强度、服务用户个数、上下行带宽和不同频率间的切换次数的因素。后者包括网页请求的URL和用户设备的MAC地址等。利用机器学习方法，他们根据信息计算出用户浏览网页的长度、放弃网页的次数和网页未完全加载百分比作为衡量用户体验的标准。在此基础上他们用机器学习的方法构建网络性能参数与用户体验标准的关系。最终，他们发现用户设备在不同频率间的切换对于用户体验有最大的影响，因为频率间的切换会导致网络中断长达几秒钟的时间，接下来影响较大的因素是信噪比。与预期相反，上下行带宽对用户体验影响不大，可能原因是网页浏览对带宽的需求较低。

## 主要工作

本文的主要工作简述为以下几个部分：

1. 实现基于OpenWRT路由器的测量工具

* 在路由器上抓取网络相关参数

为了构建网络性能与用户体验的关系，首先需要获得在用户使用无线网络时的路由器参数，包括时间戳、等待占空比、上下行带宽、物理层上下行速率、MAC层数据包重传比例、路由器当前连接设备数和每个设备的信号强度等。获得的数据传输到服务器上用于进一步分析。

* 在路由器上抓取用户访问信息

目前的Web应用基本上基于HTTP完成。为了了解用户的网页浏览状态，需要在路由器上抓取HTTP数据包记录。具体包括HTTP中GET类型数据包的GET、Host、User-Agent和Referer域，以及设备的MAC地址和时间戳。获得的数据同样传输到服务器上用于进一步分析。

1. 实现用户体验的分析工具

* 构建网站的HTTP请求的下载链

路由器上抓取的HTTP的GET数据包不仅来源于用户的网页浏览，也可能来源于其他Web应用。为了研究用户的网页浏览体验，首先要过滤掉其他来源的数据包。因为在路由器上抓取的记录是以数据包为单位的，在过滤掉无关数据之后，还需要在服务器上利用抓取的数据包的关系，恢复出用户对网站的完整访问流程。

* 区分用户请求与网页嵌入对象的自动请求

网页浏览的HTTP的GET请求来源分为两种。一种来自于用户点击一个超链接或者在浏览器中输入一个网址导致的主动请求。另一种来源于浏览器解析网页时，针对嵌入的对象如图片、JavaScript和CSS等自动发出的请求。前者与用户网页浏览的密切程度关联紧密即与用户体验直接相关。后者主要用于呈现网页的内容，并不直接反应用户体验。因此，需要对两种不同的请求进行区分。

* 构建网络参数与用户体验的关系模型

在区分出用户请求之后，将其作为衡量用户体验的标准。把用户请求与测量参数匹配，分析测量参数与Web用户体验的相关程度。在此基础上挑选出具有主要影响力的参数用于构建关系模型。

1. 为了能够将带有测量工具的路由器分发开展实际测量，首先需要完成测量数据与服务器间的传输工作。然后，实际部署测量路由器进行实验。最后，根据采集到的数据进行分析，进一步完善测量程序。

## 论文结构

本文第一章介绍了相应的研究背景、已有的工作和文章的主要工作内容。

本文第二章主要介绍不同的测量Web用户体验的方法。

本文第三章主要介绍系统的架构及实验的部署。

本文第四章主要介绍数据的分析以及模型的构建。

本文第五章对文章进行总结，探讨未来的后续工作。

# Web用户体验的测量

本文的主要目的是构建路由器上网络测量参数与Web用户体验的关系模型，准确有效地获得Web用户体验数据是构建模型的前提与基础。测量的方式也决定了需要收集的数据和实验框架的构建。

本章节首先根据测量方式的分类具体介绍了几种不同的测量方式，然后详细介绍了实际采用的一种适于大规模部署的基于路由器被动测量的方法。

## 实验室环境下的控制实验

一些关于用户体验的实验邀请实验者到实验室中完成规定的任务，通过实验后的问卷调查了解用户体验。DF Galletta[5]请196名学生在实验室环境下完成对三个网站的搜索任务。三个网站保存在光盘上，用以严格控制访问时间。在用户完成访问不同的访问任务时，电脑加入人工设置的延迟时间模拟不同网络性能。在用户完成实验后，他们被要求填写问卷评估自己完成任务的目的性和感受。

实验室环境下的控制实验的优点在于调查问卷获得的用户反馈最真实直接地反应用户体验，是最准确的标准。然而这种实验方式的缺点也非常明显，实验需要在实验室环境中才能完成，对于被试者非常麻烦，不可能进行大规模实验。另外，由于地点受到限制，被试者的挑选也受到限制，普遍性相对较差。

## 客户端的测量

目前常用的一种测量方法是在用户的设备即客户端上安装专门的程序捕捉用户信息或者用户与设备直接的交互情况以此来推测用户体验。网页加载分析插件WProf[8]被实验者植入于不同种类的浏览器的引擎中（图 2‑1）。该插件可以将浏览器加载网页的每一个事件记录下来并根据发生时间和相互影响构建它们的依赖关系。在客户端上进行测量的优势是可以获得客户端上独有的数据，比如网页的渲染时间和浏览器对JavaScript的计算时间。这些因素构成了网页加载时间的重要部分，直接影响着用户体验。在客户端进行测量的最大的缺陷是需要让用户额外安装测量软件，这种不便使得该种测量方法难以被大规模部署，也就难以适用于广泛地获得大量数据进行实验。另外一个缺点是在用户安装测量软件后，用户对隐私被泄露的担心可能改变他们使用网络的习惯，不能最真实地反应实际用户使用情况。

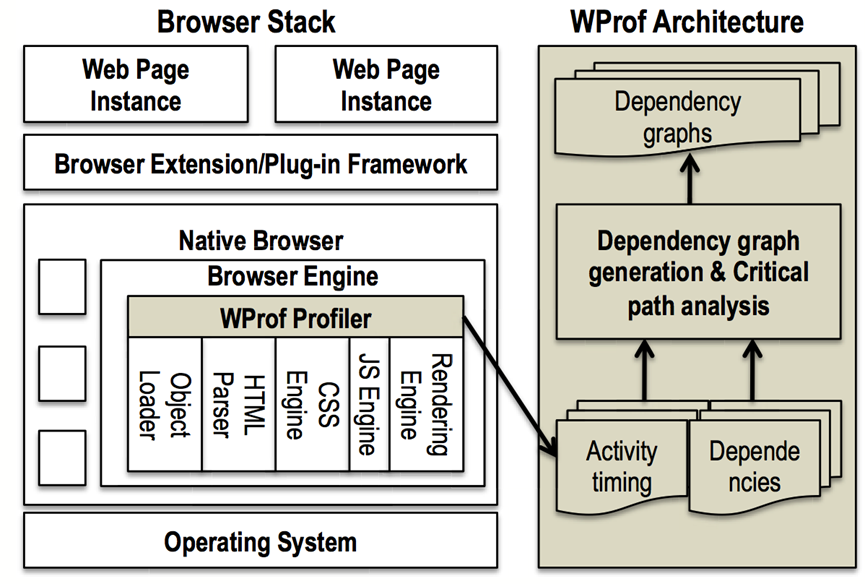


图 2‑1嵌入浏览器的WProf网页分析插件

## 中间节点的测量

鉴于本文探讨的是构建Wi-Fi环境下网络性能与Web用户体验的关系，我们将中间节点限制在普通类型的家用路由器上。路由器作为室内无线上网的主要接口，通常是室内无线上网所有流量经过的通道。同时，路由器作为接入互联网的最后一跳，它与用户设备直接相连，在路由器上收集数据使得我们能够将不同的记录针对不同设备或用户区分开，给后续研究奠定基础。从路由器之后这些数据包被汇聚在一起，弱化了客户端独有的特征。因此，路由器是部署Wi-Fi性能测量工具的最佳选择。在路由器上，采集到的Wi-Fi参数主要分为两类。一类是路由器的总的工作状态记录，例如路由器累计的工作时间、累计传输数据的时间、累计接受数据的时间等。另一类数据是与单个设备相关的，包括当前连接的设备个数、设备MAC地址、每个设备传输的数据量、每个设备接受的数据量和每个设备的信号强度等。

路由器虽然是理想的测量Wi-Fi性能参数的地点，在测量用户体验时却不够直接。因为用户体验是用户层面的一种衡量标准，更加宏观，而路由器上只能获得包级别的底层数据。这就需要一个机制能够从底层的数据包记录中恢复用户体验的相关数据。

我们选择的评价Web用户体验的指标是会话长度（Session Length）[3]。会话长度指用户进入一个网站后，从该网站及相关链接中手动打开的网页个数。会话长度表示了用户的投入程度。当网络性能良好时，网页打开的过程非常流畅，用户体验良好，因此会使用网页直到达到最初的目标。当网络性能不好时，网页的加载过程可能非常缓慢，有可能出现网页部分加载失败甚至整个网页无法加载的情况，这时用户的浏览体验不好就会关闭网页，导致会话长度很短。因此会话长度可以有效衡量Web用户体验。

接下来需要利用包级别的数据推测会话长度。用户手动打开网页反映在数据包记录中就是一个针对网页地址的HTTP的GET请求。然而路由器上记录的HTTP的GET请求并不仅仅来自于用户手动点击网页的访问。当浏览器解析网页的HTML文件时，会对其中的嵌入对象比如图片、JavaScript和CSS发出HTTP的GET请求，下载内容完成网页的加载，这些请求占据了GET请求的一大部分。现在主要问题转变为如何区分用户主动点击发出的请求和网页嵌入对象自动发出的请求。一种传统做法是利用请求的间隔时间（idle time）[7]。这种方法做出了以下假设，因为对嵌入对象的请求是在浏览器解析网页时自动发出的，时间间隔非常短，请求一般都在网页加载时发出。相反，用户的点击表示打开新的页面，一般在浏览原有网页之后，相隔时间较长。因此，通过对比两次相邻链接之间的时间差，如果大于预定阈值就可以认为这个请求是用户点击产生的，反之就是浏览器发出的。

然而，如今网页的复杂度逐渐增加，需要动态加载的内容越来越多，加载时间的跨度也逐渐增加，利用一个固定的阈值已经难以区分网页嵌入对象自动发出的请求与用户点击的请求。本文采用了由Athula Balachandran[3]等人提出的一种更加有效的基于文本分类的用户点击识别方法。观察从一个网站中发出的所有请求，可以发现请求的URL有明显的区分。以新浪网为例，www.sina.com.cn和sports.sina.com.cn很可能是用户点击的网站链接。而对sax.sina.com.cn或d1.sina.com.cn的请求很大可能性来源于对嵌入对象的请求。基于上述特征，文本分类的方法可以应用于区分用户点击去嵌入对象。我们分解URL的中间部分以及文件类型作为特征。以下面这个URL为例http://www.cs.tsinghua.edu.cn/publish/cs/index.html，特征集为{cs, tsinghua, edu, cn, publish, html}。分类方法上，该方法选用了在文本分类中效果很好的朴素贝叶斯方法。为了减少人工标注训练集的工作量，该方法采用了一种自动标注的方法。将从一个网站开始的全部请求按时间排序，取出前10秒的部分。如前所述，因为人在每次手动点击之间的时间是浏览网页，相对较长，我们可以认为10秒这个区间内只有第一次请求为用户点击发出的，其他请求都为浏览器自动发出的。我们可以使用这样的方法自动生成训练集。获得足够的训练集之后，用这些训练集训练朴素贝叶斯模型，然后用训练后的模型对所有请求数据进行处理，获得仅包含用户点击的列表。

## 本章小结

综上所述，为了保证测量实验可以被普遍的部署而不限制在具体的实验环境中，我们决定采用放置在中间节点即家用路由器上的测量方法。通过朴素贝叶斯法对用户的访问记录进行分析，识别其中来自于用户点击的请求。对于一个网站的请求个数即该网站的会话长度，表示着用户访问该网站的体验程度。

# 系统的架构及实验部署

本文实现了一个部署在路由器上的轻量级的测量工具，分别记录路由器与设备之间的Wi-Fi网络性能以及与用户访问网页相关的记录。除此之外，本文还在服务器端实现了一串工具管道完成两者之间关系模型的分析与构建。本章将具体介绍系统的架构设计以及具体实现，描述系统捕捉数据的规格以及功能实现的算法。

## 系统架构

整个系统主要由三个部分组成，分别是：路由器端数据包的抓取、服务器端数据的分析和结果的输出。首先，我们在路由器端的无线网卡上抓取数据包，然后根据需求对数据包进行处理，保留有用的部分。每隔一段时间，系统会自动地把保留下的数据包记录发送到服务器端。最后，我们在服务器端将这些数据汇聚起来。先通过算法将用户请求重建起来，然后通过匹配算法将用户访问记录与Wi-Fi网络参数对应起来，进而分析两者之间的关系。

## 数据的获取

### 网页请求数据的获取

在本系统中，我们利用libpcap[7]在wlan接口上抓包。wlan为路由器的无线网卡接口。因为我们研究的对象是Wi-Fi网络性能，在wlan上抓包可以避免抓到来自通过网线与路由器直接连接的设备的数据包，避免了对实验结果的影响。用户的Web访问是基于HTTP的GET请求，因此HTTP的GET请求就是我们的研究对象，而HTTP是基于TCP的。所以我们利用libpcap对每个TCP数据包的负载进行检查，如果以GET开头，则说明该TCP数据包实际是HTTP的GET请求。TCP的帧结构如图 3‑1所示。

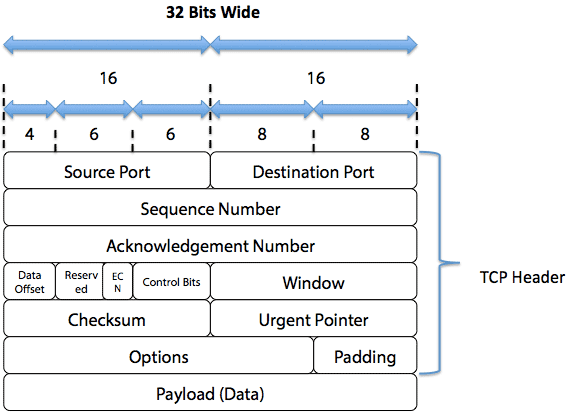


图 3‑1 TCP帧结构

HTTP的GET内容出于TCP包的负载部分，结构如图 3‑2。当我们发现一个TCP包的内容是HTTP的GET时，我们利用正则表达式匹配，将GET请求中的GET、Host、User-Agent和Referer四个域提取出来，每个域最多保留256字节，保留在我们的记录结构中。同时，我们还会保留源MAC地址并且打上系统的时间戳。整个记录的格式如图 3‑3。



图 3‑2 HTTP GET请求域示意图

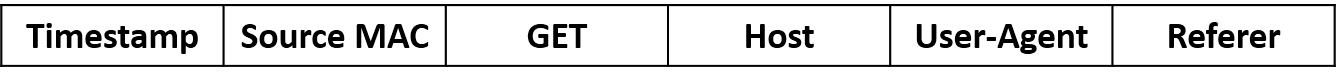


图 3‑3网页请求数据记录六元组

GET域是请求资源的地址，不包括域名。Host域为GET请求资源的主机名称。User-Agent域代表GET请求的发出者。Referer域表示当前GET请求的来源地址，如果当前请求是主动发出的，即没有来源地址，则数据包负载中没有Referer域。

### Wi-Fi网络参数的获取

TO BE CONTINUED

## 网页请求链的重建

为了分析用户访问网站的用户体验，我们需要解决如下几个问题。首先，在路由器上抓取的数据是包级别的，而会话长度相对是宏观的，需要从包级别的数据中重建出用户对网站的访问。其次，一个路由器上可能同时连接了多个设备，我们需要对每个设备进行区分才能准确匹配相应的网络参数。最后，即使一个路由器上只连接了一个设备，用户可能使用不同浏览器访问网络，或者用户电脑上其他Web应用发出的请求也会对分析造成影响。这里，我们提出一个算法从包级别的数据恢复单一用户单一浏览器发出的请求过程。

|  |
| --- |
| **算法1** 网页请求链的重建算法 |
| **程序输入：**记录网页请求的数组  **程序输出：**以请求会话为单位的存储结构record\_dict  **func** reconstruct\_visit\_record(lines)  **for** line in lines  **if** User\_Agent is not valid  continue  **if** Referer field is empty  session = Session()  new\_ref = Host + GET  session.MAC = MAC  session.url\_list.append(new\_ref)  session\_list.append(session)  record\_dict[MAC][User\_Agent][new\_ref] = session  **else**  ref = Referer  if ref not in record\_dict[MAC][User\_Agent]:  continue  session = record\_dict[MAC][User\_Agent][ref]  new\_ref = Host + GET  session.url\_list.append(new\_ref)  record\_dict[MAC][User\_Agent][new\_ref] = session  **return** record\_dict |

该算法的核心是依赖于GET的Referer域。如前所述，Referer标记着该请求的来源地址。用户在网站A中手动点击打开超链接发出的请求或者浏览器对网页嵌入对象的请求的Referer域都为网站A。利用Referer，我们就可以将单个的GET请求恢复成原先的下载链。然而，由于我们需要达到单个用户单个浏览器的精度，还需要做一些额外的处理。该算法的核心数据结构是一个三级索引的字典，关键字分别为请求源MAC地址、User\_Agent和Referer地址，返回的值是一个用户访问会话。利用TCP包中的源MAC地址可以对连接在路由器上的设备进行区分。在保证记录来自于同一个设备后，User\_Agent可以保证请求从同一个浏览器发出，避免不同浏览器访问同一网站造成的混淆。最后Referer保证这些请求属于同一个会话。算法逐行读取包级别的GET请求记录。首先针对User\_Agent进行判断，如果请求发出者不是浏览器，那么我们就忽略该条记录。然后根据Referer域进行判断，如果Referer域为空，说明该请求将开始一个新的会话。我们创建一个新的会话类，把当前的记录写入会话类中。对于每一个会话，我们都需要知道哪些请求可能从它发出的，即Referer中的哪些地址属于这个会话。对于当前的新会话，我们把它的Host和GET两个域的字符串拼接起来，即得到该请求的完整地址。将完整地址结合TCP的源MAC地址和GET的User\_Agent三者作为关键字加入字典，则可以根据当前网站发出的请求利用字典找到对应的会话。如果Referer不为空，则使用上述三个关键字在字典中寻找。如果找到会话，则把最新的请求的完整地址加入会话的下载记录中，并且指向目标会话。如果没有找到，原因可能是路由器没有完整记录所有的数据包，丢失了创建该会话的数据包，那么对于该种记录，我们不予考虑。

在将所有请求按会话归类存储在字典之后，程序会根据会话为单位遍历记录，按时间顺序输出一个会话的所有请求，输出格式如图 3‑4。

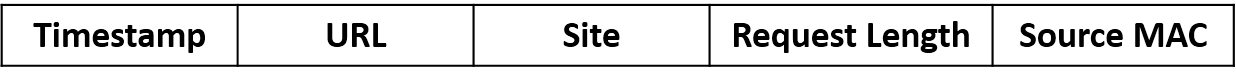


图 3‑4 按访问会话组织输出的五元组

时间戳是路由器收到该请求数据包时记录的时间，URL是该请求的资源地址。对于重建后的请求链使用在第二章中提到的方法进行分析即可得到用户的访问会话长度记录，我们以如图 3‑5格式记录。

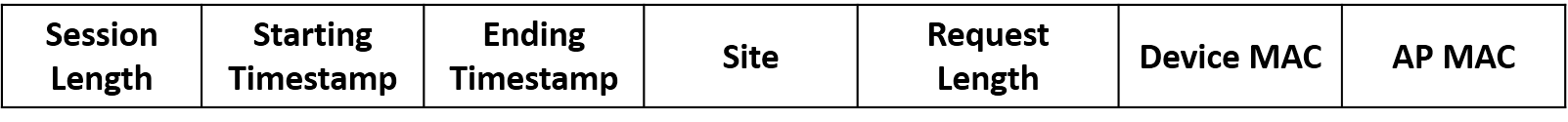


图 3‑5 会话长度记录七元组

Session Length表示用户访问网站的会话长度。Starting Timestamp和Ending Timestamp分别为这个会话的起止时间戳。Site表示该会话初始访问的网站。Request Length表示用户恢复该会话的原始记录中包含的请求个数。Device MAC是用户设备的MAC地址。AP MAC是用户连接的路由器的MAC地址。

## 数据的匹配

因为路由器两类Wi-Fi参数记录的时间戳含义相同，因此可以使用相同的算法进行匹配。对于每一个用户会话记录，分别获得它的开始和终止时间戳。利用二分查找，在Wi-Fi参数记录中分别找到不超过起始和终止时间戳的最大时间。然后取这两个时间中的所有Wi-Fi记录作为与查询会话的匹配结果。在恢复用户的访问会话之后，需要把用户访问过程与该过程发生时的Wi-Fi参数匹配起来，从而分析两者之间的关系。用户请求的时间戳来自于路由器抓取到GET数据包时记录下的系统时间，而Wi-Fi参数则是固定地每10秒采集一次，两者不是精确匹配的。因此，我们决定对每次用户访问会话寻找能够覆盖它的Wi-Fi参数记录区间作为近似。路由器的Wi-Fi记录分为两类，一类是路由器工作信道的数据，另一类是路由器与客户端相关的数据。路由器工作信道的记录如图 3‑6。

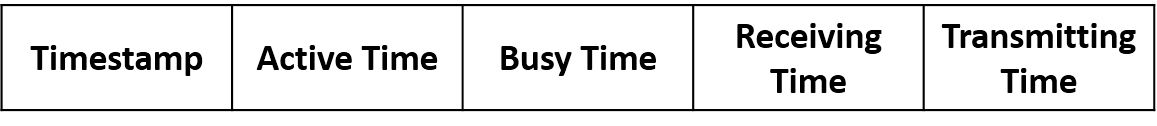


图 3‑6 路由器工作信道记录五元组

Timestamp是记录该组数据的时间戳。Active Time是路由器从开机启动开始累计的时间。Busy Time是路由器从开机启动开始累计的繁忙时间，包括传输数据、接收数据以及等待其他造成干扰的路由器传输数据。Receiving Time是路由器自启动之后累计接收各客户端数据的时间。Transmitting Time是路由器自启动之后累计向客户端传输数据的时间。

路由器与客户端设备之间的数据记录格式如下。

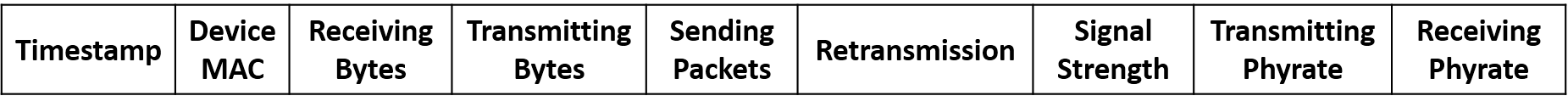


图 3‑7 路由器与客户端相关的数据记录九元组

Timestamp是记录该组数据的时间戳。Device MAC是客户端设备的MAC地址。Receiving Bytes是路由器自启动之后累计接收该客户端的字节数。Transmitting Bytes是路由器累计给该客户端发送的字节数。Sending Packets代表路由器累计给该客户端发送的包的个数。Retransmission代表路由器累计给该客户端重传的包的个数。Signal Strength代表该客户端接受到的路由器的信号强度，单位为dB。Transmitting Phyrate代表路由器给客户端发送数据包时物理层的发送速率。Receiving Phyrate代表路由器物理层接受客户端发送的数据的速率。利用时间戳进行匹配的算法伪代码如下：

|  |
| --- |
| **算法2** 数据的匹配算法 |
| **程序输入：**查询的文件、待查询的会话的起止时间、设备MAC地址和路由器的MAC地址  **程序输出：**与待查询会话时间匹配的Wi-Fi参数记录  **func** find\_records(file\_name, start\_time, end\_time, device\_MAC, AP\_MAC):  file\_path = '../wifidata/%s/%s' % (AP\_MAC, file\_name)  f = open(file\_path, 'r')  lines = f.readlines()  f.close()  start\_index = bisect(lines, start\_time) // binary search  end\_index = bisect(lines, end\_time) // binary search  **return** lines[start\_index - 1 : end\_index + 1] |

因为路由器两类Wi-Fi参数记录的时间戳含义相同，因此可以使用相同的算法进行匹配。对于每一个用户会话记录，分别获得它的开始和终止时间戳。利用二分查找，在Wi-Fi参数记录中分别找到不超过起始和终止时间戳的最大时间。然后取这两个时间中的所有Wi-Fi记录作为与查询会话的匹配结果。

## 本章小结

本章对实验系统的架构进行了介绍，特别详细介绍了在路由器上的数据抓取部分，给出了相应的算法代码以及每个记录参数的具体含义。在路由器上抓取的数据是分析并构建模型的来源，也是整个实验的基石。

# 数据分析与模型构建

在前面我们已经采集了多种Wi-Fi参数，并且利用时间戳将Wi-Fi参数与用户体验做了匹配。在本章，首先利用多种统计方法分析Wi-Fi参数与用户体验之间的相关性。然后，我们从中挑选出对用户体验影响较大的参数构建与Web用户体验的关系模型。最后，我们针对构建的模型进行讨论和评估。

## 分析的网络参数

### 实验环境

本实验在实验室环境下进行。在进行测量的过程当中，为了避免其他不可控的干扰，我们选择在假期的时候对各种情形进行干扰值和基准值的测量。同时，根据对实验室环境的评估，我们使用了无人使用的第三信道作为主路由器和干扰路由器采用的通信信道。

### 实验设备

本实验中，主路由器为一台WZR-HP-AG300H型号的Buffalo品牌路由器，其上搭载Openwrt系统。干扰路由器为两台极路由1S型号路由器，使用原厂的操作系统。客户端为三台ThinkPad笔记本电脑。

## 散点图

对于每个我们关心的WiFi参数，我们首先绘制它和表示用户体验的会话长度的散点图，以便直观地发现两者之间是否有明显的关系。

Busy time和Receive time与会话长度的散点图形状相似，数据聚集成了两个峰，没有单调趋势。数据聚集成两个峰的原因可能是路由器的忙碌时间或接收客户端数据的时间由两部分组成，一部分对应当前会话所属的客户端，另一部分对应其他客户端。当路由器服务当前会话客户端的比例越高，对用户体验的影响为正。当路由器服务其他客户端的比例越高，对当前会话的用户体验影响为负。

图表 1 Busy Time与Session Length散点图

图表 2 Receiving Time与Session Length散点图

出于上述原因，我们考察路由器的等待占空比。Air utilization数值越高说明当前路由器因为其他路由器的存在等待的时间比例越高。观察图XX，随着Air utilizaiton的升高，会话长度中长会话的点越少。虽然在Air utilization数值小时也有短会话，但是当Air utilization大时基本没有长会话，说明两者为负相关。

图表 3 Air Utilization与Session Length散点图

观察图XX，随着Resend packet的上升，会话长度呈减小的趋势。Resend packet越大暗示网络条件不好，导致数据包需要被不断重传，可能对用户体验有负的影响。

图表 4 Retransmission与Session Length散点图

观察图XX和XX，路由器物理层的数据接收速率与传输速率与会话长度存在正相关趋势，但不明显，数据大约分成两个成三角形的上升区域。

图表 5 Receiving Physical Rate与Session Length散点图

图表 6 Sending Physical Rate与Session Length散点图

观察图XX，随着客户端接受路由器信号的增强，会话长度呈上升趋势。路由器的信号增强使得网络性能更好，因此更容易有好的用户体验。

图表 7 Signal Strength与Session Length散点图

图XX显示，在Session duration较短时，与会话长度呈现正相关。当Session duration更长时，会话长度反而减少，这似乎与直觉不符。对于Session duration非常长的部分，用户可能是在非常仔细地浏览网页，比如查阅资料，或者观看视频，因而导致Session duration远远长于一般情况，与会话长度的关系就不那么明显了。

图表 8 Session Duration与Session Length散点图

对于图XX，因为路由器连接的客户端数目一般只有几个取值，大量的数据点重叠在一起，导致这张图看不出点的分布或者趋势。

图表 9 Device Number与Session Length散点图

从散点图中，我们可以看到Air utilizaiton，Signal Strength和Resend packet与会话长度有比较明显的关系。对于其他因素，虽然散点图中并没有给出非常明确的单调关系趋势，但不代表这些因素与会话长度就没有关系。接下来我们将用其他方法进行分析这些因素与用户体验的关系。

## 实验结果

### 数据处理

对一个数据帧，由于发送队列的存在，从其到达路由器到准备被发送（到达发送队列队首），总存在一定的时间延迟。同理，一个数据帧从进入路由器，到被发送，也存在一个延迟。我们将该延迟称之为队列延迟。

通过数据包匹配的过程，我们能得到帧到达路由器的时间，不考虑重传的情况，匹配的帧对之间的时间差即是队列延迟。下文中的图4-4反映了部分帧的队列延迟情况。

考虑到有重传帧的情况下。由于匹配算法对于时间的有序性，因此匹配得到的帧对总是先匹配到第一次发送的帧。在算法中，被匹配的数据将不再参与后续的匹配过程，故重传的数据帧不会被匹配（这里的重传特指路由器无线端的重传，不包括本身有线接口下载的重传帧），因此该队列延迟的计算是有效的。

图4-4中，横轴为帧的编号，纵轴为队列延迟的时间，单位为。如上图所示，我们在匹配后的数据当中选择了10000个数据点，绘制了上述图像。可以看出，取值为正数的数据点主要分布在0-40000之间，平均值大约在15000-20000之间；同时，取值为负数的数据点则非常接近于坐标轴，其值的量级在几十。

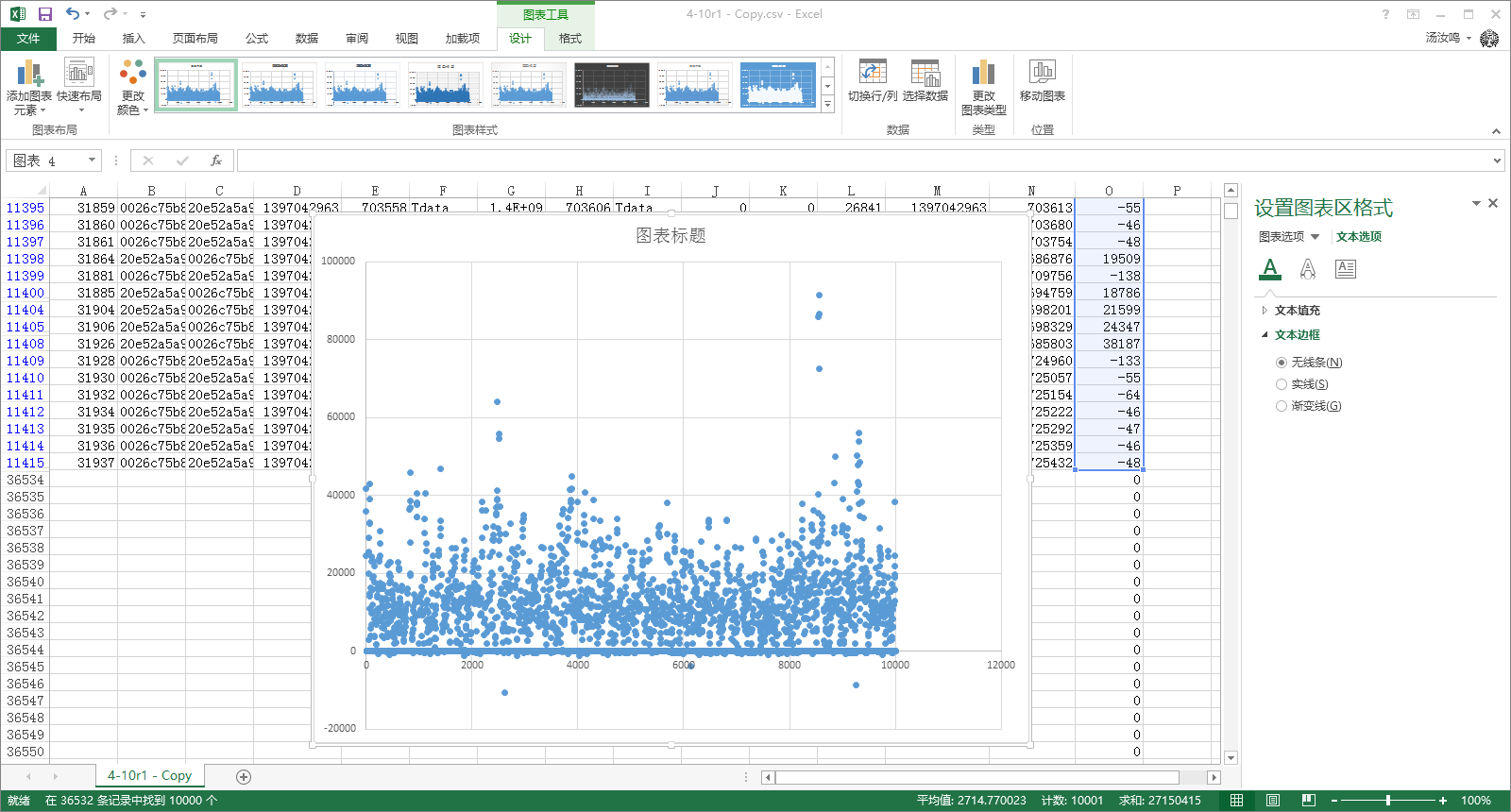


图 4‑3 部分帧的队列延迟分布散点图

图4-4中，取值为正数则表示数据帧先到达有线端口，再到达无线端口，也即下载的数据帧；反之，取值为负数则代表上传的数据帧。

从图中我们可以看出，上行的数据帧近乎没有等待就开始发送了，这说明了在实际过程当中，上传流量一般很小，导致了上传的发送等待队列也比较空闲。而下载的数据帧则与之相反。大部分下载的数据帧均在进入队列之后有10ms数量级的延迟，该量级的时间相对于一个发送回合来说非常长。这说明了下载的等待队列一般不为空。

通过前文的叙述，我们知道通过到达时间和上一帧的结束时间，可以推断出当前帧的准备发送时间。而现在我们知道大部分下载的数据帧，在进入队列后总需要等待一段时间。

因此，在研究下载数据帧时，可以直接利用上一帧来得到准备发送时间。通过这种方法，我们不需要得到当前帧的到达时间，也即我们不需要进行数据包匹配的过程。在特殊的研究情况下，该方法可以有效地简化程序的复杂性，提升实验的效率，降低硬件要求，有助于实验的大规模部署。

### 隐藏终端情形

实验操作如前文叙述。通过Iperf和路由器脚本，产生稳定的背景流量并记录具体的吞吐率数据，如图4-5和图4-6所示。之后，将计算得到的干扰值与基准值进行误差分析，结果如图4-7所示。

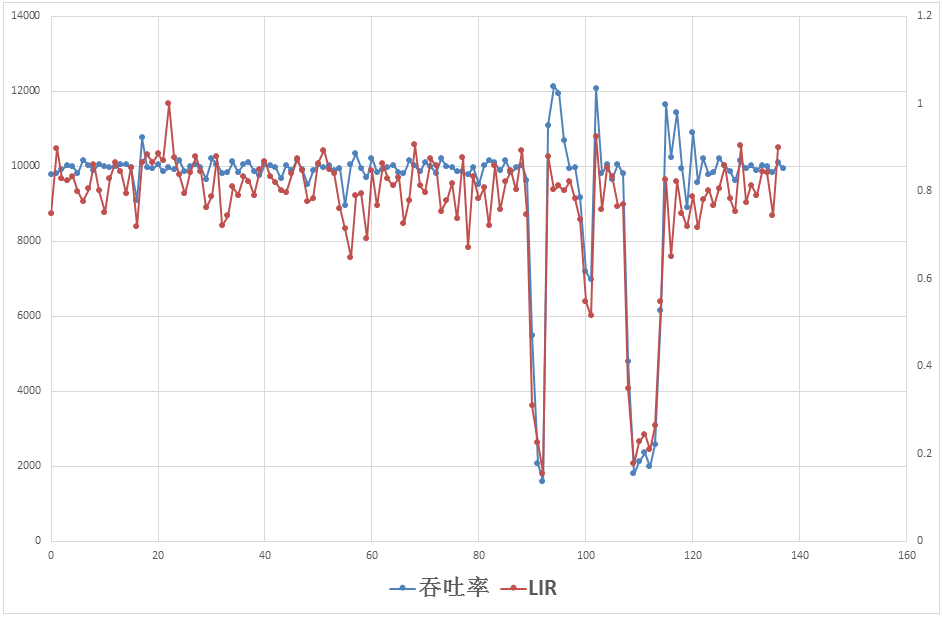


图 4‑4 隐藏终端情形的吞吐率与LIR值

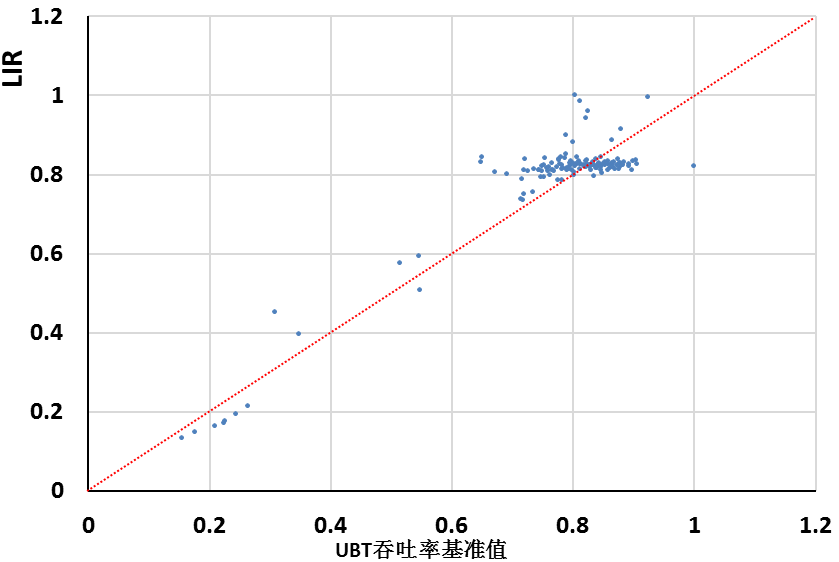


图 4‑5 隐藏终端情形的吞吐率与LIR值拟合图（归一化后）

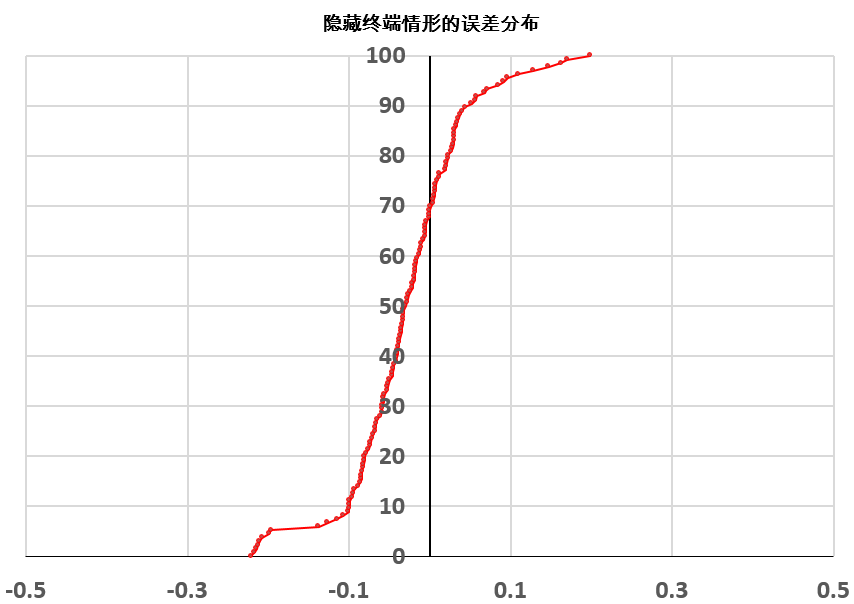


图 4‑6 隐藏终端情形的误差分布

由于隐藏终端位置的不确定性，我们在一定范围内的不同位置进行了多次重复实验。实验结果显示误差范围不超过，约85%的误差的范围之内。该实验结果与PIE[3]的结果很接近，比PIE的结果略差。上述结果说明，在隐藏终端的情形下，不确定性更大，误差也更大。但整体上该计算方法仍然是有效的。

### 载波侦听情形

实验操作如前文叙述。与上一情景类似，通过Iperf和路由器上的脚本，我们可以得到具体的吞吐率数据，以此为基准值与计算得到的干扰值进行误差分析，误差分布结果如图4-8所示。

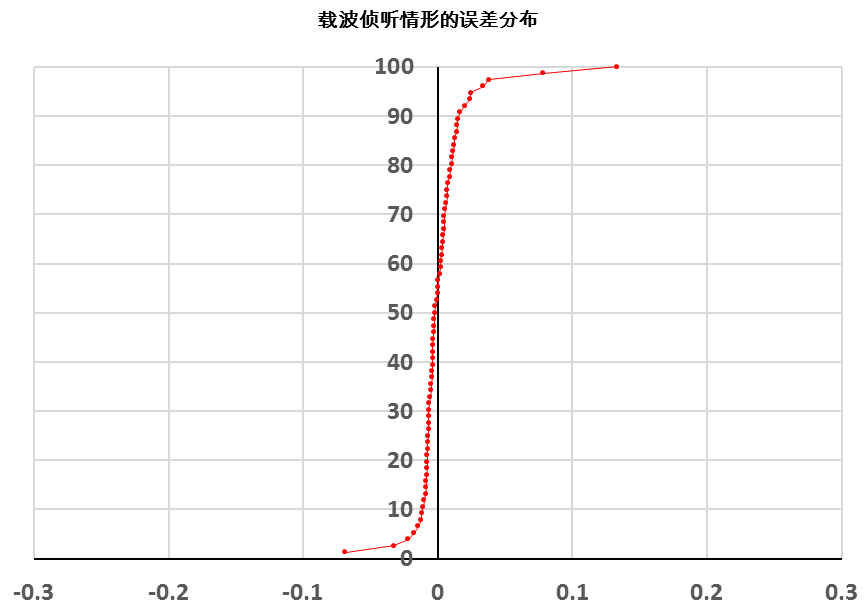


图 4‑7 载波侦听情形的误差分布

可以看出，约99%的误差均保持在的范围之内。说明在载波侦听的实验情形下，该干扰值的计算方法是有效的。

### 多干扰者混合情形

在多干扰者存在的情形下，我们在主路由器上计算各自的干扰值，并通过多个Iperf客户端程序和干扰路由器上的Iperf服务端来记录UBT吞吐率数据。误差分析的结果如图4-9和图4-10。

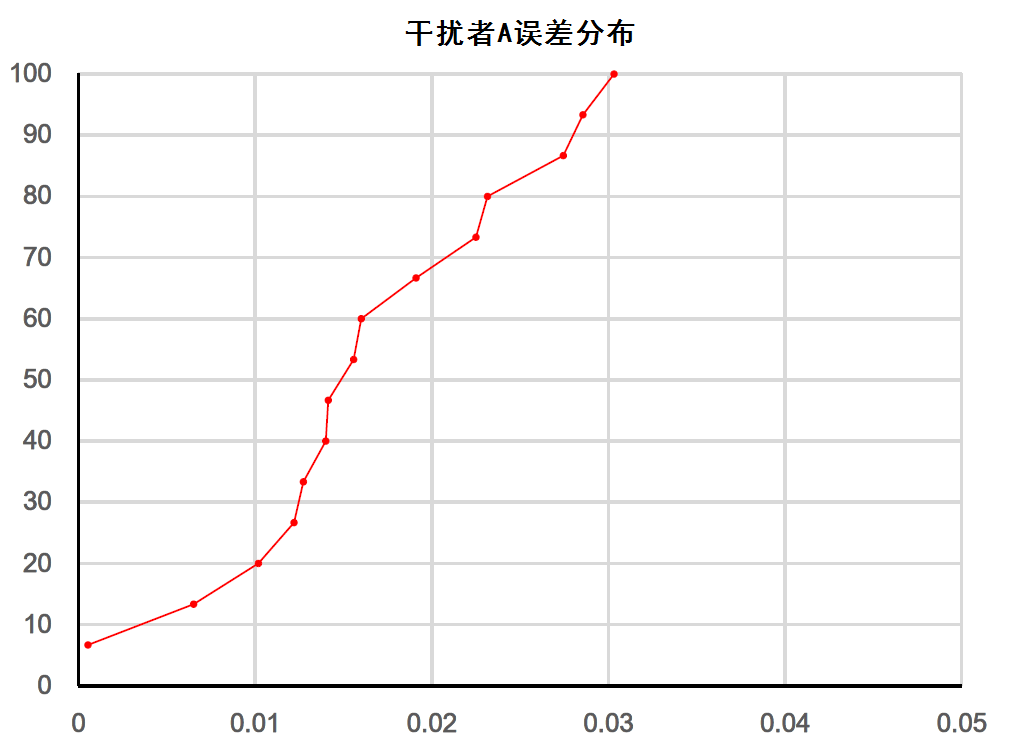


图 4‑8 干扰者A的误差分布

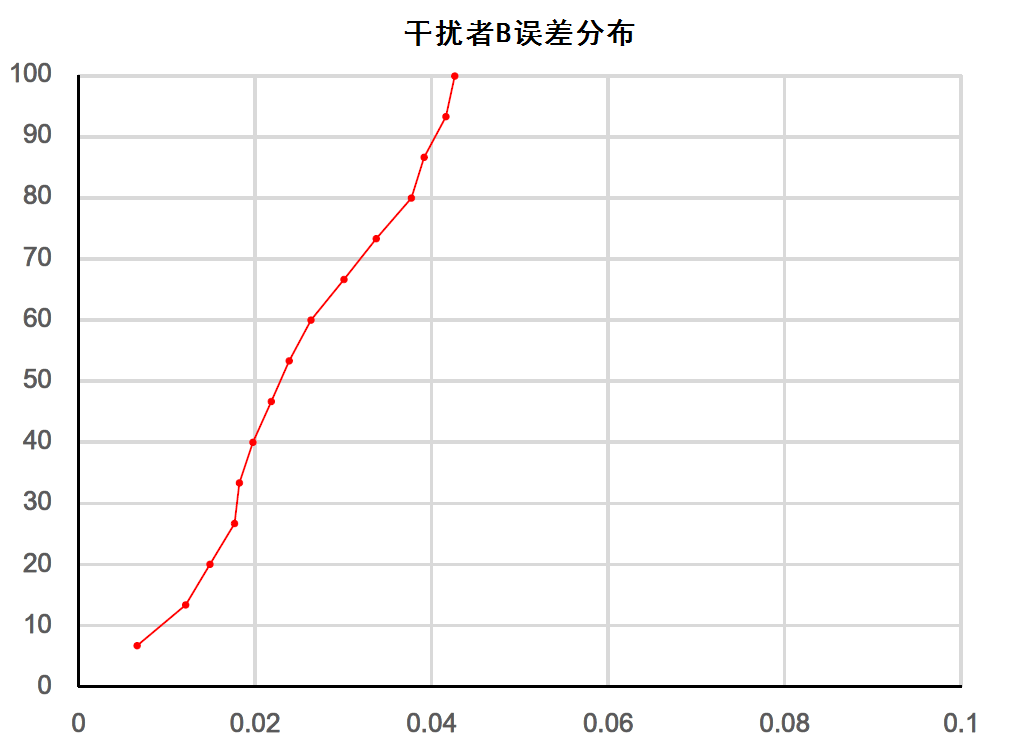


图 4‑9 干扰者B的误差分布

如图中所示，两台干扰路由器的误差分布均被控制在0.05的范围之内。在PIE[3]的实验当中，这一数值大约为，说明了我们采用的计算方法是有效的、准确的。

## 系统开销

由于我们的部分程序在路由器上运行。考虑到路由器本身的资源有限，我们需要保证造成的额外的计算不会对路由器本身产生较大的影响。否则的话，若计算过程（在路由器上，仅有抓包和预处理过程）影响了路由器的正常使用，一方面对整个背景流量产生影响，会导致被动测量不准确。另一方面也会使得无线网络性能下降，这就与我们的根本目标背道而驰了。

我们在主路由器上，使用简单的shell脚本来监测CPU及内存的使用情况，直接调用系统命令来实现。其结果如图4-11和图4-12所示。

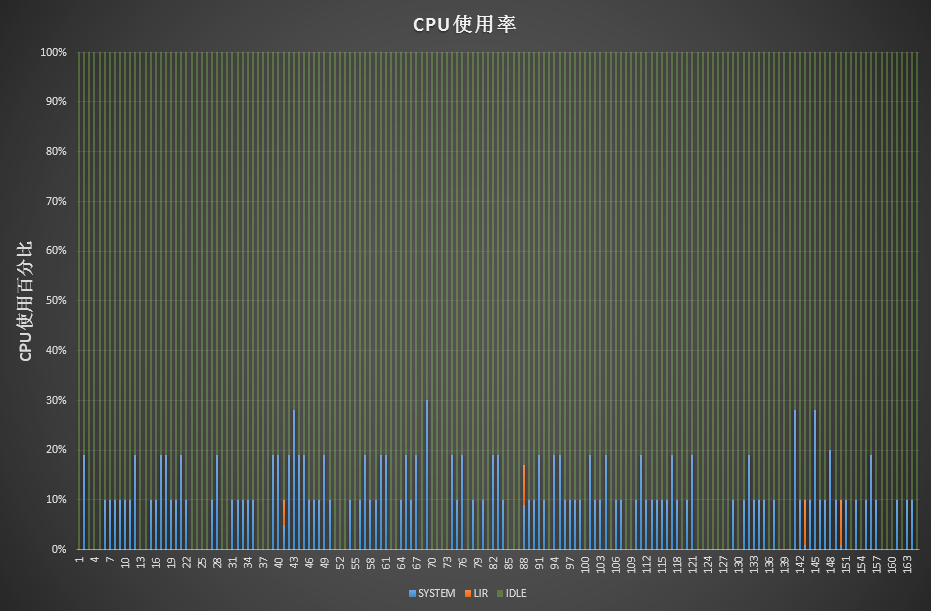


图 4‑10 路由器CPU使用情况

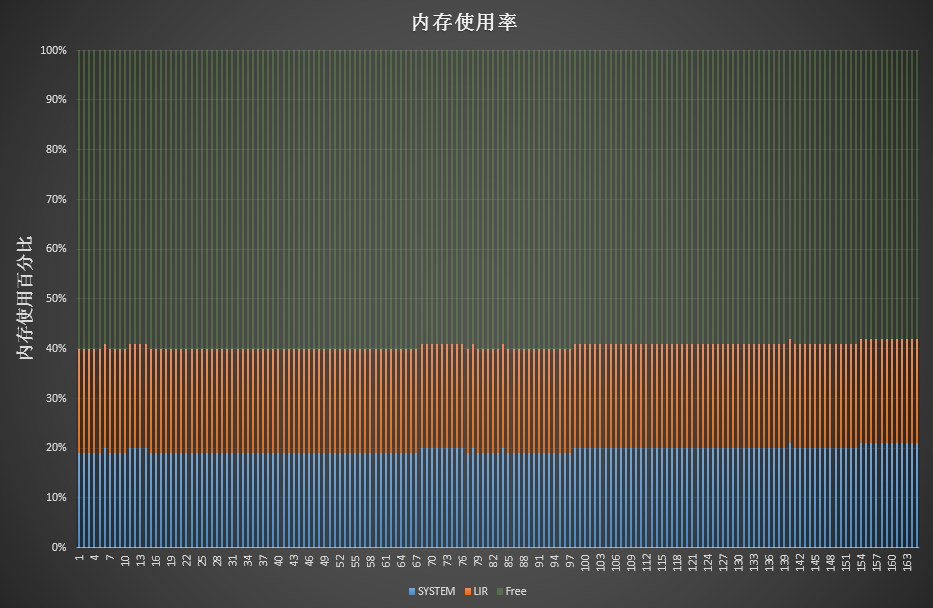


图 4‑11 路由器内存使用情况

由CPU的使用情况可以看出，在大部分的时间，计算程序的CPU使用率都接近于0%，在极少数情况下使用率将达到10%（例如，背景流量激增时）。

而另一方面，内存的占用率表现地非常稳定。这是由于计算程序所占用的内容主要是由抓包程序导致的。在抓包程序当中，会使用固定大小的结构体数组来缓存数据帧以用于预处理和最后的传输。因此使用的内存大小基本上是固定的。在我们的实验场景中，这一值为20%（考虑到实际使用的路由器，具体大小为20M）。在实际的部署当中，这一大小是可调节的。

综上所述，计算程序的运行并不会影响到路由器的正常使用，也不会降低无线网络的性能。

## 实际部署的干扰输出

我们将程序安装到一台实际使用的、搭载嵌入式Linux系统（Openwrt）的路由器当中，将该路由器作为某宿舍的无线网络接入点，让周围的同学使用其访问网络。我们在实际环境下测量了24小时的结果，我们使用局部累计直方图来显示干扰的大小。如图4-13所示。

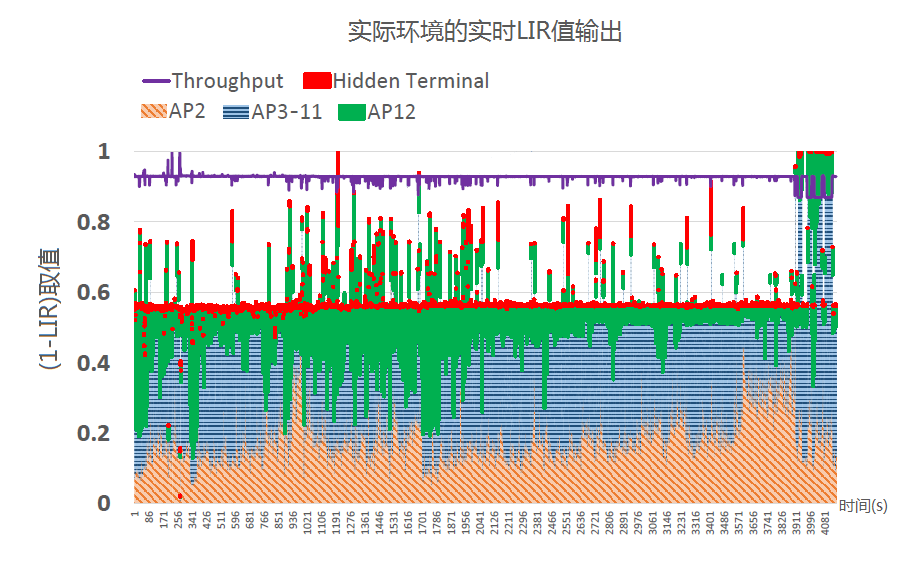


图 4‑12 实际环境下的干扰值输出图

为了更加直观的显示干扰值，我们在图中使用（1-LIR）来进行作图而非LIR本身。这种情况下LIR值越大，说明受到的干扰约严重。同时，由于（1-LIR）与干扰造成的延迟正相关，可以直观地反映出不同的干扰路由器对主路由器的干扰大小。同时该值也可以直接相加。图4-13反映了从某时刻（时间0）开始的一段时间内，不同干扰路由器的干扰大小。可以看出，有两台路由器（AP2和AP12）对主路由器干扰较大，其他路由器的干扰较小。

## 本章小结

本章中的实验结果表明，我们采用的新的干扰值计算方法，在准确性方面与以往的一些其他方法不相上下，在某些情形下的结果还要更优[3]。考虑到本实验是基于单一的路由器进行测量，不依靠额外的硬件网卡和嗅探器支持，因此在实用性方面有较大的优势。

在实验的过程当中，通过前期测量得到的结论，我们对后期的实验以及算法结构也作出了相应的调整，以提高整体性能和效率。另外，在实验中发现的某些实验结果，可能在某些特定的场景下提高效率（例如4.3.1节中的简化方式）。这对于今后的其他研究也可能有一定的帮助。

综上所述，我们采用的新的干扰值计算方法及其整个系统的架设是可行的，在正确性方面有较高的保证。同时，在实际的部署方面也更为简单，说明了本文所提出的实验目标是有实际意义的。

# 总结与展望

## 主要结论

在本文中，我们采用了一种新的轻量级的、用户层的、基于时间延迟的被动测量的干扰值计算方法。我们的测量系统能够在嵌入式Linux路由器上运行，并将与时间戳对应的干扰值信息作为输出。与以往的工作相比较，本文采用的方法可以再单一的路由器上运行，不需要额外的硬件网卡、路由器或者嗅探器的参与。因此，在实际的实验中和今后的大规模部署实验当中，会有一定的优势。

通过实验室环境的实测，本文所采用的LIR计算方法，在正确性方面得到了验证。不论是在载波侦听情形、隐藏终端情形还是多干扰的混合场景下，都显示出了较高的准确率。与以往相似的工作相比[3]，在正确性方面保持一致，并在某些特定的场景当中略有提升。

## 工作展望

本文所采用的计算方法，可以不需要多路由器和控制器的协同，独立地对自身受到的干扰情况进行测量。因此在实际的部署中有很大的优势。参考以往的研究工作，例如Bismark[10]、NAT[6]、WiSe[2]等。这些工作都需要多网卡或者多路由器支持，同时研究的实验环境中网络接入点也较为稀疏。考虑到我国的实际环境，居民楼的人口密度、路由器密度都很高，网络环境复杂，但另一方面网络带宽不是很高，背景流量不会特别大。在这种网络条件下，没有充分的实验来进行测量。而在我国，使用多网卡、多路由器的测量方式也不太可行，而大部分的用户和厂商也较为排斥主动测量的方法。因此，基于单路由器的被动测量方式是十分重要的、有意义的。本文所提出的这一测量方法，也正是立足于当前存在的实际问题。

除了前文描述的系统架构，本文中还提出了一些可行的优化方法，这在今后的研究工作当中，也可以作为参考的方向。

同时，目前也有一些研究工作集中在网络环境的优化上[9,16,17]。本文计算得到的干扰值，以带时间戳的形式，可以作为一些优化工作的输入[14,18]。利用已收集的数据，对路由器的参数进行调节，以提升自身的无线网络性能。更近一步地，在多台路由器的数据基础上，合理调节参数，让整体的性能达到最优。

在本文的工作基础上，可以进一步的改进和优化。例如，将数据的处理、计算过程由后台的服务器移至前端的路由器设备上。这种方式会加大路由器的计算负担，但可以有效地降低网络传输的开销。同时，也可以考虑使用多线程等更为高效的算法来减少计算量增大导致的性能下降。另一方面，本文主要侧重于无线端的干扰和延迟。在今后的研究当中，也可以考虑对有线端的性能进行分析[6,10]，并将二者结合到一起。

随着无线网络越来越普及，市面上也不断有“智能路由器”问世。可以预见，在今后的发展中，路由器的性能优化会受到越来越多的关注，而我们目前的工作，也是将来无线网络不断发展的重要的基础。

插图索引

[图 1‑1 家用路由器网络环境示意图 1](#_Toc390174097)

[图 1‑2 WiSe APs 监测的周围AP的信道使用状况[2] 3](#_Toc390174098)

[图 2‑1 四种载波侦听关系图[3] 8](#_Toc390174099)

[图 2‑2 包级别时间流程示意图 9](#_Toc390174100)

[图 2‑3 简化的无线干扰示意图 11](#_Toc390174101)

[图 3‑1 干扰值计算系统流程图 13](#_Toc390174102)

[图 3‑2 TCP帧格式[15] 15](#_Toc390174103)

[图 3‑3 有线端口输出八元组 15](#_Toc390174104)

[图 3‑4 802.11MAC帧结构 16](#_Toc390174105)

[图 3‑5 无线端口输出17元组 16](#_Toc390174106)

[图 3‑6 简化的无线协议FSA模型 20](#_Toc390174107)

[图 3‑7 改进后的简化FSA模型 21](#_Toc390174108)

[图 3‑8 干扰值输出结果示意图 25](#_Toc390174109)

[图 4‑1 载波侦听实验示意图 28](#_Toc390174110)

[图 4‑2 隐藏终端实验示意图 29](#_Toc390174111)

[图 4‑3 移动距离与吞吐率关系示意图 29](#_Toc390174112)

[图 4‑4 部分帧的队列延迟分布散点图 31](#_Toc390174113)

[图 4‑5 隐藏终端情形的吞吐率与LIR值 32](#_Toc390174114)

[图 4‑6 隐藏终端情形的吞吐率与LIR值拟合图（归一化后） 32](#_Toc390174115)

[图 4‑7 隐藏终端情形的误差分布 33](#_Toc390174116)

[图 4‑8 载波侦听情形的误差分布 34](#_Toc390174117)

[图 4‑9 干扰者A的误差分布 35](#_Toc390174118)

[图 4‑10 干扰者B的误差分布 35](#_Toc390174119)

[图 4‑11 路由器CPU使用情况 36](#_Toc390174120)

[图 4‑12 路由器内存使用情况 37](#_Toc390174121)

[图 4‑13 实际环境下的干扰值输出图 38](#_Toc390174122)

表格索引

[表 3‑1 802.11无线帧的部分类型 17](#_Toc390174124)

参考文献

1. ABI Research. https://www.abiresearch.co m/press/over-5-billion-wireless-con nectivity-chips-will-sh. Accessed: 2015-01-20.
2. Balachandran, Athula, et al. "Developing a predictive model of quality of experience for internet video." *ACM SIGCOMM Computer Communication Review*. Vol. 43. No. 4. ACM, 2013.
3. Balachandran, Athula, et al. "Modeling web quality-of-experience on cellular networks." *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014.
4. Cisco, I. "Cisco visual networking index: Forecast and methodology, 2013–2018." CISCO White paper (2014): 2013-2018.
5. Galletta, Dennis F., et al. "Web site delays: How tolerant are users?." *Journal of the Association for Information Systems* 5.1 (2004): 1.
6. J. Brutlag. Speed matters for Google Web search. http://services.google.com/fh/files/blogs/google\_delayexp.pdf, June 2009.
7. Jacobson, V., C. Leres, and S. McCanne. "libpcap, Lawrence Berkeley Laboratory, Berkeley, CA." *Initial public release June* (1994).
8. Mah, Bruce. "An empirical model of HTTP network traffic." *INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution., Proceedings IEEE*. Vol. 2. IEEE, 1997.
9. Sundaresan, Srikanth, et al. "Measuring and mitigating web performance bottlenecks in broadband access networks." *ACM Internet Measurement Conference*. 2013.
10. S. Souders. Velocity and the bottom line. http://radar.oreilly.com/2009/07/velocity-making-your-site-fast.html, July 2009.
11. Wang, Xiao Sophia, et al. "Demystifying Page Load Performance with WProf."*NSDI*. 2013.

致 谢

非常感谢实验室的裴丹老师、赵有健老师、徐恪老师对我的指导和帮助。各位老师在选题、系统设计和实验思路等方面都给了我很大的启发和帮助。通过与老师的讨论，我及时地发现了系统设计的不合理之处和算法上的一些漏洞并做出了相应的改进。在实验遇到了一些问题和困难时，老师积极地帮助我分析原因，使得我在问题处理方面得到了锻炼。

在实验的过程当中，实验室的师兄师姐们也给了我很大的帮助。非常感谢实验室的裴昶华师兄、隋楷心师姐等对我的帮助，给我提供了笔记本电脑、路由器等硬件设备帮助我完成实验。在平时遇到问题时，师兄师姐们都积极热心地解答我的问题。特别是在节假日时，我也经常能见到师兄师姐们的身影。再次感谢在百忙之中解答我的问题的实验室前辈们。

在此我也要感谢我的家人，他们虽然不在我的身边，但却一直在支持我、鼓励我，帮助我度过各种的困难。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日期：\_\_\_\_\_\_\_\_\_\_\_\_

附录A 外文资料的调研阅读报告

**Measurements of the Wi-Fi performance**

**Estimation of home wireless network**

With the rapid development of the Wi-Fi, more and more people pay close attention to the availability of wireless network. Most of the requests for wireless access often occur in narrow scenario like house, cafe, classroom, dormitory and small office. All these places share the features like single router, numerous devices and massed neighbouring routers or rogue access points. As the infrastructure of the home network, residential router plays an important role in the wireless network.

NAT[1] presents an empirical study of home networks. It presents three primary questions, the connection between home networks and the Internet, the connection performance of different devices in the home network scenario, and how do users use the Internet. However, it is surprisingly difficult to gather information about the home networks widely. As the view from the global Internet stays opaque due to network technologies such as network address translators( NATs). [1] points out that, in most occasions, as the only gateway, the residential router, a.k.a, the home wireless access point takes an irreplaceable position of the entire home network. Measuring from the gateway can provide both information inside and outside the home networks. Compared to previous studies, the measurement can present longitudinal quantitative results which see all devices coverd by the NAT as well as the last-mile statistics. [1] presented the first measurement study of home network on a large scale, it collected data from about 130 homes across 19 countries with both active and passive measurements.

BISmark[2] is a project provided by Georgia Tech. It presents a programmable gateway that performs continuous measurements of all the terminals in the home network. Since deploying measurement at the gateway devices, it can take advantage of the direct measurement of the last-mile info such as the ISPs’ access link performance and the effect of the wireless traffic. Furthermore, it provides longitudinal measurements which show the effects of the traffic patterns.

**Measurements of the last-mile**

Measurements from the gateway devices provide a direct access of the Ethernet. For most families, cable or optical fiber still cannot be abaddoned. Worsestill, people cannot access the Internet by themselves, which means the ISPs still stay on the stage. The performance of the ISPs’ service plan has a critical impact on the user experience.

Based on the prior work by [2], [3] focus on the web performance in broadband access networks. In that paper, they used a tool called *Mirage* to measure the page load time which can divide the whole load time into detailed time slices. The overall load time is divided into several parts, including DNS lookups, TTFB and download time. *Mirage* is easy to access and use. It can be deployed on home Aps which benefits for the large-scale measurement. After deployed on over 5,000 homes across the U.S., the large-scale home router-based web testing characterized the performance bottlenecks as two primary factors, the downstream throughput and the latency. The increases in the throughput lead to the decreases in download time of page objects. However, when the downstream reach a high value, because of the rising proportion of other parts of the overall load time, the improvement in throughput becomes insignificant. Meanwhile, the last-mile latency has an essential effect for the page load time since all transmissions will suffer from the delay caused by ISPs. Worsestill, small objects have more TCP overhead. [3] presents three possible optimizations for web performance, content caching, DNS caching and TCP connection caching. All three methods try to optimize latency but focus on different parts. The home router-based experiments proved the effectiveness of the optimizations when using popularity-based prefetching.

**Estimation of the link interference ratio**

Due to the increasing number of the wireless access points or hotspots, the wireless conflicts have an impact on the performance of the wireless network. In narrow scenario such as dormitories, the uncontrolled neighboring Aps will lead to the poor performance of the user’s router. And in school or company, the enterprise routers will be cumbered by the rogue Aps around.

PIE[4] presents a passive measurement of the link interference ratio, based on the previous work in [6], which uses active measuring on enterprise routers based on two-way TCP throughput . PIE presents two ways of estimation of the interference, one is based on the MAC-level loss rate, the other is based on packet-counting. They evaluate the accuracy of the estimation using controlled experiments with real-time traffic traces as the ground truth in [7]. PIE also employ their measurement in two different scenarios, one is the carrier-sense and the other hidden-terminal scenario. And the result of the interference can enhance the performance of other interference mitigation mechanisms.

WiSe[5] categorizes their deployment into three types. They put 30 Aps in three different apartment scenarios and kept them running for a few months. The access point platform is based on Openwrt with two mini-PCI Wi-Fi NICs. WiSe also designs a simple metric to describe the wireless performance called *Witt*, which is short for Wi-Fi based estimated TCP throughput. *Witt* uses the *link-experience* model as candidate which is related to the airtime utilization, local contention and effective rate. It also uses controlled experiments to discover the causes of poor performance by analyzing different indicators like airtime utilization, loss rate, signal strength and so on.

People have presented many methods for the estimation of the link interference ratio. Meanwhile, more and more commodity residential routers begin to pay close attention to the performance and the optimization. Unlike the enterprise routers with Ap-controllers, residential Aps are often controlled by the remote servicer running by the Ap productor. Through all the test of estimation, all the work of deploying and measurement, will finally lead to the birth of a well-performance autonomous wireless system.

调研阅读报告对应的原文索引

1. Grover S, Park M S, Sundaresan S, et al. Peeking behind the NAT: an empirical study of home networks[C]//Proceedings of the 2013 conference on Internet measurement conference. ACM, 2013: 377-390.
2. Sundaresan S, De Donato W, Feamster N, et al. Broadband internet performance: a view from the gateway[C]//ACM SIGCOMM Computer Communication Review. ACM, 2011, 41(4): 134-145.
3. Sundaresan S, Feamster N, Teixeira R, et al. Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks[C]//ACM Internet Measurement Conference. 2013.
4. Shrivastava V, Rayanchu S, Banerjee S, et al. PIE in the sky: online passive interference estimation for enterprise WLANs[C]//Proc. of NSDI. 2011.
5. Patro A, Govindan S, Banerjee S. Observing home wireless experience through WiFi APs[C]//Proceedings of the 19th annual international conference on Mobile computing & networking. ACM, 2013: 339-350.
6. Padhye J, Agarwal S, Padmanabhan V N, et al. Estimation of link interference in static multi-hop wireless networks[C]//Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005: 28-28.
7. Manweiler J, Franklin P, Choudhury R R. RxIP: Monitoring the health of home wireless networks[C]//INFOCOM, 2012 Proceedings IEEE. IEEE, 2012: 558-566.

综合论文训练记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学生姓名** |  | **学号** |  | **班级** |  |
| **论文题目** |  | | | | |
| **主**  **要**  **内**  **容**  **以**  **及**  **进**  **度**  **安**  **排** | **指导教师签字：**  **考核组组长签字：**  **年 月 日** | | | | |
| **中**  **期**  **考**  **核**  **意**  **见** | **考核组组长签字：**  **年 月 日** | | | | |
| **指**  **导**  **教**  **师**  **评**  **语** | **指导教师签字：**  **年 月 日** | | | | |
| **评**  **阅**  **教**  **师**  **评**  **语** | **评阅教师签字：**  **年 月 日** | | | | |
| **答**  **辩**  **小**  **组**  **评**  **语** | **答辩小组组长签字：**  **年 月 日** | | | | |

**总成绩：**

**教学负责人签字：**

**年 月 日**