文章编号: 1006-4710(2013)03-0279-06

# 基于倒排索引的 MR 定位算法

周红芳[1], 周扬[1], 钱钢[2]
(1. 西安理工大学 计算机科学与工程学院,陕西 西安 710048;
2. 西安高压电器研究院有限责任公司,陕西 西安 710048)

**摘要**: 为了实现手机测量报告 MR 的实时、精确定位,提出一种基于倒排索引的定位算法。该算法以路测数据作为定位数据库,对于不含路测数据的区域用哈塔传播模型估算电平信息,之后用射线模型修正定位数据库。在执行定位算法时,通过为 MR 矢量化,将定位问题转化为求最大相似度问题。实验结果表明,该算法可以实现海量 MR 的实时定位,并且提高了定位精度。

**关键词**: 测量报告; 定位; 实时; 倒排索引; 相似度
**中图分类号**: TP311 **文献标志码**: A

## An Algorithm of MR Positioning Based on Inverted Index

ZHOU Hongfang[1], ZHOU Yang[1], QIAN Gang[2]
(1. Faculty of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. Xi'an High Voltage Apparatus Research Institute Co. Ltd, Xi'an 710048, China)

**Abstract**: In order to determine location of measurement reports in real time with high accuracy, a location algorithm based inverted index is proposed. This algorithm takes the road test data as location database. For areas without containing road test data, level information is estimated by hata model. And then the location database is modified by ray model. In implementing location algorithm, measurement reports are vectorized and the problem of fixing position of measurement reports is transformed into the problem of getting the maximum similarity. The experimental results show that the new algorithm can indicate the positions for massive measurement reports with optimal precision.

**Key words**: measurement report; positioning; real time; inverted index; similarity

随着移动网络的发展,手机用户数量几乎处于饱和状态,运营商间的竞争已经从市场转移到服务。如何提升网络质量、提高服务质量,已经成为运营商掌握核心竞争力的前提。传统模式下提高网络质量主要是通过性能统计等优化手段来完成,但是随着移动市场竞争的加剧,运营商对网络质量的提升已经开始由网络层面转移到用户层面——网络质量的好坏是依据用户感知评价的,而非仅仅是网络性能指标。在此背景下,借助 MR(Message Report)消息对网络进行优化并处理用户投诉,可以较准确地模拟用户感受,对提升用户感知度有重要的帮助。MR 中并不包含位置信息,利用 MR 模拟用户感受,最关键的技术是定位用户所在的位置、周围环境(无线环境)等空间信息,因此 MR 定位算法至关重要。

目前的定位技术主要有以下三种方法。①质心定位法:根据 MR 中电平信息、服务小区和邻区的位置信息计算 MR 位置。质心算法速度快且适用任何地型,但定位精度低。②"指纹"定位法:建立庞大的定位数据库,定位时在定位数据库中搜索和 MR 最相似的记录信息(含经纬度)作为定位结果。其优点是在定位数据库覆盖密集时定位精度较高。③基于 RSSI(Recept Signal Strength Indication,接收信号强度指示,简称 RSSI)的定位算法:根据信号损耗模型计算基站与 MR 的距离,再根据与多个基站的

**收稿日期**: 2013-03-26
**基金项目**: 国家自然科学基金资助项目(61172124); 陕西省教育厅科学研究计划基金资助项目(12JK0739); 西安理工大学校特色研究计划项目(116-211302);西安市碑林区科技计划项目(GX1308);西安市科学计划项目(CXY1339(5))。
**作者简介**: 周红芳,女,博士,副教授,主要研究方向为数据仓库与数据挖掘、知识发现、粗糙集等。
E-mail: zhouhf@xaut.edu.cn。

不同距离计算 MR 的位置。基于 RSSI 的定位算法[2,5-7]在空旷地区精度较高,一旦在建筑物较多、无线网络环境复杂的城市地区,此类定位算法的精度会明显降低。

本文提出的基于倒排索引的 MR 定位算法 MRII(an algorithm of MR positioning based on Inverted Index)基于以下原理:在一定时间内(如一周),某地的无线网络情况相对稳定。即在一定时期内,相同的地点,不同用户上报的 MR 电平和邻区信息相同或有较小差异。算法以路测数据作为定位数据库,在 MR 矢量化后,将定位问题转化为求最大相似度的问题。在求最大相似度时,借鉴搜索引擎中倒排索引的思想,实现快速定位。实验结果表明,该算法可以满足海量 MR 的实时定位,并且对于道路上 MR 的定位精度可以达到90%。

## 1 相关工作

基于 RSSI 的定位算法是较经典的算法。RSSI 算法是根据无线传感器接收到的信号指示强度,计算该信号在传播中的损耗,根据信号理论传播模型将信号强度转换成距离。RSSI 算法的模型如图 1 所示。

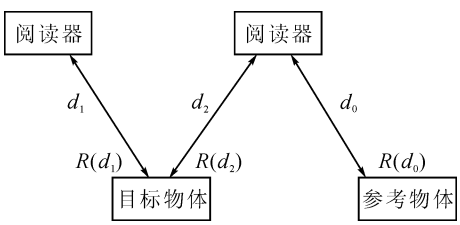


图 1 基于 RSSI 的测距模型

Fig. 1 The location model based on RSSI

图 1 中 $R(d_2)$ 表示目标物体在距离阅读器 $d_2$ 处的信号强度,$R(d_0)$ 表示参考物体在距离阅读器 $d_0$ 处的信号强度,根据信号损耗的理论模型,可以得到公式(1):

$$R(d_2) = R(d_0) - 10n\ln(\frac{d_2}{d_0}) \tag{1}$$

式中 $n$ 是信号衰减指数,一般根据具体的环境情况而设定,属于经验值,一般在 2 ~ 5 之间取值。

根据公式(1),可以求出距离 $d_2$:

$$d_2 = d_0\exp(\frac{R(d_0) - R(d_2)}{10n}) \tag{2}$$

参考物体的坐标为 $(x_0, y_0)$,两个阅读器(型号相同)的坐标为 $(x_1, y_1)$ 和 $(x_2, y_2)$。设目标物体的坐标为$(x, y)$,可得到以下方程:

$$\sqrt{(x_1 - x)^2 + (y_1 - y)^2} = d_0\exp(\frac{R(d_0) - R(d_1)}{10n}) \tag{3}$$

$$\sqrt{(x_2 - x)^2 + (y_2 - y)^2} = d_0\exp(\frac{R(d_0) - R(d_2)}{10n}) \tag{4}$$

根据等式(3)、(4)可求出目标物体的坐标值。

基于 RSSI 的定位算法原理简单,成本低廉,但受地理环境和天气的影响较大。本文提出的 MRII 算法,不受地理环境和天气的影响,具体定位精度和定位效率见实验部分。

## 2 定位数据库

使用中国移动自动路测工具获得路测数据,从中取出 MR 的经纬度、服务小区的 CGI(全球小区识别码,Cell Global Identity)和电平、6 个邻区的 CGI 和电平,之后建立定位数据库。以文件形式存储定位数据库,文件格式为:MR 经度、MR 纬度、服务小区 CGI、服务小区电平、第 1 邻区 CGI、第 1 邻区电平、第 2 邻区 CGI、第 2 邻区电平……第 6 邻区 CGI、第 6 邻区电平。对定位数据库建立倒排索引前,给出以下定义。

**定义 1** MR 的模:设某一 MR 为 $a$,将 $a$ 服务小区电平和 6 个邻区电平矢量化,即:

$$\boldsymbol{a} = (level_0, level_1, level_2, level_3, level_4, level_5, level_6)$$

则 $\boldsymbol{a}$ 的模为:

$$\mathrm{mod}(\boldsymbol{a}) = \sqrt{\sum_{i=0}^{6}(level_i)^2} \tag{5}$$

式中 $level_0$ 表示服务小区的电平,$level_1$ 至 $level_6$ 表示第 1 至第 6 邻区的电平,此后,文中的 MR 指矢量化后的 MR。

**定义 2** CGI 偏移量:在定位数据库中,某 MR 距索引文件开始处偏移的字节数为该 MR 中各 CGI 的偏移量。

根据 CGI 偏移量的定义,可以以 $O(1)$ 的时间复杂度确定 CGI 所在的 MR。以文件形式存放倒排索引,倒排索引文件名为 CGI,文件内容为所有该 CGI 所在 MR 的 CGI 偏移量、模、电平,如图 2 所示。

### 2.1 定位数据库栅格化

若所有路测 MR 都入选定位数据库,会导致定位数据库过大。现实情况是在一个 100 m × 100 m 的区域中约有 30 条路测 MR,假定某地区的面积为 50 km × 50 km,那么指纹库中将有 750 万条路测 MR。要更新和维护如此庞大的定位数据库极其困难,并且过大的定位数据库会严重影响算法效率。所以要对定位数据库栅格化,栅格的大小可以根据

实际情况选择。在一个栅格中,选择 CGI 信息和电平信息较完整的路测 MR 作为定位数据库的数据。每栅格中路测数据记录太少会影响定位精度,太多则无助于定位精度的提高,反而会影响定位算法的效率。定位数据库的栅格化定量分析见实验部分。

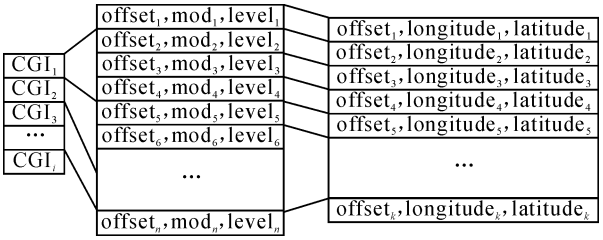


图 2 倒排索引示意图

Fig. 2 Schematic diagram of inverted index

## 2.2 栅格填充

路测数据中 MR 的位置都在道路上,即在非道路的栅格中没有路测数据。选用哈塔传播模型对非道路上的栅格进行填充。定位数据库中的记录除了经纬度还有服务小区的 CGI、电平和 6 个邻区的 CGI、电平,估算栅格中 CGI 和电平的具体方法如下。

1) 服务小区的 CGI 和 6 邻区的 CGI 分别为和该点距离最近的 7 个小区的 CGI。

2) 用哈塔传播模型估算服务小区的电平,用下节提出的射线模型对估算的电平进行修正。

3) 对全体路测数据的 7 个电平分别求算术平均值。分别对每个平均值加一个常数(保证其为正)后作为权重,根据各个权重和估算的服务小区电平求得各邻区的电平。数学表达为:设 $Alevel_0$, $Alevel_1$, …, $Alevel_6$ 分别为全体路测数据的 7 个电平算术平均值,$R$ 为常数(如 110), $Elevel_0$ 为步骤 2 估算和修正的服务小区电平, $Elevel_1$, $Elevel_2$, …, $Elevel_6$ 分别为 6 个邻区的电平,则:

$$Elevel_i = \frac{(R + Alevel_i)Elevel_0}{R + Alevel_0}, \quad i = 1,2,\cdots,6 \tag{6}$$

## 2.3 射线模型

用哈塔传播模型估算没有路测数据的栅格服务小区的电平。实际的地型、环境对哈塔传播模型的准确度有很大影响。所以根据已有的路测数据,用下文介绍的射线模型对估算的电平进行修正。具体步骤如下。

步骤 1 以基站为圆心,以某一宽度 $D$ 向周围区域扩展圆环 $R_1$, $R_2$, $R_3$ …当定位数据库 MR 的服务小区和邻区不含该基站对应的小区时扩展结束。

步骤 2 从内环向外环寻找到含有路测数据且未修正的环 $R_i$。若未找到(如图 3(d)),开始对下一个基站进行射线模型修正,直到所有基站遍历完毕,算法结束;若找到,求出环 $R_i$ 内各路测点哈塔估算电平和实际电平的差值(称为增益)。将环内各路测点按和正北方向的夹角从小到大排列。若两个路测点 $A$、$B$ 之间的栅格没有路测数据,则该栅格电平的增益按 $A$、$B$ 之间的角度和增益的比例关系求得。举例说明,设路测点 $A$ 和正北方向的夹角为 0°,设 $A$ 点增益为 10 db,设路测点 $B$ 和正北方向的夹角为 60°,设 $B$ 点增益为 20 db。若在 $A$、$B$ 所在环上某栅格 C 与正北方向的夹角为 30°,那么依照角度和增益的比例关系可求得栅格 C 的增益即为 15 db。重复此过程,修正该环内所有没有路测点的栅格的电平。

步骤 3 若 $R_i$ 所有内环都已修正(见图 3(a), $i=1$),向外寻找下一个包含路测数据的环,若未找到(见图 3(c), $i=2$)则执行步骤 4,若找到则执行步骤 2。若 $R_i$ 含有未修正的内环(见图 3(b), $i=3$),对环 $R_i$ 未修正的内环依据环 $R_i$ 的路测数据进行修正,修正方法见步骤 2。全部修正完成后,返回步骤 2。

步骤 4 对环 $R_i$ 之外的所有栅格依据环 $R_i$ 的路测数据进行修正,修正方法见步骤 2。全部修正完成后则对下一个基站执行射线模型修正,返回步骤 1,若所有基站遍历完毕,算法结束。

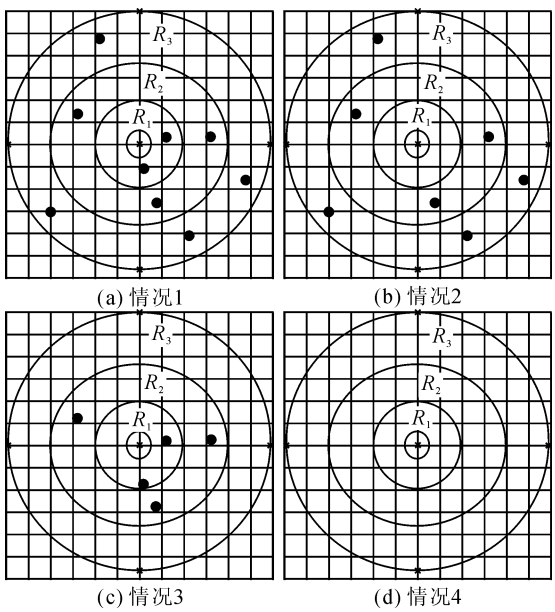


图 3 射线模型的 4 种情况

Fig. 3 4 situations of ray model

## 3 MR 定位算法

MR 中仅含服务小区和各邻区的 BSIC(基站识别码,Base Station Identity Code)和 BCCH(广播控制信道,Broadcast Control CHannel 信息,不含 CGI 信息,要根据配制列表和邻区列表回填服务小区和各邻区的 CGI,具体的回复方法见参考文献[12]。回填 CGI 后对其在倒排索引中查找相似度最大的 *topN* 个路测数据中的 MR,再对 *topN* 个路测数据中 MR 的经纬度求加权平均值,得到待测 MR 的经纬度。

### 3.1 相似度度量方法

将待测 MR 的 7 个电平作为一个 7 元组(矢量),可以利用余弦距离(公式7)的思想来衡量待测 MR 和路测 MR 的相似度。

$$Similarity(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{7}$$

设一待测 MR 为 $\boldsymbol{w}$,$\boldsymbol{w}$ 服务小区和 6 个邻区的 CGI 分别为 $\boldsymbol{w}CGI_0$,$\boldsymbol{w}CGI_1$,…,$\boldsymbol{w}CGI_6$,$\boldsymbol{w}$ 服务小区和 6 个邻区的电平分别为 $\boldsymbol{w}Level_0$,$\boldsymbol{w}Level_1$,…$\boldsymbol{w}Level_6$。

设一路测 MR 为 $\boldsymbol{a}$,$\boldsymbol{a}$ 服务小区和 6 个邻区的 CGI 分别为 $\boldsymbol{a}CGI_0$,$\boldsymbol{a}CGI_1$,…,$\boldsymbol{a}CGI_6$,$\boldsymbol{a}$ 服务小区和 6 个邻区的电平分别为 $\boldsymbol{a}Level_0$,$\boldsymbol{a}Level_1$,…,$\boldsymbol{a}Level_6$,则 $\boldsymbol{w}$ 与 $\boldsymbol{a}$ 的相似度为:

$$Similarity(\boldsymbol{w},\boldsymbol{a}) = \frac{\sum_{wCGI_i = aCGI_j} \boldsymbol{w}Level_i \cdot \boldsymbol{a}Level_j}{\mathrm{mod}(\boldsymbol{w})\,\mathrm{mod}(\boldsymbol{a})} \quad (i,j = 0,1,\cdots,6) \tag{8}$$

### 3.2 算法描述

MRII 的详细步骤如下。

步骤 1 对于一待测 MR,在倒排索引的 CGI 索引中找到其服务小区的 CGI。假定该 CGI 索引有 $k$ 个索引键,用服务小区的电平分别乘以 $k$ 个索引键的电平。用哈希表保存 $k$ 个乘积,哈希表的键为 CGI 偏移量,哈希表的值为该偏移量对应路测 MR 的模。

步骤 2 在倒排索引的 CGI 索引中找到其第 1 邻区的 CGI,假定该 CGI 索引有 $m$ 个索引键,用第 1 邻区的电平分别乘以 $m$ 个索引键的电平,得到 $m$ 个乘积。对于 $m$ 个索引键,若偏移量在哈希表中存在,则在哈希表中更新此偏移量的值,将新得到的乘积和原来哈希表中偏移量对应的值相加(MR 模不变);若偏移量在哈希表中不存在,则在哈希表中增加一项,键为此偏移量,值为新得到的乘积。

步骤 3 对第 2、3、4、5、6 邻区依次执行步骤 2。

步骤 4 对哈希表中的每一个键值除以路测 MR 的模和该待测 MR 模之积,得到 $n$ 个相似度。

步骤 5 找出 *topN* 个相似度最高的键值对应的键。在定位数据库中根据 *topN* 个偏移量找到 *topN* 条路测 MR。对 *topN* 条路测 MR 的经、纬度分别求加权平均值,即认为是该待测 MR 的经、纬度。

### 3.3 算法分析

假定有 $N$ 个小区,定位数据库中有 $M$ 条路测 MR,平均每个索引 CGI 对应 $K$ 条路测 MR。

在查找 CGI 索引时利用 B-树,时间复杂度为 $O(N\log N)$,对服务小区和 6 个邻区都需要查找 CGI 索引,所以第一步的时间复杂度为 $O(7N\log N)$。

步骤 2、3 中的更新哈希表理论上并不产生时间复杂度,求 7 个邻区的内积时间复杂度为 $O(7K)$。在求 *topN* 个最大相似度时利用堆排序的思想,仅确定前 *topN* 个元素,不完成整个排序过程,在最极端的情况下时间复杂度为 $O(topN * \log N)$。

综上所述,时间复杂度应小于:

$$O(7N\log(N) + 7K + topN * \log(M))$$

步骤 2、3 要维护一个哈希表,最坏情况是 $M$ 条路测数据和待测 MR 都有相似度,此时哈希表的长度为 $M$,此种情况,在步骤 5 要维护一个长度为 $M$ 的线性结构,以求得 *topN* 个最大相似度,综上所述空间复杂度应小于 $O(2M)$。

索引方面,假定共有 $N$ 条路测数据,每个栅格中选取 $e$ 条路测数据。

在建立倒排索引时,要遍历所有路测数据,并且要对路测数据栅格化。其中遍历路测数据时间复杂度为 $O(N)$,路测数据栅格化时每个栅格要缓存 $e$ 条路测数据,因此空间复杂度为 $O(N/e)$。

## 4 实验与分析

将大同地区 2010 年 8 月份的中国移动自动路测信息作为实验数据。以 8 月 10 日至 15 日的数据作为定位数据库,利用 MRII 去计算 8 月 16 日至 20 日路测 MR 的位置,再将计算结果和真实路测 MR

的位置作比较,即可评价算法的精度与性能。

将定位误差在 100 m 以内的 MR 数量在总数量中的比例定义为精度。

在 2.1 节中定性讨论了定位数据库中每栅格 MR 的数量与定位精度的关系,利用本文提出的算法计算 10 000 条 MR 的位置,得到如图 4 所示结果。

从图 4 可看出,当定位数据库中每栅格路测 MR 数量大于 5 时,再增大定位数据库中每栅格路测 MR 的数量,定位精度不但没有明显提升,反而引起算法速度降低。所以在以下实验中每栅格皆取 5 个路测 MR 入选定位数据库。

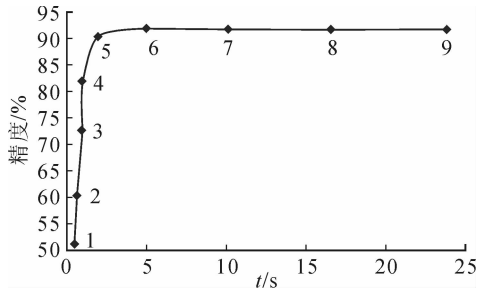


图 4 栅格中路测点数与精度的关系

Fig. 4 The relationship between precision and the number of MR in a grid

图 5 反映了倒排索引对算法性能的提升情况。横坐标代表定位数据库中路测 MR 的数量,以每栅格 5 个路测 MR 计算,数据量为 50 000 的定位数据库约覆盖 10 000 个栅格。

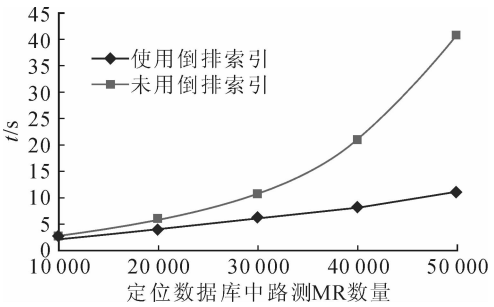


图 5 是否使用倒排索引的性能比较

Fig. 5 The comparison in performance between use and nonuse of inverted index

如图 5 所示,不使用倒排索引,直接在定位数据库中计算相似度,随着定位数据库的增大,算法消耗时间几乎是指数增长。在使用倒排索引的情况下,随着定位数据库的增大,算法消耗时间简单地线性增长,对海量 MR 定位的实现有重要意义。

文献[1]提出的算法基于基站位置和 MR 电平,相当于质心算法(Centroid Algorithm)的改进。当基站密度较小时,算法精度不高。文献[6]提出的算法是基于 RSSI 技术,实际地形对 RSSI 理论影响较大。

表 1 为利用各算法定位 10 000 条 MR 精度和时间的统计。MRII 算法对于在道路上 MR 的定位精度可以达到 90%,此精度已经接近理论值。

与道路上的定位精度相比,非道路上定位精度明显较低。原因是估算的电平和邻区信息还是有一定误差。这也是笔者下步研究的重点内容。

表 1 不同 MR 定位算法性能精度比较

Tab. 1 The comparison in performance between different algorithms

| 测试条件 | Improved Centroid[1] | | RSSI[6] | | MRII | |
|---|---|---|---|---|---|---|
| | 精度/% | 时间/s | 精度/% | 时间/s | 精度/% | 时间/s |
| 非道路 | 56.01 | 1.84 | 72.46 | 1.97 | 73.67 | 2.01 |
| 道路 | 56.43 | 1.92 | 72.61 | 2.32 | 90.32 | 1.96 |

## 5 结 语

本文提出的基于倒排索引的 MR 定位算法,以路测数据为数据库,把 MR 矢量化后将定位问题转化为求最大相似度的问题。

在求最大相似度时,借鉴搜索引擎中倒排索引的思想,实现快速定位。

实验结果表明,在道路上定位误差在 100 m 以内的 MR 可以达到 90%。

如何填充非道路上的栅格或改进射线模型以提高非道路上的定位精度是下一步工作的重点。

**参考文献:**

[1] 李婷婷. 基于 MR 的手机定位系统的设计与实现[D]. 济南:山东大学,2011.
Li Tingting. Mobile phone location system based on MR [D]. Ji'nan: Shandong University, 2011.

[2] 柴亦飞. 基于位置指纹的定位系统的实现与改进[D]. 上海:复旦大学,2007.
Chai Yifei. Realization and improvement of spotting system based on position of fingerprint[D]. Shanghai:Fudan University, 2007.

[3] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, et al. Intorduction to algorithms[M]. Third Edition. Massachusetts: MIT Press,2009:214-220.

[4] 韩斌杰,杜新颜,张建斌. GSM 原理及其网络优化[M]. 北京:机械工业出版社,2010:105-109.

[5] 饶兰兰, 马俊涛. 基于 MR 和传播模型的移动台定位算

法[J]. 武汉理工大学学报:信息与管理工程版, 2011, 33(3):371-374.
Rao Lanlan, Ma Juntao. A mobile station location algorithm based on measurement report and propagation model on GSM network[J]. Journal of WUT(Information & Management Engineering), 2011,33(3):371-374.

[6] 王振强,朱义胜. 改进的基于RSSI差值的定位算法[J]. 通信技术,2011,44(7):78-80.
Wang Zhenqiang, Zhu Yisheng. A modified mobile location algorithm based on RSSI difference[J]. Communications Technology, 2011,44(7):78-80.

[7] 郭瑞星. 基于ZigBee的无线传感网络RSSI定位算法的改进与实现[D]. 太原:太原理工大学,2011.
Guo Ruixing. The improvement and implementation of RSSI location algorithm in wireless sensor network based on ZigBee[D]. Taiyuan: Taiyuan University of Technology, 2011.

[8] Jennifer Yick, Biswanath Mukherjee, Dipak Ghosal. Wireless sensor network survey[J]. Computer Networks, 2008, 52(12):143-152.

[9] 臧建魁,卿粼波,何小海. 基于RSSI和LQI的分段距离估计改进算法[J]. 通信技术,2011,44(11):100-102.
Zang Jiankui, Qing Linbo, He Xiaohai. Subsection distance estimation method based on RSSI and LQI [J]. Communications Technology, 2011,44(11):100-102.

[10] Sayed A H, Tarighat A, Khajehnouri N, et al. Network-based wireless location: challenges faced in developing techniques for accurate wireless location information[J]. IEEE Signal Processing Magazine,2005,22(4):24-40.

[11] Liu Bo Chieh, Lin Ken Huang. Distance difference error correction by least square for stationary signal-strength-difference-based hyperbolic location in cellular communications[J]. IEEE Transactions on Vehicular Technology, 2008,57(1):227-238.

[12] 3GPP Committee 3GPP TS 48.058[DB/OL] http://www.3gpp.org/ftp/Specs/html-info/48058.htm. 2012-09-16.

[13] 孟欣,李郁侠. 基于数据融合算法优化的GM(1,1)负荷预测模型[J]. 西安理工大学学报, 2012,28(4):449-452.
Meng Xin, Li Yuxia. The optimized GM (1, 1) load forecasting model based on data fusion algorithm [J]. Journal of Xi'an University of Technology, 2012,28(4):449-452.

(责任编辑 王卫勋)