LETTER

Efficient RFID Data Cleaning in Supply Chain Management

Hua FAN[†a)], ***Student Member***, **Quanyuan WU**[†], ***and*** **Jianfeng ZHANG**[†], ***Nonmembers***

SUMMARY Despite the improvement of the accuracy of RFID readers, there are still erroneous readings such as missed reads and ghost reads. In this letter, we propose two effective models, a Bayesian inference-based decision model and a path-based detection model, to increase the accuracy of RFID data cleaning in RFID based supply chain management. In addition, the maximum entropy model is introduced for determining the value of sliding window size. Experiment results validate the performance of the proposed method and show that it is able to clean raw RFID data with a higher accuracy.

key words: *data cleaning, RFID technology, Bayesian inference, maximum entropy model, supply chain management*

1. Introduction

With the development of low-cost passive RFID (Radio Frequency Identification) tags and vigorous RFID standardization efforts, RFID technology has become an indispensable technology in the modern supply chain management and logistics industry [1], [2]. Meanwhile, RFID technology is still facing a lot of technical challenges in actual applications. The special way in which RFID device gets data brings more uncertainty to the raw RFID data sets, particularly the phenomenon of miss read and ghost read. For example, the simultaneous reading of a large number of tags, transient exhaustion of power for the instability of the current signal and tag blocked by metal or other materials will make the tag data inaccessible to the reader. So it will inevitably lead to inaccuracy of the query results, if there is no effective cleaning and pretreatment to the raw RFID data.

Yu et al. [3] proposed a series of models and algorithms about association degree maintenance and data cleaning by analyzing the group changes based on the defined association degree and dynamic clusters. Jeffery et al. [4] proposed an adaptive smoothing filter SMURF for RFID data cleaning. Chen et al. [5] proposed a Bayesian inference based approach, which takes full advantage of data redundancy, for cleaning RFID raw data. Fan et al. [6] proposed a behavior based RFID data smoothing approach, but it is only appropriate for a scenario with a single reader.

According to the problem in RFID technology based applications, we present a Bayesian inference based approach for cleaning RFID raw data in this letter. And we take full advantage of the moving path of tags in order to recover the true information as much as possible.

Manuscript received January 15, 2013.
Manuscript revised March 13, 2013.

[†]The authors are with the School of Computer, National University of Defense Technology, Changsha, 410073 China.
a) E-mail: huafan@nudt.edu.cn

DOI: 10.1587/transinf.E96.D.1557

2. Problem Statement

In the RFID based supply chain management system, when the product with RFID tag moves through or stays in the detection region of the reader, it will be detected by the reader and a record will be generated in the form of (*tag_id*, *reader_id*, *timestamp*), where *tag_id* and *reader_id* refer to EPCs which universally unique identify the tagged item and the RFID reader (as well as its location), and the *timestamp* is the time when the reading occurred [7].

The readers are installed at every entrance and exit of all warehouses in supply chain, including retailers, manufacturers, wholesalers, commodity distribution centers, and so on. The readers at the entrance and exit of warehouse are used to detect the incoming and outgoing of tags respectively, and the readers in the warehouse are used to periodically monitor the tag state of existence.

In addition, in order to facilitate supply chain management, warehouses usually have several exits, and the products which are sent to different destinations will be arranged to corresponding exits (the destinations and the exits are in one-to-one correspondence). In Fig. 1, the nodes *A*, *B*, *C*, *D* and E mean several warehouses in supply chain. As for warehouse *B*, there are one entrance *a* and three exits *c*, *d* and *e* corresponding to three destination *C*, *D* and *E* respectively.

RFID readers provide a mean to observe the products in supply chain, but the observation may be incomplete since some products may not respond to reader queries for technological reasons. There is a RFID reader at every entrance and exit, and the miss reading likely occurs to readers at all locations. In this letter, we mainly concern the miss reading which occurs to the readers at the exits of warehouses. The readers at the entrance and inside of warehouses may also have miss reading, but it is easily resolved. We will not discuss these cases in this letter due to the space limitations.

If a tag was miss read when it leaves the Warehouse, the item information about its next destination would not be

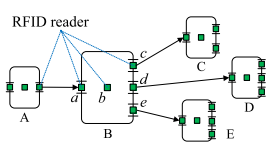

Fig. 1 Detection region of readers in warehouse.

real-time obtained by users until they reach the next destination. However, in many real-time logistics tracking applications, it is very important to users. In fact, in the other RFID based applications related to the path tracking, we will also face similar problems, such as the exhibition hall, park real-time monitoring system, intelligent transportation systems, and so on.

3. RFID Data Cleaning

3.1 Bayesian Inference Based Decision Model

Bayesian inference is a method of inference in which Bayes' rule is used to update the probability estimate for a hypothesis as additional evidence is learned, which can be represented as $P(h \mid e) \propto P(e \mid h)P(h)$.

Suppose there are m exits in a warehouse W, each exit with a reader to detect the next destination of tags. We use a random vector $H_{iW} = (h_{i1}, h_{i2}, \ldots, h_{im})$ to represent the exit through which tag T_i leaves W, the vector element h_{ij} is a Boolean variable. There is one and only one element whose value is 1 in H_{iW}, and the other elements are all 0. When $h_{ij} = 1$, it means that tag T_i left W from exit E_j and H_{iW} can be denoted as H_{iW_j}. For example, the vector $H_{iW\_3} = (0, 0, 1, 0)$ not only shows that W has four exits, but also shows that tag T_i was transported to its next destination through exit E_3 of W. We use random vector $D_{iW} = (d_{i1}, d_{i2}, \ldots, d_{im})$ to represent the raw data from the readers at all exits of warehouse W. If the reader at exit E_j of W has detected the tag T_i, the corresponding variable d_{ij} will be set to 1, otherwise it is 0. Therefore, in general, there is only one element with the value of 1 in the raw data vector D_{iW} at most. However, if the tag T_i has been miss read when it went through the exit of the warehouse W, the value of raw data vector would be $D_{iW} = (0, 0, \ldots, 0)$. In this case, the Bayesian inference can be represented as:

$$P(H_{iW\_j} \mid D_{iW}) \propto P(D_{iW} \mid H_{iW\_j})P(H_{iW\_j}) \tag{1}$$

And the main purpose of this paper is just to show how to fill up the incomplete data when the tag T_i has left the warehouse W but $D_{iW} = (0, 0, \ldots, 0)$. As a matter of convenience, when $D_{iW} = (0, 0, \ldots, 0)$, we denote D_{iW} by D_{iW0}. Now Eq. (1) can be further represented as:

$$P(H_{iW\_j} \mid D_{iW0}) \propto P(D_{iW0} \mid H_{iW\_j})P(H_{iW\_j}) \tag{2}$$

In the previous paragraph, we have introduced that the different values of H_{iW_j} means that the tag T_i has left the warehouse W through different exit. Consequently, we just need to calculate the values of $P(H_{iW\_j} \mid D_{iW0})$ corresponding to each H_{iW_j} by Eq. (2), and we can determine the most likely next destination of the tag through the further comparison of each $P(H_{iW\_j} \mid D_{iW0})$. Before we can derive the posterior of each sample based on Eq. (2), we must derive $P(D_{iW0} \mid H_{iW\_j})$ and $P(H_{iW\_j})$ first. So the major difficulty in computing the posterior of each sample lies in how to accurately estimate the probability $P(D_{iW0} \mid H_{iW\_j})$ and $P(H_{iW\_j})$. Essentially, $P(D_{iW0} \mid H_{iW\_j})$ means the miss read probability of the tag T_i which goes through the exit E_j. And it can be estimated by Eq. (3), where M_j is the number of tags that have been miss read at exit E_j and O_j is the total number of tags that go through the exit E_j. But the calculation of $P(H_{iW\_j})$ is more complex, it will be discussed in the next section.

$$P(D_{iW0} \mid H_{iW\_j}) = M_j / O_j \qquad (1 \le j \le m) \tag{3}$$

3.2 Path-Based Detection Model

According to Eq. (2), the value of $P(H_{iW\_j})$ and $P(D_{iW0} \mid H_{iW\_j})$ can decide the value of $P(H_{iW\_j} \mid D_{iW0})$, which will directly affect the choice of the final result.

To understand intuitively, $P(H_{iW\_j})$ is the probability of tag T_i which exits from warehouse through exit E_j. Therefore, there is a simple way to calculate it:

$$\hat{P}(H_{iW\_j}) = C_j \Big/ \sum_{j=1}^{m} C_j \qquad (1 \le j \le m) \tag{4}$$

Where, C_j is the tag number that exit from E_j. However, the experiment illustrates that the result got by the Eq. (4) is not accurate enough. In order to further improve the accuracy of data cleaning, we have proposed a path-based detection model.

Suppose the graph in Fig. 2 is the connection relationship between warehouses in a supply chain. It can be modeled as a directed acyclic graph for the products in a supply chain that are normally not transported to a specific position twice and the transportation of products is unidirectional from manufacturers to retailers. Each node in it represents a warehouse in supply chain, and the directed edges represent the transport routes between warehouses. Therefore, the trajectory of each tag will correspond to a path in the graph. Table 1 shows a set of raw RFID data produced by the readers in the graph and Table 2 shows the corresponding trace records of some tags that generated from the data

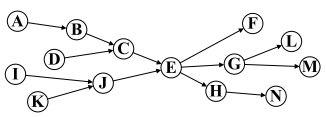

Fig. 2 An example of connection relationship in supply chain.

Table 1 A set of raw RFID data.

Raw data (*tag_id*, *reader_id*, *timestamp*)
(T1, A1, 1) (T3, K1, 1) (T10, K1, 1) (T4, I1, 2) (T1, B0, 2) (T2, I1, 2) (T5, I1, 2) (T8, D1, 3) (T6, D1, 3) (T3, J0, 3) (T2, J0, 3) (T10, J0, 3) (T5, J0, 3) (T9, D1, 3) (T7, D1, 3) (T4, J0, 3) (T9, C0, 4) (T5, J1, 4) (T8, C0, 4) (T1, B1, 4) (T3, J1, 4) (T10, J1, 4) (T4, J1, 4) (T6, C0, 4) (T7, C0, 4) (T2, J1, 4) (T1, C0, 5) (T7, C1, 6) (T1, C1, 6) (T8, C1, 6) (T9, C1, 6) (T6, C1, 6) (T1, E0, 8) (T8, E0, 8) (T6, E0, 8) (T9, E0, 8) (T7, E0, 8) (T10, E0, 9) (T9, E3, 9) (T4, E0, 9) (T2, E0, 9) (T3, E0, 9) (T7, E3, 9) (T5, E0, 9) (T4, E1, 10) (T5, E1, 10) (T1, E1, 10) (T8, E1, 10) (T6, E1, 10) (T2, E1, 10) (T6, F0, 11) (T8, F0, 11) (T1, F0, 11) (T9, H0, 11) (T7, H0, 11) (T5, F0, 11) (T10, E2, 11) (T3, E2, 11) (T2, F0, 11) (T4, F0, 11) (T10, G0, 12) (T3, G0, 12) (T9, H1, 13) (T7, H1, 13) (T10, G1, 13) (T10, M0, 14) (T3, G1, 14) (T3, L0, 15) (T9, N0, 16) (T7, N0, 16)

Table 2 Path information.

Path	Count	Tags
A->B->C->E->F	1	T1
D->C->E->F	2	T6, T8
D->C->E->H->N	2	T7, T9
I->J->E->F	3	T2, T4, T5
K->J->E->G->L	1	T3
K->J->E->G->M	1	T1

in Table 1.

Assume that the tag T_i has been miss read at the exit of E. We can get the results as shown in Fig. 3 (a), if we calculate $P(H_{iW\_j})$ using Eq. (4) and only consider the data produced at E in Table 1. In practice, however, there may be some potential relationship between different nodes in the graph, and we could get some valuable information from it. For example, in Fig. 2, node E may be a large clothing distribution center, K is the manufacturer of a certain brand of clothing and G is the designated agent of this brand of clothing, but F and H are the sales agents of other brands. So, all the clothes that are transported to the distribution center E and produced in K will be then transported to agent G, but F or H. So making full use of this potential information can greatly improve the accuracy of $P(H_{iW\_j})$, which can further improve the cleaning accuracy of the RFID data. The mining and discovery of such the prior knowledge need introduce a large number of product attribute data and huge computation overhead, which will not be tolerated. To this end, we introduce a compromise solution, mining path information of the tags instead of the analysis of the attribute relationship between tags and the destination, which needs not to introduce mining tag attribute data, so it can improve the accuracy of the results without bring huge computational overhead.

Definition 3.2.1 (Path). Suppose n is a natural number, $v_0, v_1, \ldots, v_n$ are nodes in a directed acyclic gragh G_0, and there is an edge e_i between all adjacent nodes v_{i-1} and v_i, then the sequence $v_0 v_1 \ldots v_n$ is called the path from v_0 to v_n in the gragh G_0 and it is denoted as *path*. The number n is the length of *path*, denoted by $Length(path) = n$.

We introduce a path-based detection model to calculate the probability of tags from different paths respectively by taking prior knowledge into account. We denote the tag number that had passed through a certain path *path* as C_{path}. So, $P(H_{iW\_j})$ can be calculated by:

$$P(H_{iW\_j}) = C_{path\_Ej} \Big/ \sum_{l=1}^{m} C_{path\_El} \tag{5}$$

Where, *path_Ej* is the path combined by *path* and the next destination corresponding to the exit *Ej*, and $Length(path\_Ej) = Length(path) + 1$.

For example, a tag T_i which had passed through the path *K->J->E* have been miss read when it leaves the node E. According to the data provided in Table 1, if we calculate $P(H_{iW\_j})$ by Eq. (4), we will get the results of $P(H_{iW\_j})$ shown in Fig. 3 (a) and the node F will be mistakenly considered to be the next destination of T_i. In contrast, the result calculated by path-based detection model (Eq. (5)) is shown in Fig. 3 (d), and the node G is considered to be the next destination of T_i. So, we can effectively remove the interference of the tag data from other paths to provide a more accurate approach for RFID data cleaning in supply chain management.

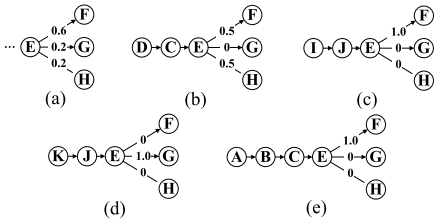

Fig. 3 Decision results based on the raw data in Table 1.

3.3 Sliding Window Model

RFID data stream is a infinite data series, and it is produced continuously. It is not possible to transmit all data in the memory for processing. What is more important is that only recent data may be important for RFID data cleaning. So we can make decisions based only on recent data rather than running computations on all of the RFID data seen so far. For the reasons above, we use the sliding time window model in our system. In the calculation of the relevant variables, we only analyze the data produced in the sliding window w rather than all of the data. It will also reduce memory requirements.

If we did not consider the paths that tags have pass through before reached node W and denote the mean value of the probabilities that tags leave node W from the exit E_j as $\hat{P}(H_{iW\_j})$, the estimate of the likelihood can be represented in Eq. (6).

$$\hat{P}_w(H_{iW\_j} \mid D_{iW0}) = \alpha P_w(D_{iW0} \mid H_{iW\_j}) \hat{P}_w(H_{iW\_j}) \tag{6}$$

where α is the normalizing constant, and the subscript w means that the value of corresponding parameters are calculated only by the data in the window w. We proposed an adaptive method for determining the most appropriate value of w using maximum entropy model. First, according to the Eq. (6), we can calculate the entropy of the distribution of missing tags' next destination as:

$$\begin{aligned} &H_w(H_{iW\_j} \mid D_{iW0}) \\ &= -\sum_{j=1}^{m} \hat{P}_w(H_{iW\_j} \mid D_{iW0}) \log(\hat{P}_w(H_{iW\_j} \mid D_{iW0})) \end{aligned} \tag{7}$$

Generally, a more accurate data cleansing result will lead to systems with bigger entropy. $w_{\max}$ and $w_{\min}$ are the threshold values of w. So, we select the value that maximizing the entropy H_w in $[w_{\max}, w_{\min}]$ as the most appropriate value of w.

4. Evaluation

To evaluate our method, we have implemented a prototype

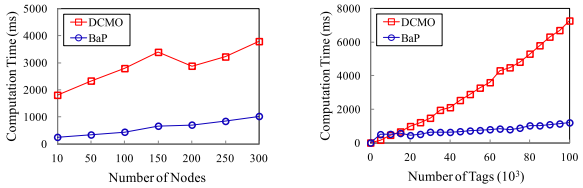

Fig. 4 Performance comparison of different methods. (a) The number of nodes means the number of warehouses in the whole supply chain. (b) The number of tags means the number of tags in system in current time.

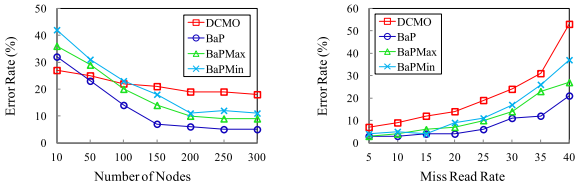

Fig. 5 Accuracy comparison of different methods.

of Path-Based RFID data cleaning system in Java. A synthetic RFID data generator is used to simulate the raw RFID data stream required by the experiment. The generator can generate randomly the corresponding data according to the distribution and connection relationship of each node in the supply chain. In this letter, we consider a comparative performance and accuracy analysis of the RFID data cleaning model based on dynamic clusters of monitored objects [3], denoted as DCMO, and the proposed method denoted as BaP. All algorithms are carried out on PC with Intel 2.66 GHz dual-core CPU and 4 GB of Memory.

We first analyze the performance of the two methods by recording the average computation time of the cleaning of 2,000 tags. DCMO has to maintain the group relationship among all the tags in real time. The high time overhead of the maintenance causes a low overall performance of DCMO. Compared with that of DCMO, the time cost of the proposed model is low, and the adoption of the sliding window model further reduces the computational cost of BaP. As shown in Fig. 4 (a), the computational cost of DCMO remains much higher than BaP. We plot the average computation time of DCMO and BaP with different number of tags in Fig. 4 (b), and BaP is also faster than DCMO, which is known to be fast, only when the number of tags is very small.

To compare the accuracy of different methods and analyze the influence of the window size w, the methods of BaPMax and BaPMin have been added to the experiments, which are the same to BaP except that their window size w are statically set to $w_{\max}$ and $w_{\min}$, respectively. Information on other tags in the same group is the only basis of DCMO for data cleaning. Thus, it is not applicable for the data cleaning of tags with scattered movements. The proposed RFID data cleaning model, BaP, is based on the path information of tags, and it does not matter whether the tags move together or not. As shown in Fig. 5, the accuracy of BaP is higher than that of DCMO, especially when the number of nodes in the supply chain exceeds 150 or the miss read rate exceeds 30%. With the increase of miss read rate, the error rate of DCMO will increase dramatically while the error rate of BaP keeps in a lower level. The accuracy of BaP is always higher than that of BaPMax and BaPMin, because that the parameter w in BaP will be adjusted to the most appropriate value, adaptively, while the window sizes of BaPMax and BaPMin are static. An appropriate w can bring out a higher accuracy to the method.

5. Conclusion

In this letter, we present two effective models, a Bayesian inference-based decision model and a Path-based detection model, to address the miss reading problem in RFID based supply chain management. In addition, the maximum entropy model is introduced for determining the most appropriate value of window size. Experiment results validate the performance of the proposed method and show that it is able to clean raw RFID data with a higher accuracy.

Acknowledgments

This research is supported by the National High Technology Research and Development Program of China (No. 2010AA012505, No. 2011AA010702, No. 2012AA01A401 and No. 2012AA01A402).

References

[1] L. Chun-Hee, "Rfid data processing in supply chain management using a path encoding scheme," IEEE Trans. Knowl. Data Eng., vol.23, no.5, pp.742–758, 2011.
[2] W. Roy, "The magic of rfid," ACM Queue, vol.2, no.7, pp.40–48, 2004.
[3] Y. Ge and G. Yu, "Efficient RFID data cleaning model based on dynamic clusters of monitored objects," J. Software, vol.21, no.4, pp.632–643, March 2010.
[4] R.J. Shawn, G. Minos, and J.F. Michael, "Adaptive cleaning for RFID data streams," Proc. 32nd international conference on Very large data bases, VLDB Endowment, pp.163–174, 2006.
[5] C. Haiquan, K. Wei-Shinn, W. Haixun, and S. Min-Te, "Leveraging spatio-temporal redundancy for RFID data cleansing," Proc. 2010 International Conference on Management of Data, pp.51–62, 2010.
[6] H. Fan, Q. Wu, and Y. Lin, "Behavior-based cleaning for unreliable rfid data sets," Sensors, vol.12, no.8, pp.10196–10207, 2012.
[7] EPC global. http://www.gs1.org