Comment on "Process Mining: Process Model from Event Logs"

Vijay Wadhwa Dennis. K .J. Lin [1]*

Abstract

As companies move from piloting to deploying RFID technology, they have to be prepared to deal with a variety of data management issues. RFID will generate huge amount of data. Process mining is a technique to analyze the stream-data from the viewpoint of building or detecting process. Process mining helps in understanding the actual process from event logs. In this paper we present a case study on application of Process Mining to RFID data collected from a pilot study. Through the application of modified α algorithm we discover the underlying process for flow of products from Warehouse/DC to the retail store. This is a discussion of the work by Lin, Chen and Kuo (LCK).

Keywords: α-algorithm, Process Mining, RFID, Supply Chain Management.

[1] Department of Statistics, Pennsylvania State University, University Park, USA and Department of Statistics, Fudan University, Shanghai, China.

1. RFID

Radio frequency identification (RFID) serves as the basis for an automated data collection system. In a generic sense, it refers to the use of radio frequency waves for identifying individual items or stock keeping units. RFID is a generic term for technologies that use radio waves to uniquely identify an object, animal, or person, where radio waves mean the radio frequency portion of the electromagnetic spectrum. Interest in RFID technology is particularly acute for Retail, consumer- goods and packaging companies. For e.g. Wal-Mart has mandated that all of its suppliers be RFID enabled i.e. all the suppliers will have to embed an RFID tag on all pallets and cases traveling to Wal-Mart facilities. This mandate will eventually affect remaining industry sectors.

1.1. RFID Advantages

The two primary advantages of RFID are supply chain visibility and theft reduction, which are both achieved due to increased information availability. RFID can enable storage and relay of item-level data associated with a products life cycle. It can provide item level inventory visibility in real-time. The cost of implementing RFID will exceed barcodes for many years to come but still the advantages of RFID can outweigh the difficulties.

1.2. Challenges to Information Systems

To realize the full potential of these two and host of other benefits (Wadhwa and Lin, 2008), the data obtained from RFID has to be analyzed in a timely and efficient manner. Many large businesses are already overwhelmed by a flood of RFID data; most of it is poorly integrated and barely managed. RFID tags which may be placed on each pallet or item will generate huge amount of data. A single product may have many attributes, including but not limited to physical specs such as form, color, weight, point of origin and date shipped, as well as associated data including promotions, ads etc. Multiply these attributes by the large number of items in a supermarket, then multiply that number by the

all the stores in a national chain (For example Target, Wal-Mart), and we have huge number of attributes to track.

By itself RFID has limited use, but coupled with an application that can convert this data into useful information, a solution can be designed that potentially can provide benefit to an organization. In today's business environment where companies are seeking leaner and at the same time more responsive supply chains, RFID in many cases may provide the needed functionality to support these objectives. RFID implementation can provide benefits like; cost reduction, increased accuracy, improved workforce efficiency, streamlined business process while also improving the company's ability to integrate business functions.

Our primary objective in this paper is to develop novel RFID-data mining technique to detect abnormality in processes, improve process efficiency, or both. In this paper we illustrate the use of Process Mining algorithm on RFID data collected from a pilot study. Through the application of modified α algorithm we discover the underlying process for flow of products from Warehouse/DC to the retail store.

In Section 2, we discuss process mining, which is a technique to analyze event logs. Furthermore, we also present how RFID data can be analyzed using process mining. In section 3 we present literature review on analyzing RFID data. In section 4, we introduce the RFID data collected from a major US retailer. Section 5 contains the application of modified α-algorithm on event logs of RFID data. Finally Section 6 contains some conclusions and future work.

2. PROCESS MINING AND RFID

Process mining is a technique used to analyze the stream-data, it has recently received a great deal of attention in many fields, such as software engineering and workflow management. It analyzes the data from a process viewpoint and presents the result with a

process model. The resulting model helps in understanding the actual process from the stream-data point of view.

For the basic utilization of a process model, Humphrey *et al.*(1989) introduced four desirable properties, including enabling effective communication regarding the process, facilitating process reuse, supporting process evolution and facilitating process management. They suggested that a process model should be able to (1) represent the way the work is actually preformed; (2) provide a flexible and understandable, yet powerful, framework for representing and enhancing the process; and (3) be refinable to whatever level of detail is needed. Modern information technologies allow us to collect complete global stream-data in an efficient manner. Process mining helps in understanding the actual process from these stream-data.

Process mining is gaining more importance as the advancements in IT systems have resulted in improvement of data quality and more and more information about processes collected in the form of event logs. In this paper we analyze real RFID data provided by a leading retail organization, which we shall call as X, the true identity of the organization shall remain anonymous for obvious reasons.

An RFID data set is a collection of paths, one per distinct tag in the system; for our example we assume our EPC tag contains the most basic information, namely, location & time; where location refers to the location of the reader and time refers to the time the tag is scanned.

3. LITERATURE REVIEW

Due to a general lack of RFID data there haven't been many papers been published based on real RFID data. Delen et.al (2006) analyzed the movement of a single EPC from distribution center to a retail outlet. They also present how this information can be useful to both the retailer and the supplier. Lin and Wadhwa (2008) present a comparison of RFID & barcode in supply chain; they illustrate how various supply chain entities might be

impacted by adapting RFID. Wadhwa and Lin (2008) illustrate the challenges to RFID data mining & present a real life case study of RFID implementation at IBM. Gonzalez et.al. (2006) have developed a model for the design and construction of a highly compressed RFID workflow that captures the main trends and important deviations in item movements in an RFID application.

McFarlane et.al. (2003) discuss the implications of RFID for manufacturing shop floor control. They put forward the concept of an intelligent product; a product whose information content is permanently bound to its material content and is able to influence decisions made about the product. Twist (2005) discusses the impact of RFID on supply chain facilities and discusses the potential implications of implementing RFID. Angeles (2005) discusses the implementation issues in supply chain and has also presented several case studies of successful RFID implementation. Karkkainen (2003) discusses method to improve supply chain efficiency for perishable goods using RFID tagging.

4. PROCESS MINING AND RFID

Data from X is obtained on five different products over a 6 month time frame; there are various points during the flow of the product where the RFID tag is scanned. The data obtained from X contains the following information:

- EPC Number: EPC number is a unique identifier which identifies each and every item in the supply chain
- UPC (Universal Product Code): UPC or Bar-code is also used on products together with RFID.
- EPC Time Stamp: EPC time stamp contains the time stamp of the when the tag was read by a RFID reader.
- Store Number: Store number is the store/Warehouse in which the product was scanned.

- Location: Location is the RFID reader location, there are 7 distinct locations

There are 7 unique locations (RFID Readers) where the tags are scanned. The locations where the tags are scanned include:

- 100: Receiving dock door.
- 102: Sorting conveyor system
- 104: Back room
- 105: Box crasher
- 112: Sales floor.
- 117: Shrink-wrap machine
- 125: Layaway indicator

4.1. Methodology

Process mining techniques allow for extracting information from event logs. The process flow can be used to discover the models describing the processes. Moreover it is possible to use the process monitoring to monitor any deviations e.g. comparing the observed events with predefined models. Every product that flows into and out-of X has a unique process flow, for e.g. the product first goes to receiving door then to sales floor and then to a box crasher and so on. Our first objective in this analysis is to analyze the event logs to understand the flow of products in X.

4.2. Data Cleaning

Since RFID technology hasn't matured there are still a few problems with data reads, two most commonly found issues with RFID data scans are multiple reads & missing reads.

- Multiple reads: multiple reads happen when a reader scans the tag more than one once, consecutively; an example of multiple read is as shown below.

100 100 100 100 100 100 100 100 100 112 105

For each RFID read we first identify whether the read is a legitimate or not; there are two ways a location is classified as unique:

- We sort the data by date and time; if the previous location scanned is not same as the current location then those two locations are unique.
- If the time difference between two consecutive locations (which are same) is greater than a user specified value than those two locations are classified as unique. For each EPC we vary the user specified time difference to see the process flow obtained. For example if we declare the user specified time to be 100 seconds and for EPC 1 location 100 was scanned on Dec 8th 9:40 A.M and Dec 8th 9:41 A.M; since the time gap between the two scans is less than 100 seconds we classify the two as duplicates; and eliminate one of them.

- Missing Reads: Although one of the potential advantages of RFID over barcode is faster & more accurate scans, organizations haven't mastered the art of placing readers and tags accurately which leads to lower than expected read accuracy. There are several reason for which there can be a missing read, some of them are:
  - Box on which tag is placed is taken from store by customer or associate after product is stocked on the shelf or
  - Tag may have dislodged while stocking a product
  - Accidental double read at crusher etc.

Once we have cleaned data we analyze the event log to obtain the most frequent event logs. Table 1 shows the most common event logs extracted from the cleaned data.

Table 1. RFID Event Logs

Case	Event Log
1	100,102,104,117,125,117,125,117,105
2	100,104,102,117,125,117,105
3	100,112,117,125,117,125,117,105
4	100,102,117,125,117,105
5	100,102,117,125,117,125,117,105

After obtaining the most common process logs we apply the modified α-algorithm developed by LCK to obtain the process.

5. MODIFIED α-ALGORITHM FOR PROCESS MINING ON RFID DATA

Modified α-algorithm is based on petri nets. A petri net consists of a series of transitions, places and arcs. Transitions and places are connected through directed arcs in such a way that

- Places & transitions have atleast one incident edge
- In every path, transitions and places alternate.

The basic components of the algorithm (as seen in LCK, Figure 2 in Section 3) are:

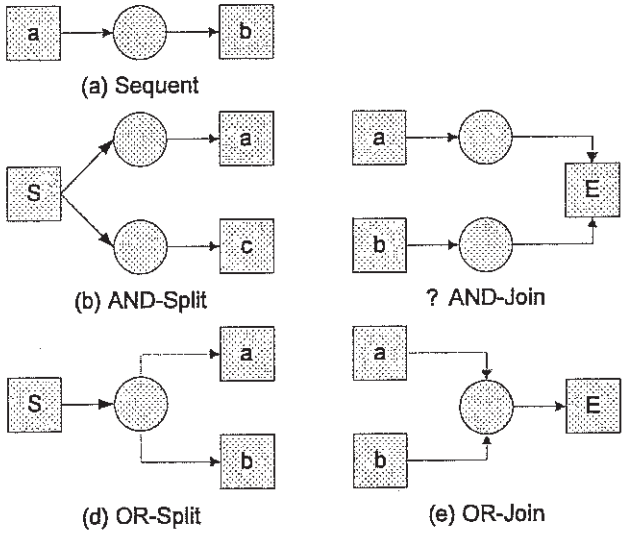

Figure 1. Basic components in the α algorithm

The ultimate goal is to use the algorithm to develop a process model describing all the workflows using RFID. Fig (1) shows the process model developed on the data obtained through RFID

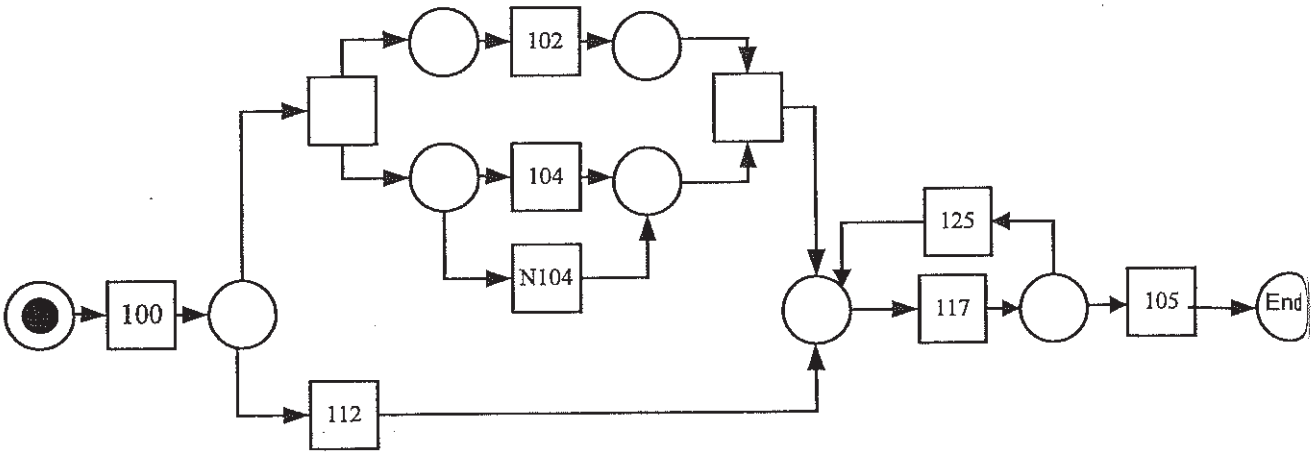

Figure 2. RFID Process Model

In the process model we observe that Reader 100 & 105 are chosen to be the starting & the ending activities. Based on the concurrence ({102},{104}); a fictitious activity No-104 is created since 104 did not occur for some cases when 102 occurs, and thus ({102}{104}) is modified to ({102}{104, N104}). A detailed description of α algorithm can be found in LCK.

6. CONCLUSION AND FUTURE WORK

Based on our analysis we can see that X is still along way from making RFID work. There are no standardized process-flow for different products from DC to a warehouse, the true potential of RFID can only be realized when the movement of each and every part with-in the supply chain can be tracked from beginning to end, hence there is need to establish standardized processes for parts flowing between different supply chain entities.

Moreover in this analysis we worked with a very limited data set (5 products, 6 months) it is possible that the finding through this data set cannot be extended to other products in X. This analysis should be extended to include more products.

REFERENCE

1. Angeles, R, (2005). RFID technologies: Supply-Chain Application and Implementation Issues.
2. *Information systems management*, 22(1), 51-65.
3. Delen, D., Hardgrave, B., Sharda, R (2006). RFID for better supply chain management through enhanced information visibility. *Forthcoming in Production and Operations Management*
4. Gonzalez, H , Jiawei Han , Xiaolei Li (2006), Mining compressed commodity workflows from massive RFID data sets, *Proceedings of the 15th ACM international conference on Information and knowledge management*, November 06-11, 2006, Arlington, Virginia, USA
5. Humphrey WS, Kellner MI. Software process modeling: principles of entity process models. *ACM* 1989; 331-342.
6. Karkkainen, M, (2003). Increasing efficiency in the supply chain for short shelf life goods using RFID tagging. *International Journal of Retail & Distribution Management*, 31(10), 529-536
7. Lin, Dennis K.J. and Wadhwa, Vijay (2008) "Efficiency Gets a New Identity: Barcode and RFID in the Supply Chain," *Quality Progress*, 41, 20-24.
8. McFarlane, D., Sheffi, Y. (2003), "The Impact of Automated Identification on Supply Chain Operations", *International Journal of Logistics Management*, 14(1), 1-17.
9. Twist, D.C, (2005). The impact of radio frequency identification on supply chain facilities. *Journal of facilities management*, 3(3), 226-239.
10. Wadhwa, Vijay and Lin, Dennis K.J. (2008) "Radio Frequency Identification: A New Opportunity for Data Science," *Journal of Data Science*, 6, 369-388.