A Graphic-based Data Mining Approach in RFID Sensor Networks

Lianjun Yi

School of Science and Technology; Gannan Normal University, Ganzhou, China
yilianjungnsy@163.com

Abstract

As the wide use of radio frequency identification (RFID) technology, such wireless sensor network has created enormous data in supply chain management, manufacturing sites, retailer sections, as well as health care domain. In order to discover the invaluable information and knowledge from such massive RFID data, this paper introduces a data mining approach based on graphic theory to pick out the useful information and knowledge from logistics RFID data. This approach considers the complex relationship of different RFID data and uses graphic model to build up the logistics path. That means the logistics information stored in RFID data is converted by graphic diagram, which could be mined by the proposed approach over OLAP. The inquiry efficiency has been greatly improved through experimental results.

Keywords: Graphic Theory; Data Mining; RFID; Sensor Network; Apriori Algorithm

1. Introduction

As the wide increasing use of radio frequency identification (RFID) technology in logistics, retailing, manufacturing, etc, large number of RFID data has been generated [1, 2]. It is significant to mine the invaluable knowledge from such massive RFID data. RFID technology is an automatic identification technology which uses the radio magnetic waves to communicate between readers and tags [3]. Since the tag ID is an electronic product code (EPC), the tracked items could be unique so that it could be used for tracing and tracking various materials, products, and so on.

RFID technology uses the communication between readers and tags for transferring information or data [4]. Thus, it is able to create a world that everything is connected with each other. Every object in the system is labeled, categorized so that they could be traced and tracked. However, there are some challenges to address the RFID data since this technology generates lots of data within a second. For example, RFID data are usually dynamic, redundant, spatio-temporal, and real-time. Take RFID in specific application area for example, supply chain management (SCM) with RFID often carries out enormous data about product movement [5]. Walmart is a large international supermarket which has been used RFID technology for controlling its SCM and logistics flow in item level. If a manager in such enterprise wants to know the information about a specific product like a camera, such data related to manufacturer, international supply hub, district distribution center, retailer warehouse, supermarket shelf, and finally cashier could be picked up so as to analyze the movement of this product from beginning to the end. Such information forms a closed-loop information cycle. Every product could be identified by RFID readers in every location. The identification information will be recorded in database in the form of a piece of record (EPC, Location, Time). Where, EPC is the code for this product, location is the place of reader, time is the exact date when carrying out the read operation. Assume that if a Walmart can sale over
ten thousand products per day, every product will be scanned by RFID reader about several hundreds times. This shop may generate over 1 TB data per day. This data is enormous and carries invaluable information and knowledge which could be mined through different trajectory density level and different dimensions of each product.

Business intelligence (BI) is widely used for addressing with the enormous data [6]. The technical architecture of BI contains three key components, which are Data Warehouse, OLAP, and Data Mining. Through these three components, the massive data could be integrated, processed, analyzed, and visualized. RFID system is able to collect large number of trajectory information about the moving objects. Data mining approach could be used for analyzing and managing the frequent path which is useful for guiding the decision-making [5]. However, there are several challenges. Firstly, in data capture layer, the data cleansing is low efficiency and effectiveness due to the complexity and logic relationships among various RFID data. Secondly, in transaction processing layer, the reasoning rules are uncertain and dynamic due to the concentration focused by different enterprises. Thirdly, in enterprise application layer, the RFID data should be qualified to be safe and private so that the security concerns could be dealt with.

In order to address the above challenges, many researchers have worked on associated dimensions. Qu et al proposed an approach on how to establish massive and enormous RFID dataset in data warehouse, in where simple data analysis could be carried out [7]. Logistics network is full of moving objects which are equipped by RFID devices. The movement is characterized in the initiative phase. Large number of objects is trend to move at the same time. At later phase, the moving path is trend to be divided into separate trajectory. Therefore, this paper uses the concept of “grid” to record the package of moving objects so as to reduce the path quantity in the data warehouse. Zhong et al introduced a concept of FlowCube which is able to access the RFID data warehouse quickly [8]. Data Cube technology aims to access the data warehouse efficiently and quickly. It is the main object of OLAP. Decision makers are able to access the data with high effectiveness and pick up the useful knowledge from massive data. But the evaluation degree of Cube is algebraic such as price, etc. The aim of OLAP on RFID data is to excavate the path information about different moving objects/products. It is challengeable for data cube for doing so. This paper proposes the FlowCube to use the path information as the evaluation degree. Based on some specific algorithms, it is able to mine the frequent path, meeting the requirements from end-users. In order to compress the record, Saygin et al. introduced a manner to compress the path support degree [9]. After twice scanning of the dataset on path, the compressed workflow could be obtained. During the mining of frequent trajectory, path support degree could be got through the compressed workflow, instead of scanning the dataset. That greatly reduces the time spent on mining the targeted path. Zhong et al. introduced a method which is able to mine the closed frequent trajectory [10]. This method categorizes the moving path into two aspects. One is location set, the other is time set. The current mining approach is used for excavating the closed location set. After that, the mined results are used for mining the time set. This approach has been approved to be an excellent one through experimental evaluations. Based on the complexity of RFID data, Zhong et al. proposed an architecture for discovering the invaluable information and knowledge from data warehouse, in which records are independent [5]. However, in most of the cases, records are closed related to each other.

This paper uses the graphic theory into data mining approach for address the RFID logistics data which is complex, massive, enormous, and high dimensional. The RFID data carrying with lots of path information will be visualized through graphics. Based on the graphics, frequent path mining algorithms are proposed. There are several key concerns in this paper. First of all, RFID data is compressed by using the data warehouse approach so as
to reduce the storage space. Secondly, compress graphic set is used for mining the invaluable information so as to reduce the costs and time on sub-graphics, improving the speed and efficiency. Finally, a path categorized method is investigated to cluster the RFID objects so as to speed up the clustering of objects. Through these steps, the familiarity of different records could be worked out by using the dynamic modeling approach. That aims to use the layered clustering method to pick up with the moving mode hidden in the massive and enormous RFID data and to predict the behavior of moving objects.

2. RFID Data Warehouse

RFID data warehouse is the essential of keeping and mining the logistics path [11]. This section introduces the key components and RFID data records before discovering the invaluable information and knowledge.

2.1. RFID Logistics Dataset.

RFID dataset is consisted by a set of elements \((EPC, location, time)\), where \(EPC\) represents the electronic product code which is used for identifying a product. \(location\) indicates where the readers deploy at. \(time\) means the actual date when the event carries out. If all the elements related to a specific product are able to integrated together according to the sequence of time stamp, a moving path could be figured out. After the data cleansing, the path contains the program \((location, time_{in}, time_{out})\). In order to study the flow mode, the duration time is important, thus, the transaction could be converted into \((Location, duration)\), where \(duration = time_{out} - time_{in}\).

A path dataset is consisted of several dimensions \(<d_1, ..., d_m, (l_1, t_1)...(l_k, t_k)>\), \(d_1, ..., d_m\) represents the dimension without relationship with path, which is used for describing the characteristics of product such as price, manufacturer, purchaser, etc. \((l_i, t_i)\) mainly represents the duration time \(t_i\) at location \(l_i\).

There are some specific characteristics of RFID data warehouse. First of all, the RFID data warehouse has the simple data. The data generated from RFID applications could be regarded as elements stream like \((EPC, location, time)\). In such expression, RFID data cannot carry much information. In order to convert the raw data into useful information that could be used for enterprise decision-makings, layer-based reasoning must be carried out. Secondly, RFID data warehouse is massive and enormous. RFID applications may create large number of data. Take Walmart for example, the data generated by RFID every three days is as large as the size of the council library in USA. It is challengeable to address such massive data. As a RFID development project, TB level of data may be created per day since every labeled item will hand upon data about EPC, location, and time. Thirdly, the inaccurate data in RFID data warehouse is always highlighted due to the data stream generated by RFID readers are not accurate enough. In real-life applications, the accuracy of data may be around 60%~70%. Unfortunately, the inaccuracy may cause some disturbances for high level decision-makings. Therefore, the data should be cleansed before using for various applications. Finally, the RFID data involves time and space dimensions. The RFID in various applications generates data dynamically. The RFID data carry the information about status change. Thus, in RFID data management, it is necessary to model the data through a data model which suits the interactions in application layer in terms of traceability and supervision. Under the consideration of RFID readers and tags are
2.2. Architecture of the RFID data warehouse.

Assume that, the cleansed RFID data could be expressed by the table which contains several attributes \((EPC, \text{location}, \text{time}_{\text{in}}, \text{time}_{\text{out}} : \text{measure})\). Data cube is able to calculate the possible groups from the table through the records assembled on the possible dimension's group. If we use the amount as a degree, it is easy to get the production quantity at specific location with a month. The clustering does not consider the relationship among different records. For example, if we want to know the product logistics information from Chicago distribution center to the stores in Urbana. However, it is hard to do so. We may get the quantity in every location. But we do not know how many products moved from one location to another one. Thus, it is necessary to build up an architecture to cluster the data while keeping the similarity of path.

The architecture of the RFID data warehouse used in this paper has some specific tables for addressing the challenges mentioned previously. Firstly, there is a reality table \(stay\) which contains the RFID records after cleansing. Secondly, there is a table \(info\) which stores the information about each product and its path data. That is SKU information such as manufacturer, batch ID, color, etc. Thirdly, there is a \(map\) table which is to connect different records from \(stay\) so as to generate the path. These considerations include the batch moving, data heterogeneity, partial path integration and extension. RFID path data are compressed through three dimensions in the architecture which is realized by the RFID Cubiod.

The differences between RFID data warehouse and database come from several aspects. Firstly, the \(map\) table is able to connect to the records in \(stay\) so as to keep the original structure of the data. Secondly, the RFID data warehouse realizes the operation of RFID data cubic path selection-oriented approach, which is to analyze and reach the OLAP on data warehouse. Finally, the architecture of RFID data warehouse has the query operation of the path given the data compression and efficiency during massive data select, insert, and update.

3. RFID Frequent Path

The RFID frequent path is significant to make the logistics decisions. Thus the delivery costs could be predicted and reduced if the optimal logistics path could be obtained. In order to figure out the frequent path, several definitions are proposed in this paper.

**Definition 1**: sub-path and super path. Assume that two path \(p_1 = (l_1, d_1)\) and \(p_2 = (l_2, d_2)\), if \(p_1\) is the sub-path of \(p_2\), that could be expressed as \(p_1 \subseteq p_2\). When and only when \(l_1 = l_2, d_1 = d_2\) or \(d_1 = *\) the expression will be established. Assume there are two paths \(p_1 = (a_1)(a_2)...(a_n)\) and \(p_2 = (b_1)(b_2)...(b_m)\), where \(a_i\) and \(b_j\) are path phase, \(i = 1, 2, ..., n\), \(j = 1, 2, ..., m\). If there is a group of integer \(1 \leq j_1 < j_2 < ... < j_n \leq m\) that makes \(a_1 \subseteq b_{j_1}...a_n \subseteq b_{j_n}\). Therefore, \(p_1\) is the sub-path of \(p_2\). \(p_2\) is the supper path of \(p_1\), which is expressed by \(p_1 \subseteq p_2\).
Definition 2: frequent path. If a path \( p \) is a sub-path of the \( path \) of a RFID element \( <EPC, (a_1, a_2, ... a_n), (m_1, m_2, ... m_k), path> \). The element includes the path \( p \). The amount of element which contains the \( p \) in the RFID data warehouse is termed as the support factor of \( p \), marking as \( count(p) \). The support degree of path \( p \) is \( support(\ p \ )=\ count(p) / |DB| \), \(|DB|\) is the total number of elements in RFID data warehouse. We can set up a threshold \( min\_support \), if \( support(\ p \ )\geq min\_support \), then path \( p \) is a frequent path.

Definition 3: closed-loop frequent path. If path \( p \) is a frequent path, there is no path \( p' \), s.t. \( p \subseteq p' \), \( support(\ p \ )=\ support(\ p' \ ) \), thus, \( p \) is a closed-loop frequent path. There is a characteristic that if the support degree of \( p \) is less than \( min\_support \), that means \( p \) is not a frequent path. Thus, any path included \( p \) is not a frequent path. The characteristic is called Apriori feature.

3.1. Flow cube

Flow cube is used in this research for building up the warehouse. In OLAP, cube is consisted by different granularity level of Cubiods from aggregated product stream [12]. Flow cube has two key differences comparing with data cube. Firstly, the evaluation of each unit cell is not the sum of the value, but using the product flow chart instead. Product flow chart is used for describing the information of moving trends and deviations of the trends. Secondly, each flow chart is able to adjust the granularity level, aiming to observe the moving information from different levels. In this paper, there are three steps to carry out the Cube calculation:

1. The calculation of path aggregated execution is carried out in the all levels;
2. Using the product lattice and path lattice to reduce the support degree of the path;
3. Remove the non-frequent unit lattice and that which could be obtained from the higher level of granularity in the product steam, aiming to compress the cube.

The flow cube in this paper is used for presenting the RFID path data. There are several definitions.

Definition 4: flow chart. An elements set \((V, D, T, X)\) is used for expressing the possibility working flow. That is called flow chart. Where, \( V \) is the set of nodes, each of which associates with a unique path in the data warehouse. \( D \) is a multi-distributed set, each element is corresponding to a node in \( V \). Its value is calculated by the possibility of the possible lasting time of this node. \( T \) is a multi-distributed set, each of whose element associates with a node in \( V \). Its value is the possibility of transfer from this node to others. \( X \) is a set that keeps the emergency situations of the node transferring and the distribution of lasting time.

Definition 5: product lattice. Let \((L_1, L_2, ... L_m)\) denotes the granularity level of product in path data warehouse, where \( L_i \) represents the granularity level which is not associated with \( d_j \). The minimal granularity level of the product lattice is a path data
warehouse. All the granularity levels of products form product lattice. If the granularity level of \(n_1\) is larger than \(n_2\), That means the layer of \(n_1\) is higher than \(n_2\).

**Definition 6**: path lattice. As same as the relationship between products and granularity level, paths also associate with specific location and time. Path lattice is expressed as \(\langle v_1, ..., v_k, t_i \rangle\), where \(v_i\) is the granularity level of path location, \(t_i\) is the granularity level of time.

**Definition 7**: flow cube. Based on the above definitions, flow cube is a set of the unit lattice \((I_i, P_j)\). The Cubiod in unit lattice is consisted by the records aggregated from the reality table. Each unit lattice is aggregated according to the granularity level \(I_i\), and the path is aggregated by path granularity level \(P_j\). The evaluation degree of unit lattice is based on the flow chart from path calculation.

### 4. Data Mining Approach based on Graphic Theory

Graphic theory-based data mining approach is used for figuring out the path through RFID data due to its capacity of dealing with the complex structure of data. There are some definitions related to the approach.

**Definition 8**: marked graph. A marked graph is presented by 
\[ G = (V(G), E(G), L(V(G)), L(E(G))) \]
where \(V(G) = \{v_1, ..., v_n\}\) is a set of the nodes in marked graph \(G\). \(E(G) = \{e_{ij} = (v_i, v_j) | v_i, v_j \in V(G)\} \subseteq V \times V\) is the set of edges. The marks of the node in \(G\) are expressed as \(L(V(G)) = \{lb(v_i) | \forall v_i \in V(G)\}\). The marked edge set is \(L(E(G)) = \{lb(e_{ij}) | \forall e_{ij} \in E(G)\}\).

**Definition 9**: omorphism of marked graph. There are two marked graphs \(G\) and \(G'\). If there is a mapping function \(f : V(G) \rightarrow V(G')\) meets
\(1\) \(\forall v_i \in V(G), lb(v_i) = lb(f(v_i))\); 
\(2\) \(\forall (v_i, v_j) \in E(G), (f(v_i), f(v_j)) \in E(G')\), \(lb((v_i, v_j)) = lb(f(v_i), f(v_j))\), then \(G\) and \(G'\) is omorphism graph, expressed as \(G \cong G'\).

**Definition 10**: sub-graph of \(G\). Assume there are two marked graph \(G\) and \(H\). If \(H\) has a sub-graph \(H'\) which is an omorphism graph of \(G\). Then, \(G\) is a sub-graph of \(H\) \(G \subseteq H\).

**Definition 11**: graphical support and frequent sub-graph. If there is a graph repository \(GR = \{G_1, ..., G_n\}\) and a graph \(G'\). The support of \(G'\) in \(GR\) is \(sup(G') = \{H | H \in GR, G' \subseteq H\}\). Let a minimal support \(minsup\), in \(GR\), the sub-graphs with supports larger than \(minsup\) are called frequent sub-graph. If \(G'\) is a frequent sub-graph, meeting \(G' \subseteq G\), \(G'\) is the maximal frequent sub-graph.

Based on the above definitions, the data mining approach uses Apriori frequent sub-structure algorithm for this purpose. The algorithm is similar with the frequent item data mining manner using Apriori. Frequent graph searching starts from small-sized
graph, following bottom-up mechanism which generates the extra node, edge/path in backup graph. The algorithm is detailed as follows:

Input: RFID graph data set \( D \), minimal support threshold \( \text{min}\_sup \).

Output: frequent sub-structure set \( S_k \).

Single frequent element in \( D \rightarrow S_i \)

Procedure AprioriGraph(\( D,\text{min}\_sup, S_k \))

\[
\begin{align*}
S_{k+1} & \leftarrow \text{NULL}; \\
& \text{For each } g_i \in S_k \text{ do} \\
& \quad \text{For each } g_j \in S_k \text{ do} \\
& \quad \quad \text{Integrate } g_i \text{ and } g_j \text{ to form a } g \text{ with scale of } k+1; \\
& \quad \text{For each } g \text{ do} \\
& \quad \quad \text{If } g \text{ is a frequent graph in } D \text{, and } k \leq S_{k+1} \text{ then} \\
& \quad \quad \quad g \rightarrow S_{k+1}; \\
& \quad \text{End; End; End;} \\
& \text{If } S_{k+1} \neq \text{NULL} \text{ then} \\
& \quad \text{AprioriGraph(} D,\text{min}\_sup, S_{k+1} \); \\
& \text{Return;} \\
\end{align*}
\]

This algorithm integrates the characteristics of AGM, FSG and path connection approach. These approaches use adjacency matrix to build up on the vertex and edges in the graph so as to get the frequent sub-graph. The bottom-up principle is used for obtaining the frequent sub-graph in this algorithm.

5. Experiments and Discussion

This section reports on the experiments carried out for evaluating the algorithm proposed in this paper to mining the RFID sensor data from the network to figure out the frequent path. In these experiments, the RFID data is regarded as a set that \( D = \{G_1, G_2, \ldots, G_N\} \), where \( G_i = \{I_{1,i}, I_{2,i}, \ldots, I_{k,i}; G_i\} \); \( I_{1,i}, I_{2,i}, \ldots, I_{k,i} \) are the attributes of the RFID data. \( G_i = (V_i, E_i) \) is individual graph, in which every vertex or edge has its attribute.
Table 1. RFID Data for Experiments

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Time</th>
<th>Amount</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2001</td>
<td>300</td>
<td>y, z, n</td>
</tr>
<tr>
<td>A</td>
<td>2002</td>
<td>700</td>
<td>y, s, h</td>
</tr>
<tr>
<td>B</td>
<td>2001</td>
<td>200</td>
<td>y, z, n</td>
</tr>
<tr>
<td>C</td>
<td>2002</td>
<td>115</td>
<td>y, z, h</td>
</tr>
<tr>
<td>C</td>
<td>2006</td>
<td>100</td>
<td>y, z</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>B</td>
<td>2008</td>
<td>230</td>
<td>y, z, h, n</td>
</tr>
<tr>
<td>D</td>
<td>2004</td>
<td>1000</td>
<td>y, z, h, s</td>
</tr>
</tbody>
</table>

Table 1 demonstrates the RFID data from a warehouse with all the product movement information from 2001 to 2008. Within the data, there are non-path dimension and transportation path. Based on the RFID data, Figure 1 shows the graphs from the data sources. The set information includes (Product Type, Time), the node attribute is (Location, Amount), edge is the movement of products, The attribute for edge is the transportation times.

Figure 1. Graphs Established from RFID Data

Using the above data, the experimental environment is Pentium IV, 2GHz CPU with 4G memory. The operation system is Windows XP. The data comes from 2001 to 2008 which has been used for building up 9230 graphs. In this experiment, the algorithm proposed in this paper compares with the FFSM algorithm in terms of running time and performance.
From Figure 2, under the support threshold, the comparisons of running time from FFSM and proposed algorithm (RMine) are observed that the proposed algorithm in this paper outperforms than FFSM. This algorithm is able to avoid the direct omorphism testing among graphs, using algebraic approach to deal with the omorphism sub-graph efficiently. This algorithm uses the minimal code to represent a marked graph so that the amount of graphs is greatly reduced. That means the calculation basis is cut down. As a result, the running time could be significantly saved.

In order to verify the feasibility of the algorithm on RFID massive data, experiments are carried out on 6 groups of different number of dimensions. 10,000 patterns are used for this testing. It is observed that as the increasing of dimensions $Q$ given by the users, the targeted graphs will be greatly reduced. When carrying out the query, the backup graphs are cut down as well. That results on great reduction of searching time. The frequent graphs obtained are also reduced. The output of the graph will be decreased, removing some useless graphs from the raw RFID data sources. It also observed from this experiment, that, the algorithm is able to perform well for mining the frequent graphs from RFID data sets in terms of cutting down the mining time as well as saving the storage space.

6. Conclusions

In order to address the massive and large RFID logistics data sets, current methodologies focus on the OLAP using the data Cubic for mining the frequent path. Such methodologies are not suitable for RFID-enabled logistics data due to its efficiency and effectiveness. This paper proposes a graph theory-based mining approach for working out the frequent logistics path from RFID data. The approach is able to provide multi-dimensional views of logistics path after the establishing of graphs from RFID logistics data. Based on some definitions, the minimal support threshold is defined to figure out the frequent graphs. This algorithm is compared with FFSM in terms of running time and feasibility. This algorithm outperforms in the time and space perspectives, having high efficiency and effectiveness.

Future research directions will be carried out through the following aspects. First of all, when doing data compress, the data update in RFID dataset will be carried out frequently. During this process, large number of disturbances will be occurred. Studies on minimal support threshold and other thresholds will be concerned so as to minimize the errors. Secondly, how to define the minimal support threshold by using a mathematical model scientifically needs further investigation as well. Finally, the optimization of real-life applications will be further considered. The algorithms in this
paper maybe used for a decision support system, which is able to assist different users for making logistics and supply chain decisions. After that, the algorithms and approaches from this paper could be implemented optimally in real RFID logistics industry.

References


Authors

Yi Lianjun

He received bachelor's degree in physics in the School of Science and Technology; Gannan Normal University. He is current a lecturer in Gannan Normal University. He is studying master degree on computer application technology in School of Information Engineering, Jiangxi University of Science and Technology. His research interests are computer networks and data mining. He has been published four papers on journals and conferences.