# 基于超算平台的公共 Wi-Fi 无线网络无痕信息获取与舆情分析系统研究

杨 明 舒明雷 顾卫东 郭 强 周书旺
(山东省计算中心 济南 250014)

**摘 要** 提出一种利用国家超级计算济南中心的千万亿次计算平台对整个城市范围内的公共 Wi-Fi 无线网络进行信息获取和舆情分析的系统,它基于非介入式的无线数据包捕获技术、Web 页面还原与容错修复技术、多种文本挖掘技术和海量数据处理技术,可对公共 Wi-Fi 无线网络中的各种非法行为进行取证,对网络舆情进行准确分析和预测,可为相关部门的网络舆论导向工作提供全面准确的参考。

**关键词** 超级计算,无线网络,信息获取,舆情分析
**中图法分类号** TP393 **文献标识码** A

## Research on Non-intervention Information Acquisition and Public Sentiment Analysis System for Public Wi-Fi Wireless Networks Based on Supercomputer Platform

YANG Ming SHU Ming-lei GU Wei-dong GUO Qiang ZHOU Shu-wang
(Shandong Computer Science Center, Ji'nan 250014, China)

**Abstract** An information acquisition and public sentiment analysis system for the city public Wi-Fi wireless networks was presented, which uses the petaflops computing platform in National Supercomputer Center in Ji'nan. Based on the non-intervention wireless packets capture technology, Web page recovery and fault-tolerant reassembly technology, multiple text data mining technology and mass data process technology, the system can implement the functionality of wireless network forensics, public sentiment analysis and prediction, and provide overall and accurate references for the guidance of public sentiment for the government.

**Keywords** Supercomputer, Wireless network, Information acquisition, Public sentiment analysis

## 1 引言

伴随中国"无线城市"建设的加快,在整个城市范围内实现无线网络覆盖和服务,向公众提供随时随地的无线网络接入已成为现实。基于 Wi-Fi(Wireless Fidelity,无线保真)技术标准的 WLAN(Wireless Local Area Network,无线局域网)具有覆盖范围广、传输速度高、建设费用低等优点,已成为当前最普及的无线网络形式之一。在许多城市的商业区、住宅小区、高校、交通枢纽、酒店、休闲娱乐等公共场所均已实现 Wi-Fi 信号覆盖,携带支持 Wi-Fi 的终端即可接入互联网。

然而,公共 Wi-Fi 无线网络在为社会生活提供极大便利的同时,其开放性所导致的安全问题也变得尤为突出。一方面,由于公共无线网络用户的流动性和复杂性,各种非法信息越来越多地通过开放性网络传播;另一方面,由于公共无线网络中的舆情具有更高的突发性,这更广泛地反映了社会热点和焦点信息,若缺乏及时有效的分析和正确的引导,将对社会稳定带来很大的负面影响。

因此,相关部门迫切需要一种有效的公共 Wi-Fi 无线网络监管系统,它既可对无线网络中的非法信息进行无痕捕获和取证,又可快速处理无线网络中的海量数据,实现对网络舆情的准确分析和预警。

## 2 系统的关键技术分析

常规的互联网舆情分析系统[1],主要是基于网络爬虫工具,按某种策略对互联网 Web 页面进行遍历和下载,经预处理后再综合运用多种文本挖掘方法,从多方面对网络舆情进行分析和预警。而本文所研究的公共 Wi-Fi 无线网络无痕信息获取与舆情分析系统,其舆情挖掘对象为整个城市范围全部公共 Wi-Fi 无线网络内实际捕获的海量数据,也即全部 Wi-Fi 用户所实际浏览的海量 Web 页面,因此所得的网络舆情更准确、更真实地反映出该地区的舆论趋势;同时,该系统通过空中链路捕获无线数据,对无线网络环境无任何改变和影响,真正符合无痕信息获取的需求。涉及的关键技术包括无线数据包捕获、Web 页面还原和文本数据挖掘等。

无线捕包是进行信息还原与文本挖掘的基础和前提。常规方法是将无线网卡设置为射频监听模式,基于通用捕包函

到稿日期:2012-10-19 返修日期:2012-12-29 本文受山东省科学院青年科学基金项目(科基合字 2011 第 10 号)资助。

杨 明(1981—),男,博士,助理研究员,主要研究方向为无线通信与网络,E-mail:yangm@keylab.net;舒明雷(1979—),男,博士生,助理研究员,主要研究方向为移动通信、无线传感器网络;顾卫东(1970—),男,硕士,研究员,主要研究方向为超算、信息安全、云计算;郭 强(1975—),男,博士,副研究员,主要研究方向为移动通信、无线传感器网络;周书旺(1985—),男,硕士,研究实习员,主要研究方向为无线通信与网络。

数库 LibPcap/WinPcap 所提供的接口实现数据嗅探[2]。其缺陷在于，一方面，在吞吐量较大的宽带无线网络(如基于 IEEE 802.11n 标准的 Wi-Fi 网络可达 300Mbps)中，LibPcap/WinPcap 的丢包率将随单位时间内网卡接口所到达数据包的增加而迅速升高[3]；另一方面，在电磁环境非常复杂的公共网络中，若捕包网卡的位置未经优化部署，捕包效果也将严重恶化[4]。

Web 页面还原是对所捕获的无线数据进行可视化解析，为文本挖掘提供数据支撑的关键。常规方法是基于 TCP/IP 协议族对数据包进行逐层解析，然后根据不同的 Internet 应用进行信息重建。其缺陷在于仅能处理相对较完整的捕包数据，如在有线网络中通过对集线器(HUB)[5]或路由器[6]进行镜像而获取的数据，或在丢包率较低的低速无线网络中获取的数据[7]。而针对丢包严重且因无痕获取需求而无法要求丢包重传的捕获数据，常规方法便无法对 Web 页面进行修复和部分还原。另外，常规方法和常规计算平台亦无法处理海量捕包数据的 Web 页面还原。

文本挖掘是基于超级计算平台进行海量数据处理，最终提供网络舆情分析服务的核心技术。当前对文本挖掘的研究和应用主要集中于热点话题发现[1]、敏感话题发现[8]、舆情倾向性分析[9]和舆情趋势预测[10]几方面。话题发现技术的难点在于对海量数据和动态舆情进行文本聚类时的优化处理，以及对话题热度和敏感度的准确评价。当前的舆情倾向性分析技术主要包括基于语言学知识的方法和基于机器学习的方法，其不足之处在于对情感词典的依赖性较大，而且文本模型中缺乏对语义信息的描述。对于舆情趋势分析，当前技术主要集中在基于时间序列模型的短期预测方面，而无法预测舆情发展的长期趋势。

综上，当前对于高吞吐量复杂无线网络中高效率的无痕捕获技术、丢包情况下的信息还原容错技术以及海量数据的文本挖掘与舆情分析技术的研究都尚不成熟。

## 3 系统的核心算法设计

首先，从无线捕包和 Web 页面还原方面对无痕信息获取的核心算法进行设计；然后，从页面预处理和多种文本挖掘技术的优化改进方面对网络舆情分析的核心算法进行设计。

### 3.1 无线数据包捕获的算法设计

#### 3.1.1 基于多网卡数据融合机制的无线捕包算法

针对单网卡的无线接收范围有限和在公共网络中受电磁干扰影响较大的问题，本文提出基于多网卡数据融合机制的捕包算法，以从硬件方面提高对无线数据包的捕获性能。以 $Sniffer_i$ 表示公共网络内的第 $i$ 个捕包网卡，各网卡部署于不同位置，则融合算法可表述为：

Step1 以 $SNR(x,y)$ 表示公共网络内坐标 $(x,y)$ 处的无线信号信噪比。对于主要捕获 AP(无线接入点，Access Point)发送至用户的下行数据的网卡，其部署位置应在信噪比最大的位置，即

$$(x,y):\max_{(x,y)} SNR(x,y) \tag{1}$$

而主要捕获用户发送至 AP 的上行数据的网卡，其部署位置应在用户密集且信噪比抖动最小的位置，即

$$(x,y):\min_{(x,y)} \partial[SNR(x,y)]/\partial x\partial y \tag{2}$$

Step2 以 $p_i=(p_1^i,p_2^i,\cdots,p_{N_i}^i)$ 表示 $Sniffer_i$ 所捕获的数据包序列，这里 $N_i$ 表示 $Sniffer_i$ 的捕包总数。对于 $Sniffer_i$ 未能捕获或接收错误的数据包，部署于不同位置的其它捕包网卡可能以一定概率成功捕获，因此将各 $p_i$ 融合并滤除冗余数据包，即可获取较完整的捕包数据 $p$：

$$p=\bigcup_i p_i=\bigcup_i (p_1^i,p_2^i,\cdots,p_{N_i}^i) \tag{3}$$

#### 3.1.2 基于优化的驱动机制的无线捕包算法

网卡每捕获一个数据包都会触发中断，当网络流量较大时，CPU 将陷入过度频繁的中断状态而无法处理已接收的数据包；而从捕包到处理数据包，每次都需经过网卡至用户程序的多次复制，耗费了大量系统资源。本文提出数据包响应机制和传输机制的优化算法，以从驱动和软件方面提高对无线数据包的捕获性能。

(1)基于中断与轮询相结合的数据包响应机制：在批量数据包中的首个包到达时，以中断方式唤醒捕获进程，然后以轮询方式读取数据包，可有效降低中断开销，使更多资源用于数据包解析和处理。

(2)基于内核缓冲过滤和内存映射的数据包处理机制：在系统内核的环形缓冲中，对符合捕获条件的数据包进行过滤；将系统内核的缓冲区映射到用户缓存，可使用户进程直接对此内存访问，大大提高了数据包的处理效率。

### 3.2 Web 页面还原的算法设计

#### 3.2.1 TCP 数据流组装和数据流匹配的快速算法

Web 页面还原过程中，将无线数据包组装为 TCP 数据流的处理是其主要瓶颈，可占用 70%以上的处理时间[6]。常规算法以源、目的 IP 和端口对($srcIP$, $srcPort$, $dstIP$, $dstPort$)作为 TCP 数据流标识，其缺陷在于处理 HTTP1.1 协议的持久性连接时，需引入额外的处理模块才可解析同一 TCP 数据流中的多个页面文件和页面多媒体文件。对同一 TCP 连接中的各页面文件和页面多媒体文件，其请求信息、响应信息所对应的 TCP 数据流间都满足以下关系：

$$\begin{cases} SEQ_{\text{请求流末端数据包}}+Length_{\text{请求流末端数据包的应用层}}= \\ \quad ACK_{\text{响应流}} \\ SEQ_{\text{响应流首个数据包}}=ACK_{\text{请求流}} \end{cases} \tag{4}$$

可用六元组 $R_{HTTP}$ 唯一索引每对 HTTP 请求及响应数据流：

$$R_{\text{HTTP}}:(clientIP,\ clientPort,\ serverIP,\ serverPort,\ getACK,resACK) \tag{5}$$

式中，$getACK$ 和 $resACK$ 分别表示 HTTP 请求流和响应流所对应的 TCP 确认号。而对于每个数据包，可用五元组 $R_{packet}$ 唯一索引：

$$R_{packet}:(srcIP,srcPort,dstIP,dstPort,ACK) \tag{6}$$

基于以下 3 方面分析：多数 HTTP 请求流仅含 1 个数据包；丢包情况下式(4)中的前者具有更高的可靠性；$R_{HTTP}$ 中各元素对于唯一索引每对 HTTP 的贡献度排序为：

$$clientPort>serverIP>clientIP \tag{7}$$

可获得 TCP 数据流组装以及将请求流、响应流进行匹配的快速算法，如图 1 所示。

Step1 从数据包队列读取一个数据包五元组 $R_{packet}$

Step2 从 HTTP 请求及响应索引表读取一个六元组 $R_{HTTP}$

Step3 若为响应数据包：

if $R_{packet}$. dstPort = $R_{HTTP}$. clientPort 且 $R_{packet}$. srcIP = $R_{HTTP}$. serverIP 且 $R_{packet}$. dstIP = $R_{HTTP}$. clientIP

if $R_{packet}$. ACK=$R_{HTTP}$. resACK

将数据插入该响应流，且按 SEQ 排序

elseif $R_{HTTP}$. resACK=0 且 $R_{packet}$. ACK=SEQ(请求流末端数据包)+Length(请求流末端数据包)

将数据包插入该响应流，且按 SEQ 排序

跳转 Step1，读取下一个数据包 $R_{packet}$

else

跳转 Step2，读取下一个 HTTP 索引 $R_{HTTP}$

若为请求数据包：

……

Step4 if 到达 HTTP 索引表末尾

利用该数据包 $R_{packet}$，新建一个 HTTP 索引 $R_{HTTP}$

图 1 TCP 数据流组装和匹配的快速算法

3.2.2 不完整 Web 页面的容错修复算法

常规的 Web 页面还原算法并未考虑丢包情况下对 HTTP 报文中的块(Chunk)编码和 Gzip 压缩数据的容错处理。针对此问题，本文提出下述的容错修复算法。

Step1 基于 Chunk 编码理论和传输格式，对损坏的数据块进行定位和识别，将坏点之前和之后的内容进行融合，再进行融合块的解码。

Step2 基于 LZ77 压缩算法和 Huffman 编码树，对 Gzip 压缩数据中的坏点进行定位和识别，将坏点前后的内容进行分段解压。

### 3.3 Web 页面预处理的算法设计

基于 Html 语言的 Web 页面，相比常规文本数据具有文字与图片、脚本等相混合，正文内容与用于页面格式控制的“Tag 标签串”相混合，元数据信息缺乏，文字编码多种多样等特点[11]，更加不利于计算机处理。本文从 Web 页面去噪、中文分词和文本向量化表示方面，对 Web 页面预处理算法进行设计。

(1)针对网页去噪和主题正文提取的问题，基于改进的 DOM(Document Object Model，文档对象模型)算法，对 Web 页面进行预处理：

单纯利用 DOM 算法，以 Html 树状语法结构进行 Web 页面解析和去噪具有可扩展性较差的缺点，因此可将由 Html 换行标志划分的连续文本作为基本内容单元，在语义相关性较好的单元块内，通过定制启发规则，对 DOM 树中的节点信息进行分类而将噪声节点滤除。

(2)针对中文文本的分词和向量化表示问题，本文采用下述算法：

基于词典匹配的分词算法具有效率高、易实现的优点，在当前的中文自动分词中应用最为广泛，但其受词典完备性和规则一致性的影响较大；而基于词频统计的算法可较好地体现分词的可信度和文字间的互信息，具有较好的歧义消除能力。因此，可将词典算法和统计算法相结合。以提高中文分词的效率和准确性。

向量空间模型是文本表示的常用数学模型之一，可将包含 $n$ 个词条项的文本 $\boldsymbol{D}$ 表示为：

$$\boldsymbol{D}=\boldsymbol{D}(t_1,w_1,\cdots,t_n,w_n) \tag{8}$$

式中，$t_i$ 和 $w_i$ 分别表示第 $i$ 个词条项及其权重值，则文本相似度可用向量距离或夹角进行衡量。但该模型的缺点在于具有较高的维数和稀疏性[11]，因此需采用降维算法，从特征空间搜索最优子空间，在更高层次更有效地表征原始的数据分布，以进行后续的文本挖掘处理。

### 3.4 基于文本挖掘和预测模型的网络舆情分析算法设计

热点话题、敏感话题、舆论倾向性和舆论发展趋势预测是网络舆情的主要内容。基于对文本挖掘中的文本聚类、文本模式匹配、文本分类算法和时间序列预测模型的综合优化，具体的舆情分析算法设计如下。

(1)针对传统文本聚类技术在处理动态、海量网络信息方面的不足，将多种聚类算法相结合，首先基于密度聚类法发现聚类中心，然后通过划分聚类法划分子簇而降低数据复杂度，再利用层次聚类法对多种形状的子簇进行处理，以有效提高对文本主题聚类的准确度。以文本主题的“有效时间”、“出现频率”和“主题内相似度”3 个量对主题热度进行量化，从而对热点话题进行发现和识别。

(2)在配置、建立敏感词库的基础上，基于多模式文本信息匹配技术对敏感话题进行发现和识别。针对当前常见的敏感信息隐蔽方式，分别采用谐音字处理、拆分字处理和通配符处理技术，将模糊文本匹配转化为精确文本匹配[8]；然后利用“Wu-Manber”多模式匹配算法对文本 $\boldsymbol{D}$ 中的各词条项 $t_i$ 进行搜索并发现敏感话题；基于敏感词的词频计算敏感度，依据设定的敏感阈值，对敏感话题进行识别。

(3)将基于机器学习和基于语义的文本倾向性分析算法相结合，采用语义分析方法对特征词语进行识别和提取，将语义信息加入文本表示[9]，也即将 $n$ 维文本向量 $\boldsymbol{D}$ 中的词条项 $t_i$ 表示为三元组：

$$\boldsymbol{R}_{t_i}=(Obj_i,SW_i,SF_i) \tag{9}$$

式中各项分别表示评价对象、情感词和修饰标记。然后利用机器学习的分类器对其处理，判别文本的情感倾向。最后计算文本的情感强度，依据设定的强度阈值，对过度偏激的文本进行过滤。

(4)对于网络舆情的短期趋势分析，可基于时间序列模型中的指数平滑模型、ARIMA 模型等进行预测。而对于舆情的长期趋势分析[10]，需在对舆情事件进行分类、时间序列平滑和周期化处理的基础上，对各类事件按周期建立模型库；然后基于待预测事件的已知时间序列与各类模型的匹配度，选取均方误差和最小者作为最优模型，实现对新舆情事件的长期预测。

## 4 基于超算平台的系统实现模型与运行性能分析

在上述的系统核心算法基础上，本节基于超算平台，对系统的具体实现模型和运行性能进行分析。

### 4.1 “分布式捕获+集中式处理”的系统实现模型

本文所提系统由前端的无痕捕获终端和后端的超算平台组成，如图 2 所示。

(1)无痕捕获终端基于便携式手持设备实现，分布于城市范围多个公共 Wi-Fi 无线网络内的不同位置，可全天候、高效、无痕地捕获本地无线数据，并通过专用线路上传至超算平台。该终端的具体配置为：

外观尺寸：28×19cm；CPU：Intel Atom N2600，1.6GHz；内存：2GB；存储：64GB SSD 固态硬盘；无线捕包网卡芯片：Atheros AR9170；操作系统：Windows XP/Windows7。

(2)超算平台位于国家超算济南中心，可集中处理无线捕获终端上传的海量数据。其平台性能如下：所用的神威蓝光千万亿次系统按照万万亿次架构设计，装配 8704 片 16 核的申威 1600 处理器，其 LINPACK 效率超过 73%，单机仓组装密度 1024CPU，综合水平处于世界领先行列。其软件架构自下向上可分为 4 层，如图 2 所示。

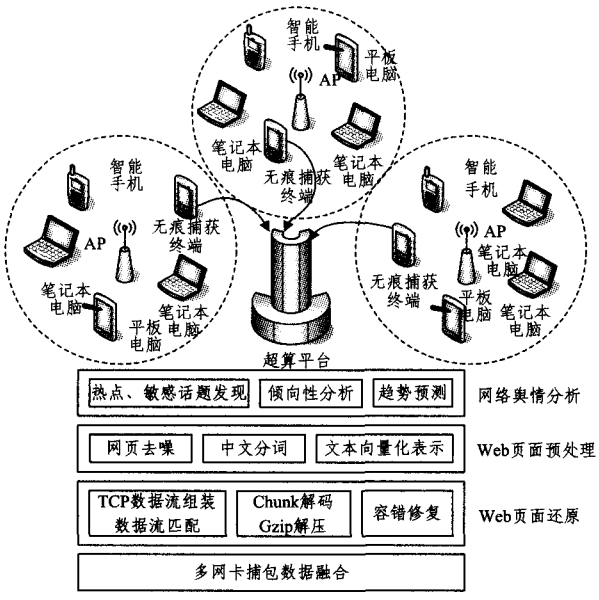


图 2 “分布式捕获＋集中式处理”的系统实现模型

多网卡捕包数据融合层，可基于分布式终端的捕获数据，获取整个城市范围各公共 Wi-Fi 无线网络的全信道、全部用户传输的较完整的无线数据。

Web 页面还原层，可基于 TCP 数据流组装与匹配模块、Chunk 解码与 Gzip 解压模块以及页面容错修复模块，完成对海量网页信息的还原。

Web 页面预处理层，可基于网页去噪模块、中文分词模块和文本向量化表示模块，实现对海量非结构化 Web 页面的文本信息提取。

网络舆情分析层，可基于热点发现模块、敏感话题发现模块、文本倾向性分析模块和舆情趋势预测模块，提供及时准确的网络舆情分析服务。

### 4.2 超算平台上该系统的运行性能分析

常规程序在超算平台上的优化方式包括模块流水化和模块并行化。针对模块流水化在多核平台易造成计算资源分配不均衡的问题，该系统采用将各功能模块在多核处理器上并行运行的模式。

当前公共 Wi-Fi 无线网络的主流协议标准仍为 IEEE 802.11g，其物理层理论速率为 54Mbps，实际吞吐量约为 20Mbps。网络流量中主要包括 HTTP 流量和 P2P 流量，而无线网络中 P2P 的平均使用量仅为 20%，由此可得单位时间内传输的 HTTP 数据约为 20/8 * 80% = 2MB。以整个无线城市范围内包含 1000 个公共 Wi-Fi 无线网络计，若每天对各网络的监控时间为 8 小时，则每天捕获的 HTTP 数据量可达：

$$2 * 60 * 60 * 8 * 1000 = 57.6\text{TB}$$

基于前期研究实验统计，当每个捕包文件的大小设为 2MB 时，对其进行 Web 页面还原的效率可达最高，因此每天获取的捕包文件数可达 57.6TB/2MB=2880 万个。

基于普通 PC(CPU 酷睿 i5，2.3GHz)平台，对每个捕包文件进行 Web 页面还原所耗费的平均时间约为 0.6 秒，因此对 57.6TB 海量数据(2880 万个捕包文件)的处理时间可达 0.6 秒 * 2880 万 = 4800 小时，即 200 天，这在实际运行中是无法接受的。而若基于 10 万核的超级计算平台，即使其单核运算能力与普通 PC 近似，海量数据也可在 2.88 分钟内处理完成。因此，基于超算平台的海量数据处理具有重要的实用价值。

**结束语** 本文所提基于超算平台的无痕信息获取与网络舆情分析系统，可有效协助相关部门捕获公共 Wi-Fi 无线网络中的各种非法活动，为无线网络取证工作提供强有力的技术支持；并可辅助相关部门及时了解本地区的网络舆论情况，从而做出科学正确的决策，达到超算技术服务于社会、服务于人民的目的。

## 参 考 文 献


[1] 李忠俊. 基于话题检测与聚类的内部舆情监测系统[J]. 计算机科学，2012，39(12)：237-240

[2] 沈辉，张龙. 基于 WinPcap 的网络数据监测及分析[J]. 计算机科学，2012，39(10)：15-19

[3] Fusco F, Deri L. High Speed Network Traffic Analysis with Commodity Multi-Core Systems [C]// Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC2010). New York, USA: ACM, 2010: 218-224

[4] Jihwang Y, Moustafa Y, Ashok A. A Framework for Wireless LAN Monitoring and Its Applications [C]// Proceedings of the 3rd ACM Workshop on Wireless Security (WiSE04). Philadelphia, USA: ACM, 2004: 70-79

[5] Ye Fei-yue, Wang Wen-jing, Du Jia-yong, et al. Research on Sensitive Information Discovery [C]// Proceedings of 2011 International Conference on Computational and Information Sciences (ICCIS2011). Chengdu, China: IEEE, 2011: 382-386

[6] Yang Fan, Dou Yi-nan, Lei Zhen-ming, et al. The Optimization of HTTP Packets Reassembly Based on Multi-Core Platform [C]// Proceedings of 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content (ICNIDC2010). Beijing, China: IEEE, 2010: 530-535

[7] Wan Ming, Yao Nan, Liu Ying. A Fast Information Reproduction Method for HTTP in WLAN[C]// Proceedings of 2010 International Conference on Wireless Communications, Networking and Information Security (WCNIS2010). Beijing, China: IEEE, 2010: 365-369

[8] 刘蔚琴. 网络敏感信息监控系统研究[D]. 广州：广东工业大学，2008

[9] 朱杰. 基于评价对象及其情感特征的中文文本倾向性分类研究[D]. 上海：上海交通大学，2010

[10] 高辉，王沙沙，傅彦. Web 舆情的长期趋势预测方法[J]. 电子科技大学学报，2011，40(3)：440-445

[11] 孟宪军. 互联网文本聚类与检索技术研究[D]. 哈尔滨：哈尔滨工业大学，2009