An effective and adaptive data cleaning technique for colossal RFID data sets in healthcare.

A. ANNY LEEMA

Assistant Professor, Department of Computer Applications,

B.S.Abdur Rahman University, Vandalur, India

Research Pscholar , Karpagam University, Coimbatore, India.

annyleema@gmail.com

&

DR.HEMALATHA.M

Head, Dept of Software Systems,

Karpagam University, Coimbatore, India

hema.bioinf@gmail.com

*Abstract***:** Radio frequency identification (RFID) technology has seen increasing adoption rates in applications that range from supply chain management, asset tracking, Medical/Health Care applications, People tracking, Manufacturing, Retail, Warehouses, and Livestock Timing. This technology is used in many applications for data collection. The data captured by RFID readers are usually of low quality and may contain many anomalies. Data quality has become increasingly important to many organizations. This is especially true in the Medical/health care field because minute errors in it can cost heavy financial and personal losses. In order to provide reliable data to RFID application it is necessary to clean the collected data. SMURF is a declarative and adaptive smoothing cleaning technique for unreliable RFID data. However it does not work well when tag moves rapidly in and out of reader's communication range. The errors need to be cleansed in an effective manner before they are subjected to warehousing. Factors such as inter tag distance , tag-antenna distance, number of tags in the read range of antenna, reader communication range, velocity of tag movement affect the data cleaning result. Our proposed algorithm considers these factors and also the missing tag information, tags that are mistakenly read as present dynamically in determination of the size of slide window. Simulation shows our cleansing approach deals with RFID data more accurately and efficiently. Thus with the aid of the planned data cleaning technique we can bring down the health care costs, optimize business processes, streamline patient identification processes and improve patient safety.

Keywords: RFID technology, significance of data quality, RFID Architecture, RFID middleware systems, cleaning methods, Data cleaning approaches

1 Introduction:

Radio Frequency Identification (RFID) is a means to identify and track objects using radio frequency transmission. An RFID system consists of readers and tags. Readers use radio signals to communicate with the tags. Tags may be active (battery powered) or passive (powered by the reader's signals). A tag consists of a microchip applied to a miniaturized antenna, which carries a "unique tag ID" but can be programmed with other information. The reader captures the signals, decoded it and sends it to host computers. In both the reader and tag, the antenna can be sized and shaped in different ways. Because of the small size of the tag, RFID tag can be designed to fit almost any situation. Placement of the tag and reader is now no longer critical since there is no contact or line of sight is required; the RFID system allows great freedom of movement. RFID includes hardware, middleware and software components.

This technology is widely used in diverse application such as supply chain automation, asset tracking, medical/Health Care applications, people tracking, Manufacturing, Retail, Warehouses, and Livestock Timing. Of these, Medical/Health care applications are of more importance because minute errors in it can cost heavy financial and personal losses. For hospitals and healthcare systems, increasing the operational efficiency is the primary target. It is a tough task to keep up the effectiveness and monitor each and every patient [1]. However, utilization of RFID (Radio Frequency Identification) technology in addition to reducing the health care costs facilitates automating and streamlining patient identification processes in hospitals and use of mobile devices like PDA, smart phones, design of health care management systems etc.,

2 RFID in Healthcare: The Advantages

Healthcare management system has several RFID applications. Drug faking applications give a unique code and identity to medicine by placing RFID tags [8]. Fake medicines can be easily distinguished in case of illegal distribution. By using RFID tags, the expired medicines can also be kept under control. In order to have stocks in mass for emergency necessities RFID tags can facilitate distributors to monitor on the availability of medicines in inventorying and Stocktaking application. The efficiency of the suppliers and the distributors can be improved by this as it saves time.

The information about the status (cleaned/sterilized/unsterilized) of medical devices can be obtained by supplying those devices with RFID tags and placing RFID readers at the entrance of storing chambers [14]. Thus, assists to avoid devastating errors that may occur on utilizing contaminated instruments. The RFID tags can help saving time by providing the exact location of the medical devices inside the rooms/chambers. The seal and packaging of medicines may be uncertain. Read/write RFID tags can be extremely useful in resolving these problems. By placing the information related to the seal time, packaging date etc [7] on the RFID read/write tags they can be examined whenever required. It is also used to improve the security of a hospital by issuing tags for both employee and patient so that when a restricted area is entered an alarm would be triggered.

The RFID tags can be used to directly automate data management about the patient's health (report tests, medicines etc.) maintained on a database by means of the Patient Tracking and Information system [9]. By monitoring and recording the blood pressures and heart beat rates, the RFID tags can also convey information about the health status of patients [12]. This can assist the nurses/caretakers to have a better control on the patient's health and also enable the faster recovery of patients. RFID technology can be formed as an essential part of healthcare. The risk factors of human error are reduced by RFID in the healthcare industry [13].

The number of fatalities reported in American hospitals is far more than realistic. It has decreased by the use of RFID tags. There are even more difficulties in hospitals to match the newly born babies with their parents. RFID tag can be locked on baby's ankle. This tag will

track their position in the hospital. It will help to accurately match the babies with their parents. It will prevent the theft of new born babies. Often after a surgery, there is always the fear of leaving surgical tool in patient's body. A small RFID tag on every tool will help doctors to track every piece of instrument taken to the operating room will eliminate the fear of leaving surgical tool in patient's body. However, huge quantities of data [20] are created by these devices at high production rates because of the streaming nature of RFID readings [15]. So, an important problem with great potential benefits in the health care sectors is the warehousing and mining of RFID data.

3 Warehousing RFID data:

Data warehousing is the only viable solution for providing strategic information. In fact, they are very large repositories that integrate data coming from operational databases of several enterprise sectors for decisional analysis (illustrated in figure [1]. Data warehouses are implemented like databases and they are not designed for online transaction processing. Cross-system data integration is a relevant characteristic of data warehousing. Figure.1. depicts data warehouses which support decision support applications integrate data from multiple heterogeneous information sources by transforming them into a multidimensional representation [16, 17]. Time-dependency, dynamic changes, huge quantities and large numbers of implicit semantics are some of the characteristics of the RFID data. Data loading in a data warehouse is often a complex process involving data cleaning and transformation because quality of data in the warehouse is a primary goal in order to respond efficiently and accurately to unforeseeable queries at any moment of time [3].

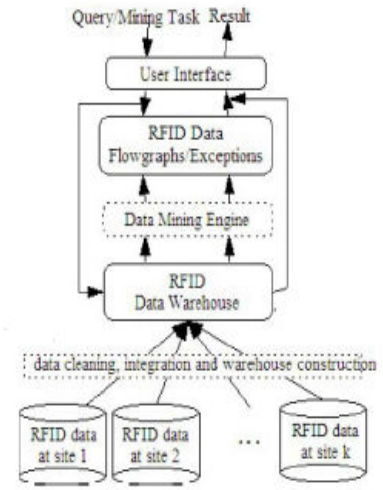

Fig.1. Data Warehousing

Figure.2. shows the three areas of software research into management of RFID data (as illustrated in figure2). The first area deals with the secure fetching and management of online tag associated information. The second area is related to removal of errors that are present in RFID because of error inaccuracies. The third area is concerned with providing OLAP operations over huge RFID data sets by constructing suitable multi-dimensional warehouse [19]. The major challenge of data warehousing is data cleaning [11] and our work focus towards it. Data cleaning is necessary for improving the quality of data so that it becomes "fit for use" by users. It improves the documentation and presentation of data by reducing the errors present in it. Auditing data for discovering discrepancies, selecting transformations to correct these, and employing the transformations on the dataset are the three components present in data cleaning [10].

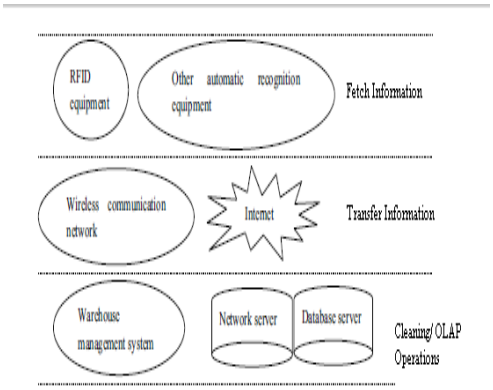

Fig.2. shows RFID data management

4 Significance of Data Quality:

Data quality has become increasingly important to many firms as they build data warehouses and focus more on customer relationship management. For health care organizations, data is central to both effective health care and to financial survival. Data quality was concerned with accuracy, precision and timeliness. It can best be defined as "fitness for use". [18] Recently the Institute of Medicine shocked the public with a report that 98,000 people die every year due to medical errors [8]. Some of the errors are the result of missing or bad information about drugs, orders and treatments. Poor data quality has adverse effects at the operational, tactical and strategic levels of an organization. At the operational level, poor data reduces customer satisfaction, increases operational costs and reduces employee job satisfaction [4]. Quality of data in the data warehouse is dependent on the quality of data in the various source systems and on the quality of the extraction, cleansing, transformation and transfer processes that make up the source- to-target transformation. Therefore data quality problems are identified and steps can be taken to monitor and improve the quality.

5 RFID Architecture:

RFID Architecture is depicted in the Figure.3. It comprises of the following:

- Firstly items-tag are scanned by reader;
- Secondly in backend transmitted data coming through antenna (RF-wave) are being recognized by RFID-based system PC. It acts as a middleware communication gateway among items, reader and system database;
- And at the end it filters out and store data in RFID-databases for checking the data fault and relevant operation.

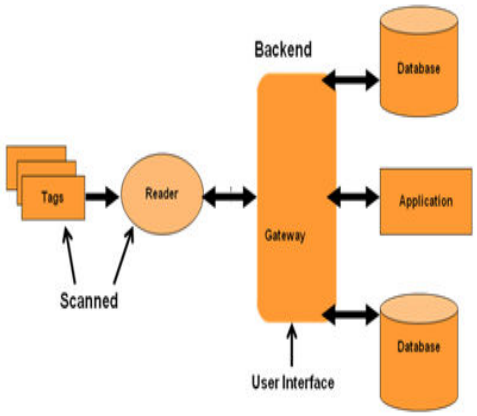

Fig.3. RFID Architecture

5.1 RFID Middleware Systems:

The accuracy of current RFID is improving, but there is still erroneous reading error, such as duplicate readings or missing readings. RFID reader antennas should have a number of characteristics such as: (i) good data transfer rate, (ii) robust, (iii) inexpensive, (iv) high radiation efficiency, (v) low side-lobe to reduce interference. We use RFID to track the objects instead of GPS for the following reasons 1) the cost of RFID Tag is cheap, 2) the size of RFID Tag is small, 3) it is easy to carry, and 4) the RFID positioning can be used in the buildings. Read reliability is defined as the probability that an RFID reader successfully detects and identifies an RFID tag when it is in the read range of one of the readers antennas.

Middleware becomes a very hot topic. The observed read rate in real-world RFID deployments is the range 60-70% and 30% of the tag readings are routinely dropped. In order

to correct reading errors, and allow data streams to meet the high level information requirements, we deploy the RFID middleware system between reader and applications. The concept behind this middleware is using a non-overlapping static window or an overlapping sliding window RFID approach which interpolates dropped readings for every tag within the time window on the collected data stream. The standard data-cleaning mechanism for today's RFID middleware systems is a temporal "Smoothing filter". The goal is to reduce or eliminate dropped readings by giving each tag more opportunities to be read within the smoothing window and then providing data to a higher level end application. While the APIs for RFID middleware systems vary, smoothing functionality can be expressed as a simplified continuous query language. For Example for a 3 sec window we have the following query to be executed. This query states that if the tag appears at least once in the window, it is considered to be present for the entire window.

SELECT distinct tag_id FROM RFID_stream [RANGE ' 3 sec']
GROUP BY tag_id

In order for RFID technology to become feasible, RFID middleware must be able to produce reliable streams describing the physical world.

6 Hospital/Healthcare takes off with RFID:

RFID is an amazing technology holds an unlimited number of benefits within healthcare industry. In this section first we describe how simulation can be done and we describe how to build an elegant hospital environment and finally the provided healthcare services.

6.1 Simulation in Healthcare:

Simulation has long been used as a decision support tool in various sectors. It 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. Experimentation of different workflows, staffing decisions and what-if analysis are all promising applications of simulation in healthcare. However, a typical simulation study requires deliberate data collection effort over a considerably long period of time.

6.2 Hospital Patient Management System:

The RFID patient tracking kit consists of RFID wristbands, a PDA handheld reader, a desktop HF reader and necessary software. Because of automated data capture, the RFID patient tracking kit brings improved efficiency. The waterproof, non-allergic wristband can be reprogrammed to enable patient information to be stored and transferred to and from RFID readers, information systems, and medical devices in hospital [2]. The Handheld RFID reader is used to receive the patient real time information just beside the beds, whereas desktop reader is used to read/write wristband's information beside computer to save time. Hospitals can use this RFID patient tracking kit to boost efficiency and accuracy while reducing costly and dangerous errors, and giving patient more privacy. [5, 6] Patients are monitored in many hospitals whether proper care is given or not. These systems tend to reduce the data-entry workload of nurses, and also let them spend more time caring for patients and automate the process of billing. Additionally, hospitals are tracking high-value assets, including gurneys, wheel chairs, oxygen pumps and defibrillators. These systems reduce the time employees spend looking for assets, improve asset utilization and enhance the hospitals' ability to performed scheduled maintenance. Patient bracelets embedded with RFID technology securely tracks patient movement from admission to discharge.

An Active Wave RFID system can be used to track patients, doctors and expensive equipment in hospitals in real time. RFID tags can be attached to the ID bracelets of all patients, or just patients requiring special attention, so their location can be tracked continuously. It also Restrict access to drugs, pediatrics, and other

high-threat areas to authorized staff. The important data (e.g., patient ID, name, age, location, drug allergies, blood group, drugs that the patient is on today) can be stored in the patient's back-end databases for processing. The databases containing patient data can also be linked through Internet into other hospitals databases [5]. The Patient Management System's administrator can issue unused tag (wristband) to every patient at registration time. Healthcare professionals (e.g., doctors, consultants) can edit/update password protected patient's medical record for increased patient and data security by clicking the Update Patient Button. This PMS can be implemented in departments (e.g., medicine, surgery, obstetrics and gynecology, pediatrics) in both public and private hospitals for fast and accurate patient identification without human intervention. Fig [4] shows our proposed health care services which are given below:

(i) Patient Identification & Tracking
(ii) Patient Record Maintenance
(iii) Work flow Management
(iv) Resource management
(v) Avoiding Equipment Theft
(vi)Prevent Infant Theft
(vii)Inventory management

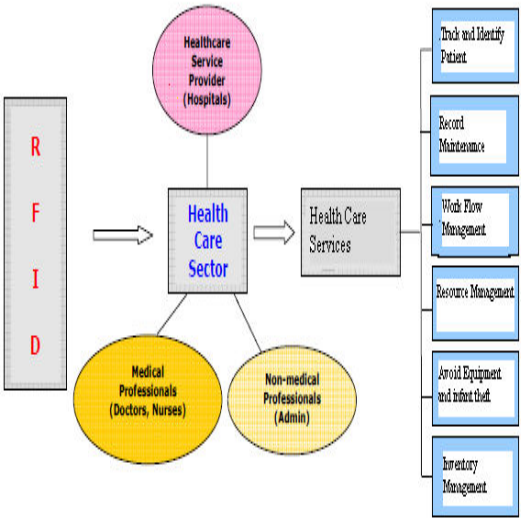

Fig.4. Health care application

In order to provide the above said health care services the followings are required:

a) Issue of medical card attached RFID tag.

b) The collection and store of much kind of data.

c) The extraction of individual ID through RFID reader for abstracting individual information from RFID tag to enter within a limited area.

d) The acquisition and verification of security information.

e) The receipt of requests through input/output interfaces.

f) The search of requested data according to extracted individual ID

g) The provision of searched data.

6.2. Health care Process Requirement:

To study the following:

1. Existing process flow in a hospital unit
2. Existing Manual data processing methodology
3. Existing record maintenance system that is in practice
4. Identify Tangible benefits of RFID implementation
5. Identify Implementation pitfalls

6.3. Sample Data Requirement:

We require the following:

1. Patient entry record data
2. Staff record maintenance
3. Patient health history report
4. Scheduling and work flow management document
5. Technical document – Department wise

7 RFID Data Cleaning:

7.1 Data Cleaning:

The major challenge of data warehousing is data cleaning. It is necessary for improving the quality of data so that it becomes "fit for use" by the users.

7.2 Issues in Data Cleaning

7.2.1 Lack of Completeness

a) RFID readers capture only 60-70% of all tags that are in the vicinity.
b) Smoothing of data is done to rectify the loss of intermediate messages.

7.2.2 Temporal Nature of data or tag dynamics

a) RFID tags are in motion and that is what makes them more difficult to handle.
b) But motion of a tag causes dropping of messages.
c) RFID data streams are very fast and are huge in number

Hence filtering is important before sending them to database

7.3 Types of Errors:

The errors occur in the process of data capture are false negatives, false positives, and duplicates. False negative is that tags are in the vicinity of a reader but not detected by it. False positive refers to that a tag is not present but captured. Besides RFID tags to be read, additional unexpected reading are generated [18]. Duplicate readings refers to tags are in the scope of a reader for a long time and are read by the reader multiple times. It also occurs due to the tags in the overlapped areas read by multiple readers

7.4 Data Cleaning Approaches:

The approaches in RFID data cleaning are studied by many authors. Table 1 depicts this. They are fix-size slide window [20], EPS (Extensible sensor stream processing) and SMURF (Statistical smoothing for unreliable RFID data). Fix-size slide window approach uses the time window to determine whether the data is right or not. In EPS (Extensible Sensor Stream Processing) the static size of the window is the limitation of the approach because large window induces false positives and small window cannot fill false negatives. ESP organizes receptor stream cleaning into a cascade of five programmable stages: Point-Smooth-Merge- Arbitrate-Virtualizes. ESP captures the context of temporal and spatial application layers by introducing the concepts of temporal and spatial granularity. SMURF is introduced to reduce the reading dropping rate on the statistical basis, and incorporating cleaning logic in applications produce more reliable RFID data stream. It dynamically adjusts the size of the window to pre-treat the RFID data. However SMURF does not work well when tag moves rapidly in and out of reader's communication range. Through an adaptive smoothing filter for cleaning raw RFID data streams, this greatly increases the complexity of the application. To ensure that all tags are in the reader's detection range are read, the smoothing window must be large enough to correct for reader unreliability. Small window sizes will easily omit some tags, causing a false negative. On the other hand, adopting large windows will limit the detection of the tag's movements (or reader's movement) resulting in a false positive (Jeffery et al., 2008).

SMURF has weakness in dealing with the size of slide window. Another weakness is determining the parameter delta. SMURF assumes that tag does not move frequently. When it frequently moves in and out of a reader's communication range, SMURF does not work well and it leads to higher error rate.

Comparative study of Data Cleaning Approach

Approach	Representative Works
Static window approach: It uses a single window for all the tags, which is unable to catch the variation.	Cleaning for RFID data streams
Adaptive window SMURF is non parametric but sampling based Adaptive data cleaning method Mechanism for per-tag and multi-tag cleaning Doesn't work well when tag moves rapidly in and out of readers communication range Increases the complexity of the application	Adaptive cleaning for RFID streams
Our framework: Good adaptation. Complexity is reduced. Good flexibility Uncertainty handling is high It can operate with or without training data. As it is made adaptive, the cleaning operation will automatically adjust itself to focus more on more erroneous data and less on less erroneous data It find the solution for false positive, missing readings and false negatives	An effective and adaptive data cleaning technique for RFID in healthcare.

8 The Proposed Methodology:

Obviously, to ensure a tag can be captured, window size must be large enough to fill missing readings. On the other hand if the window size is too large, the data is filled incorrectly and results in false positives. In real world many factors that affect the reliability of data. They are Inter tag distance, the orientation of the tags with respect to the antenna, tag - antenna distance, number of tags in the read range of antenna and the speed of the tagged objects etc. The proposed technique will consider missing tag information, tags that are mistakenly read as present and also the above said factors. The cleaning is planned to perform in an adaptive manner. As it is made adaptive, the cleaning operation will automatically adjust itself to focus more on more erroneous data and less on less erroneous data. By performing this adaptive kind of operation, the computational complexity is expected to be less. This will have benefit of spending more computational cycles on cleaning more erroneous data. Hence, an effective cleaning will be accomplished for the RFID data. This will pave the way for an effective means of data warehousing system that will keep the RFID data safe for future mining. Thus, with the aid of the planned data cleaning technique, the patients under treatment will be accurately tracked and monitored.

9 Conclusions:

RFID plays an essential role in all the sub-domains of the applications in health care applications. Among them, RFID technology dominates in tracking the patients under treatment. But, there may be errors and redundancies in the obtained RFID data from all the readers. The effectiveness in cleaning the RFID data in healthcare sectors remains a concern, even though a number of literary works are available. The probability of occurrence of the error due to the wrong consideration about the presence of the tags is high. This happens because of the electronic devices that are operating in the same RF range. The RFID readers may read an unavailable tag because of the interference generated by the malicious electronic device. This gives rise to erroneous reading by the corresponding RFID reader. 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. SMURF is one of the recognized data cleaning approaches. However it does not have good performance when tag moves rapidly in and out of reader's communication range. In this paper, we have proposed an improved algorithm which considers the factors such as reader communication range, tag - antenna distance, number of tags in the read range of antenna and the speed of the tagged objects etc. Thus the errors are cleansed in an effective manner before they are subjected to warehousing and 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] M.J. Uddin, M.I. Ibrahimy, M.B.I. Reaz and A.N. Nordin, "Design and Application of Radio Frequency Identification Systems", ISSN 1450-216X , Vol.33 No.3 (2009), pp.438-453, Euro Journals Publishing, Inc. 2009.

[2] Tseng, Vincent S., Lu, Hsueh-Chan, Tsai, Chia-Ming, Wang and Chun-Hung, "A Hybrid Data Mining Approach for Analysis of Patient Behaviors in RFID Environments", 2008.

[3]. H. Gonzalez, J. Han, X. Li and D. Klabjan, "Warehousing and Analysis of Massive RFID Data Sets", In Proceedings of 2006 Int. Conf. Data Engineering (ICDE'06), Atlanta, Georgia, April 2006.

[4]. Angela M. Wicks, John K. Visich and Suhong Li, "Radio Frequency Identification Applications in Hospital Environments", heldref publication, Vol. 84, No.3, pp.3-9, 2006.

[5]. Belal Chowdhury and Rajiv Khosla, "RFID-based Hospital Real-time Patient Management System," ICIS, pp.363-368, In proceedings of 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007), 2007.

[6]. Sudarshan S. Chawathe, Venkat Krishnamurthy, Sridhar Ramachandran, and Sanjay Sarma, "Managing RFID Data", In Proceedings of the 30th VLDB Conference, pp.1189-1195, 2004.

[7] O. Shoewu and O. Badejo, "Radio Frequency Identification Technology: Development, Application, and Security Issues", The Pacific Journal of Science and Technology, Vol.7, No.2, November 2006.

[8]. A. Lahtela, "A Short Overview of the RFID Technology in Healthcare", Systems and Networks Communications, 2009, ICSNC '09, In Proceedings of Fourth International Conference on, pp.165-169, 2009

[9]. Stephen A. Weis, Sanjay E. Sarma, Ronald L. Rivest and Daniel W. Engels, "Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems", In Proceedings of 1st Int'l Conf. Security in Pervasive Computing, springer Verlag, Vol . 2802 , pp. 201-212, 2004.

[10]. Vijayshankar Raman and Joseph M. Hellerstein, "Potter'sWheel: An Interactive Data Cleaning System", In Proceedings of the 27th International Conference on Very Large Data Bases, pp.381 - 390, 2001.

[11]. Panos Vassiliadis,Christoph Quix,Yannis Vassiliou and Matthias Jarke, "Data warehouse process management", In proceedings of the 12th international conference on advanced information systems engineering, Vol. 26,Issue 3, pp.205 - 236, 2001.

[12]. F. A. Correa, M.J. Gil, and L.B. Redín, "Benefits of Connecting RFID and Lean Principles in Health Care", Working Paper 05-4, Business Economics Series 10, 2005.

[13]. S. Shepard, "RFID Radio Frequency Identification", The McGraw-Hall Companies, Inc. USA, 2005.

[14]. J. Fisher and T. Monahan, "Tracking the social dimensions of RFID systems in hospitals", In Proceedings of International Journal of Medical Informatics, Vol. 77, Issue 3,

pp. 176 -183, 2007.

[15]. Bettina Fazzinga, Sergio Flesca, Elio Masciari and Filippo Furfaro, "Efficient and effective

RFID data warehousing", In proceedings of the 2009 International Database Engineering & Applications Symposium,pp. 251-258 , 2009.

[16]. R. Derakhshan, F. Dehne, O. Korn and B. Stantic, "Simulated Annealing for Materialized View Selection in Data Warehousing Environment," Proceedings of the 24th IASTED international conference on Database and applications, pp. 89-94, 2006

[17] Hector Gonzalez, Jiawei Han and Xiaolei Li, Mining compressed commodity workflows from massive RFID data sets", In proceedings of the 15th ACM international conference on Information and knowledge management, pp. 162 – 171, 2006.

[18]Hector Gonzalez, Jiawei Han, Hong Cheng, Xiaolei Li, Diego Klabjan, Tianyi Wu, "Modeling Massive RFID Data Sets: A Gateway-Based Movement Graph Approach," IEEE
Transactions on Knowledge and Data Engineering, vol. 22, no. 1, pp. 90-104, Jan. 2010.

[19] Robert L. Leatherier, Data quality in health care data warehouse environments, Proceedings of the 34th Hawaii International conference on system sciences -2001.

[20] Manage data successfully with RFID anywhere Edge Processing.

http://www.ianywhere.com./developer/rfid_anywhere/rfidanywhere.ed