

基于标签速度和滑动子窗口的 RFID 数据清洗算法

顾韵华 高 宝 张俊勇 杜 杰

(南京信息工程大学 江苏省网络监控中心 南京 210044)

(南京信息工程大学计算机与软件学院 南京 210044)

摘 要 为提高非匀速 RFID(Radio Frequency Identification)数据流情形下的数据清洗准确性,在传统数据清洗算法 SMURF(statistical SMOothing for unreliable RFID data)的基础上,提出了一种基于标签速度和滑动子窗口的 RFID 数据清洗方法。该方法考虑到标签速度对滑动窗口调整的影响,依据标签速度动态调整置信度 δ ,同时进一步划分滑动窗口,对子窗口中的标签数据进行统计采样,并将其与整个滑动窗口的统计采样处理结果联合起来,以及时检测出标签的跃迁现象,从而准确判断标签的运动情况。实验表明,该方法有效地降低了平均错误率和积极读现象的出现频率,提高了数据准确性。

关键词 RFID,数据清洗,标签速度,滑动子窗口,SMURF

中图分类号 TP311.13 文献标识码 A DOI 10.11896/j.issn.1002-137X.2015.1.034

RFID Data Cleaning Algorithm Based on Tag Velocity and Sliding Sub-window

GU Yun-hua GAO Bao ZHANG Jun-yong DU Jie

(Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China)

(School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract To improve the accuracy of data cleaning under the circumstances of non-uniform RFID(Radio Frequency Identification) data stream, on the basis of the classical algorithm SMURF(statistical SMOothing for unreliable RFID data), an algorithm was presented based on tag velocity and sliding sub-window to clean RFID data. The method considers the influence of tag velocity on the sliding window adjustment, adjusts the speed dynamically according to the label confidence δ , at the same time further divides the sliding window, and takes statistical sampling for tag data in sub window. The results of the statistical sampling are dealing with the data of entire sliding window together to speed up the detection of tag transition, so that the movement of tag can be determined more accurately. Experimental results show that the algorithm reduces the average error rate and frequency of the phenomenon of positive reading, thereby increasing the accuracy of the data.

Keywords RFID, Data cleaning, Tag velocity, Sliding sub-window, SMURF

1 引言

RFID(Radio Frequency Identification, 无线射频)是一种非接触的自动识别技术,它通过无线射频信号获取物体的相关数据,并对物体加以识别。然而,由于无线射频信号极易受到周围环境影响,而且相互之间的干扰也很大,若阅读器和标签的数量增多,信号之间的干扰会更加强烈,从而导致了 RFID 具有很高的不确定性,这种不确定性可以分为 3 种情况:漏读(false negative)、积极读(false positive)和冗余读(redundant reading)^[1,2]。

在理想情况下,得到的 RFID 数据流是匀速的。但是在实际应用中,RFID 数据流往往是非匀速的,其主要原因有以下两点:一是漏读,RFID 数据流中漏读数据的存在,使得数据到达的时间的间隔不确定,从而使得原始数据非匀速;二是阅

读器布局,相邻阅读器的阅读范围可以不相交,从而使得标签在阅读器之间运动时得到的数据出现非匀速的现象^[3,4]。

传统 RFID 数据清洗算法 SMURF(statistical SMOothing for unreliable RFID data),在进行数据清洗的过程中只是依据滑动窗口中标签的平均阅读率来调整窗口的大小,从而清洗 RFID 数据流^[5]。SMURF 算法针对的是理想情况下的匀速数据流,若数据流是非匀速的,该算法则难以准确检测出标签的跃迁(即标签进入和离开阅读器的阅读范围),使得算法在对漏读数据进行填补的同时出现了很多不必要的积极读数据,降低了数据的质量。为提高标签跃迁检测的准确度,改进非匀速 RFID 数据流的数据清洗效果,本文对 RFID 数据清洗进行了深入研究,提出了一种基于标签速度和滑动子窗口相结合的清洗方法,该方法将标签运动速度引入到滑动窗口的自适应调整过程中,并将滑动窗口进一步划分,以更好地检测

到稿日期:2014-02-15 返修日期:2014-04-17 本文受国家自然科学基金项目(61103142),江苏高校优势学科建设工程资助项目(PAPD)资助。
顾韵华(1965—),女,教授,主要研究方向为信息安全、RFID 技术等,E-mail: yghu@nuist.edu.cn;高 宝(1989—),女,硕士生,主要研究方向为 RFID 技术;张俊勇(1989—),男,硕士生,主要研究方向为 RFID 技术;杜 杰(1979—),男,博士,副教授,主要研究方向为智能计算、信息安全等。

标签离开阅读器的情况,从而更准确地检测标签的跃迁。本文通过相关实验,说明了该方法在提高数据准确性方面的效果。

2 相关研究

针对 RFID 数据的不确定性,国内外一些研究机构已经做了不少关于 RFID 数据清洗的研究,典型的方法有基于固定滑动窗口的清洗方法^[3,4]和基于统计平滑自适应调整滑动窗口的清洗方法 SMURF。

基于固定滑动窗口的清洗方法的主要原理是:在任意时刻,只要窗口中存在某个标签的一个阅读,则认为该标签的阅读数据存在于整个窗口中,于是将其作为结果输出。该方法简单易实现,通常以经验值设定窗口大小。滑动窗口的设置对结果有较大影响,窗口过小则不足以填补原始数据中的漏读,而窗口过大则会产生不必要的积极读。

SMURF 算法是一种自适应窗口大小的数据清洗方法,该算法将阅读器读取的 RFID 数据流作为统计学中的随机事件,从而统计样本进而建模,根据观测值自动连续调整窗口,克服了固定窗口的弊端。

对于以上经典算法,在标签运动速度快、频繁出现跃迁现象的情况下,基于固定滑动窗口的清洗方法没有检测跃迁现象的机制,因而清洗时错误率偏高;算法 SMURF 只是依据滑动窗口中标签的平均阅读率来调整窗口的大小,从而会出现滑动窗口中数据质量改变但窗口中平均标签阅读率不变的情况,导致算法仍依据平均阅读率不变而不改变滑动窗口大小,从而不能及时检测标签的跃迁,即标签已离开阅读器的阅读范围而根据 SMURF 算法机制标签仍在阅读器阅读范围内,这种情况使得清洗数据时漏读数据得到填补的同时出现了很多不必要的积极读数据,降低了数据的质量。在 SMURF 算法的基础上,文献[6-9]进行了各种讨论,从自适应时间阈值、窗口设置机制等方面对 SMURF 算法进行了改进,在一定程度上提高了 RFID 数据质量;文献[10]提出了伪事件处理机制,降低了数据处理量,提高了效率;但是这些方法均没有考虑到标签运动速度等问题对于数据清洗的影响。

本文考虑标签运动模式,将标签速度参数引入到自适应窗口的调整过程中,并通过划分子窗口的方法,提高标签跃迁检测的准确性。

3 基于标签速度和子窗口的清洗算法

3.1 SMURF 算法研究

SMURF 算法将滑动窗口中的观测值作为多次重复的随机事件的结果,统计样本进而建模,结果服从二项分布,从而根据观测值的处理结果来自动连续调整窗口。算法设滑动窗口大小为 w_i 个阅读周期,同时设定标签 i 出现在阅读器阅读范围的概率在整个滑动窗口中的每个阅读周期都是相同的,记为 p_i 。

首先将初始滑动窗口的大小设为 1,接着获取下一个阅读周期(epoch)的数据,分析当前滑动窗口中的数据情况。若当前滑动窗口内没有阅读数据,则滑动窗口的大小置为 1。若滑动窗口内有阅读数据,则根据式(1)求得当前滑动窗口满足完整性需求的窗口大小。

$$w_i \geq \left\lceil \frac{\ln(1/\delta)}{p_i^{\sigma_{\text{avg}}}} \right\rceil \quad (1)$$

式中,置信度 δ 是一个依据经验给定的常量,即无论情况如何改变,置信度始终不变。 $p_i^{\sigma_{\text{avg}}}$ 为平均读取率。

若当前滑动窗口大小不满足完整性需求,则根据式(2)调整滑动窗口大小,其中 w_i^* 为满足完整性的滑动窗口大小的最小值,即式(1)的最小值。

$$w_i = \max\{\min\{w_i + 2, w_i^*\}, 1\} \quad (2)$$

若满足了完整性需求,则基于式(3)监测标签状态的变化,其中 S_i 是标签 i 的读取次数。

$$||S_i| - w_i p_i^{\sigma_{\text{avg}}}| > 2 \cdot \sqrt{w_i p_i^{\sigma_{\text{avg}}} (1 - p_i^{\sigma_{\text{avg}}})} \quad (3)$$

若状态改变,则调整滑动窗口大小为式(4)的结果。

$$w_i = \max\{\min\{w_i/2, w_i^*\}, 1\} \quad (4)$$

若状态未变,且满足完整性需求,则滑动窗口保持不变。

3.2 基于标签速度的置信度调整

标签的运动速度是影响 RFID 数据准确性的一个重要因素。标签速度的快慢,使得标签在一个阅读器的阅读范围内,被读取的平均概率以及被读取的次数都会有所不同,因此二项分布中的统计置信度 δ 也应是不同的。而在 SMURF 算法中, δ 是一个依据经验而定的常量,即无论情况如何改变, δ 始终不变。当标签移动速度慢,变化不大时,此时标签总是会在滑动窗口中出现,而且没有积极读的情况发生,固定的统计置信度 δ 对清洗结果影响不大。但当标签运动速度变化差异较大时,若 δ 始终不变,对于标签移动迅速的情况,就会出现滑动窗口调整不当而导致积极读数据出现的情形。

通过以上分析可知,将标签运动速度引入置信度的调整中可提高 δ 值的准确性。考虑使用标签运动速度来自适应地调整置信度 δ 的大小,具体方法如下。

设标签速度为 V ,两个相邻的查询周期(标签与阅读器的一次通信过程)之间的时间间隔为 T_e ,阅读器的阅读范围的半径为 R ,当标签经过阅读器的阅读范围时,阅读器会读取标签若干次,最大阅读次数 f 由式(5)计算。

$$f = (2 * R) / (V * T_e) \quad (5)$$

为了动态调整 δ ,依置信度参数的含义,定义 δ 的计算式,见式(6)。

$$\delta = p_i^{\sigma_{\text{avg}}} / f \quad (6)$$

标签未被读取的概率与置信度 δ 间满足式(7):

$$(1 - p_i^{\sigma_{\text{avg}}})^{w_i} \leq \delta = (p_i^{\sigma_{\text{avg}}} / f) \quad (7)$$

两边取对数:

$$w_i \ln(1 - p_i^{\sigma_{\text{avg}}}) \leq \ln(p_i^{\sigma_{\text{avg}}} / f) \quad (8)$$

经变换可得,保证完整性的充分条件是滑动窗口大小为:

$$w_i \geq \left\lceil \frac{\ln(f / p_i^{\sigma_{\text{avg}}})}{\ln(1 - p_i^{\sigma_{\text{avg}}})} \right\rceil \quad (9)$$

3.3 划分子窗口

当标签运动速度快、频繁出现跃迁现象时,很容易出现标签已经离开阅读器范围,却仍被以 SMURF 算法为代表的经典算法认为其依然在阅读器范围内的情形,从而不能很好地检测标签是否离开阅读器阅读范围。

例如,有这样一组原始数据流,分别为:0.6,0,0.5,0.6,0,0.7,0.6,0.5,0.6,0,0.6,0.5,0,0,0,并设滑动窗口大小为 8 个阅读周期,如图 1 所示。由于 SMURF 算法只依据滑动窗口中标签的平均阅读率来调整窗口的大小,通过计算可知,在处理的过程中滑动窗口中标签的平均阅读率一直没有变,因此滑动窗口也一直为 8 个阅读周期。观察上述数据流,

可以看出在最后一次处理时,标签只在滑动窗口的前半个窗口中出现,后半窗口中则没有标签的出现,标签此时已经离开阅读器范围,但是按照算法 SMURF,标签依然在阅读器范围内,这就导致了大量积极读的出现(图中后4个均为积极读)。

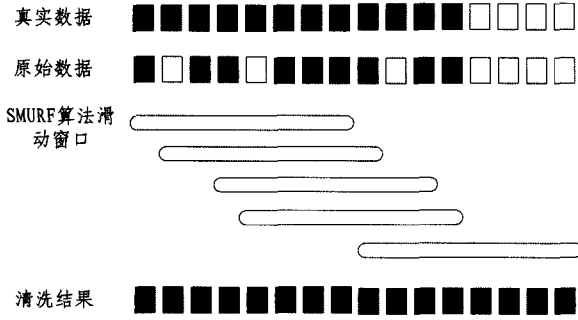


图1 跃迁检测错误导致的积极读

引起这一问题的主要原因仍是窗口设置过大,因此本文算法通过细分滑动窗口来更好地检测标签在整个滑动窗口中的运动情况,从而提高检测移动标签离开阅读器的阅读范围的准确度。

首先,通过标签在滑动窗口中出现的次数 S_i 和滑动窗口 w_i ,求得标签的出现率 P ,即 $P = S_i / w_i$ 。因研究已表明,RFID 原始数据的准确性约为 60%~70%^[11],故设定当出现率 P 大于 70%时,则认为标签不会出现跃迁现象;当出现率小于 70%时,则认为可能出现跃迁现象。

然后,根据可能出现跃迁时的标签出现率,将滑动窗口划分成两个子窗口,如图2所示。

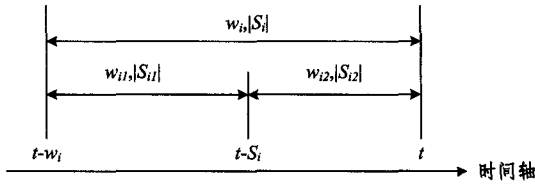


图2 划分子窗口

设标签 i 目前窗口为 w_i ,记为 $w_i(w_i, |S_i|)$;对其进行子窗口划分,分为 $w_{i1}(w_{i1}, |S_{i1}|)$ 和 $w_{i2}(w_{i2}, |S_{i2}|)$ 两个子窗口,其中 w_{i1} 子窗口大小为 $P * w_i$,即 $|S_{i1}|$ 。

标签 i 在 $w_{i1}(w_{i1}, |S_{i1}|)$ 和 $w_{i2}(w_{i2}, |S_{i2}|)$ 内如同在 $w_i(w_i, |S_i|)$ 内一样符合二项分布模型。通过对比 $w_i(w_i, |S_i|)$ 和子窗口 $w_{i2}(w_{i2}, |S_{i2}|)$ 中的标签数量,来检测标签是否离开了阅读器的阅读范围,从而动态调整窗口大小。

对标签动态性进行检测与窗口大小调整的具体方法是,获取标签在 $w_i(w_i, |S_i|)$ 中的读取次数,计算标签在 $w_i(w_i, |S_i|)$ 中的出现率 P ,若标签在 $w_i(w_i, |S_i|)$ 窗口中的读取率 P 持续下降,则表明标签正在离开阅读器的阅读范围。若检测标签正在离开阅读器的阅读范围,并且该标签没有在 $w_{i2}(w_{i2}, |S_{i2}|)$ 子窗口中被检测到,则判定标签离开了阅读范围,将 $w_i(w_i, |S_i|)$ 减小到 $w_{i1}(w_{i1}, |S_{i1}|)$ 子窗口的大小。若标签在 $w_{i2}(w_{i2}, |S_{i2}|)$ 中被检测到,但是 $|S_{i2}| < |S_{i1}|$,表明标签在 w_{i1} 中的概率更高,则 $w_i(w_i, |S_i|)$ 减小到 $w_{i1}(w_{i1}, |S_{i1}|) + 1$,旨在降低积极读的同时,使得漏读数据得到较好的清洗效果。若滑动窗口小于确保完整性的滑动窗口的大小 w_i^* ,并且期望的观测样本数量少于实际的观测样本数,即标签阅

读率低,则说明窗口设置偏小,应将窗口增加 2 个阅读周期,若仍小于 w_i^* ,则将 w_i 置为 w_i^* 。

3.4 基于标签速度和滑动子窗口的数据清洗算法描述

基于标签速度和滑动子窗口的 RFID 数据清洗算法的流程如图3所示。

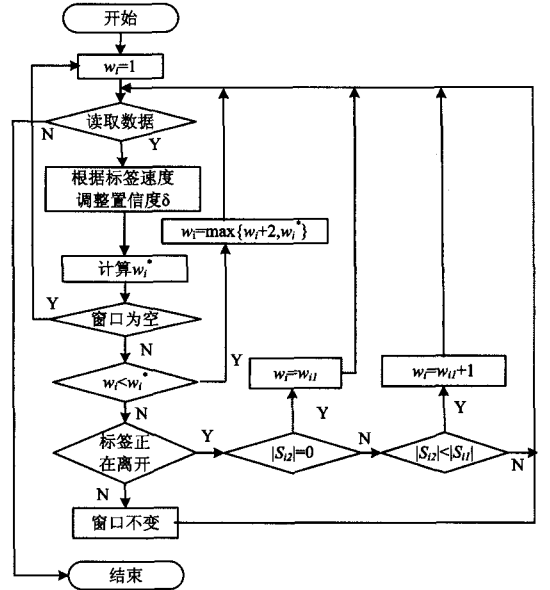


图3 数据清洗算法流程

算法初始将滑动窗口的大小设置为 1 个阅读周期 epoch。算法步骤如下：

- Step 1 读取窗口标签数据,若已无数据则算法结束。否则根据式(5)和式(6)调整统计采样所需的置信度 δ 。
- Step 2 计算满足标签完整性需求的滑动窗口大小 w^* 。
- Step 3 若当前窗口 $w_i < w^*$,则增大窗口,新窗口大小置为 $\max\{w_i + 2, w^*\}$ 。
- Step 4 若当前窗口 $w_i \geq w^*$,则采用 3.3 节中描述的方法判断标签是否正在离开,若是且 S_{i2} 为 0,即标签未在第二子窗口中出现,则将 w_i 调整为 w_{i1} ;否则,若 $|S_{i2}| < |S_{i1}|$,则将 w_i 调整为 $w_{i1} + 1$;转 Step 1。

4 实验与分析

4.1 实验方法

为建立标签运动模型,实验中选取 25 个标签在阅读范围内外进行运动。为与已有方法进行对比,设定速度取值范围为 0~3 inch / 阅读周期。初始时刻每个标签有一个任意的速度,每 100 个阅读周期标签随机改变速度。这样使得每个标签都拥有独立的运动方式。每个标签通过数据生成器生成 2000 个 epochs 的数据。

评价 RFID 数据清洗算法的性能主要采用平均错误率 $AvgErrors$ 、 $AvgNeg$ 和 $AvgPos$ 等指标,即每个阅读周期的平均错误数、平均积极读数和平漏读数。计算方法分别如式(10)~式(12)所示。

$$AvgErrors = \sum_{i=1}^n (FalseNeg + FalsePos) / n \quad (10)$$

$$AvgPos = \sum_{i=1}^n FalsePos / n \quad (11)$$

$$AvgNeg = \sum_{i=1}^n FalseNeg / n \quad (12)$$

其中, n 是阅读周期数, $FalsePos$ 和 $FalseNeg$ 分别是积极读和漏读数。

实验方法和过程如下:在原始数据中设置标识错误数据,不断改变阅读器主阅读区在整个阅读区中的百分比,获取不同清洗算法下的错误数,计算 $AvgErrors$ 、 $AvgNeg$ 和 $AvgPos$ 等指标值,将其进行性能对比。主阅读区是指在其范围内阅读器读到标签概率较大的区域,实际应用中阅读器主要获取的都是主阅读区数据。实验分为 3 部分,第 1 部分将本文算法与传统的固定窗口大小的清洗算法、SMURF 算法两类方法进行算法的 $AvgErrors$ 数据对比分析。第 2 部分将本文算法与传统的固定窗口大小的清洗算法、SMURF 算法进行算法的 $AvgNeg$ 数据对比分析。第 3 部分将本文算法与传统的固定窗口大小的清洗算法、SMURF 算法进行算法的 $AvgPos$ 数据对比分析。

4.2 实验结果与分析

实验的第 1 部分测试了主阅读区百分比从 0 变化到 1 的不同情况下,分别采用传统的固定窗口大小的清洗算法、SMURF 算法和本文算法进行数据清洗处理后的每个阅读周期的平均错误数。结果如表 1 所列。

表 1 3 种数据清洗算法的平均错误率

主阅读区百分比	原始错误数	固定窗口为 5	SMURF 算法	本文算法
0	11.1036	4.503	3.519	3.105
0.1	10.401	4.2197	3.33	2.958
0.2	9.8014	3.9	2.9451	2.7293
0.3	9.1762	3.521	2.8067	2.5287
0.4	8.3586	3.2099	2.651	2.3913
0.5	7.4871	2.897	2.42	2.2033
0.6	6.619	2.508	2.1009	1.839
0.7	5.8013	2.212	1.8801	1.6264
0.8	5.2047	1.839	1.605	1.358
0.9	4.752	1.599	1.4093	1.1892
1	3.9024	1.4683	1.182	1.033

由表 1 可见,原始 RFID 数据错误率很高,无法直接被上层应用程序所使用;对于传统的固定窗口大小的清洗算法,这里使用的窗口大小是 5(固定窗口大小中处理数据效果相对较好的窗口值^[5]),从表中可以看出数据的错误率有所降低,但是相比较于自适应滑动窗口 SMURF 算法,其错误率依然高;而本文算法相较于 SMURF 算法,数据清洗之后平均错误率有了进一步降低,是 3 种方法中平均错误率最低的。这是由于本文算法考虑了非匀速数据流的速度特征,并采用更灵活的自适应窗口调整机制,使可能导致窗口中数据质量改变后而平均标签阅读率不变的情况得到改善。

实验的第 2 和第 3 部分测试了主阅读区百分比从 0 变化到 1 的不同情况下,分别采用传统的固定窗口大小的清洗算法、SMURF 算法和本文算法在每个 epoch 中积极读出现的平均数 $AvgPos$ 和每个 epoch 中漏读出现的平均数 $AvgNeg$ 。结果分别如图 4 和图 5 所示。

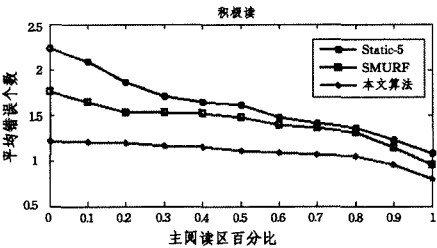


图 4 积极读平均出现数 $AvgPos$ 对比

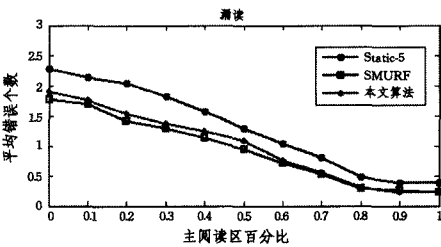


图 5 漏读平均出现数 $AvgNeg$ 对比

如图 4 所示,本文算法在 3 种方法中 $AvgPos$ 值最小,其积极读出现率比 SMURF 算法下积极读出现率明显要低较多,这是由于 SMURF 算法自身存在的缺陷而使得许多离开了的标签被错误认为是没有离开的标签,即填补了一些不需要的积极读数据,而本文算法在一定程度上弥补了 SMURF 的这个缺陷,使得数据平滑之后增加的积极读数量得以降低。这也是本文方法的主要特点。

从图 5 可以看出,本文算法相对 SMURF 来说,漏读出现率稍高。分析其原因主要是本文方法通过动态调整机制使滑动窗口变小,从而使积极读降低,但滑动窗口变小通常会带来漏读的增加。本文算法在考虑标签速度的前提下,再对滑动窗口进行划分,试图在积极读与漏读之间进行平衡,即以牺牲一点漏读来换得积极读的较大降低。

结束语 本文针对传统 RFID 数据清洗算法在非匀速数据流情形下出现的过多积极读的问题,深入研究了产生该问题的主要原因,指出了其在进行数据平滑处理时存在的问题,如标签运动速度的增长使得清洗后的数据的错误率也随之上升和标签运动速度快、频繁出现跃迁现象等,这导致填补了许多不必要的积极读数据。在此基础上提出了一种基于标签速度和滑动子窗口相结合的算法来进行数据清洗。实验结果表明,该算法解决了传统算法的过多积极读问题,提高了数据准确性。接下来的工作将考虑引入逻辑层事件处理以进一步提高数据质量。

参 考 文 献

[1] Derakhshan R,Orlowska M E,Li Xue. RFID Data Management Challenges and Opportunities[C]// IEEE First International Conference on RFID. 2007:175-182

[2] 许嘉,于戈,谷峪,等. RFID 不确定数据管理技术[J]. 计算机科学与探索,2009,3(6):562-575

[3] Bai Yi-jian, Wang Fu-sheng, Liu Pei-ya. Efficiently Filtering RFID Data Streams[C]// First International VLDB Workshop on Clean Databases. 2006:50-57

[4] Jeffery S R,Alonso G, Franklin M J, et al. A Pipelined Framework for Online Cleaning of Sensor Data Streams[C]//the 22nd International Conference on Data Engineering. 2006:140-142

[5] Jeffery S R, Alonso G, Franklin M J. Adaptive Cleaning for RFID Data Streams[C]//Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB Endowment. 2006:163-174

[6] 潘伟杰,李少波,许吉斌. 自适应时间阈值的 RFID 数据清洗算法[J]. 制造业自动化,2012,34(7):24-27

[7] Massawe L V, Kinyua J D M, Vermaak H. Reducing False Negative Reads in RFID Data Streams Using an Adaptive Sliding-Window Approach[J]. Sensors,2012,12(4):4187-4212

[8] Li Xing, Fu Wen-xiu. Efficient RFID Data Cleaning Method[J]. Telkomnika Indonesian Journal of Electrical Engineering, 2013; 1707-1713

[9] Li Ling-juan, Liu Tao, Rong Xiang, et al. An Improved RFID Data Cleaning Algorithm Based on Sliding Window[C]// IOT

(Internet of Things) Workshop. 2012; 262-268

[10] 王妍, 石鑫, 宋宝燕. 基于伪事件的 RFID 数据清洗方法[J]. 计算机研究与发展, 2009, 46(z2): 270-274

[11] 谷峪, 于戈, 张天成. RFID 复杂事件处理技术[J]. 计算机科学与探索, 2007, 1(3): 255-267

(上接第 121 页)

在实验中, 一共发现了 11 个分组。与实际的 12 个分组有点区别, 这是因为第 6 组中的队伍都是独立的队伍, 它们并没有相互比赛, 因此在本算法中, 这一组被归类到与其比赛次数最多的分组中, 因而出现了与实际分组不同的情况。由此我们可以看到本算法在橄榄球联赛数据集上的社团发现结果也是非常符合现实的, 同时不会出现多次实验结果不同的现象, 非常稳定。

本文针对两种算法在足球队比赛网络上运行 10 次的时间情况绘制了对比折线图, 如图 5 所示。为了验证算法对一个随机网络的有效性, 我们选了一个有 2000 个节点的网络进行测试, 具体运行时间对比如图 6 所示。算法的准确性如表 1 所列。从中可以看出, 我们的算法与 LPA 算法相比在时间上没有太多的增加, 且提高了准确率。

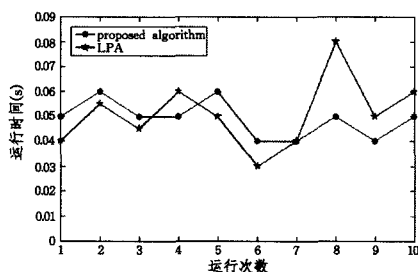


图 5 足球队比赛网络的运行时间对比

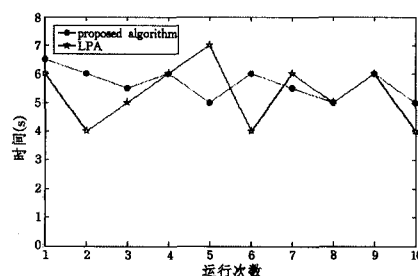


图 6 随机网络的运行时间对比

表 1 算法的准确率比较

具体的网络	LPA 算法	改进的 LPA 算法
空手道俱乐部网络	80.4%	97.2%
海豚社会网络	85.6%	96.3%
足球队比赛网络	86.1%	93.2%
随机网络	88.2%	91.3%

结束语 本文对 LPA 算法的初始条件进行了改进, 提出了基于核心图的标签传播算法, 即利用节点的 k 阶共同邻居, 找出网络中的初始核心社团。用这些初始核心社团作为标签传播的初始状态, 不仅可以显著提升算法的稳定性, 在许多情况下还能够提高算法所发现的社区结构的质量。由于对网络进行初始化, 算法会花费一些时间, 改进后的 LPA 算法在时

间上并没有太多的增加, 而且在提高了准确率的前提下, 增加运算时间是可以接受的。

参 考 文 献

[1] Zhao Y P, Levina E, Zhu J. Community extraction for social networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(18): 7321-7326

[2] Kelley S, Goldberg M, Magdon-Ismael M, et al. Defining and discovering communities in social networks[J]. Handbook of Optimization in Complex Networks, 2012, 57(2): 139-168

[3] Angeles S M, Boguna M, Sagues F. Uncovering the hidden geometry behind metabolic networks[J]. Molecular BioSystems, 2012, 8(3): 843-850

[4] Ino H, Kudo M, Nakamura A. Partitioning of Web graphs by community topology[C]// Proceedings of the 14th International Conference on World Wide Web. New York: ACM, 2005: 661-669

[5] Farutin V, Robison K, Lightcap E, et al. Edge-count probabilities for the identification of local protein communities and their organization[J]. Proteins: Structure, Function, and Bioinformatics, 2006, 62(3): 800-818

[6] Newman M E J. Modularity and communities structure in networks [J]. Proceedings of the National Academy of Science, 2006, 103(23): 8577-8582

[7] Guimera R, Amaral L. Functional cartography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900

[8] Flake G W, Lawrence S, Giles C L, et al. Self-organization and identification of Web communities[J]. IEEE Computer, 2002, 35(3): 66-71

[9] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632

[10] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[11] Raghavan U N, Albert P, Kumara S. Near linear time algorithm to detect community structure in larger-scale networks [J]. Physical review E, 2007, 76(3): 036106

[12] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 33(4): 452-473

[13] Lusseau D. The emergent properties of a dolphin social network [J]. Proceedings of the Royal Society B: Biological Sciences, 2003, 270(S2): 186-188

[14] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci., 2002, 99: 7821-7826