An AM-MCMC cleaning algorithm for multi-reader RFID redundant data

Yinju Lu*, Guoquan Shan

College of Information and Engineering, Zhongzhou University, Zhengzhou, China

Received 1 March 2014, www.cmnt.lv

Abstract

The inherent characteristics of RFID device and the environmental noise cause the uncertainty of RFID raw data and, in the RFID event detection, decrease accuracy of the query results. In this paper, the recognition model of RFID reader is defined and, by using the maximum entropy method, 3-state recognition model in this recognition model has been proved to have the optimal performance. Using Bayesian principle, the posterior probability distribution of parameters to be estimated can be got from the condition likelihood observed and prior distribution of unknown parameters. Based on adaptive sampler, a Markov Chain Monte Carlo (MCMC) simulation is proposed to do data cleaning on the redundant data from RFID multi-reader. The simulation test results, carried on a large number of simulation data, verify the accuracy and efficiency of the proposed data cleaning algorithm.

Keywords: RFID, data redundancy, data cleaning, Bayesian principle, Markova chain Monte Carlo (MCMC)

1 Introduction

Radio Frequency Identification (RFID) technology is defined as a data collection technology that uses electronic tags for storing data [1]. The tag, also known as an "electronic label," "transponder" or "code plate," is made up of an RFID chip attached to an antenna. Transmitting in the kilohertz, megahertz and gigahertz ranges, tags may be battery-powered or derive their power from the RF waves coming from the reader. RFID technology does data communication between readers and electronic tags mainly by radio frequency signals, to detect the logical position of the object affixed with RFID tags. RFID technology has been used in Business application, such as the domain like supply chain management, product pipeline tracing, etc., and has developed rapidly.

Reference [2] discusses how the RFID technology is used in Taiwan logistics industry. Since June 2003, mass consuming markets had demonstrated a significant shift toward RFID technology. This has occurred not only because of RFID mandates imposed by Wal-Mart and other stores, but also the widely used of RFID by government organizations. Its use has the potential to affect an extremely wide spectrum of the population, from technology adopters to vendors, integrators, and users [3]. According to a new report RFID production is to increase 25-fold in four years, buoyed on by the scramble by pharmaceutical manufacturers to comply with the new RFID certification program, which aims to synchronize the industry's transition to RFID technology [4]. In Taiwan, government departments, the medical and pharmaceutical sectors, and private businesses have followed the RFID trend to take advantage of this new technology to enhance their standard operation processes.

Constructing an intelligent traffic monitoring system firstly depends on automatic identification for vehicles. At present, automatic identification technology based on image and vehicle license plate is going to fall in the trap due to its low recognition rate and affection by adverse weather. Thus it is necessary to apply new technologies to solve this problem, and technologies based on Internet of Things provide a new approach for it. In [5], the author explored this issue and proposed a feasible scheme. At first, they took global unique EPC code as identity identification of vehicles in stead of vehicle license plate and utilized RFID reader to read EPC code by RF electromagnetic wave, which completely solved the problem of no all-weather operations. Secondly, they obtained positioning information of vehicles by using GPS technology. Thirdly, because GPRS provides high-speed wireless IP services for mobile users, fully supports the TCP/IP, they took wireless GPRS scheme to transmit data of mobile objects. The realization of automatic detection and transmission of data provided a fundamental guarantee for constructing an intelligent traffic monitoring system. And then, they designed and discussed in turn its network architecture, data flow analysis, hardware logic structure, software flow, as well as its intelligent decision-making module. Research and design show that it is feasible and inexpensive to construct an intelligent traffic monitoring system based on Internet of Things, and the intelligent traffic monitoring system based on Internet of Things has a number of advantages such low cost, high reliability, never affected by adverse weather, all weather operations etc. Therefore, it will have a broad applying perspective.

*** Corresponding author's** e-mail:luyinju2003@163.com

In [6], an object search solution for the Internet of Things (IoT) is proposed. It first differentiates localization and searching. Localization is to calculate an object's current location. Searching is to return a set of locations where a target object could be. It is possible that the locations of the returned set are not contiguous. Searching accuracy can be improved if the number of the returned locations is small. Even though localization technique is applicable to searching applications, a simpler and easier solution will attract more enterprise users. In [6], based on a concept called location signature, defined by a set of reference tags, an object searching method named Location Signature Search (LSS) is proposed. The study of LSS shows that the searching accuracy can be very high if a location signature is not shared by too many locations. Since location signatures are affected by the deployment of the reference tags, trade-off between searching accuracy and implementation cost is achievable. A real world experiment is conducted in this research. The results show that LSS indeed is a practical method for object searching applications.

The most notable is that the world's largest retailer Wal-Mart has installed inventory management system based on RFID in its warehouse and distribution center, and it requires all the top 100 suppliers of it to install UHF RFID tags in tray in order to improve efficiency and do tracing. But, the problem often encountered in practice is that the raw data collected by RFID readers is inherently unreliable [7]. So, middleware system is responsible to correct the raw data. Currently, most methods used to clean RFID raw data are concentrated in smoothing the data read by a group of readers. These methods suffer from limitations of three aspects:

Data redundancy. The validity of many algorithms depends on the assumption that the label object is read by one and only one reader at a time. However, in real application scenarios, spatial redundancy and temporal redundancy are ubiquitous.

Dynamic association. In practical applications, many objects do not remain relatively static, i.e. having a certain dynamic. This lead to the result that the label always belongs to not a certain reader but a different reader while it shuttles between different readers, thus, showing a certain dynamic.

Priori knowledge. The priori knowledge of object affixed with electronic tag and RFID reader (such as misreading rate and deployment of RFID readers) will help us reduce the uncertainty of the data read. However, most existing algorithms do not make good use of this priori information.

In order to solve the above problems, based on real RFID application scenario, a Bayesian probability cleaning algorithm for multi-reader RFID redundant data is proposed in this paper.

By modelling effective recognition model and the effectiveness of reader recognition model, an AM-MCMC sampling algorithm is designed to do cleaning on RFID raw data effectively. During the evolutionary progress, AM-MCMC sampling algorithm can adaptively adjust the covariance matrix, thus greatly improving the convergence rate. The model proposed in this paper can take effectively advantage of Bayesian principle to obtain posterior probability distribution of the parameters to be estimated from the condition likelihood observed and the prior distribution of unknown parameters in order to improve the accuracy of the cleaning. The main contribution of this paper is as follows:

1) The probability calculation model is proposed based on Bayesian inference to obtain posterior probability distribution of the parameters to be estimated from the condition likelihood observed and the prior distribution of unknown parameters in order to infer the position of the detected object.

2) Based on the physical characteristics of the RFID reader, an RFID reader recognition model is proposed and commented by using information entropy. Further proving shows that 3-state mode has the best performance.

3) Based on adaptive sampler, Markov chain Monte Carlo (MCMC) simulation is proposed to realize the algorithm doing data cleaning on RFID multi-reader redundant data.

4) By building a real experiment platform to get real data set, using a large number of simulation data to test the algorithm, and comparing the performance of AM-MCMC with the one of MH-LC, the efficiency and effectiveness of the proposed method is proved.

2 Related work

RFID data management has gotten more and more attention which is mainly concentrated in data-centric management modelling and event-centric high effective detection. Some progress has been made in RFID data cleaning research.

In [8], Gonzalez et al. use path information to compress redundant readings in RFID data warehouse. However, this method does not apply to online real-time RFID redundant data cleaning. Reference [9] proposed a data cleaning technology based on pipeline framework, a data cleaning strategy to assure flow quality, which select different steps for different types of dirty data such as missed reading and multiple reading.

Generally, RFID complex event detection is executed over cleaned data stream. However, RFID data cleaning is always a simple process which will cost much system resources. Obviously, event detection after data cleaning will be inefficient due to twice scan of the event streams [10]. To tackle this problem, event detection is running directly over raw RFID streams and the stream is cleaned during event detection. A framework of the clean-event processing integration method is designed. Extensive experiments verify soundness and effectiveness of the proposed methods.

Although [10] can process raw data directly by SASE, this method is not based on mathematical theory, hence, mathematical model is not proposed.

SMURF algorithm [9], based on statistical sampling theory, is proposed by University of California, Berkeley, on behalf of sliding window smoothing filter technology. This algorithm adaptively adjusts window size to fill the missed data according to statistical features of flow data. However, this probability model is limited to do aggregation cleaning on many tags on given location and it does not consider the situation that there are multiple readers detecting in space.

Although [11] has discussed problems such as multiple reading, missed reading, and disorder, this method applies only to single-reader detection. This paper will deal with multiple reader problems. Reference [12] uses set theory and method based on Bayesian derivation to deal problems that existing techniques for cleaning can not accurately restore data source information (i.e. positional information). However, these two methods rely too much on the reader deployment topology.

In summary, in this paper, a new cleaning algorithm adaptive metropolis Markov Chain Monte Carlo (AM-MCMC), is proposed, which is based on Bayesian principle, can process multi-reader reading data, take the constraints into account, and does not rely much on reader deployment topology.

3 Description of the problem

3.1 SCENE ABSTRACT

Some previous RFID data cleaning methods make inference totally depending on statistical features of RFID original data set. How to find a way to fully take advantage of the priori knowledge of reader and environment and deployment topology lies on whether a RFID application scene with general features can be abstracted from the problem. This paper set logistics warehouse as background to explain the problem.

Figure 1 is a typical RFID-based logistics and warehousing scene abstract. In this scenario space model abstraction, the warehouse target range is divided into six business location regions, location 1-6, respectively, in each center of the region, equipped with an RFID reader, namely $R_1, \cdots, R_6$. The scenario has significant spatial redundancy; the spatial overlap of the reader recognition region causes duplicate reading, which an object is in the identification range of multiple readers. For example, an object is identified by the reader both in position 2 and position 3, which makes determining the exact location of the object very difficult. Since the object, at the same time, can not appear in a plurality of regions, at least one reading belongs to spatial redundancy reading.

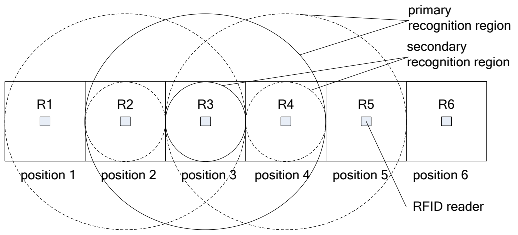

FIGURE 1 Abstract scene for redundant data

3.2 BASIC CONCEPT

Without loss of generality, suppose that there are m regions and n objects in our monitoring environment. Let O_i be the object with ID i and let h_i be position random variable of object O_i (i.e. ID of the region that O_i belongs to). For example, $h_1 = 10$ means object O_1 is in region 10. So, the possible distribution of the n objects in m regions can be expressed as random variable $\hat{H}$, and $\hat{H} = \{h_1, h_2, \cdots, h_n\}$. For the reader in region j , RFID label raw data received by it from object O_i is denoted by z_{ij} . For m regions and n objects, the raw data matrix from m readers is described as an nXm matrix $Z = [z_{ij}]$. $post(\hat{H} \mid Z)$ represents posterior probability of positional vector $\hat{H}$ in raw data Z . $post(z_{ij} \mid h_i)$ represents the value z_{ij} of object O_i reported by the reader in region j when object O_i appears in region h_i . $p(h_i)$ represents the priori probability that object O_i appears in region h_i .

The AM-MCMC cleaning algorithm for multi-reader RFID redundant data needs take full advantage of the location where these redundant readings come from and their characteristics. Some important concepts are defined as follows:

Definition 1 *RFID Data Structure*. The data structure of RFID raw data is a triple $(EPC, Reader, TimeStamp)$, which means the electronic tag with encoding EPC obtained by the reader with number $Reader$ at time $TimeStamp$. EPC is the encoding of electronic tag, $Reader$ is the number of reader, and $TimeStamp$ is timestamp.

Definition 2 *Data Element.* The reader detects and reports data in its detection range. The matrix $\vec{R}$ is used to express the raw data acquired by readers in m regions from n monitored objects. The matrix elements r_{ij} indicates whether the reader in position j has read tag O_i. $r_{ij} = 0$ means that the reader in position j has not read tag O_i ,while $r_{ij} = 1$ means that the reader in position j has read tag O_i .

Definition 3 *Reader Correlation Model.* Let R_i denote the detection range of reader i , where i is the EPC coding of reader. n readers are deployed in a space, the spatial correlation of two reader is defined as $\delta_{R_iR_j} = R_i \cap R_j / R_i \cup R_j$. When $\delta_{R_iR_j} = 0$, i.e. $R_i \cap R_j = \phi(i \neq j)$, the reader R_i and R_j is called mutually exclusive, otherwise is called compatible.

Definition 4: *Data Redundancy*. There is two types of data redundancies in RFID-related applications. One means that one marked object is recognized by multiple readers lying in its adjacent region and the other means that during a continuous time, a reader identifies an object multiple times, which happens only in the same position. For instance, z_{22} and z_{23} can not both equal to 1 at the same time because object 2 can not appear in position 2 and position 3 at the same time.

Definition 5: *Priori Knowledge*. When the priori knowledge, such as the physical characteristics of readers and tags, the reader misreading rate and deployment of readers, and the mapping between readers and business locations, is completely in conjunction with reading data, this is very valuable for data cleaning. For example, if the reader responsible for monitoring the position 3 has cross recognition region with reader responsible for monitoring the position 4, position 4 is secondary recognition region of reader R_3, and, at the same time, is primary recognition region of reader R_4.

4 Redundant data cleaning algorithms

4.1 BAYESIAN INFERENCE METHOD

Bayes theorem provides a direct method for calculating probabilities [13]. It is the foundation of Bayesian learning methods. More precisely, Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

To define Bayes theorem precisely, let us first introduce a little notation. We shall write $P(h)$ to denote the initial probability that hypothesis h holds, before we have observed the training data. $P(h)$ is often called the prior probability of h and may reflect any background knowledge we have about the chance that h is a correct hypothesis. If we have no such prior knowledge, then we might simply assign the same prior probability to each candidate hypothesis. Similarly, we will write $P(D)$ to denote the prior probability that training data D will be observed (i.e., the probability of D given no knowledge about which hypothesis holds). Next, we will write $P(D|h)$ to denote the probability of observing data D given some world in which hypothesis h holds. More generally, we write $P(x|y)$ to denote the probability of x given y . In this section, we are interested in the probability $P(h|D)$ that h holds given the observed training data D . $P(h|D)$ is called the posterior probability of h, because it reflects our confidence that h holds after we have seen the training data D . Notice the posterior probability $P(h|D)$ reflects the influence of the training data D, in contrast to the prior probability $P(h)$, which is independent of D .

Bayes theorem is the cornerstone of Bayesian learning methods because it provides a way to calculate the posterior probability $P(h|D)$, from the prior probability $P(h)$, together with $P(D)$ and $P(D|h)$.

Bayes theorem*:*

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}. \tag{1}$$

As one might intuitively expect, $P(h|D)$ increases with $P(h)$ and with $P(D|h)$ according to Bayes theorem. It is also reasonable to see that $P(h|D)$ decreases as $P(D)$ increases, because the more probable it is that D will be observed independent of h, the less evidence D provides in support of h .

In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed data D (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis. We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis. More precisely, we will say that h_{MAP} is a MAP hypothesis provided

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} = \arg\max_{h \in H} P(D|h)P(h). \tag{2}$$

Notice in the final step above we dropped the term $P(D)$ because it is a constant independent of h .

In some cases, we will assume that every hypothesis in H is equally probable a priori ($P(h_i) = P(h_j)$ for all h_i and h_j in H). In this case we can further simplify Equation (2) and need only consider the term $P(D|h)$ to find the most probable hypothesis. $P(D|h)$ is often called the likelihood of the data D given h, and any hypothesis that maximizes $P(D|h)$ is called a maximum likelihood (ML) hypothesis, h_{ML}.

$$h_{ML} \equiv \arg\max_{h \in H} P(D|h). \tag{3}$$

From above we introduced Bayes theorem by referring to the data D as training examples of some target function and referring to H as the space of candidate target functions. In fact, Bayes theorem is much more general than suggested by this discussion. It can be applied equally well to any set H of mutually exclusive propositions whose probabilities sum to one (e.g., "the sky is blue," and "the sky is not blue"). In this section, we will at times consider cases where H is a hypothesis space containing possible target functions and the data D are training examples. At other times we will consider cases where H is some other set of mutually exclusive propositions, and D is some other kind of data. In the remaining part of this section, we will show how to apply Bayes theorem as the fundament of our redundant data cleaning method.

Bayesian inference evaluates the probability of hypothesis (x) based on observed values (y). It means that the posterior probability is proportional to the product of probability and priori probability, i.e. $p(x \mid y) \infty p(y \mid x) p(x)$. By definition, Bayesian inference is described as shown in Equation (4), wherein, Z represents the assumptions of original data. So, the posterior probability of positional vector H is expressed as $post(\hat{H} \mid Z)$.

$$post(\hat{H} \mid Z) = post\left(h_1, h_2, \cdots h_n \mid \begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} \right) \infty$$
$$post\left(\begin{bmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{bmatrix} \middle| h_1, h_2, \cdots, h_n \right) \cdot p(h_1, h_2, \cdots, h_n) \quad . \qquad (4)$$

As the arbitration protocol can effectively prevent reader collision and tag collision, assuming that each reader can individually identify label of the different objects, results in Equation (5) is reached. Because of using AM-MCMC to consider the constraints, we assume that each h_i is independent from each other and recognition range of same object from each reader is independent. Then the prior distribution of each object does not depend on the one of other objects. Thus, we obtain Equation (6). Let α be a constant in Equation (6), Equation (7) is reached.

$$post(H \mid Z) \infty \prod_i p(z_{i1}, z_{i2}, \cdots, z_{in} | h_1, h_2, \cdots h_n) \cdot p(h_1, h_2, \cdots, h_n), \qquad (5)$$

$$post(\ddot{H} \mid Z) \infty \prod_i p(z_{i1}|h_i) \cdot p(z_{i2}|h_i) \cdot \ldots \cdot p(z_{im}|h_i) \cdot \prod_j p(h_j), \qquad (6)$$

$$post(\ddot{H} \mid Z) = \alpha \prod_i p(z_{i1}|h_i) \cdot p(z_{i2}|h_i) \cdot \ldots \cdot p(z_{im}|h_i) \cdot \prod_j p(h_j), \qquad (7)$$

4.2 READER RECOGNITION MODEL

Three-state recognition model is proposed, that is the reader can recognize only its own region and two adjacent regions. In Figure 1, according to this model and based on location, the reader has three location-based target regions: the main recognition region, sub-recognition region and the 0 recognition region, corresponding to the region having the same location as the reader, the region adjacent to the reader, and the region being not able to be recognized, respectively. The assessment of likelihood of 3-state model is described in Equation (8).

$$p(z_{ij} = 1 \mid h_i) = \begin{cases} r_{major} & h_i = j \\ r_{minor} & h_i \in \{j-1, j, j+1\} \\ 0 & otherwise \end{cases}, \qquad (8)$$

where r_{major} means reading rate of main recognition region of the reader, r_{minor} means reading rate of sub-recognition region of the reader, 0 means beyond the range of the reader identification.

By using 3-state recognition model, not only is duplicate reading date combined, but also it is possible to distinguish between a region and its adjacent regions of all, because they have their own different reading rates. Specifically, if the object O_i is in the area j, not only z_{ij}, but also $z_{i(j-1)}$ and $z_{i(j+1)}$ should have considerable opportunity being 1.

4.3 RECOGNITION MODEL ENTROPY ANALYSIS

Entropy is a measure commonly used in information theory, which characterizes the impurity of an arbitrary collection of examples [14]. In information theory, another commonly used statistical property, information gain, which measures how well a given attribute separates the training examples according to their target classification, can be defined from entropy.

Given a collection S, containing positive and negative examples of some target concept, the entropy of S relative to this Boolean classification is

$$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\Theta} \log_2 p_{\Theta}, \qquad (9)$$

where $p_{\oplus}$ is the proportion of positive examples in S and p_{Θ} is the proportion of negative examples in S. In all calculations involving entropy we define $0 \log 0$ to be 0. One interpretation of entropy from information theory is that it specifies the minimum number of bits of information needed to encode the classification of an arbitrary member of S (i.e., a member of S drawn at random with uniform probability). For example, if $p_{\oplus}$, is 1, the receiver knows the drawn example will be positive, so no message need be sent, and the entropy is zero. On the other hand, if $p_{\oplus}$ is 0.5, one bit is required to indicate whether the drawn example is positive or negative. If $p_{\oplus}$ is 0.8, then a collection of messages can be encoded using on average less than 1 bit per message by assigning shorter codes to collections of positive examples and longer codes to less likely negative examples.

Thus far we have discussed entropy in the special case where the target classification is Boolean. More generally, if the target attribute can take on c different values, then the entropy of S relative to this c-wise classification is defined as:

$$Entropy(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \,, \tag{10}$$

where p_i is the proportion of S belonging to class *i*. Note the logarithm is still base 2 because entropy is a measure of the expected encoding length measured in bits. Note also that if the target attribute can take on c possible values, the entropy can be as large as $\log_2 c$.

In the remaining part of this section, we will show how to apply entropy into our method to analyze the recognition model.

After invalid system state is removed using data cleaning method, the performance of a system can be measured by entropy. Let the random variable L be the true position of the object *i*, the a priori probability is assumed to be a uniform distribution, and let x be the reading rate in the primary recognition region, the reading rate in the secondary recognition region is expressed as $x/2$. Thus, according to the right side of Equation (7), the probability distribution of L is as follows:

$$p(L=l)=\begin{cases} \alpha(1-\frac{x}{2})x(1-\frac{x}{2})\beta & l=j \\ \alpha(1-\frac{x}{2})(1-x)\frac{x}{2}\beta & l\in\{j-1,j+1\} \\ 0 & otherwise \end{cases} \quad . \tag{11}$$

Of which, α represents normalization constants, β represents the priori probability of Equation (7). This gives the entropy of a probability distribution L:

$$\begin{aligned} H(L) &= -\alpha(1-\tfrac{x}{2})(1-x)\tfrac{x}{2}\beta\cdot\ln(\alpha(1-\tfrac{x}{2})(1-x)\tfrac{x}{2}\beta) - \\ &\alpha x(1-\tfrac{x}{2})^2\beta\cdot\ln(\alpha x(1-\tfrac{x}{2})^2\beta) - \\ &\alpha(1-\tfrac{x}{2})(1-x)\tfrac{x}{2}\beta\cdot\ln(\alpha(1-\tfrac{x}{2})(1-x)\tfrac{x}{2}\beta) \end{aligned} \quad . \tag{12}$$

Since the sum of all probability is 1, we can get Equation (13):

$$\alpha\beta = \frac{1}{x(1-\frac{x}{2})(2-\frac{3x}{2})} \,. \tag{13}$$

Combining Equations (12) and (13), we get:

$$H(L) = -2\cdot\frac{1-x}{4-3x}\cdot\ln\frac{1-x}{4-3x} - \frac{2-x}{4-3x}\cdot\ln\frac{2-x}{4-3x} \,. \tag{14}$$

By identifying the relationship curve between entropy and the number of states in the model, we obtain that 3-state recognition model can minimize entropy of the system, and maximize system performance. With the reading rate increases, the entropy will decrease. This indicates that the system contains more readers, while less uncertainty.

Theorem On the premise that priori knowledge and constraints are meet, the estimate of the location parameter obtained by 3-state recognition model can make the system performance better than the estimate of the location parameter obtained by other state recognition model can.

Proof: Because sum of probability of all $2n-3$ regions of *n*-state model is 1, that is:

$$\prod_{i=1}^{n-2}\left(1-\frac{ix}{n-1}\right)2x + 2\prod_{i=1}^{n-2}\left(1-\frac{ix}{n-1}\right)2(1-x)\sum_{k=1}^{n-2}\frac{\frac{k}{n-1}x}{1-\frac{k}{n-1}x} = 1 \,. \tag{15}$$

By Equations (11) and (15), we obtain the entropy of *n*-state model:

$$\begin{aligned} H(L) &= \sum p_i \ln\frac{1}{p_i} = \\ &\frac{-x}{x+2(1-x)\sum_{j=1}^{n-2}\frac{jx}{n-1-jx}} \ln \frac{x}{x+2(1-x)\sum_{j=1}^{n-2}\frac{jx}{n-1-jx}} \\ &-2\sum_{k=1}^{n-2}\left(\frac{(1-x)\cdot\frac{kx}{n-1-kx}}{x+2(n-1)\sum_{j=1}^{n-2}\frac{jx}{n-1-jx}} \ln \frac{(1-x)\cdot\frac{kx}{n-1-kx}}{x+2(n-1)\sum_{j=1}^{n-2}\frac{jx}{n-1-jx}}\right) \end{aligned} \quad . \tag{16}$$

Let $f(x,n) = x + 2(1-x)\sum_{j=1}^{n-2}\frac{jx}{n-1-jx}$ and $g_k(x,n) = \frac{kx}{n-1-kx}(k=1,2,\cdots,n-2)$, then entropy function is:

$$\begin{aligned} H(L) &= -\frac{x}{f(x,n)}\cdot\ln\frac{x}{f(x,n)} \\ &-2\left[\sum_{k=1}^{n-2}\frac{(1-x)\cdot g_k(x,n)}{f(x,n)}\cdot\ln\frac{(1-x)\cdot g_k(x,n)}{f(x,n)}\right](n\geq 3) \end{aligned} \,. \tag{17}$$

The entropy function $H(L)$ of *n*-state model is incremented with n. Suppose $x=0.95$, if $n=2$ then $H(L)|_{n=2}=1.098$, and if $n=3$ then $H(L)|_{n=3}=0.395$. Therefore, if and only if $n=3$, the entropy is minimum, and the accuracy of parameter estimation is highest.

QED.

4.4 AM-MCMC ALGORITHM

By constructing Markov process, which satisfying conditions of ergodicity, normalization and stationary, MCMC method [15] can obtain a non-periodic irreducible Markov chain, whose stationary distribution and limit distribution is probability distribution the target. When a Markov chain converges after a long enough warm-up period, its sample approximate the one of target probability distribution, which can be used to estimate the target distribution.

The key of MCMC is how to choose the recommended distribution (transfer density) to make sampling more efficient. Commonly used sampling algorithms include Metropolis-Hastings algorithm, Gibbs sampling [16], and Adaptive Metropolis(AM) algorithm [17], etc. AM (Adaptive Metropolis) algorithm is an improved MCMC sampler proposed by Haafio in 2001. Compared with traditional MH and Gibbs sampling, AM no longer needs to determine the recommended distribution of variables in advance, however, is based on the covariance of the initial sample. The recommended distribution is defined as a multidimensional normal form of the parameter space; the initial covariance can be determined by priori information. In the sampling process, the recommended density (i.e., the covariance matrix) is adjusted adaptively based on the historical sampling information of Markov chain, and parallel computing can be adopted to improve the convergence speed. In this paper, this algorithm is adopted to do sampling.

Let H be random vector of object position, whose posterior distribution is denoted by $post(\hat{H}\,|\,Z)$ and assuming that $\hat{H}_{t-1}$ is directly leading state of state $\hat{H}_t$ in Markov chain. First, by AM-MCMC algorithm, the proposal sample $\hat{H}_q$ is from proposal distribution $q(\hat{H}_q\,|\,\hat{H}_{t-1})$, that is, $\hat{H}_q$ is the random deviations of $\hat{H}_{t-1}$. This paper uses the uniform proposal distribution and the proposal sample $\hat{H}^{'}$ is expressed as $\hat{H}_{t-1}+\hat{H}_q$. Then, AM-MCMC treats $\hat{H}^{'}$ as next state $\hat{H}_t$ with the probability $post(\hat{H}^{'}\,|\,Z)/\,post(\hat{H}_{t-1}\,|\,Z)$.

Reference [17] discusses the adaptive metropolis algorithm. Suppose that at time t -1 we have sampled the states $X_0, X_1, \ldots, X_{t-1}$, where X_0 is the initial state. Then a candidate point Y is sampled from the (asymptotically symmetric) proposal distribution $q_t(\cdot\,|\,X_0, X_1, \ldots, X_{t-1})$, which now may depend on the whole history $(X_0, X_1, \ldots, X_{t-1})$. The candidate point Y is accepted with probability:

$$\alpha(X_{t-1}, Y) = \min\left\{1, \frac{\pi(Y)}{\pi(X_{t-1})}\right\}, \tag{18}$$

where we set $X_t = Y$, and otherwise $X_t = X_{t-1}$. Observe that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm. However, here the choice for the acceptance probability is not based on symmetry (reversibility) conditions since these cannot be satisfied in our case-the corresponding stochastic chain is no longer Markovian. For this reason we have to study the exactness of the simulation separately.

The proposal distribution $q_t(\cdot\,|\,X_0, X_1, \ldots, X_{t-1})$ employed in the AM algorithm is a Gaussian distribution with mean at the current point X_{t-1} and covariance $C_t = C_t(X_0, \ldots, X_{t-1})$. Note that in the simulation only jumps into S are accepted since we assume that the target distribution vanishes outside S.

The crucial thing regarding the adaptation is how the covariance of the proposal distribution depends on the history of the chain. In the AM algorithm this is solved by setting $C_t = s_d \operatorname{cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d$ after an initial period, where s_d is a parameter that depends only on dimension d and $\varepsilon > 0$ is a constant that we may choose very small compared to the size of S. Here I_d denotes the d-dimensional identity matrix. In order to start, we select an arbitrary, strictly positive definite, initial covariance C_0, according to our best prior knowledge (which may be quite poor). We select an index $t_0 > 0$ for the length of an initial period and define:

$$C_t = \begin{cases} C_0, & t \le t_0, \\ s_d \operatorname{cov}(X_0, \ldots, X_{t-1}) + s_d \varepsilon I_d, & t > t_0, \end{cases}. \tag{19}$$

The covariance C_t may be viewed as a function of t variables from $\Re^d$ having values in uniformly positive definite matrices.

Recall the definition of the empirical covariance matrix determined by points $x_0, \ldots, x_k \in \Re^d$:

$$\operatorname{cov}(x_0, \ldots, x_k) = \frac{1}{k}\left(\sum_{i=0}^{k} x_i x_i^T - (k+1)\bar{x}_k \bar{x}_k^T\right), \tag{20}$$

where $\bar{x}_k = (1/(k+1))\sum_{i=0}^{k} x_i$ and the elements $x_i \in \Re^d$ are considered as column vectors. So one obtains that in definition (1) in [17] for $t > t_0 + 1$ the covariance C_t satisfies the recursion formula

$$\begin{aligned} &C_{t+1} = \frac{t-1}{t} C_t + \\ &\frac{s_d}{t}\left(t\bar{X}_{t-1}\bar{X}_{t-1}^T - (t+1)\bar{X}_t \bar{X}_t^T + X_t X_t^T + \varepsilon I_d\right). \end{aligned} \tag{21}$$

The choice for the length of the initial segment $t_0 > 0$ is free, but the bigger it is chosen the more slowly the effect of the adaptation is felt. In a sense the size of t_0 reflects our trust in the initial covariance C_0. The role of the parameter ε is just to ensure that C_t will not become singular. As a basic choice for the scaling parameter we have adopted the value $s_d = (2.4)^2 / d$ from [18], where it was shown that in a certain sense this choice optimizes the mixing properties of the Metropolis search in the case of Gaussian targets and Gaussian proposals.

Original AM-MCMC algorithm evaluates the posterior distribution by Markov chain formed in the sampling space; however, it does not take into account the constraints. For integrating constraints while sampling, Adaptive Metropolis sampler (AM-MCMC) with position constraints is proposed. For AM-MCMC, each location region is associated with multiple variables known as resource descriptors, denoted by $Descriptor_i$, which means the number of available capacity that location region i associated with. Volume variable of object j is denoted by $Volume_j$. As long as $Descriptor_i$ is not less than 0, the proposed resource allocation is feasible. Otherwise, we have to re-sampling until a new location meets all constraints. Thus, whether a position is feasible is summarized as monitoring the value of each descriptor. The relationship between them is described as Equation (22).

$$Descriptor_i = Descriptor_i - Volume_j . \tag{22}$$

For AM-MCMC, the proposal sample is iteratively generated according to the number of dimensions. If the current allocation descriptor is less than 0, for the current deviation sample, the sample is judged to be unqualified and then abandoned. And then other value for the number of this dimension is chosen by resampling. As to proposal distribution, a random walk (Random walk) chain is constructed by selecting a uniform proposal distribution in the range of step length.

From the foregoing, the sampling mechanism of AM algorithm depends on all the historical sample information $X_0, X_1, ..., X_{t-1}$. Haario etc. prove the convergence and ergodicity of the algorithm. The specific sampling procedure of AM algorithm is as follows:

1) Initialization, $i = 0$;

2) According to the constraint by Equation(22), the initial state X_i is randomly generated and accepted;

a) Calculated covariance C_i using Equation (19);

b) Recommended variable $X^* \sim N(X_i, C_i)$ is generated;

c) According to Formula (18), calculates and accepts $\alpha(X_t, X_{t-1}) = \min\left\{1, \dfrac{\pi(X_{t-1})}{\pi(X_t)}\right\}$;

d) Produce a uniform random number $u \sim U(0,1)$;

e) If $u < \alpha$, then accept $X_{i+1} = X^*$, else $X_{i+1} = X_i$;

3) $i = i+1$,repeat 1)-5) until the number of samples meets the requirement preset in advance.

4.5 AM-MCMC CONVERGENCE RULE

An important task of MCMC sampling study is to determine whether a parallel sampling sequence converges to the posterior distribution. In theory, AM-MCMC algorithm will surely converge when $t \to \infty$. However, in practical application, we must determine the number of sampling needed by AM algorithm to converge to stable posterior distribution, that is, the convergence determination conditions are given. The convergence diagnostics is an important part of AM-MCMC sampling methods. References [19] proposed scale reduction factor to determine the convergence of multiple sequence. Calculated as:

$$\sqrt{R} = \sqrt{\frac{i-1}{i} + \frac{k+1}{k \cdot i}\frac{B}{W}} , \tag{23}$$

$$B/i = \sum_{j=1}^{k}(u_j - u)^2/(k-1) , \tag{24}$$

$$W = \sum_{j=1}^{k} s_j / k , \tag{25}$$

where, i is the evolution number of each Markov chain, B/i is the variance of parameters sample mean μ_j in the Markov chain, W is the mean of parameter sample variance s_j in the Markov chain, and $\bar{\mu}$ is the mean of μ_j. Under normal circumstances, the scale reduction factor close to 1 indicates that the algorithm get to convergence. However, in practical application, that the scale reduction factor of the evolutionary sequence closes to 1 is more difficult to achieve. Reference [19] propose to take $\sqrt{R} < 1.2$ to determine whether multiple sequence sampling algorithms converge.

5 Experiments

5.1 EXPERIMENTAL ENVIRONMENT AND DATA SET

Basic laboratory equipment include Invengo's XCRF-860 RFID UHF Reader supporting EPC Gen2 protocol/ISO18000-6C and Inlay XC-TF8029-C07. Experimental environment is Visual Studio 2012, running on Pentium Core i7 CPU of 3.4GHZ, 8GB RAM, 2TB hard drive as well as Window 7 operating system.

Simulation experiments randomly generate distribution matrix with real distribution effect matrix by real matrix generator. Noise matrix generator provides noise matrix similar to RFID raw data according to the same format. AM-MCMC and MH-LC module reconstruct distribution of each instance using the input noise matrix. Simulator generates synthetic RFID raw data with duplicate reading according to the physical characteristics of the RFID reader. The main parameters used in the experiments are shown below.

The goals of the experiment are described below:

1) Evaluate sampling efficiency of AM-MCMC and MH-LC by the calculation of reconstruction time.

2) Use the K-L divergence to evaluate sampling accuracy of AM-MCMC and MH-LC, respectively.

3) Use artificial noise to evaluate the performance of AM-MCMC and MH-LC, respectively.

5.2 EXPERIMANTAL RESULTS AND ANALYSIS

5.2.1 Reconstruction efficiency

AM-MCMC and MH-LC performance is verified in this experiment. Compared with the MH-LC, average sampling time of AM-MCMC reduce significantly with the increase of the number of qualified samples, as shown in Figure 2 For example, for 5000 qualified samples, AM-MCMC sampling spends 14.01 seconds, while the MH-LC sampling time is 200.18 seconds. This is because the AM-MCMC takes advantage of the current qualified sample to generate next qualified sample. Therefore, AM-MCMC spends less time than the MH-LC does to generate the same number of qualified samples.

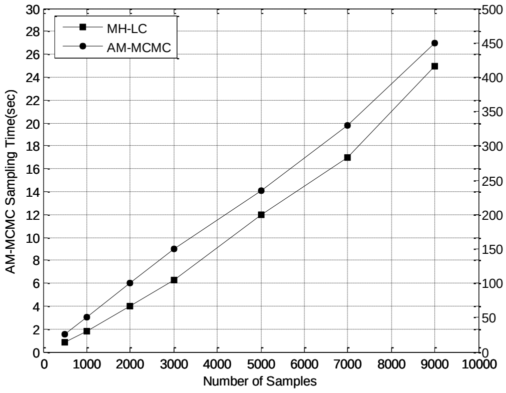

FIGURE 2 AM-MCMC versus MH-LC on sampling time

5.2.2 Reconstruction accuracy

In this experiment, a different number of qualified samples are taken to study how different number of samples affects reconstruction accuracy. First, we increase the number of qualified samples from 500 to 9000 to study how the AM-MCMC and MH-LC perform at the respect of construction accuracy, respectively. Here, the reading rate in the main recognition range is assumed to be 96%. As shown in Figure 3, as the number of qualified samples increases, K-L divergence values of the two methods are all remain reduced. However, the accuracy of AM-MCMC is always higher than MH-LC's. Especially, when we have drew 500 qualified samples, the K-L divergence value of AM-MCMC is 1.52 while the one of the MH-LC is 3.75. When we have picked up the 9000 qualified samples, the K-L divergence of AM-MCMC significantly reduces to 0.51, while the one of MH-LC is 2.52.

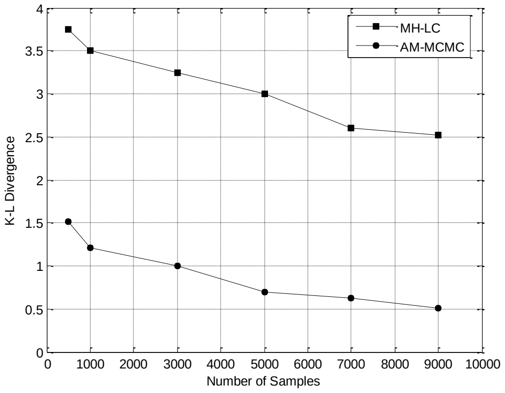

FIGURE 3 The impact of the number of qualified samples

5.2.3 How sample redundancy effects cleaning accuracy

In this experiment, different redundancies of qualified samples are taken to study how different redundancy of samples affects reconstruction accuracy. First, we increase the redundancy of qualified samples from 0.325 to 0.5 to study how the AM-MCMC and MH-LC perform at the respect of sample redundancy, respectively. As shown in Figure 4, as the redundancy of qualified samples increases, K-L divergence values of the two methods are all keep decreasing. However, the accuracy of AM-MCMC is always higher than MH-LC's. Especially, when the redundancy of qualified samples is 0.325, the K-L divergence value of AM-MCMC is 3.21 while the one of the MH-LC is 4.40. When the redundancy of qualified samples is 0.5 the K-L divergence of AM-MCMC drops to 0.87, while the one of MH-LC is 2.41.

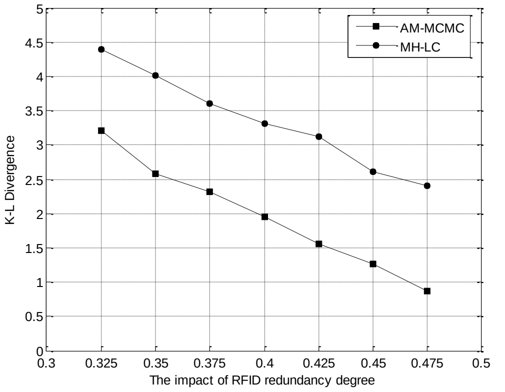

FIGURE 4 How sample redundancy effects cleaning accuracy

6 Conclusions

RFID technology has broad application prospects, but also has raised new challenges on data management. Due to RFID inherent characteristics, data cleaning problem is an important issue to consider in RFID data management. In this paper, for the uncertainty of RFID raw data, by defining the recognition model of RFID reader, adopting Bayesian principle to get the posterior probability distribution of parameters to be estimated from the condition likelihood observed and prior distribution of unknown parameters, Bayesian probability cleaning algorithm is designed to do data cleaning on the redundant data from RFID multi-reader based on adaptive sampler. The simulation test results, carried on a large number of simulation data, verify the accuracy and efficiency of the proposed data cleaning algorithm.

Acknowledgements

This work is supported by the Henan province science and technology projects (2102210247).

References

[1] Expert Barcode & RFID Inc 2009 Retrieved Oct 19, 2014 from http://www.expertbarcode.com/autotech.htm

[2] Wang Y B, Lin K Y, Chang L, Hung J C 2011 *Journal of Computers* **6**(3) 441-8 *(in Chinese)*

[3] Lahiri S 2006 *RFID Sourcebook* New Jersey:International Business Machines Press

[4] Chu W L 2006 *RFID production to increase 25-fold by 2010* Retrieved Oct 12, 2014 from http://www.labtechnologist.com/Industry-Drivers/RFID-production-to-increase-25-fold-by-2010

[5] Xiao L S, Wang Z X 2011 *Journal of Networks* **6**(6) 887-94 *(in Chinese)*

[6] Jwo J S, Chen T C, Tu M R 2012 *Journal of Software* **7**(12) 2886-93

[7] Jeffery S R, Garofalakis M N, Franklin M M 2006 *Proceedings of Vary Large Data Bases VLDB* Seoul Korea **32** 163-74

[8] Gonzalez H, Han J, Shen X 2007 *Proceedings of International Conference on Data Engineering ICDE* Istanbul Turkey 1268-72

[9] Jeffery R, Alonso G, Franklin M, Hong W H 2006 *Proceedings of International Conference on Data Engineering ICDE 2006* Atlanta Georgia USA 773-8

[10] Li Q, Chen L 2012 *Computer engineering and design* **33**(4) 1613-16 *(in Chinese)*

[11] Bai Y, Wang F S, Liu P Y 2006 *Processings of the 1st Int VLDB Workshop on Clean Databases* Seoul: Morgan Kaufmann 50-7

[12] Chen H Q, Ku W S, Wang H X 2010 *Processings of Special Interest Group on Management of Data* Indiana USA 51-62

[13] Mitchell T M 1997 Machine Learning *McGraw-Hill Science Press Chapter 6*

[14] Mitchell T M 1997 Machine Learning *McGraw-Hill Science Press* Chapter 3

[15] Xiao K, Su M C, Guo S J 2009 *Journal of North China University of Technology* **21**(3) 4-8 *(in Chinese)*

[16] Jin Y, Li J, Huang J G 2009 *Systems Engineering and Electronics* **31**(12) 2809-12 *(in Chinese)*

[17] Haar1o H, Saksman E, Tamminen J 2001 *Bernoulli* **7**(2) 223-42

[18] Gelman A G, Roberts G O, Gilks W R 1996 Efficient Metropoli jumping rules *Bayesian Statistics V* 599-608

[19] Gelman A, Rubin D B 1992 Inferenee from iterative simulation using multiple sequences *Statistical Science* **7**(4) 457-72

Authors

Yinju Lu, May 1976, Xinyi, Jiangsu Province, P.R. China.

Current position, grades: associate professor at Zhongzhou University.
University studies: MS in computer software and theory at China University of Mining & Technology in China.
Scientific interests: uncertain database, data mining, computer networks.
Publications: 15.
Experience: teaching experience of 9 years, 6 scientific research projects.

Guoquan Shan, July 1975, Hebi, Henan Province, P.R. China.

Current position, grades: associate professor at Zhongzhou University.
University studies: BS and MS at North China University of Water Resources and Electric Power in China, respectively.
Scientific interests: database theory, uncertain database, data mining, computer networks.
Publications: 7.
Experience: teaching experience of 8 years, 3 scientific research projects.