An Improved Algorithm for Reducing False and Duplicate Readings in RFID Data Stream Based on an Adaptive Data Cleaning Scheme

*^Neha Dhama *Meghna Sharma*

^ Dept. of CS&IT, ITM UNIVERSITY, GURGAON

**Assistant Professor, Dept. of CS&IT, ITM UNIVERSITY, GURGAON*

***Abstract*— RFID technology is increasing rapidly and being successfully applied to various sectors such as supply chain management, manufacturing, Retail, warehouses, wall-mart and health care applications. The data stream produced by RFID system is sometimes unreliable, inaccurate, low level. As a result, there is a heavy load of unreliable data which is useless for higher level application. Therefore, RFID data cleaning is the most important task for the successful deployment of RFID systems. Unreliable data appears in three forms of reading errors: false positive (unexpected), false negative (missed) and duplicate readings. Most common technique which is used by RFID middleware is the use of sliding window filters. In this paper, based on the individual tag reading environment, false readings and duplicate reading errors are reduced by a new adaptive data cleaning window scheme called WSTD. This technique is used along with some of the concepts of SMURF but with an improved transition detection mechanism as SMURF has some drawbacks. The result has been evaluated on RFID database. Also, comparison has been done with the existing approach i.e. SMURF. Simulation shows our approach deals with RFID data more accurately.**

***KeywordS*— RFID, data cleaning, sliding window filtering, WSTD, SMURF, RFID middleware**

I. INTRODUCTION

RFID is a wireless technology that uses radio-frequency waves that identifies and tracks objects automatically. It is emerging as a key technology for a wider range of applications such as supply chain, Retail store, access control for vehicles etc. Unlike barcode reader, RFID has a long read range. In the near future it offers a possible alternative to barcode reader. RFID exchanges data between reader and tags without line of sight. As the data capture by RFID reader wirelessly identifies tags which are sometimes not reliable. In this process of data capturing error occurs including false positives, false negatives and duplicate readings.

False positive reading: It happens when tag is present but not detected.

False negative reading: It happens when tag is not present but unexpectedly captured.

Duplicate reading: It happens when there is duplicate occurrence of tagid.

RFID middleware systems are typically deployed between reader and the applications in order to correct for dropped readings and provide “clean” RFID readings to application. The standard data-cleaning mechanism in most of the systems is a temporal “smoothing filter”: a sliding window over the reader’s data stream that interpolates for lost readings from each tag within the time window. The goal, of course, is to reduce or eliminate dropped readings by giving each tag more opportunities to be read within the smoothing window.

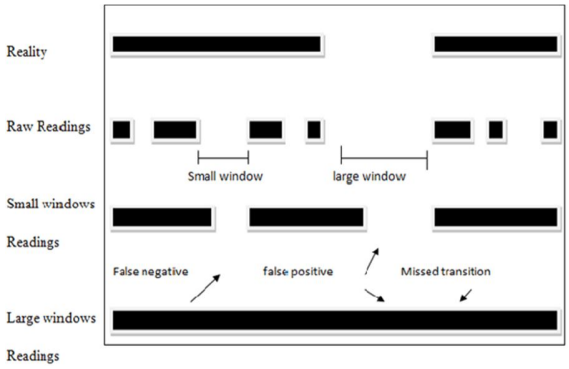


Fig3.1 shows the tension in setting the window size

II. PROPOSED WORK

A. *Data Mining Tools Applied*

In this, an assumed RFID data set is generated with attributes EPC (Electronic product code), LOCATION, TIME_IN, TIME_OUT) [1] and one discrete attribute READING ERROR and then applied this data set on data mining tools(TANAGRA AND WEKA).These tools are applied to check the results and comparison is also shown by along different attributes.

1) TANAGRA TOOL: Tanagra tool is free data mining software for academic and research purposes. It proposes several data mining methods from explanatory data analysis, statistical learning, machine learning, and database areas. This project is the successor of SIPINA which implements various learning algorithms, especially an interactive and visual construction of decision trees. Tanagra is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and non-parametric statistics, association rule, feature selection and construction algorithms. Tanagra is open source we can add any functionality we want, it is useful for research use but limitation is that we can add only one database for one operation.

2) WEKA TOOL: WEKA (Waikato environment for knowledge analysis) is a popular suite of machine learning software written in java developed at the University of Waikato, New Zealand. Weka supports several data mining tasks, more specifically, data preprocessing, clustering, regression, visualization and feature selection. All of the weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes(normally numeric or Nominal attributes, but some other type of attributes is also supported.

B. *Implementation of Proposed Algorithm Using Matlab*

SMURF which was proposed by UC Berkeley does not work well in determining the size of sliding window when tags move rapidly in and out of reader's communication range. Data cleaning based on sliding window filtering has been used by many applications but sliding window with fixed size has some drawbacks because when the size of window is too small, then many false negatives cannot be filled and when size of window is too large, then many false positives will be produced. Jeffery et al shows that setting the smoothing window size is a non-trivial task .It requires careful balancing between two opposing requirements:1) to ensuring completeness 2)to capture tag dynamics. In the physical world, SMURF views RFID streams as a statistical sampling of tags and uses techniques grounded in sampling theory to drive its cleaning process for modeling the unreliability of data. Therefore, WSTD (window sub-range transition detection) is related to transition detection mechanism. It estimates the overall tag counts to be detected and adjust the window size accordingly. Due to the fluctuation of tag-reader performance because of the environment changes WSTD is able to cope up with changes to adapt its window size .WSTD is relatively more accurate in estimating the periods of drops between readings when tag moves rapidly out of the detection range. Therefore, along with some of the concepts of SMURF but with improved mechanism WSTD, Our proposed algorithm considers all problems by producing the results more accurately.

The observed RFID readings are viewed as random samples of the population of tags in the physical world as described by [7].The missing tags imply that typically only a subset of the tag population is actually observed on every read cycle. The key insight is viewing each read cycle output as a random sampling trial and the smoothing window as repeated random sampling trials. Here, we will refer to an atomic unit of time used by one read cycle as an epoch.

Mentioned in [7,8,9] Let N_t denote the unknown size of the underlying tag population at epoch t and let $S_t \subseteq \{1,\ldots, N_t\}$ denote the subset of the tags observed("sampled") during that epoch. S_t can be viewed as unequal probability random sample of the tag population. Probability $p_{i,t}$ of selecting tag i

at epoch ***t*** can be calculated from the epoch ***t*** output information using the number of reads(tag responses) for tag ***i*** in combination with the known number of interrogation cycles(number of requests)as in (1)

$$P_{i,t} = \frac{\text{no. of requests}}{\text{no. of responses}} \quad (1)$$

Each epoch is viewed as independent Bernoulli trial (i.e. a sample draw for tag ***i***) with successful observations of tag ***i*** in the window $\mathbf{W}_i$ with $\mathbf{w}_i$ epochs (i.e. $\mathbf{W}_i = (t - \mathbf{w}_i\ t)$) is a random variable with a binomial distribution $B(\mathbf{w}_i, \mathbf{p}_i)$.In the general case, assume tag ***i*** is seen only in subset $\mathbf{S}_i$ $\mathbf{W}_i$ of all epochs in the window $\mathbf{W}_i$.

Assuming that, the tag probabilities within an approximately sized window calculated using(1),are relatively homogeneous, taking their average will give a valid estimate of the actual probability of tag ***i*** during window $\mathbf{W}_i$ .Therefore, the average empirical read rate p_i^{avg} over the observations epoch is given by(2).

$$\text{i.e.} \quad \boldsymbol{p}_i^{avg} = 1 / |\mathbf{S}_i| \sum_{t \in Si} \boldsymbol{p}_{i,t} \quad (2)$$

Also, $\mathbf{S}_i$ can be seen as binomial sample of epochs in $\mathbf{W}_i$ i.e. a Bernoulli trial probability $\mathbf{p}_i^{avg}$ for success and $|S_i|$ as a binomial random variable with binomial distribution $B(\mathbf{w}_i, \mathbf{p}_i)$.Hence, from standard probability theory the expected value and variance $|S_i|$ is given as

$$E[|\mathbf{S}_i|] = \mathbf{w}_i . \boldsymbol{p}_i^{avg}$$

$$Var[|\mathbf{S}_i|] = \mathbf{w}_i . \boldsymbol{p}_i^{avg} (1 - \boldsymbol{p}_i^{avg})$$

The above binomial sampling model is then used to set the window size to ensure that there is enough epochs in the window $\mathbf{W}_i$ such that tag ***i*** read if it does not exist in the reader's range. Setting the number of epochs within the smoothing window using (3) ensures that tag ***i*** is observed within the window $\mathbf{W}_i$ with probability $> 1 - \delta$

$$\mathbf{w}_i \geq [1 / \boldsymbol{p}_i^{avg} (\ln 1 / \boldsymbol{\delta})]$$

In order to balance between guaranteeing completeness and capture tag dynamics, the WSTD algorithm uses simple rules together with statistical analysis of the underlying data stream to adaptively adjust the cleaning window size. Similar to SMURF, transition within the window is detected of the number of observed readings is less than the expected number of readings (i.e.$|S_i| < w_i . p_i^{avg}$) and there is statistically significant variation in the tag observations using the Central Limit Theorem(CLT).

$$| |\mathbf{S}_i| - \mathbf{w}_i . \boldsymbol{p}_i^{avg} | > 2.\sqrt{\ \mathbf{w}_i . \boldsymbol{p}_i^{avg} (1 - \boldsymbol{p}_i^{avg})}$$

However, it is noted that this variation within the window can also be caused by missing tags and not necessarily only due transition.

The proposed algorithm is designed under the m-script file section of **MATLAB®.**

Algorithm :

Input: T = set of all observed tag ID
δ = required completeness confidence
Output: t = set of all present tag IDs
Initialize: for all i belongs to T, $\mathbf{w}_i$ ‹— 1
While (getNextEpoch) do
For (i in T)
Process window (W_i) —› $\boldsymbol{p}_{i,t}, \boldsymbol{p}_i^{avg}$,$|\mathbf{S}_i|$
if (tagidExist ($|S_i|$)
Output ***i***
end if
$\mathbf{w}_i^*$ ‹—requiredWindowSize ($\boldsymbol{p}_i^{avg}$, δ)
tagid = check valid ($\boldsymbol{p}_{i,t}^{S}$)
Valset=create hastable; //only for duplicate
if (tagid Λ $|S_{2i}|$ ==0)
$\mathbf{w}_i$ ‹— min{ w_i / 5, w_i^*})
elseif (tagid Λ $|S_{2i}| > 0$)
w_i ‹— min{ w_i/5, w_i^* }
elseif (tagid Λ $|S_{2i}| < 0$)
w_i ‹— min{ w_i /5, w_i^* }
elseif (ifduplicate(tagid Λ $|S_{2i}|$))// using unique hash Table
w_i ‹— min{ w_i/5, w_i^* }
end if
end for
end while

The algorithm is a description of the WSTD per tag cleaning algorithm for removing false and duplicate reading errors from an RFID data set. The problem that tag moves in and out of reader's communication range is resolved to some extent. This algorithm gives better results as compared to the previous one. Each individual tag is cleaned in its own window.

Initially all new detected tag's window is set to 1.The minimum window size is set to 5 that balance between the smoothing of the algorithm and reduces the errors. As the error may occur simultaneous, the algorithm is made common, but for demonstration purpose separate file are used for separate error. The algorithm is first detect the error based on the algorithm with comparison of the existing tag id's .If the error exist it give the graphical view of the error using matlab plot technique. Then the algorithm tries to clean the present error on the as the cleaned result is again printed on the matlab.

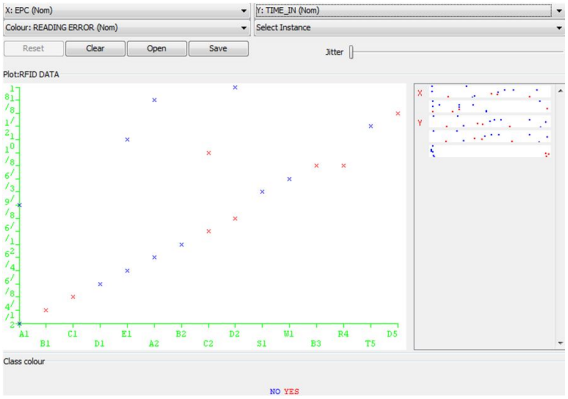
III. IMPLEMENTATION OF DATA MINING TOOLS APPLIED

1) Results of Tanagra Tool

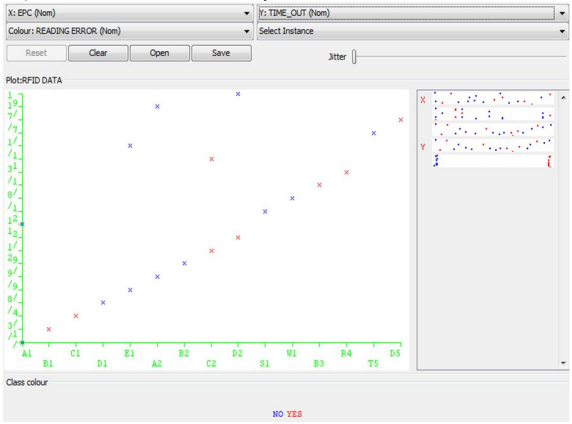
Here the result of **Tanagra tool** shows that the percentage of data without errors is 60% with counts of 12 and data with errors is 40% with counts of 8.Therefore, the error containing data is less than the data without errors.

2) Results of Weka Tool

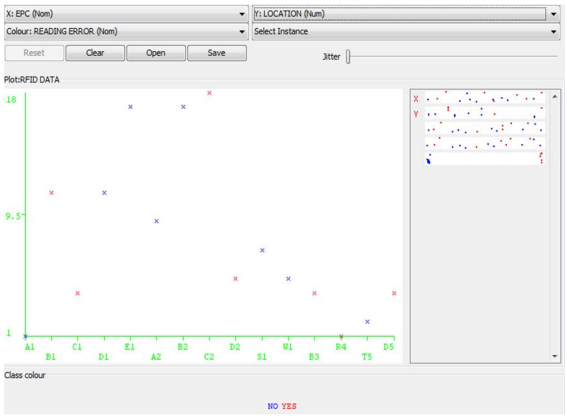
a) Visualization result of EPC and time_in



b) Visualization result of EPC and time_out



c) Visualization result with respect to EPC and location



In the all four visualization results, red * indicates data without errors and blue *indicates data with errors. Here also the result is same as Tanagra i.e. 60:40.

IV. SIMULATION RESULTS

A) Existing System

Here, the RFID data set is applied on the existing approach i.e. SMURF. By reading RFID data set, out of 75 readings 14 errors were found as shown in figure below:

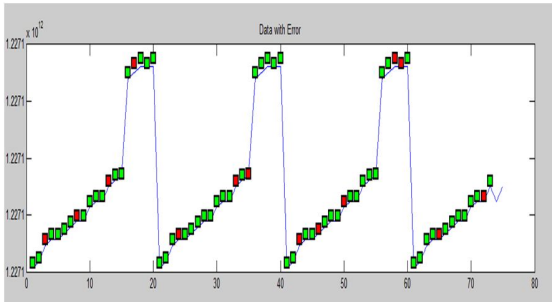


Fig1 RFID data containing errors

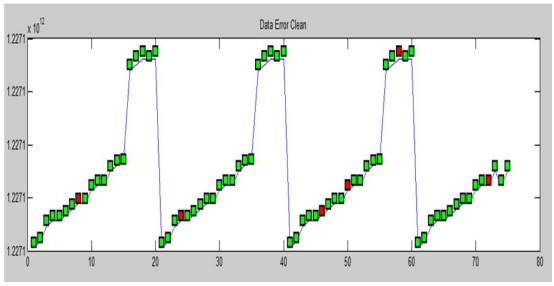


Fig2 RFID data after cleaning errors

X-axis denotes the tag-ids and Y axis denotes the timestamp. Therefore, by applying existing algorithm, there are errors which have to be reduced by an improved approach.

B. Proposed System

1) *False negative reading*:
The proposed algorithm is applied to remove false negative errors. As false negative reading errors occurs when tag is not detected by the reader but present in the database. By reading RFID data set, out of 75 sample readings 14 errors were found as shown in fig. below
X-axis denotes the tag-ids
Y-axis denotes the timestamps

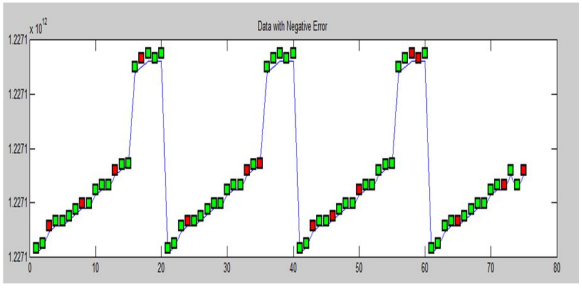


Fig3 RFID data containing false negative errors

After applying the proposed algorithm there is a vast reduction in errors by an approximation of 70% as compared to existing approach, shown in fig. below

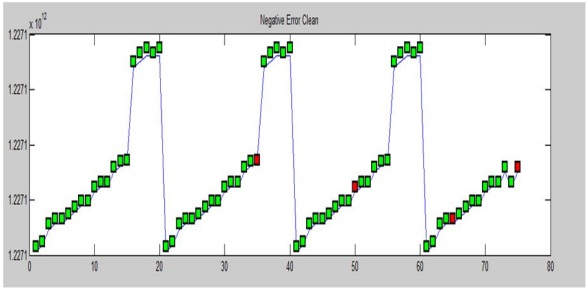


Fig4 RFID data after removing false negative errors

2) *False positive readings*:
The proposed algorithm is also applied to remove false positive errors. As false positive reading errors occur when tag is detected by the reader but not present in the database. By reading RFID data set, out of 95 sample readings 21 errors were found as shown in fig. below

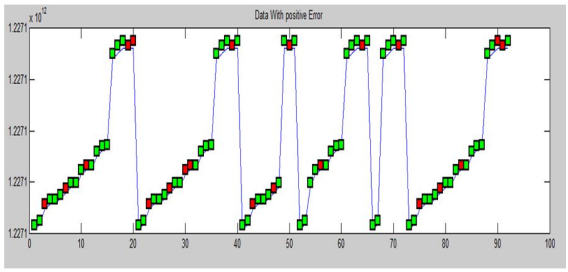


Fig5 RFID data containing false positive error

After applying the proposed algorithm there is also a vast reduction in errors in false negative readings by an approximation of 70% compared to existing approach as shown in fig. below

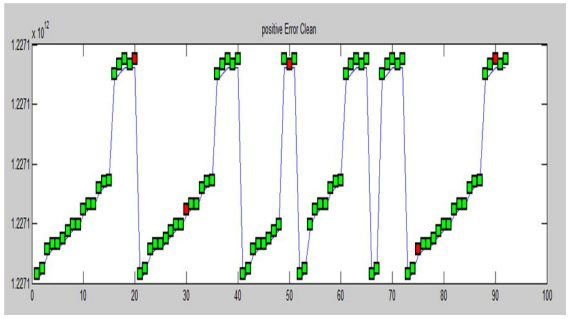


Fig6 RFID data after removing false negative errors

3) *Duplicate readings*:
The proposed algorithm is also applied to remove false positive errors. As false positive reading errors occur when tag is detected by the reader but not present in the database. By reading RFID data set, out of 75 sample readings 21 errors were found as shown in fig. below:

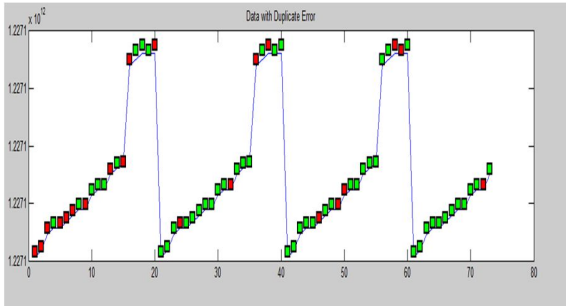


Fig7 RFID data containing duplicate readings

After applying the proposed algorithm there is also a vast reduction in errors in duplicate readings by an approximation of 70% compared to existing approach as shown in fig. below

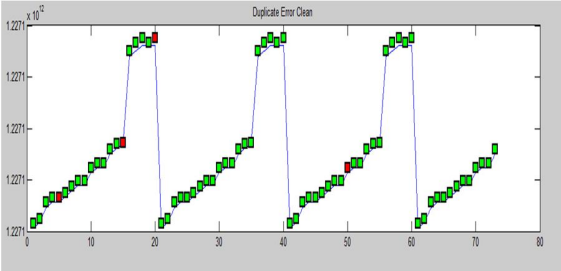
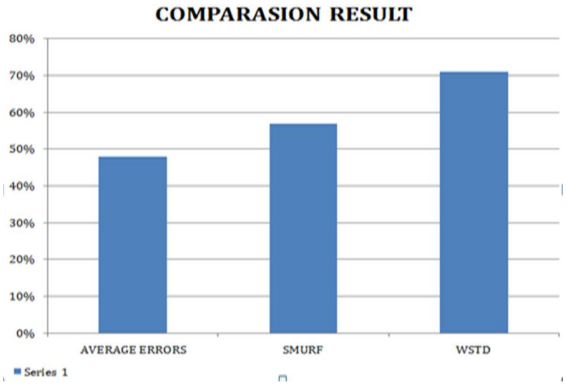


Fig8 RFID data after removing duplicate readings



Average errors = false negative + false positive

No .of tags

V. CONCLUSION AND FUTURE WORK

RFID applications are set to play an essential role in object tracking and supply chain management systems. While RFID data holds much promise, the unreliability of data produced by RFID readers is a major factor hindering large scale deployment. Specifically, RFID readers suffer from low read rates, frequently failing to read tags that are present.SMURF, which is a declarative, adaptive smoothing filter for RFID data. It does not require the application to set a smoothing window size; it automatically adapts its window size based on the characteristics of the underlying data stream. The drawback behind SMURF is that it does not well when tag moves in and out of reader's communication range. Now, for removing the errors which are produced in the RFID data stream, an improved transition detection mechanism scheme called WSTD (an adaptive sliding window based cleaning scheme) is proposed which is based on some of the concepts of SMURF but with an improved detection mechanism.WSTD uses binomial concepts or π estimators for detecting when transition occurs within the window with the help of a sliding window filtering scheme. By WSTD mechanism false and duplicate reading errors are reduced to a great extent. Simulation results shows that the proposed algorithm is better than the previous one i.e. SMURF. Future scope of the research work is to implement this algorithm with live database in a real-time RFID application. The proposed algorithm works with any fixed database of RFID. To implement this type of algorithm in real-time application, a suitable modifications needs to be done depending upon the type of application. Modifications has to be done in such a way that we get data consists of less errors to make an effective and efficient system which will be useful for higher level processing.(*WSTD: window sub-range transition detection)

AKNOWLEDGEMENT

We would like to give our sincere gratitude to our guide Ms. Meghna Sharma who guided us to pursue and complete this topic. We would also like to thank her for providing remarkable suggestions and constant encouragement. I deem it my privilege to have carried out my research work under her guidance.

REFERENCES

1) Roozbeh Derakhshan, Maria E.Orlowska and Xue Li,"**RFID Data Management: Challenges and Opportunities**, "In the proceedings of the IEEE International Conference on RFID,March 26-28,2007,pp-175-182.
2) Farahnaz Vahdati,Reza Jvaidan,Ahmad Farrahi,"**A New Method for Data Redundancy Reduction in RFID middleware**, "In the proceedings of 5th International Symposium on Telecommunications(IST'2010),pp-175-180.
3) Anny Leema.A,Dr.Hemalatha.M,"**Optimizing Operational Efficiency and Enhancing Data Reliability using Effective and Adaptive Cleaning Approach for RFID in Healthcare**, "In the proceedings published by International Journal of Computer Applications(IJCA),International Conference on Advanced Computer Technology (ICACT),2011,pp-26-29.
4) A.Anny Leema,Dr.Hemalatha.M,"**Anomaly Detection and Elimination Algorithm For RFID Data**, "In the proceedings of International Journal of Computer Applications(0975-8887),Volume 49-No.3,July 2012,pp-15-19.

5) Guoping Jin, Dong Wang, **"A Research of Time and Location Based RFID Data Cleaning,"** IEEE International C
6) Baoyan Song,Pengfei Qin,Hao Wang,Weihong Xuan,Ge Yu,"**bSpace: A Data Cleaning Approach for RFID Data Streams Based on Virtual Spatial Granularity**, "In the proceedings of 9th International Conference on Hybrid Intelligent Systems,IEEE International Conference,2009,pp-252-256
7) Shawn R.Jeffery,Minos Garofalakis,Michael J.Franklin,"**Adaptive Cleaning for RFID Data Streams**, "In the proceeding of 32nd International Conference on very large data bases(VLDB),pp-163-174,2006,Seoul Korea.
8) Lingyong Meng,Fengqi Yu,"**RFID Data Cleaning Based on Adaptive Window**, "In the proceedings 2nd International Conference on Future Computer and Communication,IEEE(2010),volume-1,pp-746-749.
9) Libe Valentine Massawe and Herman Vermaak, Johnson D.M.Kinyua,"**An Adaptive Data Cleaning Scheme for Reducing False Negative reads in RFID Data Streams**, "In the proceedings of IEEE International Conference on RFID (RFID), 2012, pp-157-164.