

一种改进的基于 RFID 中间件的冗余数据清洗算法

陈旭辉^{1,2}, 王 馨¹, 柯 铭¹

(1 兰州理工大学 计算机与通信学院, 甘肃 兰州 730050;

2 厦门理工学院 福建省高校物联网应用重点实验室, 福建 厦门 361024)

摘 要: 针对 RFID 中间件冗余数据处理问题, 提出一种改进的基本临近有序法(SNM)算法. 通过对大量 RFID 数据进行分组, 采用随机因子改变窗口大小对数据进行清洗, 并在算法的实现过程中采用循环处理方式. 经实验证明该改进算法能有效提高数据处理效率, 并在召回率和数据处理时间上均优于原 SNM 算法.

关键词: RFID 中间件; 冗余数据; 数据清洗; SNM 算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)07-0154-05

An Improved Algorithm in Redundant Data Cleaning Based on RFID Middleware

CHEN Xu-hui^{1,2}, WANG Xin¹, KE Ming¹

(1 College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China;

2 Internet of Things Lab, Xiamen University of Technology, Xiamen 361024, China)

Abstract: As the processing problems of redundant data of RFID middleware, an improved algorithm which based on the basic sorted neighborhood method (SNM) is introduced in this paper. Through grouping a large number of RFID data and using a random factor to change the window size to clean data. And then, the cycle treatment is utilized in the realizing process of the algorithm. The experiment results show that the improved algorithm can effectively improve the efficiency of the data processing. The recall rate and data processing time are superior to the original SNM algorithm.

Key words: RFID middle-ware; redundant data; data cleaning; SNM algorithm

1 引言

无线射频识别(Radio Frequency Identification, RFID)是一种非接触式自动识别技术, 典型的 RFID 系统由标签、阅读器、中间件和应用程序组成^[1], 利用无线射频通讯在 RFID 读写器和标签之间双向传递数据来达到对象自动识别的目的.

RFID 中间件的功能就是对读取到的标签数据进行过滤和整合, 把处理后的有用数据提交给需要的应用程序, 供应用系统使用^[3]. 由于 RFID 读写器的特性, RFID 数据不可靠性、冗余和时间戳乱序等

问题是 RFID 系统大规模实施的瓶颈, 因此如何有效地通过 RFID 中间件对原始数据进行冗余清理成为亟待研究的问题.

传统的冗余数据处理方法是将每个记录与读取到的其余记录逐个对比, 该方法识别精度非常高, 但处理时间让人难以忍受. 目前冗余数据清洗的算法主要采用排序—合并(删除)思想^[4], 即: 先将数据库中的记录排序, 通过比较邻近记录是否相似来检测记录是否重复. 最为常用的方法是基本临近有序法(SNM), 它有效克服了直观方法的缺点, 提高重复记录的匹配效率和重复记录清理的完成效率, 但该

收稿日期: 2012-10-30; **修回日期:** 2012-11-13

基金项目: 国家自然科学基金项目(61065007, 61263047); 甘肃省自然科学基金(0916RJZA020); 兰州理工大学博士基金(BS14200901)

算法严重依赖关键字选择和滑动窗口的大小^[5].文献[6]提出多次邻近排序法(MPN),该算法在排序后的数据集上采用不同关键字多次独立使用SNM算法,并使用较小窗口进行过滤,其缺点是容易引起误识别.文献[5]提出采用可变窗口,通过相似度与阈值的比较及时调整窗口值,减少遗漏.文献[7]提出优化队列方法(PQS),用一个重复记录的优先队列来代替固定大小的窗口,优先队列越大,算法检测精度越大,但运行时间也越长.

上述冗余数据处理方法虽各有优势,但应用到RFID中间件数据处理过程中还存在不足之处,本文针对RFID数据自身特性,在SNM算法的基础上进行改进,采用分组、分段检测及可变窗口对RFID数据进行处理,使改进后的SNM算法能够更准确、高效的在RFID中间件中运行,提高冗余数据处理速率和时间.

2 SNM算法描述及流程

每种冗余数据检测方法都需要确定是否有两个或两个以上数据表示同一实体,有效的检测方法是对每个数据都与其他数据进行比较,以此来确定是否为冗余数据.目前,常用的算法是基本临近有序法(Basic sorted-neighborhood method, SNM),但该算法需要进行 $N(N-1)/2$ 次比较(N 是所有数据的总数),算法时间复杂度大.该算法的基本思想^[8]是:确定一个关键字,根据该关键字对数据集内的数据进行排序,在排序后的数据集上滑动一个固定大小的窗口,只检测窗口内的数据,如果窗口大小包含 m 条记录,当窗口移动的时候,原窗口中的第一条数据移出,新进入窗口的数据与原窗口中的 $m-1$ 条数据进行比较,判定是否有重复.数据比较过程中的滑动窗口如图1所示.

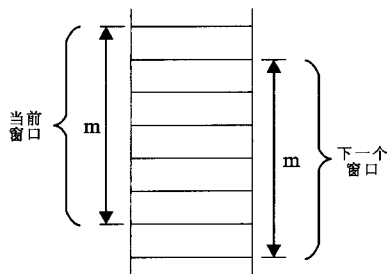


图1 滑动窗口

数据比较的过程如下:

(1)创建关键字:抽取记录属性的相关字段,构造排序关键字.

(2)用上一步产生的关键字对数据集进行排序,得到已排序的数据集.

(3)在排序后的数据集上移动一个固定大小的窗口,窗口大小为 w 条数据.

(4)比较窗口最下方的数据是否与窗口内先前的 $w-1$ 条数据重复,是转(5),否则窗口内的第一条数据移出.

(5)第 w 条数据下面的一条数据移入窗口内,转(4).

(6)对比较出的重复数据序号进行存储,用直观清洗法对重复数据进行清理.

SNM算法通过使用滑动窗口,每次比较窗口内的 w 条数据,不但提高了数据匹配效率,而且提高了数据比较速度^[9],只需进行 $w \times n$ 次比较.

主要过程流程图如图2所示.

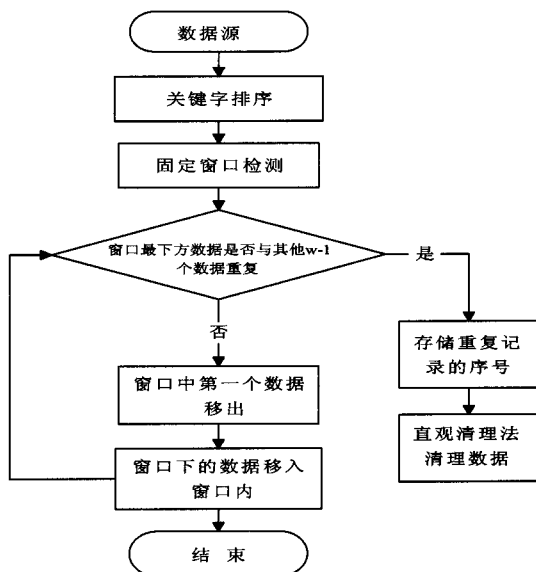


图2 SNM算法流程图

3 SNM算法改进策略

3.1 SNM算法缺点分析及优化

该算法相对于原始的合并-排序算法有效克服了直观方法的缺点,使重复数据匹配效率和重复数据清理的完成效率也有了很大提高.但是仍然存在一些问题需要改进:

(1)SNM算法识别重复数据的能力很大程度上取决于排序关键字的选择,单一的关键字并不能十分准确的表达数据特征,于是漏掉重复数据的情况就会发生.本文针对RFID数据特性,选择不同的关键字进行处理,提高数据处理精度.

(2)SNM算法及多数重复数据识别算法大多以

静态数据集为前提,如果接收到新的数据集,要与处理过的数据集合并后再过滤,时间复杂度较大.本文以时间为标准,对读取到的数据集进行分组,在各个分组里进行冗余检测和处理,有效处理同一时刻读取的相同数据,同时对数据进行分段检测,减少不必要的数据匹配过程.

(3) SNM 算法由于采用固定窗口在数据集上滑动,窗口的大小很难确定,若窗口过大,计算量也较大,性能不好;若窗口太小,会容易遗漏重复数据.文中引入随机因子,使窗口大小每检测一条数据即根据随机因子的变化进行调整.

3.2 改进后的 SNM 算法过程

面对大量 RFID 数据集,为提高检测效率,先对大数据集进行一定处理是十分必要的.首先按照一定规则检测数据集中存在的错误和不完整数据,提高数据的质量;运用分组思想,把大的数据集分割成小数据集,在小数据集集中进行冗余检测和处理.而 RFID 数据的冗余性分为两种:一是由于阅读器按照毫秒级的阅读周期来扫描标签数据,即一个阅读器可能在很短的时间内多次读取同一标签造成冗余;二是由于多个阅读器的识别范围发生重叠,一个标签在识别范围内被多个阅读器同时读取造成冗余.因此,针对该冗余特性,本文对第一种情况按照时间戳排序进行处理,然后根据阅读器 ID 再次进行排序处理来排除因第二种情况造成的冗余重复数据,提高数据精度.

具体思想如下:逐条判断各数据是否完整,若不完整则删除,若完整,则每过滤 M 条数据为一组,按时间戳对该组数据排序,窗口在排序后的数据集上滑动,窗口大小根据随机因子的变化而调整,比较窗口内数据是否有重复.主要过程如下:

(1) 按照 ALE 规范对数据集进行检测,若有不符合规则的不完整数据,则删除该数据;若没有,则执行下一步.

(2) 对检测过的完整数据进行分组,每 M 条完整数据为一组.

(3) 对这 M 条数据按照时间戳进行排序.

(4) 随机变化窗口 w_i 在排序后的数据集上滑动.

(5) 比较当前窗口内第一条数据是否与下面的 w_{i-1} 条数据重复,若有数据与其重复则转(6);若没有数据与其重复则转(7).

(6) 每检测出有一条数据与窗口内第一条数据重复,该数据的重复次数加 1,并删除旧时间戳数

据,保留最新时间戳数据.

(7) 判断窗口内第一条数据是否为最后一条,不是则移出,转(5);若是最后一条数据则移出,转(8)

(8) 判断是否有未处理数据,是则转(3),不是则结束该过程.

改进后的 SNM 算法流程如图 3 所示.

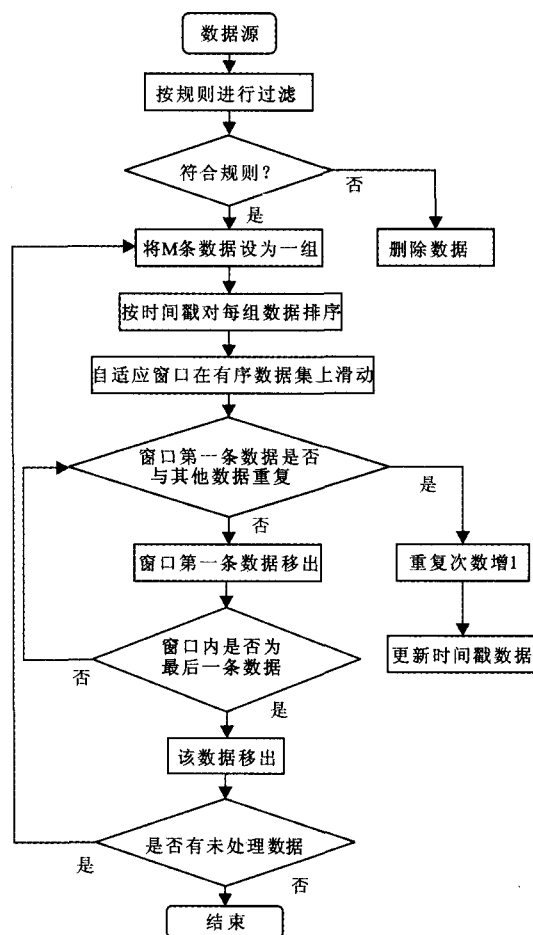


图 3 改进的 SNM 算法流程图

上述过程是按照时间戳排序,过滤重复数据的,对处理后的数据按阅读器 ID 再次排序并执行以上操作,经过这两步处理,由 RFID 数据特性产生的两种冗余数据都可以最大限度的消除,提高数据精度.

3.3 改进 SNM 算法描述

由于 RFID 数据通常为三元组 $\langle TID, RID, time \rangle$, 其中 TID 是标签 ID, 代表读写器读取对象的唯一识别标识; RID 是阅读器的唯一标识, 也即读取位置; time 是该标签被阅读器感知并读取的时间戳^[9]. 但阅读器 ID 和时间戳是变化的, 因此检测重复数据只需要对 TID 进行比较, 若两条数据的 TID 相同, 即为重复数据.

3.3.1 RFID 数据的分组排序检测

为提高检测速度, 对数据集进行分组, 分组后按

照数据时间戳进行排序,因同一时刻的同一标签可能被多个阅读器读取到,这样可把冗余重复数据尽可能排在较近的区域,以便检测.对检测出的重复数据,记录重复数据序号,并记录该数据的重复次数,每检测出有一次重复,重复次数即加 1.同时,由于数据流的规模是无限的,新到达的数据更能反映当前状况,因此只保留最新时间戳数据,删除旧时间戳数据.当所有数据都处理完后,按阅读器 ID 再次排序,以便检测同一阅读器多次读取的同一标签数据,然后按改进后的数据处理流程进行执行.

3.3.2 数据的分段检测

由于本改进算法只需比较数据中的 TID,因此通过对 TID 数据格式的分析,对数据进行分段检测,首先比较两条数据的第一段,若第一段数据不相同,则不是重复数据,最上方数据直接移出窗口;若第一段数据相同,则对第二段数据进行匹配,一直进行到各段数据检测都相同,则该数据为重复数据,记录重复数据的序号,并保留最新时间戳数据.对数据进行分段检测,可以减少不必要的数据匹配过程,提高数据过滤效率,缩短处理时间.

3.3.3 滑动窗口大小的调整

对滑动窗口,基本临近算法采用固定窗口来检测数据,窗口大小是行业专家凭经验而确定的,窗口的大小对算法的效率有很重要的影响,如果窗口很小,运算速度会很快,但检测结果不太理想;如果窗口过大,检测效果比较理想,但运算过程很长.因此本文引入随机因子 $rand(1)$,使窗口根据随机因子的变化自适应的在最大和最小窗口之间变化,其中 $rand(1)$ 产生 $(0,1)$ 之间的随机数.每当窗口中的第一条数据移出时, $rand(1)$ 即发生随机变化,窗口大小改变.

设窗口最小值 w_{min} ,窗口最大值 w_{max} ,当前窗口 w_i .每有一条数据移出时,窗口的大小为:

$$w_i = Int(w_{min} + rand(1)(w_{max} - w_{min}))$$

滑动窗口大小随着 $rand(1)$ 的变化而变化.当 $rand(1)$ 产生的随机数大时,窗口变大;当 $rand(1)$ 产生的随机数小时,窗口变小.

同时本文对数据集进行循环处理,即对检测后的数据集按照改进后的算法过程进行多次处理,减少重复数据的遗漏.

4 实验结果及分析

本实验在常配微机上,运用 Java 开发语言,进行模拟实验.实验平台为运用 EJB(Enter price Jav-

aBeans)架构搭建的 RFID 中间件,通过软件编程,根据 RFID 数据编码规则,使系统自动生成三元组 RFID 数据,即 $\langle TID, RID, time \rangle$. 本实验的数据采用该中间件系统自动生成的 3000 条记录,用 SNM 算法和改进后的 SNM 算法分别进行处理,并对处理结果进行对比.

为了衡量重复数据检测算法的效率,判断算法是否能把数据集中存在的重复数据都检测出来,常用的度量标准有召回率和查准率^[10].召回率 R (Recall)和查准率 P (Precision)分别定义为:

$$R = \frac{\text{算法正确识别的重复记录数}}{\text{数据集中实际包含的重复记录数}} \times 100\%$$

$$P = \frac{\text{算法正确识别出来的重复记录数}}{\text{算法识别出来的重复记录数}} \times 100\%$$

实验结果见下表,表中 T 表示算法处理时间,S 表示查出的重复数据,W 表示查出的正确重复数据,C 表示实际重复数据,R 表示算法召回率,P 表示查准率.表 1 和表 2 分别为原 SNM 算法和改进后算法的处理数据.

表 1 原 SNM 算法实验结果

记录数	T ₁ (ms)	C ₁	S ₁	W ₁	R ₁ (%)	P ₁ (%)
1000	357	48	51	40	83.3	78.4
2000	1456	128	132	103	80.5	78
3000	3114	151	157	127	84.1	81

表 2 改进 SNM 算法实验结果

记录数	T ₂ (ms)	C ₂	S ₂	W ₂	R ₂ (%)	P ₂ (%)
1000	121	48	53	43	89.5	83
2000	1117	128	125	106	82.8	84.8
3000	1593	151	159	131	86.7	82.4

由上述实验结果对比可知,由于过滤了不规则数据,并对其分组、分段及循环检测,因此,从处理时间(T)上来看,改进后算法处理相同数据集所用的时间平均比原算法减少了近 40%;从召回率(R)上来看,原算法召回率最大值及最小值均低于改进后算法,由原来的平均 82.6%上升到 86.3%;从查准率(P)上来看,原算法的查准率都在 79%左右,而改进后算法的查准率上升为 83.4%;从算法稳定性上来看,改进后算法也比原算法有所提高.

实验表明,改进后的算法与原 SNM 算法相比有以下优点:滤除不符合规则的数据,减少数据处理时间,在处理效率上优于原始 SNM 算法;对数据进行分段检测,减少了一些不必要的比较过程;引入随机因子使窗口自适应的进行调整,并对数据集进行

循环检测,提高了数据处理速度和精度.

5 结束语

本文介绍并分析了数据清洗领域中对数据冗余错误进行处理的基本临近有序算法,针对该算法存在的不足及 RFID 数据特性,提出一种优化算法,并通过实验证明该算法的有效性.该算法可大大缩短数据处理时间,使数据处理效率和精度得到进一步提高.

参考文献:

- [1] Sudarshan. RFID Data[C]//Proceedings of the 30th VLDB Conference, Toronto, Canada: IEEE, 2004: 1189-1195.
- [2] Chawathe S, Krishnamurthy V, Ramachandran S, et al. Managing RFID data[C]//Proc of VLDB, Toronto, Canada: IEEE, 2004: 1189-1195.
- [3] 杨大字. 基于 RFID Reader 的过滤技术的设计与实现[D]. 上海: 上海交通大学, 2006.
- [4] Bitton D, DeWitt D J. Duplicate record elimination in large data files[J]. ACM Transactions on Database Systems, 1983, 8(2): 255-256.
- [5] 李坚, 郑宁. 对基本 MPN 数据清洗算法的改进[J]. 计算机应用与软件, 2008, 25(2): 245-246.
- [6] Hernandez M, Stolfo S. Real-world data is dirty: data cleansing and the merge/purge problem[J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37.
- [7] Monge A E. An adaptive and efficient algorithm for detecting approximately duplicate database records [EB/OL]. [2012-10-30]. <http://citeseer.ni.nec.com/monge00adaptive.html>.
- [8] HEMANDEZ, M A, STOLFO S J. The merge/purge problem for large database[C]// proceedings of the ACM International Conference on Management of data. NY, USA: ACM, 1995: 127-138.
- [9] 潘巍. 一种有效的多数据源 RFID 冗余数据清洗技术[J]. 西北工业大学学报, 2011, 29(3): 435-442.
- [10] Lee M L, Ling T W, Low W L. IntelliClean: a knowledge-based intelligent data cleaner[C]//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 290-294.

作者简介:

陈旭辉 男, (1972—), 教授, 博士. 研究方向为无线传感器网络、复杂网络.

王馨 女, (1988—), 硕士研究生. 研究方向为无线传感器网络.

柯铭 女, (1978—), 副教授, 博士. 研究方向为生物信息处理、复杂网络理论等.