基于改进卡尔曼滤波的RFID数据清洗方法研究

陈静云 周 良 丁秋林
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 采用RFID(radio frequency identification)技术在阅读器和动态电子标签之间进行通信,采集到的数据通常存在大量的脏数据。为更好支持高层应用,必须对原始数据进行清洗。针对标签频繁移动这一特点,将滑动窗口技术引入卡尔曼滤波模型,给出了一种改进的卡尔曼滤波模型,进而提出了一种基于改进卡尔曼滤波的RFID数据清洗方法。该方法在保证数据清洗准确率的基础上能有效解决标签动态跃迁带来的时间延迟问题,从而更加适用于标签频繁移动的场景。实验结果表明,该算法提高了清洗效率及准确率。

关键词 卡尔曼滤波,RFID,数据清洗,滑动窗口,动态标签
中图法分类号 TP311 **文献标识码** A

Cleaning Method Research of RFID Data Stream Based on Improved Kalman Filter

CHEN Jing-yun ZHOU Liang DING Qiu-lin
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract RFID (radio frequency identification) technology has been widely applied to the communication between RFID readers and dynamic electronic labels, but the data captured by RFID readers often tends to be noisy. In order to provide a better support for high-level RFID's applications, it is necessary to clean the collected data. Considering the characteristic of frequent movement of labels, this paper put forward an improved Kalman filter model by combining the sliding window technique with Kalman filter model and then proposed a method of RFID data cleaning based on an improved Kalman filter. The method can not only guarantee the accuracy of data cleaning but also effectively solve the problem of time delay caused by dynamic electronic tags. Thus this method is more adaptable to the situation where labels are moved frequently. The experiment's result shows this approach can improve the efficiency and accuracy of the data cleaning.

Keywords Kalman filter, RFID, Data cleaning, Slide window, Dynamic label

1 引言

RFID(radio frequency identification)是一种非接触式自动识别和数据获取技术。它将无线通信、数据管理和信号处理等技术融为一体,利用无线射频信号完成阅读器与电子标签之间的数据通信,从而探测贴有标签物体的逻辑位置。RFID技术广泛应用在诸如物流装配、供应链监控、生产制造、交通管理等领域,成为研究热点。然而,RFID阅读器采集标签中的数据时,通常会受到射频信号干扰和标签周围环境等诸多因素的影响,造成采集到的数据流存在较大的不可靠性和不确定性,通常采集到的原始数据流的准确率只有60%~70%,这对高层处理应用毫无意义。为有效支持RFID高层业务逻辑,提供高质量的RFID数据,对原始RFID数据流进行清洗是非常重要的。

为得到高质量的采集数据,针对RFID原始数据清洗方法,许多学者进行了大量的研究,并提出了许多清洗算法[1,2]和清洗模型[3-5]。比较典型的有基于滑动窗口的数据清洗方法。文献[6]提出一种静态滑动窗口方法,此方法通过设置足够大的滑动窗口填补漏读数据,但是若窗口设置过大,标签在某一时刻移出后又移进阅读区,这种动态移动便不能被捕捉到。因此,不恰当的窗口大小设置将直接导致多读和漏读数据的产生。

文献[7]提出了一种适应性调整窗口大小的填补方法SMURF,该方法采用随机采样的数据统计机制,将读标签看成一个随机事件,标签读出频率即阅读率看成是事件发生的概率。该方法可根据阅读率动态调整窗口的大小,较静态窗口有明显的优势,但是,SMURF方法将每个滑动窗口的读数作为一组采样,进行状态是否发生变化的判断,然而不同的窗口划分方式会出现不同的判定结果[8]。另外,对于频繁移动的标签,阅读率变化较大,清洗效果不够理想。文献[9]将标签动态属性引入滑动窗口,增强了系统对标签频繁运动及跃迁的适应性。

文献[10]摒弃滑动窗口采用基于卡尔曼滤波的RFID数据清洗方法,该方法的优点在于考虑了观测值和估计值,根据

到稿日期:2013-05-31 返修日期:2013-09-16 本文受国防基础科研计划项目资助。
陈静云(1988-),女,硕士生,主要研究方向为系统集成,E-mail:chenjingyun074@163.com;周 良(1966-),男,博士,副教授,硕士生导师,主要研究方向为信息系统、知识工程;丁秋林(1936-),男,博士,教授,博士生导师,主要研究方向为信息系统、企业信息化。

历史数据拟合出当前时刻的估计值，再根据当前的观察值修正估计值以逼近真实值。该方法具有较高的灵敏性，在一定程度上提高了采集数据的准确性，对动态标签有了一定的探测能力，但是对于标签的频繁移动，存在跃迁时间响应延迟问题，且算法复杂度较高，处理过程需要大量的输入输出数据。

为解决上述问题，本文将滑动窗口技术引入卡尔曼滤波模型，给出了一种改进的卡尔曼模型，进而提出了基于改进卡尔曼滤波的 RFID 数据清洗算法。算法将滑动窗口应用到卡尔曼滤波的采样和处理过程中，并根据标签运动属性动态调整滑动窗口的大小来适应标签的频繁移动。该算法改善了标签发生跃迁时间响应延迟问题，提高了数据清洗的准确率，在一定程度上减少了数据的输入输出。

2 基于改进卡尔曼滤波的 RFID 数据清洗总体框架

阅读器采集的数据通常存在大量脏数据，这些脏数据包括积极读数据、消极读数据及冗余读数据。积极读数据是指标签不在阅读器阅读区域而被阅读器读取到的数据，消极读数据为标签在阅读器阅读区域而没有被阅读器读取到的数据，冗余读数据是指阅读器读取到的数据中所存在的重复数据。为去除此类脏数据，需对原始数据流进行清洗。清洗分两个层次进行：单阅读器级清洗和多阅读器级清洗。单阅读器级清洗主要解决积极读和消极读错误；多阅读器级清洗主要是排除冗余读错误以及解决标签归属问题。

清洗方法总体思路如图 1 所示。阅读器采集数据后提交到单阅读器数据清洗层进行处理：首先初始化窗口并传入参数，然后循环进行窗口设置、卡尔曼滤波处理、窗口调整直到数据处理结束，最后将清洗后的数据提交给多阅读器清洗层。经多阅读器清洗去除冗余读数据并确定标签归属后，可与传感器等其他数据源做进一步的融合处理，也可存入数据库供历史查询，或直接供上层应用访问，实现对象的实时跟踪监控。本文主要解决单阅读器数据清洗问题，即清洗结构的第一层——图 1 中的单阅读器数据清洗层。

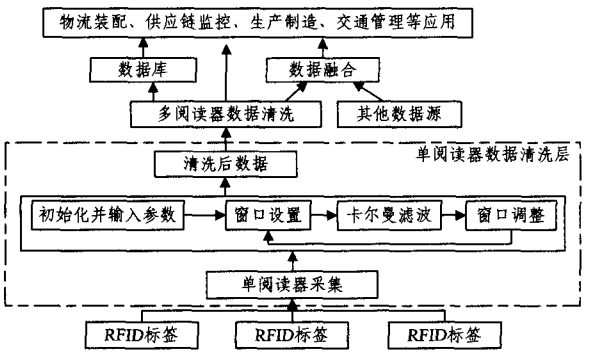

图 1 基于改进卡尔曼滤波数据清洗方法的总体框架

3 基于改进卡尔曼滤波的 RFID 数据清洗算法设计

改进卡尔曼滤波的清洗过程主要由时间更新、测量更新和窗口调整 3 部分组成。时间更新利用上一状态的估计，做出对当前状态的估计；测量更新利用当前状态的观测值修正上一时间更新获得的估计值，以获得一个更精确的新估计值，形成一个自回归的过程，以逼近真实值；窗口调整在当前数据块的数据经卡尔曼滤波处理结束后进行，根据标签的运动情况作相应的变化。图 2 为基于改进卡尔曼滤波的 RFID 数据清洗算法流程图。

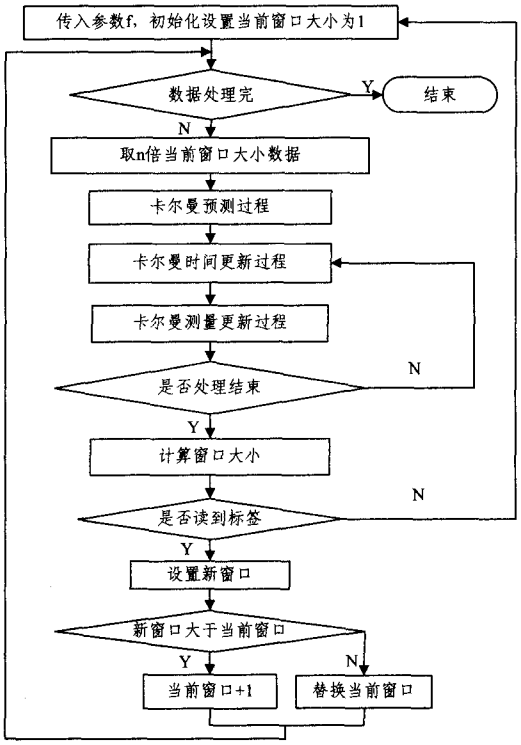

图 2 基于改进卡尔曼滤波清洗算法流程

基于改进卡尔曼滤波的数据清洗算法(KAL-RFID)具体描述为：

步骤 1 根据标签的移动特征及阅读器探测区域大小计算 f，将其作为参数传入清洗过程，并初始化窗口大小为 1。

其中参数 f 指标签运动通过阅读器通信区域，阅读器有几个探测周期来捕获标签中的信息。当标签紧贴阅读器天线移出阅读器探测区域时，阅读器可以读标签的次数为 $f=R/(V*T_{epoch})$，其中 V 表示标签的运动速度，T_{epoch} 为阅读器探测周期，R 表示阅读器的通信半径，探测周期最大值为 $2f$。

步骤 2 根据当前滑动窗口大小向数据流中取数据，转步骤 3 进行预测处理。当前窗口较大则表示系统较稳定，算法通过使用较大的滑动窗口来提高数据处理的效率；当前窗口较小表示系统变化较为频繁，算法通过使用小的滑动窗口满足清洗准确率的要求。

步骤 3 进行 KAL-RFID 预测处理。改进 KAL-RFID 算法预测处理过程包括两步：当前值的预测处理和当前值的观测处理。相应微分方程为：

预测方程：

$$x_k=x_{k-1}+b*\Delta t+w_{k-1} \tag{1}$$

观测方程：

$$z_k=x_k+v_k \tag{2}$$

式(1)中，x_{k-1} 是周期 $k-1$ 内阅读率的真实值。阅读率是在任意阅读周期内有效反馈的探测周期的次数与阅读周期包含探测周期次数的比值，本算法中阅读周期大小取滑动窗口大小。Δt 表示相邻阅读周期的时间间隔。b 是对多个阅读周期的阅读率采用直线回归分析得到的拟合直线 $y=a+bk$ 的斜率，b 的计算公式如式(3)所示：

$$b=\frac{\sum_{i=1}^{n}(k_i-\bar{k})(x_i-\bar{x})}{\sum_{i=1}^{n}(k_i-\bar{k})^2} \tag{3}$$

式(2)中，z_k 是阅读周期 k 内直接通过阅读器观测到的阅读率。w_{k-1} 和 v_k 是对应周期的观测噪声，用 Q、R 来代替，系统参数都设置为 1。

步骤 4　输入步骤 3 计算得到的 x_k、z_k，进行 KAL-RFID 更新处理。KAL-RFID 更新过程包括时间更新和测量更新两个过程，对应过程的更新方程为：

KAL-RFID 时间更新方程：

$$x_k^- = x_{k-1} + b * \Delta t \tag{4}$$

$$P_k' = P_{k-1} A' + Q \tag{5}$$

KAL-RFID 测量更新方程：

$$K_k = P_k'/(P_k' + R) \tag{6}$$

$$x_k = x_k^- + K_k(z_k - x_k^-) \tag{7}$$

$$P_k = (1 - K_k) P_k' \tag{8}$$

式(4)－式(8)中，x_k^- 表示由 $k-1$ 的过程得出的先验估计，x_k 是由测量值 z_k 得出的后验估计。K 是卡尔曼增益，用来得到最小后验估计误差的协方差。p' 是先验估计的协方差，p 是后验估计协方差。通过时间更新预测先验估计 x_k^-，并通过测量更新对后验估计 x_k 进行修正得到下一时刻的先验估计，递归式(4)－式(8)直到当前数据处理完毕转到步骤 5。

步骤 5　计算窗口大小。

将阅读器读标签看成随机事件[11]，在滑动窗口内读到标签的次数 N 满足伯努力二项分布 $B(\omega_i, p_i^{avg})$。其中 ω_i 是在滑动窗口内阅读器探测周期的次数，即重复实验的次数；p_i^{avg} 是在滑动窗口内标签的平均读出频率，即事件发生的概率。动态设置置信度 δ 使得 $\delta = p^{avg}/f$。保证标签被读到满足的充分条件：

$$(1 - p^{avg})^{\omega} < (p^{avg}/f) \tag{9}$$

由 $\ln(1-p^{avg}) < 0$，可得：

$$\omega > \ln(f/p^{avg})/p^{avg} \tag{10}$$

根据式(10)计算窗口 ω 的大小。

步骤 6　更新窗口。

判断标签是否位于阅读区。若位于阅读区即标签被读到，计算出来的 ω 大于当前窗口则将当前窗口加 1，否则替换当前窗口；若标签未被读到则窗口恢复初始值 1，然后取下一批滑动窗口数据进行处理，直到数据处理结束。使用卡尔曼滤波处理时，根据滑动窗口大小动态确定阅读周期包含的读写周期数目，然后计算各个阅读周期的阅读率，将动态变化的阅读率应用到整个处理过程。

4　实验结果及其分析

本文将对提出的清洗算法的有效性进行实验验证。实验环境为：Pentium(R) Dual-Core CPU E5800 3.2GHz 1.96G 内存处理器，Microsoft Windows XP Professional 操作系统。开发平台是 Microsoft visual studio 2010，平台开发语言是 vb.net，数据库是 oracle。本文模拟一个标签在 500 个查询周期的状态，考虑了标签的静态和动态特性，并在真实数据上加入噪声模拟阅读器产生真实数据和原始数据。

4.1　清洗结果

对实验模拟的标签频繁移动的 500 个查询周期，分别使用固定大小为 5 的静态滑动窗口(WIN-5)、改进 SMURF 方法[5]、基于卡尔曼滤波的数据清洗方法 KAL-RFID 以及本文提出的基于改进卡尔曼滤波的方法进行数据清洗。为便于比较，实验中 f 取 75，各算法的清洗结果如图 3 所示。由图 3 可以看出，WIN-5 存在大量的积极读和消极读错误。改进 SMURF 对跃迁的敏感性较 KAL-RFID 和 WIN-5 强，但清洗准确率较 KAL-RFID 低。KAL-RFID 方法的数据清洗准确率较高，但对标签跃迁响应不敏感。改进的 KAL-RFID 方法将滑动窗口技术应用到了 KAL-RFID 数据清洗过程中，在保证较好清洗准确率的基础上，提高了对跃迁响应的敏感性。下面实验将定量比较各算法数据清洗的准确率。

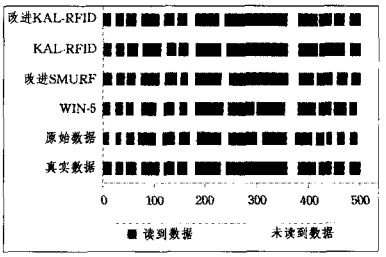

图 3　WIN-5、改进 SMURF、KAL-RFID、改进 KAL-RFID 的清洗结果

4.2　数据清洗准确率实验

为更好地比较各算法的清洗性能，对 500 个查询周期计算清洗准确率，如图 4 所示。清洗准确率 p 是查询周期内正确数据数目 t 与周期包含查询数据数目 w 的比值，即 $p = t/w$。p 值越接近 1，说明经清洗后的数据越接近真实数据。实验结果加入了原始数据的准确率，以更好地展示各清洗方法的清洗效果。由图 4 可以看出 WIN-5、改进 SMURF 和 KAL-RFID 算法的清洗准确率明显较低，而改进 KAL-RFID 算法整体有比较高且稳定的清洗效果。实验结果表明，改进 KAL-RFID 方法对于频繁移动的标签有较高的清洗准确率。

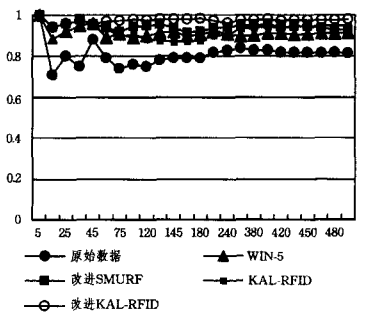

图 4　各算法数据清洗准确率

4.3　动态标签适应性实验

标签进入阅读器区域或离开阅读器区域称为一次跃迁。跃迁响应准确率是正确跃迁判定次数与真实总跃迁次数的比值，跃迁响应准确率越高，算法对标签动态的判定越准确，清洗性能就越高。跃迁响应时间是指真实数据发生跃迁的时间与清洗结果检查的跃迁时间差的绝对值。平均跃迁响应时间是指平均每次跃迁的响应时间，一段查询周期内总的跃迁响应时间与跃迁次数的比值。平均跃迁响应时间越小，清洗效果越好。从图 3 中可以看出真实数据共发生 24 次跃迁。表 1 是对这 500 个查询周期分别使用各种清洗方法后，各方法对标签跃迁的响应情况的统计结果。由于原始数据大量漏读，WIN-5 对跃迁的响应次数会比真实跃迁次数大。改进 SMURF 跃迁判定次数与真实跃迁次数相同，表明改进 SMURF 对响应标签跃迁有较好的灵敏性。KAL-RFID 对快速进出阅读区标签识别的灵敏度较改进 SMURF 的低。本文提出的改进的 KAL-RFID 可以提高系统对标签的动态变化

(下转第 227 页)

[6] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. August 2009: 248-256

[7] 黄云平，孙乐，李文波. 基于上下文图模型文本表示的文本分类研究[C]// 第四届全国信息检索与内容安全学术会议论文集(上). 2008

[8] 赵鑫，李晓明. 主题模型在文本挖掘中的应用[R]. PKU-CS-NCIS-TR2011XX. June 2011

[9] Griffiths T L, Steyvers M. Finding scientific topics[C]// Proceedings of the National Academy of Sciences. April 2004, 101: 5228-5235

[10] Griffiths T. Gibbs sampling in the generative model of Latent Dirichlet Allocation[OL]. http://people.cs.umass.edu/~wallach/courses/s11/cmpsci791ss/readings/griffithso2gibbs.pdf

[11] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines[OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

[12] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55: 77-84

[13] Blei D M, McAuliffe J D. Supervised topic models[C]// NIPS. 2007

[14] Cancho R F I, Sole R V. The small world of human language[J]. Proceedings of The Royal Society of London B: Biological Sciences, 2001, 268(1482): 2261-2265

[15] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523

(上接第 204 页)

的准确率。对比平均跃迁时间可以看出，改进 KAL-RFID 方法最少，因此，清洗效果较其他方法好。

表 1 跃迁响应表

	跃迁响应次数	跃迁响应判定准确率	跃迁响应时间	平均跃迁响应时间
WIN-5	28	83.3%	119	4.96
改进 SMURF	24	100%	37	1.54
KAL-RFID	22	91.7%	43	1.79
改进 KAL-RFID	24	100%	11	0.46

本文将标签的运动速度引入数据清洗过程，根据标签的运动速度动态设置置信度 δ。将阅读器的通信区域分成 N 个子区域，变量 S 表示一个子区域内标签被阅读器读到所需的最小读写周期数，可得到：

$$f=R/(V*T_{epoch})=S*N \tag{11}$$

其中，S 越小表示标签运动速度越快，S 越大表示标签运动的速度越慢。S 取 0 到 250，各算法清洗出错率效果图如图 5 所示。图 5 表明，标签移动越慢各算法的清洗效果越好，在标签移动速度非常快的情况下，各算法的清洗出错率仍较高。本文提出算法的出错率则较其他算法有较大的改善。

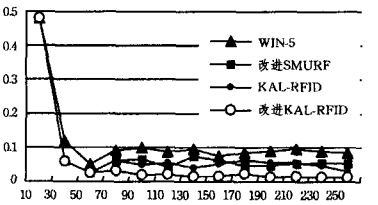

图 5 不同速度下算法的清洗出错率

4.4 空间代价分析

WIN-5 需要的存储空间最少，其算法的复杂度最低。改进 SMURF 算法滑动窗口的大小是可适应性变化的，清洗过程需要不断验证完整性条件及检测标签状态变化条件，因此输入输出数据量较 WIN-5 大，所需的空间代价相对 WIN-5 较高。KAL-RFID 通过时间更新方程和测量更新方程进行自回归逼近真实值，算法复杂度高，需要大量的输入输出数据，消耗的存储空间也较高。若不进行处理，改进的 KAL-RFID 方法与 KAL-RFID 有相当的复杂度，本算法在标签没有发生跃迁的情况下减少系统采样率，即扩大滑动窗口的大小，来减少数据的输入输出量；在检测到标签发生跃迁的时间段里通过减小滑动窗口保证预测值的准确性；因此在消耗空间代价及整体处理速度上较原始 KAL-RFID 方法有较大的改善。

结束语 本文针对现有清洗算法的不足，根据 RFID 数据的特点，提出了基于改进卡尔曼滤波的数据清洗算法。算法引入了标签的动态属性，根据标签运动情况动态调整滑动窗口大小，并将其应用到整个卡尔曼滤波的处理过程。算法动态控制采样率，及时调整阅读器读写率，使得卡尔曼预测过程使用的样本数据及更新过程作出的最近估计值都更接近真实数据。测试表明，改进 KAL-RFID 算法能很好地应用于标签频繁移动的场景。本文提出的算法主要解决数据清洗框架中单阅读器数据清洗层的问题，下一步的工作将致力于解决多阅读器数据清洗层的数据冗余问题以及判定标签的归属问题。

参 考 文 献

[1] 王霞，玄丽娟，夏秀峰. 基于时序关系的 RFID 不确定数据清洗算法[J]. 辽宁大学学报，2012，39(2)：174-178

[2] 马岩，张延园，尹方鸣. 基于滑动窗口的 RFID 数据流多标签清洗算法[J]. 科学技术与工程，2009，9(5)：1165-1171

[3] 李晓静，谷峪，吕燕飞，等. 基于动态事件概率模型的高效 RFID 数据清洗算法[J]. 计算机研究与发展，2008，45(Suppl.)：8-12

[4] 杨梦宁，赵鹏，张小洪，等. 一种基于总线模型的数据清洗方法[J]. 计算机科学，2010，37(4)：224-226

[5] Jeffrey R, Alonso G, Franklin M, et al. A pipelined framework for on line cleaning of sensor data streams[C]// Proc of ICDE, 2006. Washington: IEEE Computer Society, 2006: 773-778

[6] Bai Yi-jian, Wang Fu-sheng, Lin Pei-ya. Efficiently Filtering RFID data streams[C]// Proceedings of the 1st International VLDB Workshop on Clean Database, 2006. Seoul: Morgan Kaufmann Publishers, 2006: 50-57

[7] Jeffery S R, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]// Proceedings of the 32nd International Conference on Very Large Data Bases, 2006. Seoul: ACM, 2006: 163-174

[8] Li Xing, Fu Wen-xiu. Efficient RFID Data Cleaning Method[J]. TELKOMNIKA, 2013, 11(3): 1707-1713

[9] Meng Ling-yong, Yu Feng-qi. RFID Data Cleaning Based on Adaptive Window[C]// Proc of the 2nd International Conference on Future Computer and Communication, 2010. Wuhan, China, IEEE, 2010: 746-749

[10] Wang Yan, Song Bao-yan. Cleaning Method of RFID Data Stream Based on Kalman Filter[J]. Journal of Chinese Computer Systems, 2011, 32(9): 1794-1799

[11] Wang Fu-sheng, Liu Shao-rong, Liu Pei-ya. Complex RFID Event Processing[J]. VLDB Journal, 2009, 18: 913-931