CBADE: Hybrid Approach for Duplication Detection and Elimination

A. Anny Leema, P. Sudhakar and M. Hemalatha

Abstract Radio Frequency Identification (RFID) is an automatic data capturing technology and the dirty data stream generated by the RFID readers is one of the main factor limit the widespread adoption of RFID technology. In order to provide reliable data to RFID application, it is necessary to clean the collected data before they are subjected to warehousing. In this paper we are going to construct the elegant hospital environment using RFID and developed the cellular based approach algorithm to clean the duplication anomaly. First middleware approach is applied to deal with low complex anomalies and in next stage deferred approach is followed to deal with high complex anomalies based on business context. Simulation shows our cleansing approach for duplication removal deals with RFID data more accurately and efficiently. Thus we can bring down the health care costs, optimize business processes, streamline patient identification processes and improve patient safety.

1 Introduction

RFID (radio frequency identification) technology uses radio waves to transfer data between readers and movable tagged objects. In a networked environment of RFID readers, enormous data is generated from the proliferation of RFID readers. In RFID

A. Anny Leema (✉)
Karpagam University, Coimbatore, India
e-mail: annyleema@gmail.com

P. Sudhakar
Department of Computer Science and Engineering, M. Kumarasamy College of Engineering, karur, India
e-mail: navaladiperiyasamy@gmail.com

M. Hemalatha
Department of Software Systems, Karpagam University, Coimbatore, India
e-mail: hema.bioinf@gmail.com

N. Chaki et al. (eds.), *Computer Networks & Communications (NetCom)*,
Lecture Notes in Electrical Engineering 131, DOI: 10.1007/978-1-4614-6154-8_72,

environment, the database becomes more pervasive, therefore, various data quality issues regarding data legacy, data uniformity and data duplication arise. The raw data generated from the readers can't be directly used by the application. Thus, the RFID data repositories must cope with a number of quality issues. These data quality issues include data redundancy, false positive and false negative. Data quality has become increasingly important to many organizations [1]. This is especially true in the health care field where cost pressures and the desire to improve patient care drive efforts to integrate and clean organizational data.

The RFID technology has major benefits as follows:

- RFID technology can recognize information on multiple products at the same time as a long-sensing range.
- Mobile tracking devices can be reused or disposed, as the RFID operation requires.
- RFID does not require line-of-site communications between a receiver and a transmitter. This fact increases the range of RFID applications.
- RFID can work in the very harsh environment.
- RFID can function with low maintenance cost and without human interaction.
- Unlike barcodes, certain RFIDs can store data, allowing changes in the objects handling and processing.

2 RFID Data and its Components

Data generated from an RFID application can be seen as a stream of RFID tuples of the form (EPC; location; time) where EPC is a unique identifier code read by an RFID reader, location is the place where the RFID reader that reads the Tag and time is the reader captures the Tag's EPC. Tuples are stored based on the chronological Timestamp. An RFID reader scanning of tags can either be programmed to work at a fixed time interval or on a continuous basis.

RFID composed of three components—an interrogator (reader), passive tag(s), and a host as shown in Fig. 1. Among the types of tags—passive, active and semi passive, passive tags have much demand due to their least system cost and long life. The tag is composed of an antenna coil and a silicon chip that includes basic modulation

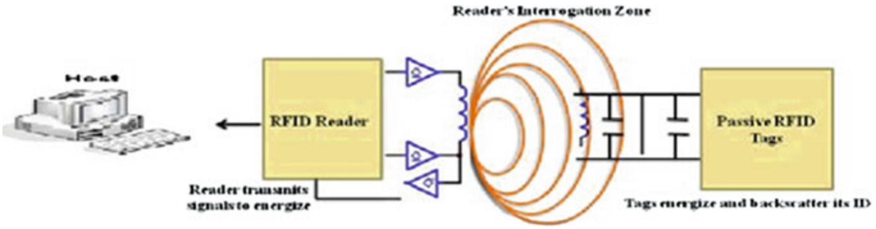

Reader's Interrogation Zone
Host
RFID Reader
Passive RFID Tags
Reader transmits signals to energize
Tags energize and backscatter its ID

Fig. 1 RFID and its components

circuitry and non-volatile memory. The tag is energized by a time-varying electromagnetic radio frequency (RF) wave that is transmitted by the reader. When the RF field passes through an antenna coil, an AC voltage is generated across the coil which is rectified to supply power to the tag. The tag using the mechanism of backscattering transmits its ID to the reader. By detecting the backscattering signal, the reader demodulates the received signal to retrieve tag's ID.

3 RFID System Design

RFID System design is depicted in Fig. 2. Raw RFID data stream is a spontaneous and very complex data to use for any analysis. The Readers are the detection nodes and are deployed in different locations across various places in the Hospital. Each detection node is identified by a unique ID that serves as the location ID. RFID tags in different locations are detected by these readers. One of the biggest challenges of the RFID data is the data volume. Sending terabyte data in to a centralized system for data cleaning requires a high performance server as well as a high speed network, which will inevitably increase the total hardware cost. Some of the data cleaning methodologies apply to data fetched by the readers, some requires an RFID middleware and others require a centralized data processing server to handle the raw data. The server level data observations include data validations, data inconsistencies and identification of anomalies before entering the enterprise application database. Data Inaccuracies are inevitable in the RFID system considering the complexity of deployment and diverse business needs it caters to.

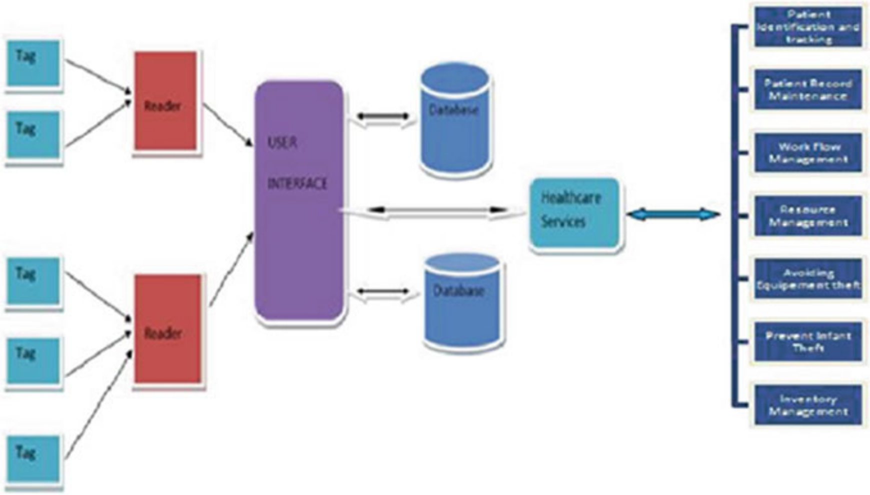


Fig. 2 RFID system design

4 Smart Hospital Environment

This paper discusses how Radio Frequency Identification (RFID) technologies were used in a hospital to construct smart environment [2, 3]. The accuracy and cost effectiveness of RFID is not guaranteed because of the issues we discovered by deploying this technology. Another issue involved managing the volume of data collected during the day. It is decided to record all the values associated with each tag event for future reference otherwise too much valid data will be lost. Finally the management can analyse the data and filter by applying business rules based on the requirement.

5 Issues in RFID Data

In the working mechanism of RFID reader and RFID tag, raw RFID readings are not always reliable [4]. Other causes of failed reads include the presence of metal in the tag vicinity, since it distorts the magnetic flux, thus weakening the energy coupling to the tag [5].

5.1 False Positive Error

An RFID reader periodically sends out RF signals to its range. When an RF tag that moves within the range of the reader receives the signals, it will send response signal along with its unique identifier code, timestamp and location ID. The reader receives the response signal and will register the data stream as one entry.

There would be some RF tags which is not supposed to be detected by the reader may be read due to the spatial divergence of RF signals sent by the reader. Such readings are termed as false positive readings.

These are predominantly caused by the RFID tags outside the normal reading scope of a reader, captured by the reader or unknown reasons from the reader or environment. The information is stored periodically by the reader through the middleware application to the database.

5.2 False Negative Error (Missing Data)

The raw RFID data streams do not provide a correct representation of the physical world that they are representing [6]. A significant number of tags which are within the reader's read range are not consistently read by the reader due to either their

orientation with respect to the reader, distance from the reader, presence of metal, dielectric or water material close to the tag and other factors [7].
Although theoretically speaking, all the tags would be read seamlessly and simultaneously on every read cycle; practically few of the tags might not be read in every cycle though present in the effective detection range. Such missed readings are deemed as false negative readings or missed readings.
These problems are very common in RFID applications and often happen in a situation of low-cost and low-power hardware [8], and the detection ability of a reader and environmental constraints. These missing tags imply that typically only a considerable percent of the tag population is actually observed on every read cycle.

5.3 Redundancy in RFID System

(i) Reader Level Redundancy

Redundancy at the reader level is the result of Cellular architecture where overlapping of reader coverage range is unavoidable. This problem occurs when a tag is in the vicinity of more than one reader at a specific time. As all the tags communicate simultaneously with the readers by sending RF signals, two are more reader reads the data from a single Tag.
Consider a scenario where readers R1, R2 and R3 are redundant since the tag T1 is read by all three readers at the same time thus responsible for reader level redundancy.

(ii) Data Level Redundancy

RFID data are usually streams of data and hence the redundancy on the data level has always been handled in the general way of dealing with data streams which might not be the optimal solution considering the uniqueness of RFID data. RFID applications handle humongous data as some readers can read hundreds of tag readings in a second. RFID data stream is considered as spontaneous and periodical in nature. The RFID data on average is highly useful than other data streams. The less useful part of RFID data is the data that are continuously reported after the initial reading.

For instance, in Hospital Management System, a tagged entity, (Say a doctor) may move to his consulting room and sit the whole day and send the data to the RFID management system constantly through the reader placed in his vicinity. But, from the management point of view, the most useful information for event detection is when the tagged entity (Say a Doctor) enters and exits his consulting room. Therefore, it is necessary to reduce RFID data redundancy before processing.

6 RFID Middleware

It serves as an interface between the RFID reader and Hospital management system database. RFID middleware serves as a platform which performs intelligent grouping of raw RFID data under predefined categories, to an extent filter raw RFID data stream based on anomalies, redundancy and preconditions [9, 10]. It is also responsible for mapping the low-level data stream from readers to a more manageable form that is suitable for application level interactions. The modules responsible for this mapping were called Savants in the original EPC work. Savants may be likened to the wrappers used in data integration systems and "edge systems" [11].

RFID middleware layer empowers healthcare providers by providing valuable data with a prompt connectivity. Application-level filtering based on exclusive process followed in each of the hospital services, validating data at different levels to ensure data consistency, monitor incoming data stream, provide real-time integration with the existing hospital management system, mapping data on to the relevant database table, redefining and executing business rule set are the various prime functions of the RFID middleware system. Figure 3 depicts this middleware concept. Data loading and extraction is the key to data intensive systems like RFID system.

6.1 Cleaning Data at the Time of Insertion

Step 1: Data insertion for source. (RFID Readings)
Step 2: Compare Input data stream with allowed data/character types. (Null, Alpha, Numeric, symbols)
Step 3: Check occurrence of similar incoming data streams for each set of streams using a for loop to identify data duplication.
Step 4: Display set of all duplicate data streams.
Step 5: Ability for the system to identify and delete redundant records with option for manual deletion.

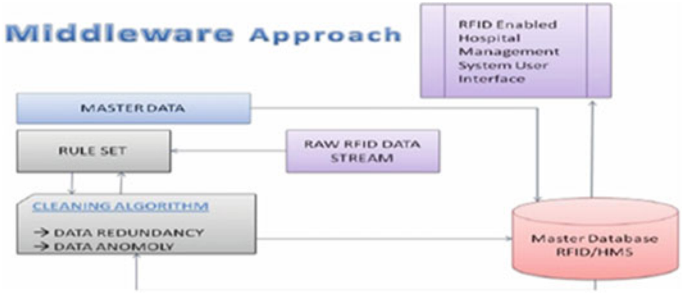


Fig. 3 Middleware approach

6.2 Identify Redundant Data Using Geometric Approach

Step 1: Get input data stream values
Step 2: Scan records in each of the rows
Step 3: Increment row index and perform search
Step 4: Automate search using for loop
The above said Geometric approach work finely to search a single value and does not suit for the huge enormous RFID data. So our approach is cellular based depicted in the Fig. 3. The diagram specifies different departments with unique Reader id which is location id.

6.3 Proposed Methodology and its Architecture

Proposed Approach is hybrid approach of middleware and deferred and the premise chosen is cellular based for detecting out of the range readings.
The RFID readers have Omni-directional antenna and hence there are possibilities for the adjacent regions to overlap with each other.
The Chosen premise to test our algorithm is depicted in the Fig. 4 and the architecture diagram of the proposed methodology is depicted in the Fig. 5.

6.4 Advantages of Proposed Approach

It is not always possible to remove all anomalies and redundancies beforehand. The rules and the business context required for cleansing may not be made available at the data loading time.

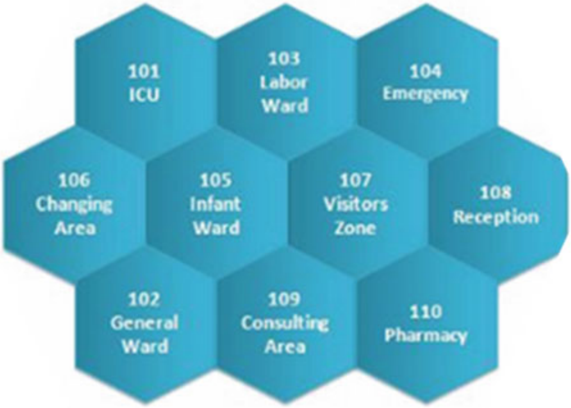


Fig. 4 Premise

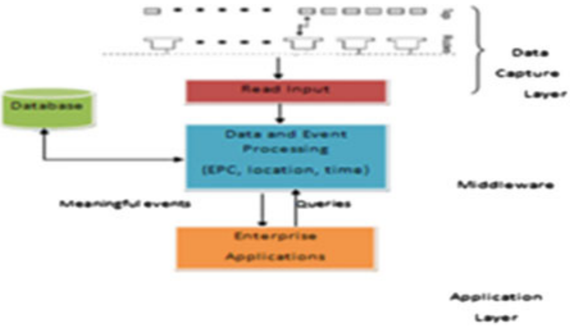


Fig. 5 System architecture

Maintaining multiple cleaned versions of RFID data are prohibitive and worthless when the rule sets defined by the application is dynamic in nature.

Known anomalies duplication are detected and corrected in the middleware approach and the meaningful data are sent to the database where processing of other anomalies is deferred until the query time.

Each application specifies the detection and the correction of relevant anomalies using declarative sequence-based rules. An application query is then automatically rewritten based on the cleansing rules that the application has specified, to provide answers over cleaned data [12].

6.5 *Proposed CBADE-Cellular Based Approach Algorithm for Duplication Detection and Elimination*

```
Algorithm CBADE ( Reader[ ], Tag [ ])
// Input: Reader R1 , Reader R2
// Input: Tag 1, Tag2, Tag3……Tag n
Begin
For (every tag in reader X (X=A, B))do
If
count (Selected Tag_Id) in all Tag_id > 1
Sub (Each Similar Tag_Id timestamp - Select TagId Timestamp)=0;
Return Duplication detected;
Select Max (Tag Id Timestamp);
Delete Other Tag_Ids;
else
return No Duplicates;
end for
End
```

7 Simulations

Simulation has long been used as a decision support tool in various sectors. There are several examples of simulation studies of healthcare facilities in the literature. Simulation is especially suited to the analysis of healthcare organizations due to its ability to handle high complexity and variability which is usually inherent in this sector [13]. It also acts as continuous quality improvement framework by integrating with the software agent developed via a database structure. Experimentation of different workflows, staffing decisions and what-if analysis are all promising applications of simulation in healthcare and it is practically infeasible in a healthcare environment. Simulation study requires deliberate data collection effort over a considerably long period of time. We have developed a simulator designed in C# with SQL Server 2008 as backend to evaluate the performance of the proposed algorithm. For simulation, the project premise chosen is given in Fig. 4 one RFID reader is assigned for each department with 100 tags in its interrogation zone is considered. The reader is modelled based on the design features of SkyeTek's M1—Mini RFID Reader. This reader operates from a Lithium rechargeable battery which has 0.48 KJ of energy. The tag to reader data rate is taken as 26 Kbps as per ISO 15693.
The following Assumptions to be made:

- Tag's antenna is never at 90°.
- RFID Readers are allowed to transmit energy until all tags are read.
- Communication from Tag to Reader is modeled as Poisson process.
- Reader has the knowledge on the number of bits present in a tag ID.
- Reader is unaware of the number of tags.
- Although tags are energized at the same time, the energy consumption is estimated only after the reading process has started.

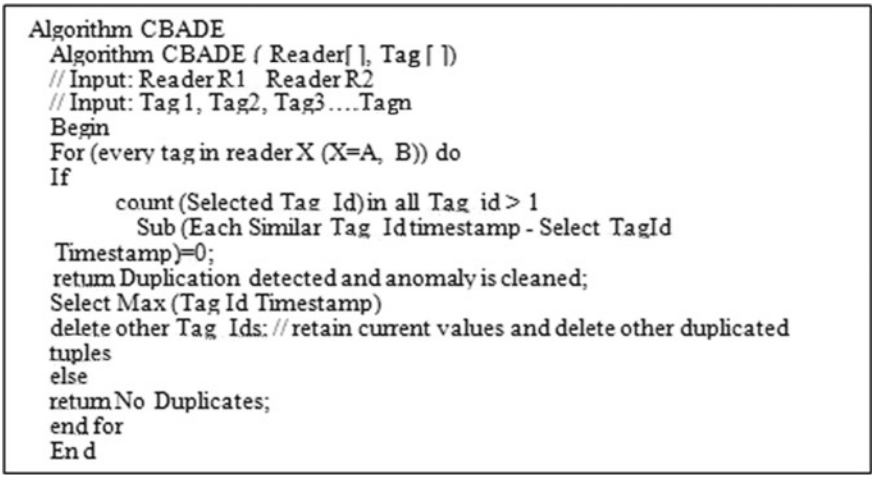
Case study: To test our algorithm the RFID readers are deployed in various departments in the hospital environment. The observed by the readers consist of anomalies. To clean the anomaly duplication the proposed algorithm is CBADE. The following table depicts the sample data observed by the RFID reader. The location assigned to the patient Hari is General ward (102) from 5.00 p.m. to 5.20 p.m. and the time reading is captured every 5 min.

Case	Patient Id Timestamp	Name Type	Ward	Reader Id	Date
1	P10004 5:00:00	Hari Normal	General Ward	102	10/4/2012
2	P10004 5:05:00	Hari Normal	General Ward	102	10/4/2012
3	P10004 5:10:00	Hari Normal	General Ward	102	10/4/2012
4	P10004 5:15:00	Hari Normal	General Ward	102	10/4/2012

Case	Patient Id	Name	Ward	Reader Id	Date
	Timestamp	Type			
5	P10004	Hari	General Ward	102	10/4/2012
	5:20:00	Normal			
6	P10004	Hari	Infant Ward	105	10/4/2012
	5:20:00	Adjacent			
Parallel					
7	P10004	Hari	General Ward	102	10/4/2012
	5:25:00	Normal			
8	P10004	Hari	General Ward	102	10/4/2012
	5:30:00	Normal			

In the above 1, 2, 3, 4, 5, 7, 8 tuples reader observed readings for P10004 are in the Allotted Location at the Allotted Time and Date. So this tuple is treated as normal and in 6th case the reading shows he is found to be in infant ward (105) which is an adjacent cell of General Ward (102) and P10004 is read by both 105 and allotted location (102) at the same time and date. Therefore the sixth case is treated as adjacent redundant (duplicate data) and with the human intervention based on business rule the duplicated data can be deleted.

7.1 Proposed Algorithm: CBADE



```
Algorithm CBADE
  Algorithm CBADE ( Reader[ ], Tag [ ])
  // Input: Reader R1   Reader R2
  // Input: Tag 1, Tag2, Tag3….Tagn
  Begin
  For (every tag in reader X (X=A, B)) do
  If
          count (Selected Tag Id)in all Tag id > 1
             Sub (Each Similar Tag Idtimestamp - Select TagId
   Timestamp)=0;
  return Duplication detected and anomaly is cleaned;
  Select Max (Tag Id Timestamp)
  delete other Tag Ids: // retain current values and delete other duplicated
  tuples
  else
  return No Duplicates;
  end for
  En d
```

7.2 Implementation

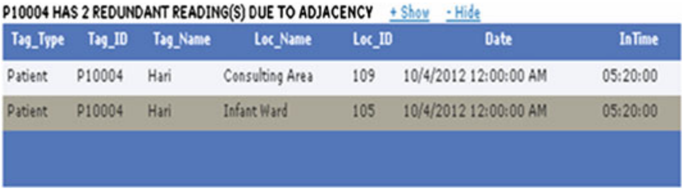
```
SqlDataReader dg15;
SqlCommand cm15;
d1 = "Update RFID_READING set Status ='Adjacent Parallel'
where EXISTS (SELECT * FROM Temp WHERE Temp.Time=
RFID_READING.Time and
Temp.Date= RFID_READING.Date ) and Tag_ID='" +
TagIDDropDown.SelectedValue + "' and Status='Adjacent'";
cm15 = new SqlCommand(d1, con);
dg15 = cm15.ExecuteReader();
dg15.Close();

protected void LinkButton11_Click(object sender, EventArgs e)
   {
Table = new DataTable();
string str = null, tag_ID = null;
tag_ID = TagIDDropDown.SelectedValue;
try
{
        str = "select * from RFID_READING where Tag_ID='" +
TagIDDropDown.SelectedValue + "' and
(Status='Adjacent Parallel' or Status='Crossover Parallel') order by Time";
SqlCommand cmd = new SqlCommand(str, con);
SqlDataAdapter ada = new SqlDataAdapter(str, con);
ada.Fill(Table);
TableGridView5.DataSource = Table;
TableGridView5.DataBind();
}
catch (IndexOutOfRangeException ex)
{
WebMsgBox.Show("Error:" + ex.Message);            }
```

8 Sample Output

Figure 6 depicts the output of the proposed algorithm CBADE. The allotted location of the person Hari is general ward and the schedule time is 5.00 to 5.30 p.m. Figure 7 depict the redundant data in RFID readings.

First Case: 109 is an adjacent cell and P10004 is read by both 109 and alloted location at the same time and date.

Second Case: 105 is an adjacent cell and P10004 is read by both 109 and alloted location at the same time and date.



P10004 HAS 2 REDUNDANT READING(S) DUE TO ADJACENCY

Tag_Type	Tag_ID	Tag_Name	Loc_Name	Loc_ID	Date	InTime
Patient	P10004	Hari	Consulting Area	109	10/4/2012 12:00:00 AM	05:20:00
Patient	P10004	Hari	Infant Ward	105	10/4/2012 12:00:00 AM	05:20:00

Fig. 6 Output of the CBADE algorithm

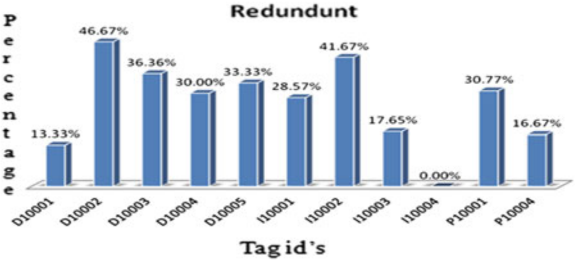


Fig. 7 Chart depict the redundant data in RFID readings

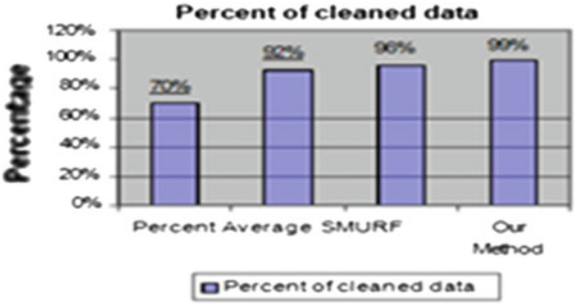


Fig. 8 Chart depict the redundant data in RFID readings

9 Comparative Studies

To show the practical relevance of our method for data duplicates elimination experimental evaluations have been carried out. We analyze the performance of different cleaning schemes. Our evaluation metric is the results of cleaned data. We simulate 1000 samples with 200 wrong Data and see results of cleaning algorithm.10 wrong data is remained. We compare our method with SMURF and other popular strategies depicted in Fig. 8. The average errors are calculated based on following equation

(falsenegatives+
falsepositives)/
NumTags

10 Conclusion

RFID plays an essential role in all the subdomains of the applications in health care applications. Necessary information is provided to patients and hospital staffs by recording and processing the medical data produced in each step, and ultimately this system can be used anytime and anywhere by managing the record of individual health-information. Additionally, these services provide for patients with medical information and guidance of hospital through I/o interfaces by reading individual ID from the medical card of patient with RFID reader and searching them in DB of control center. The effectiveness in cleaning the RFID data in healthcare sectors remains a concern, even though a number of literary works are available. To a maximum, the dirty data that are read because of these errors may even leads to patients' death. The errors need to be cleansed in an effective manner before they are subjected to warehousing. Current solutions to correct missed readings usually use time window filtering. A serious issue is that a single static window size cannot compensate for missed readings while capturing the dynamics of tag motion. An adaptive time window filtering (SMURF) cannot deal with the condition that tags are always moving. In this paper, we have proposed an improved algorithm CBADE to clean the anomaly duplication and the experimental result has proved our algorithm predicts and removes the data duplication in an effective manner compared to the existing works. Thus it will pave the way for an effective means of data warehousing system that will keep the RFID data safe for future mining.

References

1. Shepard S (2005) RFID Radio frequency identification. McGraw-Hall, New York
2. Durresi A, Merkoci A, Durresi M, Barolli L (2007) Integrated biomedical system for ubiquitous health monitoring. NbiS 4658(1):397–405
3. U.S. Government Accountability Office (2005) Radio frequency identification technology in the Federal Government, 441 G Street NW. Room LM Washington, DC 20548
4. Zhang C, Chen Y (2011) Application oriented data cleaning for RFID middleware. IEEE Trans Commun 59(1):159–168
5. Floerkemeier C, Lampe M (2004) Issues with RFID usage in ubiquitous computing applications. Pervasive 3001:188–193
6. McGlynn EA, Asch SM, Adams J et al (2003) The quality of health care delivered to adults in the United States. N Engl J Med 348(26):2635–2645
7. Shoewu O, Badejo O (2006) Radio frequency identification technology: development, application, and security issues. Pacific J Sci Technol 7(2):144–152

8. Aragonés J, Martínez-Ballesté A, Solanas A (2007) A brief survey on RFID privacy and security. In: Proceedings of the word congress on engineering WCE07. IAENG, pp 1488–1493
9. Uddin MJ, Ibrahimy MI, Reaz MBI, Nordin AN (2009) Design and application of radio frequency identification systems. Eur J Res 33(3):438–453. ISSN 1450–216X
10. Anny Leema A, Hemalatha M An effective and adaptive data cleaning technique for colossal RFID data sets in healthcare. WSEAS Trans Inf Sci Appl
11. Chawathe SS, Krishnamurthy V, Ramachandran S, Sarma S (2004) Managing RFID data. In: Proceedings of the 30th VLDB conference, pp 1189–1195
12. Rao J, Doraiswamy D, Thakkar H, Colby LS (2006) A deferred cleansing method for RFID data analytics. In: Proceedings of the 32nd VLDB conference, pp 175–186
13. Wicks AM, Visich JK, Li S (2006) Radio frequency identification applications in hospital environments. Hosp Top 84(3):3–9