A Review on Data Cleaning Technology for RFID Network

He XU[1,2,a], Jie DING[1,2,b], Peng LI[1,2,c], Wei LI[1,d]
[1] School of Computer Science & Technology/School of software, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, 210003, China
{[a]xuhe,[c]lipeng}@njupt.edu.cn,{ 1248395233[b], 577935562[d]}@qq.com

Abstract. Radio Frequency Identification (RFID) technology is a kind of automatic recognition of communication technology. With the expansion of the RFID technology application areas, it needs to constantly improve the reliability, correctness and completeness of the RFID stream data. So data cleaning is significant to the RFID system. In recent years, many experts and scholars proposed a large number of cleaning algorithms according to RFID data flow characteristics. This paper gives a review on RFID data cleaning technology and some typical data cleaning algorithms.

1 Introduction

Radio Frequency Identification (RFID) technology is a kind of automatic recognition of communication technology[1]. With the detection, identification and monitoring of the electromagnetic signals, the related data can be read and written, so that information transmission is realized without contact between recognition system and specific target. Therefore, the target can be recognized automatically. RFID technology have the features of reading over a long distance, high storage capacity and other characteristics, which is widely used in the Internet of Things in the supply chain for object tracking and tracing, and have high-profile application prospect. For example, by integrating the RFID and wireless sensors, it is possible to communication together between RFID tag and other equipment. RFID technology not only can help a company to improve the efficiency of information management significantly, also can interconnect among enterprises, sales and manufacturing enterprises, so as to obtain more accurate control and feedback, and finally realize the optimization of the whole supply chain.

With the expansion of the RFID technology application areas, the demand for reliability of business data is increasingly important. In order to reach the level of satisfying the needs of the upper application, data cleaning is essential and directly affects the correctness and completeness of the business data, so it needs to filter and handle RFID data[2].

F. Xhafa et al. (eds.), *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Lecture Notes on Data Engineering and Communications Technologies 1, DOI 10.1007/978-3-319-49109-7_35

2 The characteristics of RFID data flow

The RFID data flow has the following characteristics[3]:

Streaming: each tag always continuously produces data in great quantities, which is continuous and intensive captured by the reader in a period.

Batch: multiple tag data is always captured by certain or more readers at the same time.

Semantic: tag data represents the position and status information of the observed object in a certain observation time.

Unreliability: RFID data is obtained by RFID reader and a variety of interactive mode's electronic tag are not reliable. The types and characteristics of wrong data are not identical, and the external causes(such as environmental factors) resulting in errors are different. All those factors cause the unreliable RFID data. There are some common error types of unreliability: False Negative, False Positive and Redundancy.

False Negative: radio frequency signal is highly vulnerable to environmental disturbance. Particularly, when a large number of tags suddenly come into the read range of reader. The collision and interference of signal among RFID readers will cause miss reading some tags, and this missing read phenomenon is very common. Reader cannot read all tags data within the scope of its read range and write without error.

False Positive: reader catches the tags outside the scope of its read range and write accidentally.

Redundancy: when a reader stay in the same place and can read within a period, the tags continue to send information to the reader, and it will produce a large number of repetitive data, where these extra data lead to some problems such as low efficiency of communication, energy loss and time delay. Generally, the redundancy is divided into tag redundancy and reader redundancy.

3 RFID data cleaning process

In order to improve the reliability, accuracy and integrity of RFID data, the RFID system needs self-contained cleaning module to handle huge amounts of tag data, to make reader read tag data consistent with the original data as far as possible. Usually, the error data types in RFID system are False Negative, False Positive and data redundancy. The general RFID cleaning framework is shown in Figure 1:

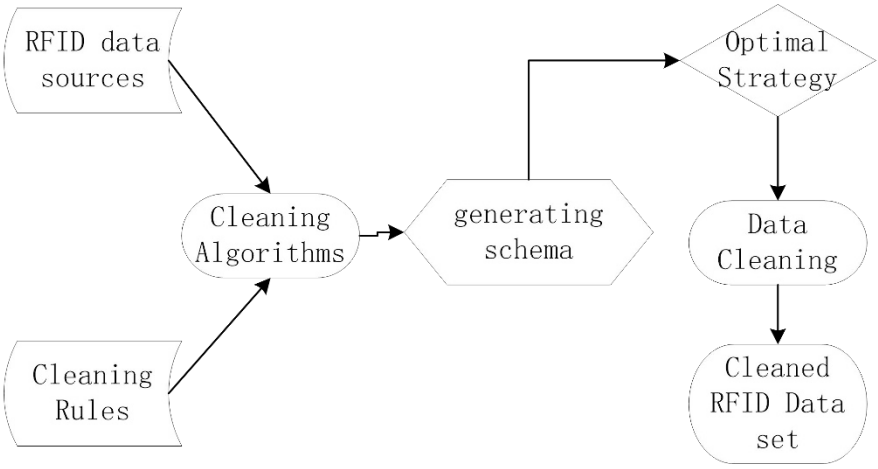

Figure 1. Generic framework of RFID data cleaning

The basic steps of data cleaning is based on the analysis of the causes of data error, process and form, then classification, using specific cleaning strategy for specific error to achieve optimized data quality and meet the requirements of the upper application. Through analysis, extract and optimize the data, it can maximize the implementation of data cleaning and make the data suitable for the upper application as much as possible.

4 Common cleaning algorithms of RFID data

4.1 The data cleaning algorithm based on fixed time window

In the original data cleaning system, the mechanism of using a slide time window technology to smooth filtering data flow is put forward by Y.Bai[4] from UCLA University and used for washing original tag data flow of RFID.

In most of the cleaning model of the RFID data, standard thought of data cleaning is equivalent to "smoothing filter based on time". The concept of "smooth window" is that refers to a window which has a certain length and sliding over time. In the whole life cycle of the system implementation, a fixed value represents the size of the time window which will continue to move forward as time goes on.

This method is simple and fast, but it don't adapt to the requirements of the dynamic environment. The window size is very difficult to set, if too small which can cause missing read (False Negative), too large which can lead to repetitive read (False Positive), thus result in the problem of data flow delay. So the ideal size of the sliding window must be set to ensure tags can be read completely and tags can be captured dynamically. Generally, although this method has shortcomings, it still has strong practicability because of its simple implementation.

4.2 Online extensible data flow cleaning framework

R.Jeffrey et al.[5] contrapose the characteristics of RFID data flow to introduce a data cleaning model based on the structure of pipeline's data cleaning model ESP through

introducing the definition of space and time granularity, which is also known as extensible data flow cleaning model. According to the characteristics of the original data using different cleaning steps, based on the time and space characteristics of RFID data, ESP uses different cleaning steps and makes them form a pipe to process the raw data. ESP can be used by all kinds of practical applications through a simple configuration. This algorithm is suitable for data leaking and repeated reading.

ESP can clean a variety of RFID data effectively, from a single reader to multiple readers, from a single space to multiple space, from a single type to multiple types. It has extensive data cleaning scope. Moreover, this model is very flexible, and can flexible select processing phase according to the different requirements. But it is very difficult to set the ideal time and space granularity parameter in ESP. Only to pick up the optimal particle size can ensure the quality of cleaning and minimize the average error rate of data cleaning. Thus, what is the most needed to deal with is to adjust self-adaptive time and space granularity.

4.3 RFID data cleaning algorithm of adaptive window size

M.Garofalakis et al.[6] has introduced the data cleaning strategy based on time correlation, which is based on probability model. Through dynamic changing the window's size, it is mainly used to solve the problem of leaking of read data. This is the first definition of self-adaptive smooth filtering of RFID data cleaning method - "SMURF".

This method puts RFID data flow as a random event in probability statistics, by using the method of probability theory to fill the data that had read missed. It has the advantages of adaptive deciding window's size according to the size of tag reading rate, and reducing the false negative and false positive which caused by unreasonable selection of window size. But for dynamic tag data, for example, when moving tags leave the reader's reading range quickly, it can make reading rate suddenly decreased. Meanwhile, the method of SMURF uses large Windows smooth, which can lead to more false positive(repeating read). And the shortcomings of sliding window are that can not avoid false negative and false positive, because SMURF is also based on sliding window.

4.4 The RFID data cleaning algorithm based on dynamic Bayesian network

H.Gonzalez et al.[7] put forward the RFID data cleaning algorithm based on Dynamic Bayesian network(DBN), through accessing data cleaning results' accuracy and balance among the cost that is needed to pay to realize the cleaning cost optimization. The fewest resources are used to clean the most original tag data under the premise of ensuring the accuracy of the cleaning's results.

DBN uses an index called implicit model, the actual value of which is the noise value, to determine the real location, and then predicts the tag data by observing historical data, finally calculates a probability value as a standard to measure tags whether exist or not.

Compared with sliding Windows, ESP and SMURF, the advantages of DBN is that do not need to record the latest tag data, because the relationship of observed value and predicted value are combined, and the weight of new data is given higher. While its shortcomings is that can not guarantee the dynamic update because

predicted values and the observed values are gained from historical data, the results of cleaning dynamic tags are not very ideal.

4.5 RFID data cleaning algorithm based on particle filter

RFID-Particle Filter Cleaning(RPFC)[8] is a kind of new tracing cleaning algorithm based on Particle filtering to realize Bayesian filter, using abstract Particle to simulate movements of tags under real situation, and get the mobile tags' location. Then original data will be converted to smooth uncertainty data flow in regard to the state of the tag location. Non-deterministic cleaning method which puts forward in RPFC starts from the two parts: semantic cleaning and misreading cleaning. Semantic cleaning which is based on uncertainly RFID data semantic uses means of adopting weighted particle swarm to track, simulate, and show changes of moving tags' location information. The effective probability value are used to express uncertainty of RFID data semantic. Misreading of cleaning consists two kinds, which are reader's false negative and false positive. And different methods are needed to start uncertainly cleaning respectively.

RPFC using RFID tag data' characteristic of uncertainly semantic to fill the false negative data from adjacent readers which is exist adjacent relation of single tag reader's mobile path (even if certain reader read data for a long time but it has not a high success rate, the missed tag data can also be filled by other readers), and that will not lead to false positive. It's very effective to solve the problem of single reader's false negative when limit error in a small range.

5 Redundancy problems

In order to ensure the tag data can be identified successfully, redundancy is inevitable. There are also many scholars who put forward several solutions, which mainly are divided into the two types of data redundancy and reader redundancy[9], to improve the accuracy of the RFID system.

5.1 Data redundancy

1.Based on Bloom Filters approximation

Metwally et al.[10] put forward using the Bloom Filters(BF) to detect redundancy, but BF do not have automatically delete function, that is, BF will fail due to fill up with data flow produced continuously for a long time. In order to solve this problem, Deng [11] put forward using Stable Bloom Filter (SBF) to process the RFID data flow dynamically, remove old data and set up the unit number corresponding to the maximum number of input data. Wang[12] proposed a cleaning algorithm to solve the distributed redundant data flow and used BF separately for distributed data flow, and achieved all redundant data filtered by sharing BF, which is not hard to find that will cause the waste of bandwidth resources, which introduced a method of cleaning RFID data redundancy based on space-time bloom filter. However, these methods will lead to the mistaken delete redundant data, which does not meet the requirements of data integrity and accuracy.

2.Sorted-Neighborhood Method(SNM)

SNM neighbor sorting algorithm[13] is a classic algorithm of data cleaning to the problem of data redundancy. SNM algorithm detects repetitive data records by comparing the sliding window, and only compares M records of windows each time. This method has improved the speed of comparison, which can improve the efficiency of matching effectively. But at the same time, the precision of repetitive records which are detected by SNM algorithm are limited by sorting of keywords, while keywords' quality directly affect the result of match. When selecting improper keywords, it is possible to make two data which are repetitive recorded far apart and never even in the same sliding window at the same time, so that means it will miss a lot of repetitive records on account of which cannot be identified. However, the selection of sliding window of size M is difficult. When M is big, comparison's times increased, which will result in a lot of unnecessary comparison. When M is small, which will lead to match missed. So these weaknesses determine the SNM algorithm need continuous optimization.

3.Improvement of SNM algorithm[14]

Full-text index technology is a kind of special functional indexes based on tags and face to the full text, which can provide full information, using meaningful words or phrase in the original as retrieval content, pointing to relevant pages or links and do not need information's indexing which can complete the retrieval.

The full-text index technology combined with the traditional algorithm of SNM forms a new algorithm of repetitive redundant data cleaning. In the process of sorting, full-text index technology can effectively make repetitive records appear in the same sliding window, reduce the error of repetitive record detection, improve the efficiency of detection, and make up for the deficiency of the SNM algorithm. When matching two data' similarity, different weight is set according to the importance of different fields, then determined whether two records are repetitive records by compared with similarity. It can reduce the number of unnecessary comparison when looking for repetitive records and does not affect the efficiency, so as to eliminate repetitive redundancy data more effectively.

5.2 Reader's redundancy

1.RRE algorithm

Purdue University and MOTOROLA laboratory put forward a method named RRE[15], which is used to solve the reader's redundant. The basic idea of this method is listed as follows: reader sends its own identification number and identified tag number to all tags recognized to write. Tag stores the largest number of reader identification number. Stored maximum tag number of reader will lock its tags, and the operation is repeated. All tags are checked that identified by each readers, which reader that has not locked any tags is redundant. This method record readers' information by means of writable tags between the reader and the tag communication layer, thus completing the reader and tag one-to-one correspondence, so each tag can only be identified by the only reader. It avoid the problem of data redundancy which is caused by the multiple readers reading the same tag, which reduced new share and

the amount of processing data. Experiments show that, compared with the greedy algorithm, this method is superior in the respect of eliminating reader's redundancy. RRE algorithm has the defect of its outstanding cleaning effect only in the case of reader is static and it does not work in majority of realistic situation.

2.LEO algorithm

Layered Elimination Optimization (LEO) algorithm[16] focuses on the situation that RRE algorithm can not remove dynamic redundancy reader accurately and puts forward the corresponding improvement, which write the lock information of tag by the reader being read first, that is, the reader will possess this tag if it is a deliver when the default tag data received information from the reader for the first time, and locking the owner of each tag at the same time. The reader which had locked tag is an effective work reader, and the probability of repetitive read tags' times are reduced, then redundant reader can be eliminated effectively.

The effect of LEO cleaning is much better than RRE algorithm, but its result of redundant data' cleaning depends on the order of reader reading data, and it may affect the overall efficiency of the algorithm if being read in order.

3.I-RRE

Literature [17] referred Improvable Redundant Reader Elimination (I-RRE) algorithm for the RFID system requirements of the algorithm which are very similar to RRE algorithm. In view of the shortcomings of the reader is evenly distributed, RRE algorithm can not identify the redundant reader correctly, and LEO algorithm rely on the reader's order. I-RRE algorithm has added a variable C-holder to record and identify the number of reader which are the second largest, according to the tags' C-holder value within the reader range to determine the priority, which can achieve the purpose of identifying the redundant reader .

Although I-RRE algorithm improves the insufficient of the RRE algorithm and LEO algorithm, when the distribution of RFID system's reader is very dense, each tag can be read by three or more readers, and the number of reader's tags in radio-frequency region are equal. Then, this algorithm will not be able to perform correctly.

4.TRRE

In the process of running RRE algorithm, when the reader updates the number of tags within the scope of reading, tags are needed to be repeatedly written in reader's identifier, which can lead to increase RRE algorithm execution time, especially tag needs more storage space when its complexity of writing operation time is larger. In view of these shortcomings of RRE algorithm, the TRRE (Two-Step Redundant Reader Elimination) algorithm[18] had been proposed. Suppose it does not need to exchange information among readers, and RFID system had solved the problem of tag conflicts before TRRE algorithm is used, then just send reader's ID information to all tags within radio-frequency region, in this way, it only needs to write reader's ID information replacing the operation that reader's ID information and identification tags' gross wrote simultaneously by tags that had been read by multiple readers. This can reduce the complexity of the write operation.

However, TRRE algorithm and LEO algorithm, both are affected by the reader's reading sequence. In addition, due to the network instability that will cause reader's reading sequence is random, the reliability of the TRRE algorithm needs to be improved.

5.MRRE

Assume that the reader can communicate with all tags in radio frequency range, and RFID middleware's tag data information which is transmitted by reader can be received and saved by RFID middleware. Literature [19] presents a cleaning algorithm MRRE based on middleware.

Relative to the above four kinds of algorithms, MRRE algorithm has obvious advantages, such as it does not need reader to write information for the tag, also does not need to know the system topology structure, in addition, it does not need reader to read and write tags frequently as well. These advantages can reduce the burden of RFID system, to some extent, it can eliminate or reduce the effects of RFID system performance, reduce the probability of the system mistakenly identify redundant readers, and improve the performance of algorithm. What's more, MRRE algorithm guarantees that it will identify individual which contains the largest number of readers at first as effective work reader when multiple reader's reading range exist cross, to minimize the amount of tags which is in cross area of reader identification range, by this way to reduce conflict rates of RFID read-write to minimum. MRRE algorithm is better than other algorithms in terms of system deployment's rationality. The algorithm by using middleware which has rich resources to store tag information that are uploaded by reader, are also applied in a situation where system can not use tags to store information.

The insufficiency of MRRE algorithm shows, in the RFID system, where readers are distributed densely and multiple reader's radio frequency range occur cross, that will result in reader conflict phenomenon. If every reader has recognized the same number of tags, the algorithm will not obtain optimal results.

A series of algorithms, such as RRE, LEO, I-RRE, TRRE, MRRE, and MXREO(RRE+LEO)[20], are used to remove redundant readers. Their insufficient mainly displays in two points: One is reader conflict appeared in the RFID system where reader are distributed densely, and when every reader has read the same amount of tags, these algorithms cannot effectively distinguish redundant readers; The other is that most algorithms depend on the reader's recognition sequence, while it is involved in algorithm's efficiency which needs to be improved. From these points of view perfect algorithms are needed to improve the recognition's effectiveness of redundant reader.

6 Conclusions

In recent years, many experts and scholars proposed a large number of cleaning algorithms according to RFID data flow characteristics. This paper gives a review on RFID data cleaning technology and some typical data cleaning algorithms. With the development of RFID technology, cleaning strategy should be improved and explored

continuously from several aspects because the demand of the data' quality is increasingly higher. At the same time, with the constant development of network technology, security has become an issue which can not be ignored. How to ensure the security and privacy of obtaining data in the dense RFID reader environment are needed to continuously explore in future work.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P. R. China (No.61373017, No.61572260, No.61572261, No.61672296, No.61602261), the Natural Science Foundation of Jiangsu Province (No.BK20140886, No.BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No.12KJA520002), China Postdoctoral Science Foundation (No.2014M551636, No.2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No.1302090B, No.1401005B), Natural Science Foundation of the Jiangsu Higher Education Institutions of China(Grant No. 14KJB520030), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX15_0381, SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks(WSNLBZY201509), and NUPTSF(Grant No. NY214060, No. NY214061).

References

1. Zoltan K. Using Rfid And Gis Technologies For Advanced Luggage Tracking[J]. SEA-Practical Application of Science, 2015,2(8): 229-234.
2. Managing and mining sensor data[M]. Springer Science & Business Media, 2013.
3. Xie L, Yin Y, Vasilakos A V, et al. Managing RFID data: challenges, opportunities and solutions[J]. IEEE Communications Surveys & Tutorials, 2014, 16(3): 1294-1311.
4. Bai Y, Wang F, Liu P. Efficiently Filtering RFID Data Streams[C]. CleanDB, 2006:1-8.
5. Jeffery R, Alonso G, Franklin M J, et al. A pipelined framework for online cleaning of sensor data streams[J]. Computer Science, 2005:1-12.
6. JefferyR, Garofalakis M, Franklin M J. Adaptive cleaning for RFID data streams[C]. Proceedings of the 32nd international conference on Very large data bases(VLDB Endowment), 2006: 163-174.
7. Gonzalez H, Han J, Shen X. Cost-conscious cleaning of massive RFID data sets[C]. IEEE 23rd International Conference on Data Engineering. IEEE, 2007: 1268-1272.
8. Want T, Zhang J. Research on RFID data cleaning method based on particle filter[J].Electronic Technology & Software Engineering, 2014(1):214-215.
9. Chen Li.Research on Data Cleaning Methods in RFID Middleware [D].China: Wuhan University of Technology, 2013.
10. Metwally A, Agrawal D, El Abbadi A. Duplicate detection in click streams[C]. Proceedings of the 14th international conference on World Wide Web. ACM, 2005: 12-21.
11. Deng F, Rafiei D. Approximately detecting duplicates for streaming data using stable bloom filters[C]. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 25-36.
12. Wang YL, Wang C, Jiang XH,et al. RFID duplicate removing algorithm based on temporal-spatial Bloom Filter[J].Journal of Nanjing University of Science and Technology, 2015,39(3):253-259.
13. Zhang JZ, Fang Z, Xiong YJ, et al. Optimization algorithm for cleaning data based on

SNM[J].Journal of Central South University (Science and Technology).2010,41(6):2240-2245.
14. Xu X,,Wang J,Yu Y.Research on full text index[J]. Computer Engineering.2002(2):101-103.
15. Carbunar B, Ramanathan M K, Koyuturk M, et al. Redundant-reader elimination in RFID systems[C]. IEEE SECON,2005:176-184.
16. Hsu C H, Chen Y M, Kang H J. Performance-effective and low-complexity redundant reader detection in wireless RFID networks[J]. EURASIP Journal on Wireless Communications and Networking, 2008(1): 1-9.
17. Jiang Y. Improvable-Redundant-Reader Elimination in RFID system[J].Computer Engineering and Applications. 2011,47(5):101-103.
18. Chen J,Yang Z.A Fast and Efficient Algorithm for Redundant Reader Elimination in RFID Application System[J].Computer Knowledge and Technology, 2009,5(16):4285-4288
19. Lv S,Yu S.A Middleware-Based Algorithm for Redundant Reader Elimination in RFID Systems[J]. ACTA ELECTRONICA SINICA.2012,40(5):965-970
20. Chen C. Research on Clean Technology of RFID Sensing Data of Manu-Facturing in IOT[D]. Guangdong University of Technology, 2014.