基于 WIFI 探针的商业大数据分析技术

Commercial data analysis technology based on WIFI probe

需求规格说明书

参赛学校：河海大学常州校区

组　　名：Super Super Handsome

指导老师：陈慧萍

队　　长：魏臻江

队　　员：丁翰雯、陶宇

目录

软件需求规格说明书

1.引言

1.1 编写目的

本需求说明书目的在于：将用户提供的需求描述系统化、精确化、全面化。从而实现：

(1) 便于用户、分析人员和设计人员进行理解和交流。

(2) 支持目标软件系统的确认。

(3) 控制系统进化过程。

预期读者：软件设计者和测试者。

1.2 项目背景

随着科学技术的高速发展，我们已步入数字化、网络化的时代。网上购物愈益流行，然而，受到产品质量检验和实际体验感的限制，线下商店当然是不可替代的，很多顾客仍然希望进入实体店亲身试用挑选。

首先开发探针设备能够采集客户唯一的定位标识，比如 MAC 地址，通过数据分析技术，采用离线计算和实时计算结合的方式，为商业环境提供科学的、全面的数据决策依据。不仅对营销能力的评估，也可以对管理上进行优化。利用探针数据的客流分析打破模式束缚，不仅仅只是提供可信的客流数据分析，同时还

还可以通过本系统随时查看店铺内的客流量情况，并根据客流高峰时段，对店内工作人员进行合理分配，提高人力资源利用率，并在一定程度上降低经营成本。

1.3 定义

WIFI 探针：WIFI 探针技术是指基于 WIFI 探测技术来识别 AP(无线访问接入点)附近已开启 WIFI 的智能手机或者 WIFI 终端（笔记本，平板电脑等），无需用户接入 WIFI，WIFI 探针就能够识别用户的信息。

大数据：指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

1.4 参考资料

[1]《Spark 快速大数据分析（第一版本)》

[2]《Hadoop 基础教程》

[3]《精通 Hadoop(第一版本)》

[4]《响应式 Web 设计：HTML5 与 CSS 实战》

[5]《JavaScript 权威指南（第 6 版）》

[6]《疯狂 Ajax 讲义（第 3 版）》

[7] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶, 周涛. 《大数据时代》[J]. 教育科学论坛, 2013, 8(7):27-31.

[8] 殷人昆,郑人杰,马素霞,白晓颖. 实用软件工程(第三版)[J]. 计算机教育, 2010(24):95.

[9]《猫酷室内行为采集系统》 **http://www.mallcoo.cn/action.html**

2.任务概述

2.1 目标

本系统旨在通过 WIFI 探针收集顾客 MAC 及与探针的距离、出现时间地点等信息，来分析门店的客流情况、精准监控客流质量、实时展示客流转化情况，从而帮助检测营销效果、发现潜在机会和改进措施，为便捷、高效精细化运营提供全方位的数据参考。

基本的技术目标包括：

(1) 用 WIFI 探针收集顾客信息，实现实时的客流量监测并实现环比与历史比较；

(2) 根据历史客流量，预测未来时刻的店内客流量，以便商家进行人员调度；

(3) 获取顾客实时入店量并予以实时展示并实现环比与历史比较，从而了解

进入店铺或区域的客流及趋势；

(4) 分析比较得出实时入店率并予以实时展示并实现环比与历史比较，从而获取进入店铺或区域的客流占全部客流的比例及趋势；

(5) 快速分析得出顾客来访周期从而实现对进入店铺或区域的顾客按照距离上次来店不同间隔实现动态归类；

(6) 顾客活跃度：按顾客距离上次来访间隔,划分为不同活跃度（高活跃度、中活跃度、低活跃度、沉睡活跃度）；

(7) 快速分析进入店铺的顾客在店内的停留时长并实现动态归类予以实时展示；

(8) 根据驻店时长来判断进入店铺后很快离店的顾客及占比(占总体客流)即跳出率，并实现实时展示与小时、日、周、月多维度环比以及历史比较；

(9) 根据顾客停留时长判定计算进入店铺深度访问的顾客及占比(占总体客流)即顾客深访率并予以实时展示与小时、日、周、月多维度环比以及历史比较

(10) 统计商家推送的营销方案使用率，辅助商家进行调整。

(11) 实现使用短信控制模块控制 WIFI 探针的开关，与通过数据获取情况简单甄别判断探针的状态予以实时监控。

2.2 用户的特点

本系统的最终用户分为管理员用户和分店用户两类，其中管理员用户拥有一定的计算机操作技术，是商场的管理人员、销售部的人员，允许他们查询每一个店铺的各项信息，包括客流量、入店量、来访周期、新老顾客、顾客活跃度、驻点时长、跳出率、深坊率、顾客访问次数、趋势预测等实时记录和历史记录；分店用户是每一位店铺老板，只需要具备简单的计算机操作能力，可以查看本店的客流量、入店量、来访周期、新老顾客、顾客活跃度、驻点时长、跳出率、深坊率、顾客访问次数、趋势预测等实时记录和历史记录。

预计本系统的使用频率为：2000 人次/天。

2.3 运行环境

运行基于 WIFI 探针的商业大数据分析系统的开发端所需的环境要求如下：

2.2.1 硬件环境

(1) CPU：Intel CoreI5 1.8GHz 及以上
(2) 内存：2G×3 及以上
(3) 硬盘：60G 及以上
(4) 探针：双核 探测距离半径>100 米频率 2.4GHz-2.5GHz

2.2.2 软件环境

(1) 服务器： Tomcat / IIS（tomcat 和 IIS 需启动 CGI 支持）
(2) 操作系统：Ubuntu 16.04 LTS
(3) 数据库：HBase 1.1.2（分布式数据库）
(4) 基本配置：Spark 2.1.0（Built for Hadoop 2.2.0）分布式环境、JDK 1.7 及以上、Scala 2.12.1 及以上
(5) 开发工具：IntelliJ IDEA 2017.1.1 及以上、Eclipse 3.6 及以上
(6) PC 端：IE6.0 及以上版本；IE 内核的其它浏览器；Chrome21.0 等
(7) 手机端：自带浏览器即可

2.4 假定和约束

基于 WIFI 探针的商业大数据分析系统使用的假定和约束主要有如下几点：

(1) 此系统有且仅有分店、管理层两类用户使用；
(2) 预测客流量数据的计算模型，是根据历史客流量数据得来的，不能保证完全正确，仅供参考；
(3) WIFI 探针通过客户手持设备的信号强度估测其所在区域，并不能实现精准定位且存在顾客手持多台设备或者设备未开启 WIFI 的情况从而数据获取存在一定误差；
(4) 目前的版本支持 IE6.0 及以上版本的浏览器，Chrome21 等，对于较低版

本的浏览器可能会出现页面错乱等现象；

(5) 条件限制不能提供 1000 台设备同时运行的测试环境从而可能发生未预知的错误。

3.数据描述

3.1 静态数据

序号	名称	描述	用途
1	历史数据	用做模型建模的数据输入	建立入店量、活跃度、驻店时长等数据预测模型和用作环比与历史对比

3.2 动态数据

序号	名称	类型	描述
1	原始数据	INPUT	采集数据
2	结构化数据	INPUT	原始数据
3	模型数据	INPUT	分析预测顾客活跃度、客流量
4	采集的数据	INPUT	模型学习趋势预测 顾客活跃度、客流量
5	模型数据	OUTPUT	模型数据文件
6	采集的数据	OUTPUT	原始数据
7	客流量的分析、预测结果	OUTPUT/INPUT	数据可视化
8	入店量的分析、预测结果	OUTPUT/INPUT	数据可视化
9	入店率的分析、预测结果	OUTPUT/INPUT	数据可视化
10	新老顾客比例的分析、预测	OUTPUT/INPUT	数据可视化
11	顾客活跃度的分析、预测结果	OUTPUT/INPUT	数据可视化
12	深访率的分析、预测结果	OUTPUT/INPUT	数据可视化
13	跳出率的分析、预测结果	OUTPUT	数据可视化
14	可视化数据	OUTPUT	管理层情况 分店情况

3.3 数据库描述

本系统的数据库为分布式的 HBase，采用键值对的形式进行数据存储。

序号	名称	描述
1	实时数据	存储所有探针采集到的实时数据
2	客流量及历史	存储当天每个时段的客流量数据
3	当前店内人数	存储每个时段店内人数，
4	入店量入店率及历史	存储当天以及过去日、周、月的的入店量入店率数据
5	顾客驻留时间及历史	存储根据驻店时长长短分类的各类人数划分阈值为 [30s,30s-1min,1min-5min,5min 以上]
6	新老顾客人数比例及历史	存储根据顾客上次来访时间判别的新老顾客（间隔一个月为阈值）的人数及比例
7	顾客访问次数及时间历史	存储顾客在规定时间内访问次数的访问次数
8	深访率及历史	存储根据驻店时长判定的深访顾客人数
9	跳出率及历史	存储根据驻店时长划分，标准为低于 30s 算作跳出顾客来判定的人数
10	活跃度及历史	存储根据顾客访问次数得出的活跃度及历史数据

4.对功能的规定

4.1 功能划分

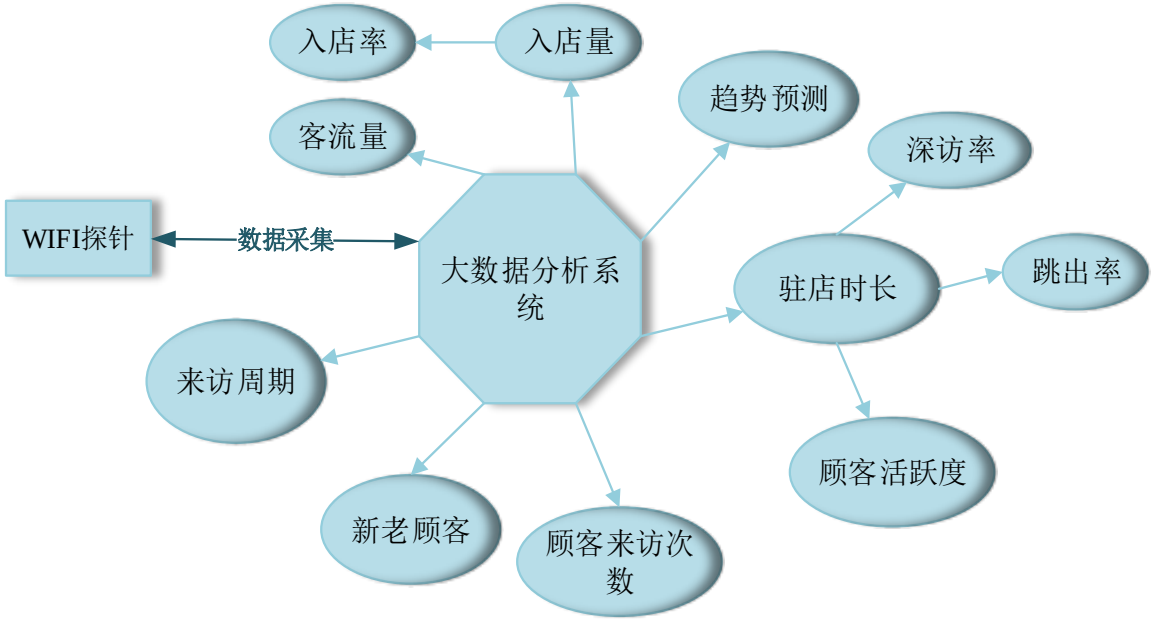
基于 WIFI 探针的商业大数据分析系统主要实现三方面的功能：一是通过探针设备采集可监测范围内的手机 MAC 地址、地理位置、与探针距离、时间等信息；二是探针采集的数据可以定时发送到服务端保存；三是利用大数据技术对数据进行人流量等指标的分析。系统应具备以下功能：

- WIFI 探针：通过探针设备采集可监测范围内的手机 MAC 地址、地理位置、与探针距离、时间信息（特别针对 ANDROID6.0 和 IOS10 版本后的移动

终端设备进行测试需能采集到 MAC 地址）。

- 数据采集：服务端接收探针定时发送的数据，将数据保存到数据分析平台待用。接收数据不能有数据丢失或者数据失真，探针每 3 秒发送一次数据，数据采集并发量即单台服务器的数据吞吐量，不得低于 1000 台设备；
- 数据分析：基本能够分析十一大指标（客流量、入店量、来访周期、新老顾客、顾客活跃度、驻点时长、跳出率、深坊率、顾客访问次数、趋势预测），支持环比和历史对比，并可以以从小时、日、周、月多维度分析，采用离线计算和实时计算技术相结合的方式。

功能需求可用如下的系统功能划分图描述：



4.2 功能描述

基于 WIFI 探针的商业大数据分析系统功能需求包括 WIFI 探针设备、数据采集、数据分析（十一个指标）。

4.2.1 WIFI 探针设备

WIFI 探针设备是通过淘宝购买，可以进行服务器的相关配置（服务器 IP、端口、路径、发送时间间隔），能够采集 MAC 地址、地理位置信息、与探针的大概距离、采集时间等信息（特别针对 ANDROID6.0 和 IOS10 版本后的移动终端设备进行测试需能采集

到 MAC 地址）。

4.2.2 数据采集

在本系统中，单独写一个服务器，用来接收探针定时发送的数据，将数据保存到数据分析平台待用，文件系统使用 HDFS 分布式文件系统，保存在 HBase 数据库中，设置探针每三秒发送一次数据，采集的原始数据为 JSON 结构。以下为数据流的定义：

数据流名称：原始数据
描述：探针采集到的数据 组成：mac 地址+时间戳+信号强度+手机距离嗅探器的距离+经纬度+地址 来源：WIFI 探针 终点：接收服务器

数据流名称：结构化数据
描述：系统分类统计的数据 组成：结构化 mac 地址+结构化时间戳+结构化手机距离+经纬度 来源：接收服务器 终点：存储至 HBase

4.2.3 数据分析

系统基本能够分析以下十一个指标，支持它们的环比和历史对比，并可以以从小时、日、周、月多维度分析，同时采用离线计算和实时计算技术相结合的方式。

序号	指标	分析处理
1	客流量	描述：店铺或区域整体客流及趋势 输入数据流：原始数据 输出数据流：结构化数据 存取的数据库：接收服务器 HBase 处理逻辑：while(探针检测到原始数据){ 统计 MAC 数 存入数据库

		原始数据表中添加对应记录 }

2	入店量	描述：进入店铺或区域的客流及趋势
		输入数据流：原始数据
		输出数据流：结构化数据
		存取的数据库：接收服务器　HBase
		处理逻辑：if（到某一探针的距离小于阈值且到另外一个探针的距离小于阈值）{
		if(该顾客不在这个店的入店记录中)
		结构化数据添加到对应记录
		}
		else{
		if(该用户在这个店的入店记录)
		删除对应记录并添加该顾客的访问时间
		}

3	入店率	描述:进入店铺或区域的客流占全部客流的比例及趋势
		输入数据流：结构化数据
		输出数据流：结构化数据
		存取的数据库：HBase　HBase
		处理逻辑：while(得到客流量和入店量){
		入店率=入店量 / 客流量
		}

4	来访周期	描述：进入店铺或区域的顾客距离上次来店的间隔 输入数据流：原始数据 输出数据流：结构化数据 存取的数据库：接收服务器　HBase 处理逻辑：if（到某一探针的距离小于阈值且到另外一个

		探针的距离小于阈值） if（该顾客不在来访记录中） 结构化数据添加到对应记录中 else 来访间隔=这次访问时间-上次访问时间
5	新老顾客	描述：一定时间段内首次/两次以上进入店铺的顾客 输入数据流：结构化数据 输出数据流：结构化数据 存取的数据库：接收服务器、HBase HBase 处理逻辑：if（到某一探针的距离小于阈值且到另外一个探针的距离小于阈值） if（该顾客不在访问记录中） 结构化数据添加到对应记录中 新顾客 else 老顾客
6	顾客活跃度	描述：按顾客距离上次来访间隔,划分为不同活跃度（高活跃度、中活跃度、低活跃度、沉睡活跃度） 输入数据流：结构化数据 输出数据流：结构化数据 存取的数据库：接收服务器、HBase HBase 处理逻辑：if（到某一探针的距离小于阈值且到另外一个探针的距离小于阈值） if（该顾客在来访记录中） 来访间隔=这次访问时间-上次访问时间 划分活跃度
7	驻点时长	描述：进入店铺的顾客在店内的停留时长 输入数据流：结构化数据

		输出数据流：结构化数据
		存取的数据库：接收服务器、HBase　HBase
		处理逻辑：if（到某一探针的距离大于阈值且到另外一个探针的距离大于阈值）{
		if(该用户在这个店的入店记录)
		驻地时长=出店时间-进店时间
		}

8	跳出率	描述：进入店铺后很快离店的顾客及占比(占总体客流) 输入数据流：结构化数据 输出数据流：结构化数据 存取的数据库：接收服务器、HBase　HBase 处理逻辑：if（驻点时间<某阈值） 跳出顾客+1 跳出率=跳出顾客数 / 客流量
9	深坊率	描述：进入店铺深度访问的顾客及占比(占总体客流) 输入数据流：结构化数据 输出数据流：结构化数据 存取的数据库：接收服务器、HBase　HBase 处理逻辑：if（驻点时间>某阈值） 深坊顾客+1 深坊率=深坊顾客数 / 客流量
10	访问次数	描述：一段时间内顾客来访的次数 输入数据流：结构化数据 输出数据流：结构化数据 存取的数据库：接收服务器、HBase　HBase 处理逻辑：if（到某一探针的距离小于阈值且到另外一个探针的距离小于阈值） if（该顾客不在来访记录中）

		访问次数=1 else 访问次数+1

11	趋势预测	描述：根据历史数据预测之后一段时间的客流量、入店量等
		输入数据流：采集的数据和模型数据
		输出数据流：趋势预测数据
		存取的数据库：原始数据、模型数据　HBase
		处理逻辑：while（采集到数据）{
		对比历史客流量、入店量等
		用模型计算未来客流量、入店量等
		}

5.对性能的规定

5.1 数据精确度

探针输入数据精确度要求：小数点后保留 4 位有效数字

探针输出数据精确度要求：小数点后保留 4 位有效数字

传输过程中的精确度要求：小数点后保留 4 位有效数字

数据存储的精确度要求：小数点后保留 4 为有效数字

5.2 数据存储容量

基于 WIFI 探针的商业大数据分析系统预期的数据存储容量是 HBase 的存储记录数目大于等于一亿条。

5.3 时间特性要求

(1) 响应时间：查询、计算分析数据的响应时间控制在 3 秒内，数据库相关操

作的响应时间也必须控制在一定范围内：基本的信息变更验证：写卡时间控制在 0.5 秒之内；数据库访问：应控制在 2 秒之内（对数据表的遍历控制在 0.8 秒之内，对数据表的删除控制在 1 秒之内）

(2) 更新处理时间：在网络无故障的情况下的局域网数据库中，对数据执行增删查改操作时，数据库的操作响应时间控制在 0.02 秒/条之内。

(3) 数据的转换和传送时间：在拨号网络连接通后，交换数据以数据单元形式进行，所有数据交换过程控制在5分钟内。

(4) 运行时间：程序启动个初始化时间控制在3秒之内；

5.4 灵活性

(1) 操作方式上的变化：该系统客户端适用于任一 Windows xp 及以上、Ubuntu 16.04 LTS 操作系统，同时也能在手机端通过网页访问。

(2) 运行环境的变化：要求数据计算平台使用 Linux 平台（Ubuntu 16.04 LTS），要求离线计算和实时计算平台必须是分布式环境（Spark 2.1.0 Built for Hadoop 2.2.0），此外该系统可以运行在 Ubuntu 16.04 LTS 及以上的操作系统。

(3) 同其他软件的接口的变化：可以满足 B/S、C/S 两种类型。操作简单，好用、易用。

(4) 系统升级历史数据的变化：升级后的系统会自动保留用户数据。

(5) 精度和有效时限的变化：可以根据实际情况自行设置。

6.其他的规定

6.1 数据管理能力要求

需要管理的记录个数：大于 1 亿。其中分为多个表，其大小规模为：1 万左右，记录的总个数每天将增长 10%~20%，存储硬盘应大于 500GB。

6.2 故障处理要求

发生错误时，保证数据完整，对于数据库发生故障时要能够进行故障恢复，以保证

数据的一致性同时也要定期进行数据备份。

序号	名称	处理方式
1	硬件传输数据故障	由硬件负责人进行故障排查和修复
2	模型欠学习或过拟合	重新训练模型
3	推送或者接入 WIFI 故障	重启
4	服务器故障	重启
5	数据库无响应	重启
6	断电	挂起任务
7	网络故障	硬件负责人进行故障排查

6.3 其他专门要求

6.3.1 可维护性

维护人员会在定期进行维护和检验，利用可靠的密码技术，掌握特定的记录或历史数据集，便于维护。

6.3.2 安全保密性

所有用户的信息经过加密后存储在数据库，如需查询有关信息，必须通过 严格的身份验证，以防数据泄露。

6.3.3 可使用性

该系统界面友好，功能详细简单，方便使用。

7.运行环境规定

7.1 用户界面

按照网页用户界面的规范来设计，使用窗体为主的用户界面，搭配柱状图、饼状图、折

线图、数据表等形式，便于用户使用和查看。

7.2 硬件接口

平台工作涉及的接口 WIFI 探针与外部服务器的通信接口，个人计算机与外部服务器通信接口，个人计算机与应用服务器通信接口，应用服务器与数据库服务器通信接口和应用服务器与 FTP 服务器通信接口。

用户使用个人计算机，访问外部网络，并从外部的服务器中获取 WIFI 探针分析处理后的数据，并进行远程操作。个人计算机通过带有防火墙安全设置的网络连接到应用服务器，并向应用服务器发送数据和操作请求。应用服务器与 FTP 服务器和数据库服务器直接相连，其根据个人计算机发送的请求，返回来自数据库服务器和 FTP 服务器的内容，或对数据库服务器和 FTP 服务器上的数据信息进行读写。

7.3 软件接口

本系统通过网络提供服务，用户通过浏览器访问服务器，向服务器发出服务请求。因此，需要使用 TCP/IP 网络协议，作为标准的通信控制接口。