A New Method for Compressing Massive RFID Data to Achieve Efficient Mining

L. Hafezi, M. H. Saraee, and M. A. Montazeri

***Abstract*—Radio Frequency Identification (RFID) technology has been used for many purposes and has had effective results. This technology eases and accelerates many applications, but it has proposed a challenge, and that is the production of such a volume of data. The volume is so enormous that disusing the system comes into consideration. However different methods have been applied for solving this problem. In this paper, we propose a new method for data compression without missing any data. By applying it to our system, a production line of car engines, we can get our data to 1/50 times the basic data. We can also return to the basic data. We apply data mining to estimate the accuracy of our method and we consider our method error rate is 16 percent that is acceptable.**

***Index Terms*—Data compression, data mining, RFID technology.**

I. Introduction

Radio Frequency Identification technology uses radio frequency waves to read a unique identifier that is attached to a tag by RFID reader from a distance and without a line of sight. RFID is fast, reliable, and does not require contact between reader and tagged objects. RFID technology makes it possible to i) collect large amount of data for tracking and identifying physical objects and ii) real time monitor physical objects and their environment for monitoring applications [1]. This technology eases applications, but it has proposed an important challenge and it involves a lot of data. Interpreting the massive amount of data is a big problem.

Large companies have already begun carrying out RFID systems for easing their work. Walmart is one of these companies that applied RFID for supply chain management application, but it has predicted that when tags are used at the item level, this company will generate around 7 terabytes of data every day [2]. After this closure, many researchers have started to provide solutions for this problem.

In this paper, we also try to present a new and robust method for compressing massive RFID data. The data compression problem is to transform the input data into an output data with a reduced data volume but with no loss of information. Such compression requires the knowledge of what data is redundant and thus can be safely discarded [3]. In our method, we consider a definite and deterministic sequence among different phases in the production line of car engines. Then we combine these phases at different levels. Therefore, low level data have converted to higher and more meaningful levels.

This method is applicable in all the systems that do a series of deterministic and repeated operations over data.

The rest of the paper is organized as follows. In Section II, we describe the related work. Then, Section III describes RFID data. In Section IV, we propose a new compression method. Section V presents mining RFID data. Section VI reports the experimental and performance results. We conclude our study in Section VII.

II. Related work

Our work presents a new method for data compression. In this paper, we explore in details the new method. Below, we investigate some of the papers that are in this area.

A new warehousing model has been proposed in [2]. It preserves object transitions while providing significant compression and path-dependent aggregates for supply chain management application based on the following observations: (1) items usually move together in large groups through early stages in the system (e.g., distribution centers) and only in later stages (e.g., stores) do they move in smaller groups, and (2) although RFID data is registered at the primitive level, data analysis usually takes place at a higher abstraction level. Techniques for summarizing and indexing data, and methods for processing a variety of queries based on this framework are developed in this study.

A data interpretation and compression substrate has been presented in [4] over RFID streams to address these challenges in enterprise supply-chain environments. This paper mention, the data compression problem is to transform the input stream into an output stream with a reduced data volume but with no loss of information. Such compression requires the knowledge of what data are redundant and, thus, can be safely discarded. This work has used interpretation to obtain such knowledge and generate an output stream that (i) augments the input stream with additional, likely information about objects, and (ii) has a significantly reduced volume of data. The paper has described a data capture technique to construct a time-varying graph model from the raw stream.

Algorithms have been presented in [5] for temporal and spatial aggregation of RFID data streams, as a means of reducing their volume in an applicational controllable manner. It has proposed algorithms of increased complexity that can aggregate the temporal records indicating the presence of an RFID tag using an application-defined storage upper bound. It has further presented complementary techniques that exploit the spatial correlations among RFID tags. Its methods have detected multiple tags that are moved as a group and replaced them with a surrogate group ID, in order to further reduce the size of the representation.

A method has been offered in [6] to construct compressed

Manuscript received June 17, 2012; revised August 3, 2012.

The authors are with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran (e-mail: l.hafezi@ec.iut.ac.ir, Saraee@cc.iut.ac.ir, Montazeri@cc.iut.ac.ir).

probabilistic workflows that capture the movement trends and significant exceptions of the overall data sets, but with a size that is substantially smaller than that of the complete RFID workflow. Compression is achieved based on the following observations: (1) only a relatively small minority of items deviates from the general trend, (2) only truly non-redundant deviations, i.e., those that substantially deviate from the previously recorded ones, are interesting, and (3) although RFID data are registered at the primitive level, data analysis usually takes place at a higher abstraction level. Techniques for workflow compression based on non-redundant transition and emission probabilities are derived and an algorithm for computing approximate path probabilities is developed.

All the above papers have been customized to work with supply chain management, focusing on grouping tags that move together, and not suitable for many applications including our case study.

III. RFID Data

Data are generated from a RFID system that has been produced as a stream of RFID tuples and in large flood and with low accuracy. To use these data, first they have to be cleaned from every kind of noise [7].

Data that are used in this paper produce a production line of car engines. Different phases that are needed for completing an engine are performed in sequence, in every phase an engine can have one of the three statuses: phases may do correctly, may do badly but they can be fixed, or may do badly but they cannot be fixed and engines have to leave. In this production line 326 phases exists. Every one of them has recorded its status separately. The phases with some status can combine at a meaningful level.

IV. New Compression Method

As cited earlier, in this paper, we present a new structure for data compressing without missing any data. In fact, in this new method, we can return data to basic data before compressing. In this method, for different phases that are done over a car engine, we consider a deterministic sequence. Then, by combining meaningful phases from level 0 we call them level 1, and the engines that have the same status at level 0 combine and come to level 1. Then combining meaningful phases from level 1, we define level 2. Then we continue the same action for phases at level 2 and define level 3. In fact, level 3 is the highest level that we consider for our system.

In the following table, we show the number of data decreasing. As it shows, the data at level 3 have reduced to 1/50 times the basic data.

TABLE I: Reduced Data Volume By New Structure

Level No	0	1	2	3
Records No	75,127,352	23,999,840	2,477,132	1,542,268

By saving the combination of different levels in the table in database, we can decompose the data at every level and return to level 0 again that is a main advantage of our method; that is to say, no data would be missed in compression. Below, we show the pseudo code of our presented method. This pseudo code shows the general algorithm that we used.

Pseudo code of our structure:

Input: n items at level 0 and their status

Output: a compressed data set with k items ($k<<n$)

```
For every level ($i$)
    For every item ($j$)
        If all the statuses of item [$j$] are same
            Then combine them and create that item at level $i$+1
        Else
            They cannot combine
    End for
End for
```

V. Mining RFID Data

To estimate the accuracy of the presented method, we decided to use a technique of data mining. For this purpose, we used classification for predicting engine status in every phase.

We have a three-label classification because engine status has three labels; it may be ‘Good Product’, ‘Fixable’ or ‘Unfixable’.

We use a decision tree for the classification and through different kinds of decision tree algorithms we selected c4.5.

A. c4.5[1]

C4.5 is one system that learns decision-tree classifiers. C4.5 uses a divide-and-conquer approach to growing decision trees. Only a brief description of the method is given here.

The following algorithm generates a decision tree from a set D of cases:

1) If *D* satisfies a stopping criterion, the tree for *D* is a leaf associated with the most frequent class in *D*. One reason for stopping is that *D* only contains cases of this class, but other criteria can also be formulated (see below).
2) Some test T with mutually exclusive outcomes T_1; T_2; : : : ; T_k is used to partition *D* into subsets D_1;D_2; : : :;D_k, where *Di* contains those cases that have outcome *Ti*. The tree for *D* has test *T* as its root with one subtree for each outcome *Ti* that is constructed by applying the same procedure recursively to the cases in *Di*.

Provided that there are no cases with identical attribute values that belong to different classes, any test *T* that produces a non-trivial partition of *D* will eventually lead to single-class subsets as above. However, in the expectation that smaller trees are preferable (being easier to understand and often more accurate predictors), a family of possible tests is examined and one of them chosen to maximize the value of some splitting criterion. The default tests considered by C4.5 are:

3) *A*=? for a discrete attribute *A*, with one outcome for each value of *A*.

A<=*t* for a continuous attribute *A*, with two outcomes, true and false. To find the threshold t that maximizes the splitting criterion, the cases in *D* are sorted on their values of attribute

[1] This section has been brought to this paper from [8]

A to give ordered distinct values $v_1, v_2,: : : ,v_N$. Every pair of adjacent values suggests a potential threshold $t = (v_i + v_{i+1}) =2$ and a corresponding partition of *D*.1 The threshold that yields the best value of the splitting criterion is then selected.

The default splitting criterion used by C4.5 is gain ratio, an information-based measure that takes into account different numbers (and different probabilities) of test outcomes. Let C denote the number of classes and $p\ (D, j)$ the proportion of cases in D that belong to the jth class. The residual uncertainty about the class to which a case in D belongs can be expressed as

$$\text{Info}(D) = -\sum_{j=1}^{c} p(D,j) * logp(D,j) \qquad (1)$$

and the corresponding information gained by a test T with k outcomes as:

$$\text{Gain}(D, T) = \text{Info}(D)\text{-} \sum_{i=1}^{k} \frac{Di}{D} * Info(Di) \qquad (2)$$

VI. Performance Study

In this section, we perform newly suggested compression structure and show the results. The first result shows that the execution time of data mining algorithm is decreased by using this structure. Table II shows the execution time at level 0 and 3. The data mining execution time for level 3 is 33 seconds and it has added execution time of the compression procedure that is 7:51. So, the total execution time is 8:24. These results show a good performance in the execution time.

All the experiments were conducted on an Intel ® core 2 duo CPU, 2.10 GHz system with 2 GB of RAM.

TABLE II: Reduced Execution Time By New Structure

Level No	0	3
Time	19:51	33(*s*)+7:51=8:24

We consider 70 percent of the data for training and 30 percent for testing, and we use 10-fold cross-validation for acquiring accuracy of our method. Then we use (3) to acquire the error rate.

$$\text{Error rate}=\frac{\text{Total missclassified test examples}}{\text{total number of test examples}} \qquad (3)$$

Results show that the error rate is nearly 16 percent. We believe it shows good accuracy and is acceptable. So our method can be used for effective compressing.

VII. Conclusion

RFID systems are applied for many purposes; these systems show efficient and effective results. They ease and accelerate many applications and they offer many possibilities that have not been at hand so far. But they have a big challenge with themselves, and it has produced massive data. So to perform query processing and data mining, we must use techniques to reduce the number of data.

In this paper, we have used a real RFID system, therefore the challenges are real. To solve this problem, we have suggested a new compression structure. We have reduced the data to 1/50 times the number of basic data by applying this structure. One of the main advantages of our structure is that we can return to the basic data. The accuracy of our structure is also acceptable. Every system that does a series of deterministic and repeated operations over data can use this structure.

References

[1] Y. Bai, F. Wang, P. Liu, C. Zaniplo, and S. Liu, “RFID Data Processing with a Data Stream Query Language,” *Proc. 23rd International Conference on Data Engineering, IEEE press*, 2007, pp.1184-1193.

[2] H. Gonzalez, J. Han, X. Li, and D. Klabjan, “Warehousing and Analyzing Massive RFID Data Sets,” *Proc. 22nd International Conference on Data Engineering, IEEE press*, 2006, vol. 2, pp. 83-93.

[3] R. Cocci, T. Tran, Y. Diao, and P. Shenoy, “Efficient Data Interpretation and Compression over RFID Streams,” *Proc. 24th International Conference on Data Engineerin, IEEE press*, 2008, pp. 1145-1147.

[4] R. Cocci, T. Tran, Y. Diao, and P. Shenoy, “Efficient Data Interpretation and Compression over RFID Streams,” *Proc. 2008 IEEE 24th International Conference on Data Engineering, IEEE press*, 2008, pp. 1445-1447.

[5] D. Bleco and Y. Kotidis, “RFID Data Aggregation,” *Proc. 3rd International Conference on GeoSensor Networks*, 2009, pp. 87-101.

[6] H. Gonzalez, J. Han, and X. Li, “Mining Compressed Commodity Workflows from Massive RFID Data Sets,” *Proc. 15th ACM international conference on Information and knowledge management*, 2006, pp. 162-171.

[7] R Derakhshan, M. E. Orlowskaand, and X. Li, “RFID Data Management: Challenges and Opportunities,” *Proc. IEEE First International Conference on RFID Gaylord Texan Resort*, 2008, pp. 175-182.

[8] J. R. Quinlan, “Improved Use of Continuous Attributes in C4.5,” *Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

Leila Hafezi was born in iran, September 1983. Leila got her MSc degree in software engineering from Isfahan University of Technology, Isfahan, Iran in Febreury 2011. She is head of R&D department of Pishgaman Kavir Co. Her area of interest is Data Mining, Data Warehousing and Optimizing Algorithms.

Dr. Mohammad Hossein Saraee is currently assisstant professor in Isfahan University of Technology, Isfahan, Iran. His research interests are Intelligent Database Systems (Temporal, Spatial and Multimedia), Data Mining and Medical Informatics.

Dr. Mohammad Ali Montazeri received his PhD from Deparment of Computational of University of Manchester Institute of Science and Technology, Febreury 1996. He is currently assisstant professor in Isfahan University of Technology, Isfahan, Iran.