一种基于时间序列的 RFID 供应链数据分析方法

高 昕[1,3],赵 文[2,3],叶 蔚[1,3],张世琨[2,3],王立福[2,3]
(1.北京大学信息科学技术学院,北京 100871; 2.北京大学软件工程国家工程研究中心,北京 100871;
3.北京大学信息科学技术学院软件研究所高可信软件技术教育部重点实验室,北京 100871)

摘 要: 通过挖掘海量 RFID(Radio Frequency Identification)数据来优化供应链已经成为一个研究热点.本文针对供应链流通中出现的若干周转异常并且难以发现的问题,提出了一种基于时间序列的 RFID 供应链数据分析方法.将供应链的 RFID 数据统一成反映各环节周转状况的时间序列格式,然后通过分段趋势分解方法分解提取的时间序列数据,并根据分解后的随机项建立阈值来判断数据是否异常,从而建立相应的时间序列分析模型;最后基于模型检测数据异常.通过多样本和多数据集的实验检测,结果表明这种方法有效并具有较高的效率.

关键词: 无线射频识别 (RFID); RFID 数据集; 供应链; 时间序列
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2010) 2A-026-07

A Data Analysis Method for RFID Supply Chain Based on Time Series

GAO Xin[1,3], ZHAO Wen[2,3], YE Wei[1,3], ZHANG Shi-kun[2,3], WANG Li-fu[2,3]
(1. *School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871, *China*;
2. *National Engineering Research Center for Software Engineering, Peking University, Beijing* 100871, *China*;
3. *Key Laboratory of High Confidence Software Technologies (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871, *China*)

Abstract: To optimize Supply Chain system by mining mass RFID (Radio Frequency Identification) data has been an important research area. In this paper, we provide a data analysis method for RFID supply chain based on time series for the exceptions like non-effective transportation and so on, which are hard to be detected in the circulation of supply chain. This method first turns RFID data in each transportation phase or storage phase into the unified form of time series which can reflect the circulation situation of each phase; then carries time series analyze on the RFID data by the method of subsection tendency analyze, builds the threshold by the random items and builds the corresponding time series analysis model; at last checks the RFID data if they are abnormal based on these models. Through multi-sample and multi-dataset experiment, the result shows that our method is effective and efficient.

Key words: radio frequency identification(RFID); RFID dataset; supply chain; time series

1 引言

在 RFID 供应链系统中,存在大量的 RFID 标签和众多的读写器,因此会产生海量的数据.而在 RFID 供应链系统中,这些海量的 RFID 数据包含物品本身的相关信息和物品在流通中的时间与空间信息,通过有效的组织将为供应链系统用户(例如,生产商、物流企业、零售商等)提供众多有价值的信息[1].例如,物品在运输过程中的路径和运输时间、特定时间的位置和运输效率等信息.同时,供应链系统用户更加关注物品在整个流通过程中出现的不符合相关企业的非功能性需求,例如运输延迟、异常行为、物品目录异常和物品丢失等.这些问题会在流通过程中的数据上得到体现,同时相应的数据分析也要面对数据跨多个企业所带来的问题.

因此如何有效地分析和挖掘这些信息就成为 RFID 供应链系统必须要解决的问题.这方面主要有两类方法,一类是通过机器学习等数据挖掘的方式分析 RFID 数据,另一类是通过定义具体业务规则分析 RFID 数据.后者更加有效,可以明确地找出异常数据,但由于不是所有的供应链系统环节都可以提供具体的业务规则,所以应用范围比较有限;前者分析数据特征,应用范围较广,但往往存在效率和误差方面的问题.所以在实际的 RFID 数据分析中,往往要结合这两方面的优势.依靠后者的业务规则来提取数据分析的业务需求和分析

收稿日期:2009-08-31;修回日期:2010-01-10
基金项目:国家高技术研究发展计划(863)(No.2006AA04A119,No.2006AA04A121);国家重点基础研究发展计划(973)(No.2009CB320706);国家自然科学基金(No.60803014)

的角度、规则等,同时基于这些信息设计出相应的数据挖掘方法,从而达到既符合实际业务要求又有较大应用范围的目的.

在本文中,通过分析供应链中的业务需求,总结并给出了具体的 RFID 数据相关业务规则,并将这些信息转化为具体的时间序列模型.然后,基于时间序列模型检测出物品在供应链中流通的异常情况.具体的分析过程是通过收集供应链系统各运输阶段和仓储阶段中的 RFID 数据,建立学习样本,并基于这些学习样本对供应链系统中各阶段 RFID 数据进行时间序列分解,获得相应的数据流模型,从而给出一种基于时间序列的 RFID 数据分析方法来检测反映流通中异常情况的 RFID 数据.

2　相关研究

RFID 数据处理有众多的研究方面,研究主要集中在两方面,包括对从读写器得到的 RFID 编码进行处理生成相应的 RFID 数据集以及 RFID 数据挖掘.

2.1　RFID 数据集

当帖有 RFID 标签的物品进入到 RFID 读写器的有效范围内时,RFID 读写器就会读取标签,生成相应的 RFID 数据集.从这些 XML 数据中我们可以抽取出 RFID 的原始数据集 *RawDataSet*,包括物品的 *ID*、事件和地点,可以用一个三元组来表示(*ID*, *Location*, *Time*)[2].这样供应链系统的数据可以用这种格式的数据集表示,例如一批数量为 n 的货物在经过 m 个地点的供应链流程中收集的数据将如表 1 所示.

表 1　原始 RFID 数据集

RFID RawDataSet(*ID*, *Location*, *Time*)
$(i_1,l_1,t_1)(i_2,l_1,t_1),\cdots,(i_n,l_1,t_1)$
$(i_1,l_2,t_2)(i_2,l_2,t_2),\cdots,(i_n,l_2,t_2)$
… …
$(i_1,l_m,t_m)(i_2,l_m,t_m),\cdots,(i_n,l_m,t_m)$

在实际过程中,往往由于操作等多种原因,同一货物在相同地点可能被读取多次,造成数据冗余,所以需要对原始数据进行清洗.这样形成 RFID 数据的四元组(*CID*, *Location*, *Time_In*, *Time_Out*)的形式,其中 *Time_In* 和 *Time_Out* 分别代表物品进入和物品离开某地的时间,如表 2 所示.

表 2　清洗处理后的 RFID 数据集

RFID DataSeAfterCleaning(*CID*, *Location*, *Time-In*, *Time-out*)
$(i_1,l_1,t_1,t_2)(i_2,l_1,t_1,t_2),\cdots,(i_n,l_1,t_1,t_2)$
$(i_1,l_2,t_2,t_3)(i_2,l_2,t_2,t_3),\cdots,(i_n,l_2,t_2,t_3)$
… …
$(i_1,l_m,t_m,t_{m+1})(i_2,l_m,t_m,t_{m+1}),\cdots,(i_n,l_m,t_m,t_{m+1})$

从数据中可以看出,一批物品在供应链流通中的相同地点和相同时间会有众多的 RFID 数据,而且对应实际运输中物品数量的变化,这些数据也会发生变化.所以对收集到的相同时间和相同地点的 RFID 数据进行聚合,并通过聚合表(*SID*, *Location*, *Time_In*, *Time_Out*)和映射表 Map 表示,从而可以挖掘出代表某次物品流通状况的 RFID 数据集.从而可以将跨多个企业的数据统一成一致的格式进行处理.

2.2　RFID 数据挖掘

RFID 数据的管理与普通数据管理相比,具有一些明显的特征.RFID 数据的海量,冗余,不准确,连续性,实时性等特点都给管理策略提出了很大的挑战.这方面得到众多研究者的关注,文献[3,5]从反映企业业务逻辑的复杂事件的角度在 RFID 数据中挖掘各种复杂事件的相关信息.文献[4]关注基于 RFID 数据的收集、转换和重组从而更有效地管理供应链的物品流通.文献[13]利用时态实体关系模型管理供应链中的 RFID 数据.文献[6]通过一个 RFID 部署模型分析业务路线和用户行为.文献[7]为了管理海量的 RFID 数据,提出了路径和工作流两种数据模型来存储和挖掘 RFID 数据.文献[9]针对供应链中的运输效率低和一些欺诈行为给出相应的异常数据挖掘方法.

从数据采集层面上来说,数据的预处理是一个必要的环节.数据预处理主要包括数据清理、数据集成和数据规约.对于 RFID 数据,其中最主要的工作就是数据清理.其实,数据清理对于一般的数据挖掘任务来说都是很重要的一个环节.而 RFID 的数据预处理与其它普通任务的区别在于它基于流数据,所以 RFID 数据具有连续性并要求数据清理策略具有实时处理的能力.在此要求之上,数据清理主要解决三个问题:阅读中丢失数据,阅读中不可靠数据以及数据冗余.

然后需要对这些经过预处理的数据进行分析和挖掘.在这个阶段中,机器学习方法扮演了重要的角色.RFID 数据是一种特殊的流数据,经过前面的数据预处理和管理,对于上层的分析算法来说,可以当成流数据来处理,而机器学习的有监督学习方法和无监督学习方法成为重要的解决方案.所谓监督学习,是指在学习过程中,每一个训练样本都被赋予了一个标记,学习的目标是从训练样本中归纳出标记的概念,从而能够正确预测未遇见过的样本的标记,典型的监督学习任务有分类和回归;而在非监督学习中,所有样本均无标记,学习是为了发现样本集中的内部结构,例如发现样本的本征维度,聚类就是属于非监督学习.而本文正是基于有监督学习以时间序列分析的方式来建立对 RFID 数据的异常检测模型,从而从海量 RFID 数据中分析出异常的数据.

3 供应链 RFID 数据分析方法

我们通过抽象 RFID 的业务需求建立相应的数据分析规则,并转化为相应的时间序列分析规则,从而给出了基于时间序列的数据分析方法.

3.1 RFID 数据分析规则

3.1.1 RFID 供应链业务需求

本文中主要关注的供应链系统业务需求[11]如下所示:

(1)速度一致性

供应链系统中通常需要对物品的流通速度进行限制.在供应链环节中,物品的流通速度设有最大速度和最小速度.速度限制是为了防止运输过程中超出运输机制的许可,而往往出现超出限制的情况是由于流通中出现了差错,例如当重复的标签出现在异地会导致物品的流通速度过快;同时,如果物品流通速度过慢,会导致物品流通的延迟,所以需要对速度做出限制.由于 RFID 数据代表着物品的流通状况,所以其反映流通速度快慢的时间属性也有相应的限制.

(2)停留时间一致性

在供应链中,物品一般都需要尽快地送到目的地,从而减少物品保质期的损耗,这样就需要设置在仓储中的最大停留时间,例如效率低的周转过程往往导致更长的停留时间;同时由于物品在仓储周转中所必需的调配时间,所以也会有相应的最小停留时间,如果出现更小的停留时间往往是由于该环节出现了不合理的调配.

(3)流通效率一致性

在 RFID 供应链系统中,在周转时间的约束内尽可能的保证周转效率趋于一致,这样才有助于管理.例如在运输中偏长的运输时间代表着运输效率的低下,而偏短的运输时间又代表着油料等运输成本的上升.

从这些业务需求中,可以发现实际上是对供应链过程中仓储和运输环节的时间约束,因为实际操作过程中不可能随时检测到流通的速度,但可以将速度量化成时间,从而对时间提出具体的限制(最大周转时间和最小周转时间).这样我们可以把供应链各环节统一为节点序列:

$$P=\{P_1,P_2,\cdots,P_m\}$$

并为 P 中各元素设定具体的时间约束 SI(静态时间间隔):

$$\forall\, p\in P, SI(p)=[SEFT(p),SLFT(p)]$$

其中,$SEFT(p)$代表环节 p 的最小周转时间,$SLFT(p)$代表环节 p 的最大周转时间.从而依据 SI 可以方便的检测出异常数据.

而在实际供应链应用中,我们往往只能给出的都是大粒度运输环节的流通速度限制和个别的仓储的停留约束,而无法对所有仓储和运输过程给出具体的时间约束.同时,虽然在实际供应链系统中不符合时间约束的异常数据很容易被检测到,但很多反映流通效率不一致的数据却无法被检测到,所以需要统一的、细粒度的可以全面衡量整个供应链物品流通过程的数据分析方法.这里我们基于挖掘 RFID 数据集形成的时间序列,形成统一的数据评测模型,从而判断供应链物品流通中的各环节是否符合业务需求.

3.1.2 RFID 时间序列分析规则

通过上面的分析,供应链中的仓储和运输环节都可以有相应的时间约束,而同时这些环节的 RFID 数据也可以表示成代表相应环节用时的时间随机变量.这样一次供应链中物品流通的过程就可以表示成一组按时间次序排列的随机变量序列:$\{X_t\}=X_1,X_2,\cdots,X_n$.同时通过从 RFID 数据集中收集到的一次流通的观测值:$x_1,x_2,\cdots,x_n$ 代表对供应链时间序列的一次实现.这样我们将收集到的代表供应链流通的 RFID 数据集按时间序列的方式进行组织,从而 RFID 数据分析的任务变成分析每次观察值 x_i 是否符合随机变量 X_i 的约束.由于供应链中各环节用时不同,所以该时间序列各元素之间的间隔也不同.这样的时间序列不利于分析,所以我们将围绕供应链中单个环节的不同时段建立的时间序列进行分析.

对供应链中的单个环节(仓储或者运输)建立时间序列,将一天之中有效工作时间等分成 m 段,用时间序列$\{Y_t\}=Y_1,Y_2,\cdots,Y_m$ 代表,其中 $Y_i(1\leqslant i\leqslant m)$表示第 i 段时间内某次物流周转时间.通过随机抽样,可以形成 RFID 数据样本矩阵 $\boldsymbol{M}$,如下所示.其中 $y_{i,j}(1\leqslant i\leqslant n,1\leqslant j\leqslant m)$代表该环节在第 j 个时间段里第 i 次物流的周转时间.

$$\boldsymbol{M}=\begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,m} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,m} \end{pmatrix}$$

基于 RFID 数据样本矩阵 $\boldsymbol{M}$ 便可以对时间序列$\{Y_t\}$进行分解,时间序列分解关键是将时间序列的随机变量分解成三部分的叠加 $Y_t=T_t+S_t+R_t$[12],其中$\{T_t\}$是趋势项,$\{S_t\}$是季节项,$\{R_t\}$是随机项,这里我们采用分段趋势分解方法.首先算出趋势项,这里将趋势项$\{T_t\}$的估计定义为单天各时段物流周转时间的平均值.这样就得到:

$$\overline{T_{1,1}}=\overline{T_{1,2}}=\cdots=\overline{T_{1,m}}=T_1$$

$$\overline{T_{2,1}}=\overline{T_{2,2}}=\cdots=\overline{T_{2,m}}=T_2$$

$$\cdots$$

$$\overline{T_{n,1}}=\overline{T_{n,2}}=\cdots=\overline{T_{n,m}}=T_n$$

然后,利用原始数据$\{x_t\}$减去趋势项$\{T_t\}$得到的数据将只包含季节项和随机项.操作这些数据,用第 k 时段的平均值作为季节项 $S(k)$,$1\leqslant k\leqslant m$ 的估计,即

$$S(k)=\frac{1}{n}\sum_{i=1}^{n}(y_{i,k}-\overline{T_{i,k}}),\ 1\leqslant k\leqslant m$$

最后,随机项$\{R_t\}$便可以通过原始数据$\{x_t\}$减去趋势项$\{T_t\}$和季节项$\{S_t\}$得到,如下所示:

$$R_{i,j}=y_{i,j}-T_i-S_j,\ 1\leqslant i\leqslant n;1\leqslant j\leqslant m$$

这样就完成了对时间序列$\{Y_t\}$的分解,可以看出趋势项$\{T_t\}$和季节项$\{S_t\}$在特定时段基本固定,对于数据的判断没有太大的意义,而随机项$\{R_t\}$具有变化性.所以统计出特定时段随机项 R_j 的最大值 $R\max_j$ 和最小值 $R\min_j$,其中:

$$R\max_j=\max_{1\leqslant i\leqslant n}(R_{i,j}),R\min_j=\min_{1\leqslant i\leqslant n}(R_{i,j})$$

所以对于要检测的 RFID 数据 y,可以通过如下方法求出对应的随机项.

$$\overline{T}=\frac{1}{n}\sum_{i=1}^{n}T_i$$

$$R=y-\overline{T}-S_j,\ 1\leqslant j\leqslant m$$

然后检测随机项 R 与区间$[R\min_j,R\max_j]$是否相符.如果 R 不在区间内,则表示该数据异常.

3.2 RFID 数据分析过程

根据上面得到的 RFID 数据分析规则,本节给出具体的数据分析过程,包括 RFID 数据预处理、时间序列分析处理和供应链数据检测.

3.2.1 RFID 数据预处理

我们对原始数据的预处理主要进行数据清洗和数据聚合.其中数据清洗是为了消除冗余数据以及包括物品丢失和物品目录异常在内的数据一致性处理,具体处理过程如算法 1 所示.

算法 1　原始数据集清洗算法

输入:原始数据集 RawDataSet(*ID*, *Location*, *Time*)
输出:清洗后数据集 DataSetAfterCleaning(*CID*, *Locaion*, *Time_In*, *Time_Out*)
方法:
Table, PreTable: Empty HashTable
//Table 记录一个物品在某一位置的进入和离开时间
//PreTable 记录在流通过程中上一个位置的物品流通状况
FOR each Record in RawDataSet
　IF Record is not in PreTable
　　Report Reocrd is missing; Continue;
　ENDIF
　IF Record is not the same as the corresponding item in PreTable
　　Copy the attributes of item to Record
//这里 attributes 指那些描述物品本身的属性,不包括流通相关属性
　ENDIF
　IF *Record*.*Time* > Table[Record.*ID* + *Record*.*Location*].Max*Time*
　THEN Table[*Record*.*ID* + *Record*.*Location*].Max*Time* = *Record*.*Time*//记录离开时间
　ELSE IF *Record*.*Time* < Table[*Record*.*ID* + *Record*.*Location*].Min*Time*
　THEN Table[*Record*.*ID* + *Record*.*Location*].Min*Time* = *Record*.*Time*//记录进入时间
　ENDIF
ENDFOR
FOR each Record in Table
　Add(*Record*.*ID*, *Record*.*Loaction*, Min*Time*, Max*Time*)
　To DataSetAfterCleaning//生成清理后数据集
ENDFOR

而数据聚合则是进一步精简数据,找出代表一批物品在供应链中的流通状况的数据,具体处理过程如算法 2 所示.

算法 2　数据聚合算法

输入:清理后数据集 DataSetAfterCleaning(*CID*, *Location*, *Time_In*, *Time_Out*)
输出:聚合数据集 AggregateDataSet(*AID*, *Locaion*, *Time_In*, *Time_Out*)
方法:
Table: Empty HashTable
//Table 记录相同地点、相同进入时间和离开时间的 CID 集合
FOR each Record in DataSetAfterCleaning
　Add Record to Table[*Record*.*Time_In* + *Record*.*Location* + *Record*.*Time_Out*]
ENDFOR
FOR each Record in Table
　Generate a corresponding AID
　Add(*AID*, *Record*.*Location*, *Record*.*Time_In*, *Record*.*Time_Out*)
　To AggregateDataSet
ENDFOR

3.2.2 时间序列分析处理

由于前面的 RFID 四元组数据主要突显了物品在供应链各仓储地点的状态,但与时间序列分析模型所关注的各环节用时还有差距,所以需要在 RFID 预处理聚合的基础上加工出物品在运输过程中的数据,即将四元组数据(i_k,l_q,t_l,t_{l+1})和$(i_k,l_{q+1},t_{l+2},t_{l+3})$处理成三个三元组(*ID*, *Time_in*, *Time*)数据,其中 *Time* 代表物品在供应链中某流通环节(运输或仓储)用时,具体处理过程如下:

$$(i_k,ti_u,t_e),\ t_e=t_{l+1}-t_l,\ ti_u=t_l$$
$$(i_k,ti_{u+1},t_{e+1}),\ t_{e+1}=t_{l+2}-t_{l+1},\ ti_{u+1}=t_{l+1}$$
$$(i_k,ti_{u+2},t_{e+2}),\ t_{e+2}=t_{l+3}-t_{l+2},\ ti_{u+2}=t_{l+2}$$

这样通过算法 3 我们可以得到某次物品在供应链特定路线的流通数据,其形式为$(i,ti_1,t_1)(i,ti_2,t_2)$,

…,(i, ti_n, t_n)的流数据,而这正是我们时间序列分析的基础.

算法 3 数据重组算法

输入:聚合数据集 AggregateDataSet(*AID*, *Locaion*, *Time_In*, *Time_Out*)
输出:重组数据集 DataSet(*AID*, *Time*)
方法:Table, DataSet: Empty HashTable
FOR each Record in AggregateDataSet
  Add Record To Table[*Record.AID*]
ENDFOR
FOR each Record in Table
  Sort elecmet in Table[*Record.AID*] by element.Time_In
  FOR each element in Table[*Record.AID*]
    $T1$ = *element.Time_Out* − *element.Time_In*
    $TI1$ = *element.Time_In*
    Add(*AID*, $TI1$, $T1$) to Dataset
    IF there are next *element in* Table[*Record.AID*]
      $T2$ = *nextelement.Time_In* − *element.Time_Out*
      $TI2$ = *element.Time_Out*
      Add(*AID*, $TI2$, $T2$) to Dataset
    ENDIF
ENDFOR

经过对 RFID 数据的预处理,我们可以得到大量的关于供应链各环节(运输环节或仓储环节)的周转时间数据.接下来就是要对这些数据进行时间序列分析,从而建立该环节的数据模型以检验流通中在该环节的周转是否正常.这个过程主要包括后台的抽样建模和前台的模型检测两部分.抽样建模首先将收集到的众多的该环节的数据按照一天之中有效工作时间等分成 m 段,每个时间段随机抽取 n 个样本,其中 m 和 n 根据该环节的具体情况而定.而建模过程就是基于这些样本数据求出各时间段的趋势项、季节项以及随机项对应区间[$R\min$, $Rmax$].具体过程如算法 4 所示.

算法 4 时间序列建模算法

输入:某环节样本数据集
  SampleDataSet(*ID*, *Time_In*, *Time*)
输出:时间模型数据集 TimeDataSet(*Time_Num*, T, S, *Rmin*, *Rmax*)
方法:Table, DataSet: Empty HashTable
T, S, $R\min$, $R\max$:0 ; RandomSet: Set
Sort Record in SampleDataSet by *Record.Time_In*
FOR each Record in SampleDataSet
  Find the corresponding *Time_Num* to *Record.Time_In*
  Add *Record to Table*[*Time_Num*]
ENDFOR
FOR each Record in Table
  Find random n elements from Record to RandomSet
  FOR each element in RandomSet
    T = T + *element.Time*
  ENDFOR
  T = T/n //求出趋势项
  FOR each element in RandomSet
    S = S + (*element.Time-T*)
  ENDFOR
  S = S/n //求出季节项
  FOR each element in RandomSet
    Temp = *element.Time-T-S*
    IF *Temp* < $R\min$ $R\min$ = *Temp*
    ELSE IF *Temp* > $R\max$ $R\max$ = *Temp*
    ENDIF
  ENDFOR //通过随机项求出阈值区间
  Add(*Time_Num*, T, S, $R\min$, $R\max$) to TimeDataSet
ENDFOR

3.2.3 供应链 RFID 数据检测

通过时间序列分析处理可以得到供应链中各环节的检测模型,所以供应链 RFID 数据检测的过程就是将流通中收集的 RFID 数据按照前面的预处理过程进行处理,然后分别判断在相应环节中所属的时间段,从而找到相应的检测模型.根据平均趋势项和该时间段的季节项,求出随机项,并判断随机项是否在相应区间[$R\min$, $R\max$]中.

4 实验和工具

4.1 检测实验

对于本文提出的关于 RFID 数据的异常检测方法是否有效,关键在于是否存在漏报和误报,以及相应的严重程度.漏报是指算法没有检测出对于应当报告异常的数据;误报是指将属于正常范围内的数据报告成异常.两者从不同角度描述了数据检测算法的精确程度,通过这两方面的结论就可以判断出方法的有效性.

由于我们的方法是基于有效数据建立检测模型来评价收集到的 RFID 数据,相应的检测边界不会超出正常最大边界,所以不存在漏报情况,同时通过了相应的实验也证明了这条结论,这也是方法的优势,所以实验将主要集中在误报情况上.由于我们的方法将供应链中的运输和仓储表示成统一的数据序列进行处理,所以我们只要围绕任何一段供应链过程即可检测方法的误报情况.所以在误报实验中我们首先提取一段过程的数据建立学习样本,建立时间序列检测模型;然后模拟该过程的 RFID 数据,建立测试数据集;最后将通过模型在测试数据集中检测到的异常数据作为误报情况,从而观察方法的误报程度.

我们对供应链系统 RFID 数据检测进行了模拟实验,实验的硬件环境为 P4 2.4GHz 的 CPU 和 1GB 内存,操作系统为 WindowsXP 专业版,java 运行环境为 JDK1.6.0-03.实验数据是人工生成的,通过随机函数对各节点中流通数据区间加权从而模拟供应链中各节点

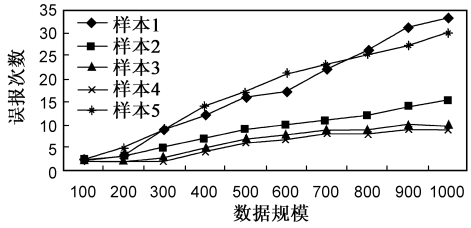
的物品流通,给出一定数量的不同物品在节点不同时段流通的数据集合.其中学习样本的数据都属于节点数据区间内部的正常数据,而测试数据要包含一部分异常数据,从而评价检测方法的在误报方面的效果.具体包含一个 5 千条 EPCIS 事件集合的学习样本和一个 5 万条的 EPCIS 事件集合的测试数据集.实验主要分下面两步:

(1)收集规模为 100 的 5 组不同样本和规模为 1000 的 5 组不同测试数据集来进行误报实验,分别进行多样本对单数据集的误报实验和单样本对多数据集的误报实验.部分结果如图 1(a)、(b)所示.其中图 1(a)代表 5 组样本分别对数据集 1 的误报情况,可见在对数据集的误报测试中,误报比率没有超过 3.5%,多数样本(样本 2,3,4,5)的误报比率集中在(1%,3%)之间,而只有样本 1 误报比率较高.图 1(b)代表样本 1 对 5 组数据集的误报情况,可见误报趋势基本一致,并且偏差不超过 1%.

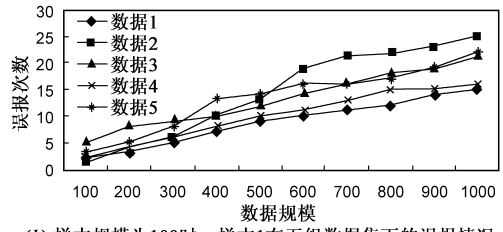
通过图 1(a)的结果和五组样本对其他数据集的实验结果表明方法的误报情况都处于比较低的状况(低于 5%),各样本的误报情况基本稳定,虽然也有个别样本的误报较大,但整体反映了方法对不同样本的有效性.图 1(b)的结果和其他样本对五组数据集的结果反映了同一样本对不同数据的误报情况,不同的数据集的误报情况相差不超过 1.5%,由于这些数据集的分布基本相同,所以可以看出同一样本对有效范围内的误报情况也是稳定的;同时对比相同数据集的误报情况,我们发现不同样本对不同数据集的反映是基本一致的(样本 1 中误报较高的数据集在其它样本中也较高,反之一样).所以我们的方法在样本和数据集不同匹配的情况下具有稳定性,说明了方法是基本有效的.

(2)我们收集规模为 200 的 5 组不同的样本,并延用上面规模为 1000 的 5 组不同的测试数据集来进行相同的误报实验,部分结果如图 1(c)、(d)所示,分别和上面的图 1(a)、(b)对应.图 1(c)中反映的误报情况基本低于 1%,与图 1(a)相比,其结果得到很大的改进,相比样本规模增加了一倍,误报降低了超过了一倍.可见当样本规模增大,将有效的提高方法的有效性.图 1(d)和图 1(b)相比,同一样本对不同数据集的误报情况相互之间偏差不超过 0.5%,降低了一倍.可见增大样本规模,建立的检测模型对不同数据的误报情况更加稳定,同时误报情况会极大的下降.

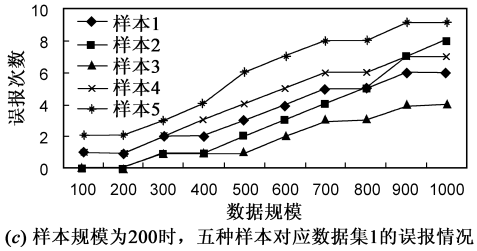
总之,通过上面的分析可以得出,我们的方法的精确程度受到学习样本的影响,不同的样本导致检测模型有不同的精确度,但是通过增加样本规模可以整体极大的提高检测模型的精确度.而且这在实际中是可行的,系统可以不断的收集到海量的有效范围内的数据来供我们抽取样本,所以我们的方法符合实际要求并且有效.



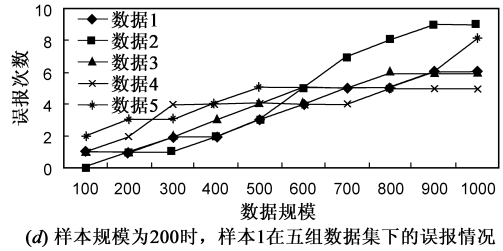
(a) 样本规模为100时，五种样本对应数据集1的误报情况



(b) 样本规模为100时，样本1在五组数据集下的误报情况



(c) 样本规模为200时，五种样本对应数据集1的误报情况



(d) 样本规模为200时，样本1在五组数据集下的误报情况

图1　误报分析图

4.2　图形化工具

针对本文的 RFID 供应链数据分析方法,我们给出了相应的图形化工具,主要描述物品在供应链中的流通路线以及在流通过程中通过数据分析方法检测到的异常.

首先,我们基于 google map[8] 的基本地理界面和叠加层可视化的描述了供应链中的物品流通,通过 RFID 的数据收集,我们可以得到物品所经过的地点.所以首先在地图上给出了流通的描述(圆心点标记表示流通经过的地点的仓储,粗折线表示连接仓储之间的运输).如图 2 所示,我们的图形化工具可以根据("广州","厦门","温州","上海","无锡")形式的路线描述给出相应的展示.

其次,描述了检测到的异常情况.如果某段运输的

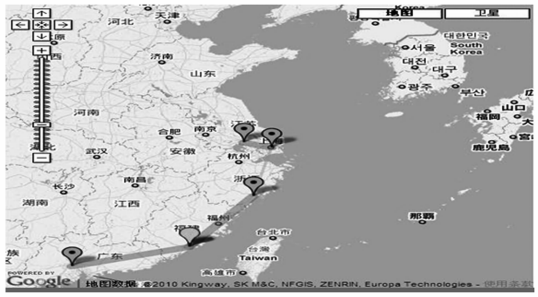


图2 供应链流通异常的图形化显示

数据出现异常,该段的粗折线将被标成细折线;同样,当某地的仓储周转数据出现异常,该地点的图标将变成实心点.图 2 中描述了在厦门出现仓储周转异常,以及在温州到上海的运输出现异常.

通过 google 地图这种实时、形象的载体可以有效的体现数据分析的结果,而且可以将数据分析方法封装成服务,让关心供应链流通的人们可方便的从网上了解到流通的状况.同时也可以很方便的集成到相关各企业的系统中.

5 总结

本文依据实际中供应链在时间和空间方面的 RFID 业务规则将 RFID 数据统一规约为时间序列格式,并给出了相应的基于时间序列的数据分析方法,从而提供了一种检测供应链系统中异常数据的方法.通过实验,结果表明了该方法是实际可行并且有效的,最后为了更好的描述供应链中数据检测,我们给出了相应的图形化工具,可视化的描述了供应链流通过程以及其中出现的异常情况.在研究过程中,我们发现还有更多的业务规则可用于 RFID 数据的挖掘,所以下一步工作将围绕特定规则的数据分析方法展开.

参考文献:

[1] Adam Melski, Lars Thoroe, Matthias Schumann. Managing RFID data in supply chains[J]. Internet Protocol Technology, 2007,2(3/4):176 – 189.

[2] 顿海强,赵文.一种基于 RFID 数据集的物品工作流挖掘方法[J].电子学报,2008,36(12A):86 – 93.
Dun Hai-qiang, Zhao Wen. A commodity workflow mining approach based on RFID data sets[J]. Acta Electronica Sinica, 2008,36(12A):86 – 93.(in Chinese)

[3] Jin Xingyi, Lee Xiaodong, Kong Ning. Efficient complex event processing over RFID data stream[A]. Seventh IEEE/ACIS International Conference on Computer and Information Science [C]. Washington: IEEE Computer Society, 2008.75 – 81.

[4] Adam Melski, Lars Thoroe, Matthias Schumann. Managing RFID data in supply chains[J]. Int. J. Internet Protocol Technology, 2007,2(3/4):176 – 189.

[5] Guangming Wang, Gonglian Jin. Research and design of RFID data processing model based on complex event processing[A]. International Conference on Computer Science and Software Engineering[C]. Washington: IEEE Computer Society, 2008. 1396 – 1399.

[6] I-En Liao, Wei-Chih Lin. Shopping path analysis and transaction mining based on RFID technology[A]. RFID Eurasia, 2007 1st Annual[C]. Washington: IEEE Computer Society, 2007. 1 – 5.

[7] Jiawei Han, Hector Gonzalez. Warehousing and mining massive RFID data sets[A]. ADMA 2006[C]. Heidelberg: Springer Berlin, 2006. 1 – 18.

[8] Google Map API[OL]. http://code.google.com/intl/zh-CN/apis/maps/.

[9] Elio Masciari. A framework for outlier mining in RFID data [A]. IDEAS 2007[C]. Washington: IEEE Computer Society, 2007.263 – 267.

[10] 袁崇义. Petri 网原理与应用[M].北京:电子工业出版社,2005.

[11] Alexander Ilic. Thomas Andersen, Florian Michahelles. Increasing supply-chain visibility with rule-based RFID data analysis[A]. Internet Computing, IEEE[C]. Piscataway: IEEE Educational Activities Department, 2009.31 – 38.

[12] 何书元.应用时间序列分析[M].北京:北京大学出版社,2007.

[13] Fusheng Wang, Peiya Liu. Temporal management of RFID data[A]. Proceedings of the 31st VLDB Conference[C]. Trondheim: VLDB Endowment, 2005.1128 – 1139.

作者简介:

高 昕 男,1983 年出生,北京大学博士研究生,主要研究方向为软件工程和 Internet 环境下应用集成的相关技术.
E-mail: gaoxin54@126.com

赵 文 男,1967 年出生,博士,副研究员,主要研究领域为软件工程、工作流技术和 RFID 相关技术.