2012 2nd International Conference on Computer Science and Network Technology

BBS: An Effective Approach for RFID Data Cleaning

Fan Hua, Hui Zhao, Jin Zhaoyan, Tan Shuang
School of Computer
National University of Defense Technology
Changsha, China
huafan@nudt.edu.cn

***Abstract*—In RFID systems, miss reading is a common phenomenon, so accurate data cleaning become an essential task for the implementing of an RFID-based system. BBS is a behavior-based unreliable RFID data smoothing system, and it takes advantage of kinematic characteristics of tags to assist in RFID data cleaning. However, in some situations, its assumptions may not be tenable. In this paper, we present an improved behavior-based unreliable RFID data smoothing system. In order to widen the scope of this interesting method, we extended BBS to be applicable in more general conditions. And we validate our solution with a common RFID application and demonstrate the advantages of our approach through extensive simulations.**

Keywords- data cleaning; RFID technology; nonlinear smoothing

I. Introduction

RFID (Radio Frequency IDentification) technology uses radio-frequency wave to transfer data between reader and movable tagged object, and it allows objects, place, or person to be automatically identified at a distance without a direct line-of-sight, using an electromagnetic challenge/response exchange [1,2]. RFID is gaining broader adoption in many areas. It promises to revolutionize the way we track items in supply chain, retail store, and asset management applications [3].

In recent years, since Wal-Mart originated and applied RFID technology in supply chain management, RFID has been widely used in many different fields [3]. However, one of the challenges in implementing an RFID-based system is the unreliability of the data streams produced by RFID readers [4,5]. We must clean the data streams before feeding our system with such unreliable data, and related research has been an important research topic in the last few years [6].

The main task of RFID data cleaning is to eliminate the erroneous readings, especially to correct for dropped readings and provide clean RFID readings to application logic. There are several challenges in dealing with RFID data. In this paper, we focus on the challenge of dealing with miss readings in RFID applications. Chen *et al.*[8] proposed a Bayesian inference based approach, which takes full advantage of data redundancy, for cleaning RFID raw data. Rao *et al.*[9] presented a deferred approach for detecting and correcting RFID data anomalies by utilizing declarative sequenced-based rules. Gonzalez *et al.*[10] proposed a cleaning framework that takes an RFID data set and a collection of cleaning methods, with associated costs, and induces a cleaning plan that optimizes the overall accuracy adjusted cleaning costs by determining the conditions under which inexpensive methods are appropriates, and those when more expensive methods are absolutely necessary. Hua *et al.*[11] proposed a behavior-based unreliable RFID data smoothing system. It takes advantage of kinematic characteristics of tags to assist in RFID data cleaning and improves the accuracy of RFID data cleaning. However, it has strict assumptions and its assumption may not be tenable in some situations, so it may not apply to the general situation. In this paper, we proposed an improved behavior-based unreliable RFID data cleaning method and extended the BBS method to be applicable in more general conditions.

II. Relationship Between Reade Rate and Distance

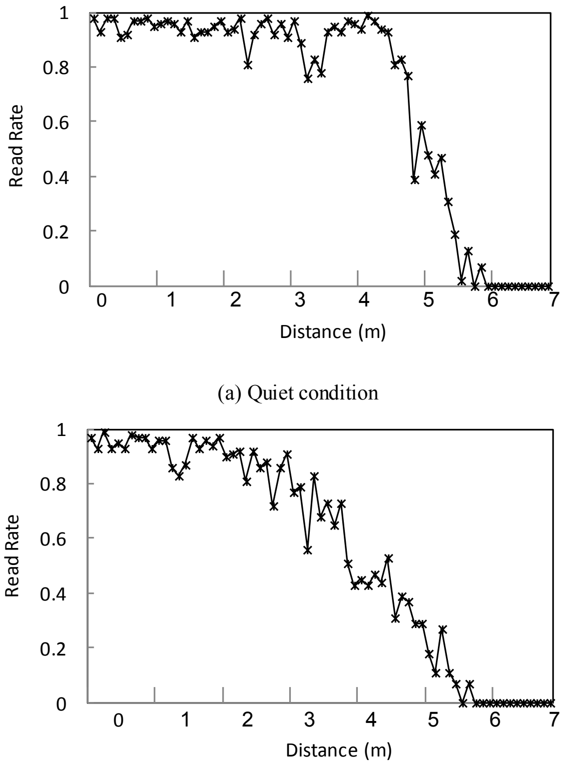
There are two distinct regions in the detection range of each reader, major detection region and minor detection region[5,12]. The major detection region have high detection probabilities (close to 100%), which is directly in front of the reader. The minor detection region, extending from the end of the major detection region to the edge of the reader's full detection range, where there is a linear relationship between read rate and distance[7]. For specific readers, the full detection range S is a constant. To confirm this conclusion, we have carried out similar experiments and the conclusion shown in Figure 1.

Epoch is the basic unit of RFID data streams. Filling the missed epoch information is the main task of RFID data cleaning. The information of RFID data stream that we get includes tag ID, the number of interrogation responses for each tag in an epoch and the time of the epoch, in the form of (*ID*, *Response number*, *time*). In our model, each epoch is viewed as an independent Bernoulli event[7]. An epoch may be specified as a number of interrogation cycles or a unit of time, and its typical range is 0.2-0.25 seconds. For each epoch, the reader keeps track of all the tags that have been identified, and additional information such as the number of interrogation responses for each tag and the last time the tag was read. Assuming, there are n interrogation cycles in an epoch, the number that tag_i is monitored is m_i. We can get read rate of tag_i at the moment by $p_i = m_i / n$.

The tag's movement in the detection region of the reader follows the law of the kinematics. And the change of kinematic

978-1-4673-2964-4/12/$31.00 ©2012 IEEE

parameters such as displacement and velocity follows the law of the kinematics, too. So we can establish the conversion relationship between RFID data and kinematic parameters of tags. The related research introduced in detail in our previous article[11].



(a) Quiet condition

(b) Noisy condition

Figure 1. Read rate of tags in different conditions

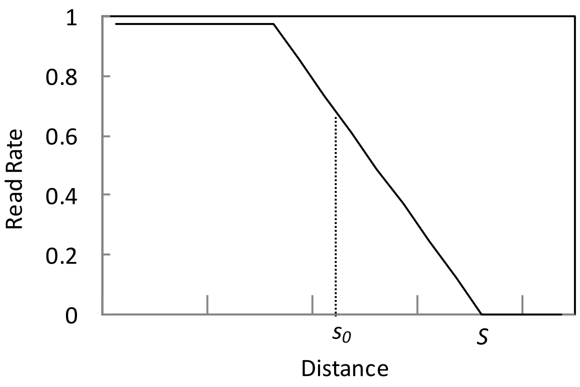


Figure 2. Relationship between read rate and distance

As shown in Figure 1, there is a linear relationship between reade rate and distance(between tag and reader), and it can be abstracted as shown in Figure 2. So, the linear relationship can be indicated as:

$$p = \begin{cases} 0 & ks+b<0 \\ ks+b & 0 \le ks+b \le 1 \\ 1 & ks+b>1 \end{cases}$$

S is a constant, so we can get $b=-kS$. k is the slope of the line, and above equation can be further written as:

$$p = \begin{cases} 0 & s>S \\ k(s-S) & 0 \le k(s-S) \le 1 \\ 1 & k(s-S)>1 \end{cases} \tag{1}$$

III. Behavior-Based Nonlinear Smoothing for Unreliable RFID Data

There are three variables (p, k and s) in Equation 1, and the detection range S is a constant. The read rate p can be calculated through *Response number*, but the distance s can't be obtained directly. In practice, the detection region of each reader is generally not very large, from a few meters to tens of meters. Therefore, the movement through the detection region for persons, vehicles and goods on the conveyor belt and other tagged items can be approximately considered as uniform linear motion. In our previous study[11], we assume the movement of the tag is uniform linear motion that through the center of the detection area, as trajectory A in Figure 3(a). However, in some situations, this assumption may not be tenable, as the trajectory B and C in Figure 3(a). Therefore, we must remove this constraint and make it applicable to more general cases. In addition, even if the direction of the trajectory C in Figure3(a) has obviously changed in this process, we can also break down it, and approximately consider each short process as uniform linear motion.

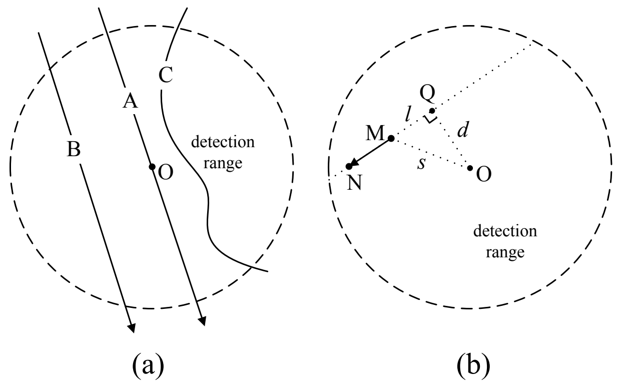


Figure 3. The trajectory of RFID tag

Equation 1 shows the relationship between the read rate p and the distance s. If we consider l is the displacement of the tag and l_0 is the original displacement of the tag, Well known, the speed v of uniform linear motion satisfies the equation $l = l_0 \pm v\Delta t$. From Figure 3(b), we can get the relationship between displacement and distance $s^2 = l^2 + d^2$. So, the Equation 1 can be further written as:

$$p = \begin{cases} 0 & kQ - B < 0 \\ kQ - B & 0 \le kQ - B \le 1 \\ 1 & kQ - B > 1 \end{cases} \qquad (2)$$

Where, $Q = \sqrt{(l_0 \pm v\Delta t)^2 + d^2}$ (It take the negative sign when the value of p increases, otherwise take the positive sign.), and $B=kS$.

The read rate that is calculated by response number will be unstable in practice, because readers are usually interfered by the surroundings including the signal reflection and obstruction or sudden current gain, etc. The results from directly treating the raw data may differ from the actual movement characteristics. The same as our previous methods[11], a weighted moving average approach is also used to smooth the initial read rate sequences.

Filling the missed epoch information is the main task of RFID data cleaning. There are two possibilities, when the read rate of the epoch is $p_i' = 0$. In the first case, the tag is indeed outside the detection range; the other is missing read, i.e., the tag is in the detection region but not captured for interference factors. It is necessary for accurate data cleaning to distinguish these two cases clearly. We should analyze its movement feature in the adjacent time. The movement of tags is approximately uniform linear motion and satisfies Equation 2, so we can calculate the read rate p_i of the moment by the values of k, l_0, v, d, S and the read rate in the adjacent time p_{ia}, to further determine it is a true value or a missing read. S is a constant. For simplicity, we assume that v, the speed of tag, is a constant too. In order to solve the coefficients k, l_0 and d, we denote $epoch_j = \{\Delta t_j, p_j\}$, where Δt_j and p_j are the time and read rate of $epoch_j$ respectively, and $TS = \{epoch_{i-l} \mid p_{i-l} \neq 0, 1 \le l \le m\}$. So the coefficients k, l_0 and d can be solved by the method of nonlinear regression on the trainings set of TS, which estimates the best-fitting line as the one that minimizes the error between the actual data and the estimate of the line.

We denote $p = f(\Delta t, k, l_0, d)$. $k^{(0)}$, $l_0^{(0)}$ and $d^{(0)}$ are the initial values of k, l_0 and d, so

$$\begin{aligned} k &= k^{(0)} + \Delta k \\ l_0 &= l_0^{(0)} + \Delta l_0 \\ d &= d^{(0)} + \Delta d \end{aligned} \qquad (3)$$

Then do the Taylor Series expassion of $f(\Delta t_i, k, l_0, d)$:

$$f(\Delta t_i, k, l_0, d) \approx f_{i0} + \frac{\partial f_{i0}}{\partial k}\Delta k + \frac{\partial f_{i0}}{\partial l_0}\Delta l_0 + \frac{\partial f_{i0}}{\partial d}\Delta d \qquad (4)$$

Where, $f_{i0} = f(t_i, k^{(0)}, l_0^{(0)}, d^{(0)})$

We can get the following equations:

$$\begin{aligned} b_{11}\Delta_k + b_{12}\Delta_{l_0} + b_{13}\Delta_d &= B_1 \\ b_{21}\Delta_k + b_{22}\Delta_{l_0} + b_{23}\Delta_d &= B_2 \\ b_{31}\Delta_k + b_{32}\Delta_{l_0} + b_{33}\Delta_d &= B_3 \end{aligned} \qquad (5)$$

Where,

$$b_{11} = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial k} \cdot \frac{\partial f_{i0}}{\partial k}$$

$$b_{12} = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial k} \cdot \frac{\partial f_{i0}}{\partial l_0}$$

$$\cdots$$

$$b_{33} = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial d} \cdot \frac{\partial f_{i0}}{\partial d}$$

$$B_1 = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial k}(p_i - f_{i0})$$

$$B_2 = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial l_0}(p_i - f_{i0})$$

$$B_3 = \sum_{i=1}^{m} \frac{\partial f_{i0}}{\partial d}(p_i - f_{i0})$$

We can get the values of Δk, Δl_0 and Δd by solving Equation 5. Further, the values of k, l_0 and d can be calculated using Equation 3. Then, the new values of k, l_0 and d are taken as the initial values again, repeating the above operations until convergence. The results are taken as the final values of k, l_0 and d.

For each tag, we firstly smooth its read rate sequence with the method of a weighted moving average of order n. Then, while the read rate of the current epoch is 0 and before is not, we calculate the new values of k, l_0 and d by the method of nonlinear regression and predict the read rate of the current epoch by Equation 2. When both the read rate of the current epoch and before is 0, we can directly predict the read rate by Equation 2.

IV. Experiment Evaluation

To evaluate our method, we have implemented a prototype of improved BBS in Java. We present an analysis of the performance of our method on several data sets and compare its accuracy with other cleaning methods. Our data include both the real collected data and simulation data. The laboratory equipments used for collecting data include Invengo XCRF-860 RFID UHF reader with 902-928MHz frequency range, Invengo XCAF-12L antenna and XCTF-8101A tag. The simulation data for our experiments were generated by a synthetic RFID data generator that simulates the operation of RFID readers under a wide variety of conditions. We simulate various movements of tags with different missing read rate, from 5% to 80%. In the experiment, we compare the error rate of data filled respectively by improved BBS, BBS and two kinds of sliding-windows methods (with 5 epoch and 35 epoch window size) under different missing read rate. As shown in Figure 4, almost in all cases the error rate of improved BBS is lower than that of other methods.

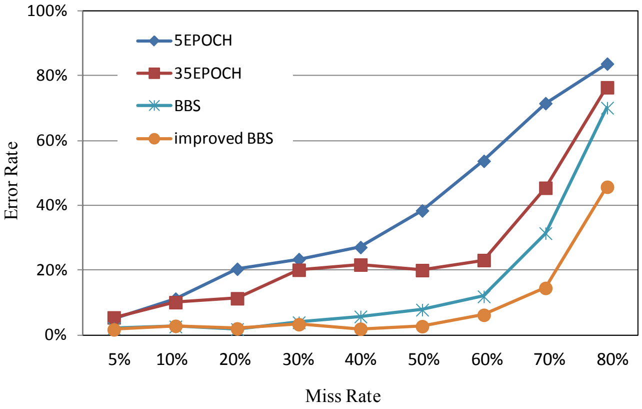


Figure 4. Accuracy compare

V. Conclusions

BBS is a behavior-based unreliable RFID data smoothing system, which can take advantage of kinematic characteristics of tags to assist in RFID data cleaning. However, in some situations, its assumptions may not be tenable. In order to widen the scope of this interesting method, we extended BBS to be applicable in more general conditions in this paper. Then, we validate our solution with a common RFID application and demonstrate the advantages of our approach through extensive simulations.

Acknowledgment

This research is supported by the National High Technology Research and Development Program of China (No. 2010AA012505, No. 2011AA010702, No. 2012AA01A401 and No. 2012AA01A402), the National Natural Science Foundation of China (No. 60933005), the Support Science and Technology Project of China (No. 2012BAH38B04) and the Information Safety Plan of China 242 (No. 2011A010).

References

[1] W. Evan, B. Leilani, C. Garret, G. Kayla, R. Kyle, R. Samuel, B. Magdalena, and B. Gaetano, "Building the Internet of Things Using RFID: The RFID Ecosystem Experience," IEEE Internet Computing, vol. 13, pp. 48-55, 2009.

[2] W. Roy, "The Magic of RFID," Queue, vol. 2, pp. 40-48, 2004.

[3] R. Derakhshan, M. E. Orlowska, X. Li, and Ieee, RFID data management: Challenges and opportunities. New York: IEEE, 2007.

[4] S. C. Sudarshan, K. Venkat, R. Sridhar, and S. Sanjay, "Managing RFID data," in Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, ed. Toronto, Canada: VLDB Endowment, 2004, pp. 1189-1195.

[5] R. J. Shawn, G. Minos, and J. F. Michael, "Adaptive cleaning for RFID data streams," in Proceedings of the 32nd international conference on Very large data bases, ed. Seoul, Korea: VLDB Endowment, 2006, pp. 163-174.

[6] Q. Sheng, S. Zeadally, Z. Luo, J.-Y. Chung, and Z. Maamar, "Ubiquitous RFID: Where are we?," Information Systems Frontiers, vol. 12, pp. 485-490, 2010.

[7] R. J. Shawn, J. F. Michael, and G. Minos, "An adaptive RFID middleware for supporting metaphysical data independence," The VLDB Journal, vol. 17, pp. 265-289, 2008.

[8] C. Haiquan, K. Wei-Shinn, W. Haixun, and S. Min-Te, "Leveraging spatio-temporal redundancy for RFID data cleansing," in Proceedings of the 2010 international conference on Management of data, ed. Indianapolis, Indiana, USA: ACM, 2010, pp. 51-62.

[9] J. Rao, S. Doraiswamy, H. Thakkar, and L. S. Colby, "A deferred cleansing method for RFID data analytics," presented at the Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea, 2006.

[10] H. Gonzalez, H. Jiawei, and S. Xuehua, "Cost-Conscious Cleaning of Massive RFID Data Sets," in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, 2007, pp. 1268-1272.

[11] H. Fan, Q. Wu, and Y. Lin, "Behavior-Based Cleaning for Unreliable RFID Data Sets," Sensors, vol. 12, pp. 10196-10207, 2012.

[12] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose, "Mapping and localization with RFID technology," in Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on, 2004, pp. 1015-1020 Vol.1.