Data Reduction Method in Massive Data Sets

Gecynth Torre Namo and Hong-Won Yun, *Member, KIMICS*

Abstract—Many researchers strive to research on ways on how to improve the performance of RFID system and many papers were written to solve one of the major drawbacks of potent technology related with data management. As RFID system captures billions of data, problems arising from dirty data and large volume of data causes uproar in the RFID community those researchers are finding ways on how to address this issue. Especially, effective data management is important to manage large volume of data. Data reduction techniques in attempts to address the issues on data are also presented in this paper. This paper introduces readers to a new data reduction algorithm that might be an alternative to reduce data in RFID Systems. A process on how to extract data from the reduced database is also presented. Performance study is conducted to analyze the new data reduction algorithm. Our performance analysis shows the utility and feasibility of our categorization reduction algorithms.

Index Terms— RFID data, Data reduction, Data management, Categorization reduction.

I. INTRODUCTION

RFID (Radio Frequency Identification) is a technology that uses tags and readers to capture data and be able to translate these data into useful information in purposes necessary for the users of different organizations. Tags are small chips attached to objects, whether in pallets or just an individual item, which transmits information to readers. Readers then transmit data to computer system, which in turn decodes data, and saves processed data to the database.

Readers or also known as interrogators are responsible for sending and receiving radio signals through antennas. They capture data stored from RFID tags and send them to computers for processing. As a whole, RFID technology is very powerful thus; many companies nowadays integrate RFID Systems into their business processes such as Walmart, Harley Davidson, Ford Motor Company, etc. In the advent of RFID, some are skeptical in embracing this technology because of the risks involved.

Such huge amount of data and various kinds of queries pose big load to traditional relational database. Big market such as Walmart generates around 7 terabytes of data every day. The key point then becomes how these companies can manage the huge volume of data. A variety of physical storage structures and indexing techniques have been proposed for relation database, but the support for physical deletion termed vacuuming. This paper presents a reduction method for the vacuuming of relational databases. The main focus is to suggest a method to reduce data size for effective data management against vacuumed databases.

The rest of the paper is organized as follows: Section II tells about the basic characteristics of RFID data and data reduction techniques. Section III introduces a new algorithm in data reduction and the Section IV discusses the performance study on the new data reduction algorithm. Section V concludes the paper.

Manuscript received received October 1, 2008; revised January 18, 2009.

Gecynth Torre Namo is a Graduate Student of Silla University.

Hong-won Yun (Corresponding Author) is with the Department of Information Technology, Silla University, Busan, 617-736, Korea (Tel: +82-51-999-5065, Fax: +82-51-999-5657, Email: hwyun@silla.ac.kr)

II. RELATED WORK

A. Characteristics of RFID Data

The first step to achieve a clear view of what RFID System is all about is to identify different basic information about its components. RFID data is one basic component of RFID technology that needs to be dissected and clearly understood in order for an individual to move to the next step of understanding RFID technology and be able to identify problems and as well as its corresponding solutions. The following are the characteristics of RFID data:

Simple. RFID system captures simple information

in a form of a tuple with three data: EPC, location and time. EPC is an acronym for Electronic Product Code which is a unique identifier for a certain item. This Electronic Produce Code is introduced by EPC Globals, an organization set up to achieve world-wide adoption and standardization of Electronic Product Code (EPC) technology. Location definitely refers to the place where the reader scanned the items and time refers to the time when the reader scanned the item.

Dynamic and Temporal. In RFID applications, especially in application like retail and as well as supply chain management, almost everything is transferred from one place to another place thus changing its' locations. Along with the changes in location, objects such as cans of soda are loaded to a box or container thus changing its status from just being an individual item to a part of the box. This is what you call as containment relationship and this can change from one location to another location as the items are shipped to their final destination, the consumers. Since RFID have readers that scan items in prescribed intervals, the information captured by these readers are associated with timestamps. So as time passes, information associated to a certain item can become obsolete. Thus, RFID data is very temporary.

Implicit semantics. As items' information are captured and read by RFID readers, there are some implied changes in the information regarding the item's state, location and relationship to other items. For example, an item's information is scanned but after a minute, the item is loaded to a container wherein the aggregation information (containment relationship) is changed. This should be detected by the system or else, there will be confusions. Another implicit semantics of RFID data is that items are shipped from one place to another place. With these movements, locations of the items should be properly traced by the system to avoid problems.

Inaccuracy. Even though the read rate accuracy of RFID is already high, but still, there are erroneous information being captured in the system, in the form of missed readings or redundant readings. Sometimes, an item can remain in the same place for such a long time with no changes at all such as items in retail stores that are not yet released to the consumers. In this scenario, readers read the item's information for several times that multiple and redundant information are stored in the database.

Large volume. RFID system scans data from items in a particular site in a certain interval of time. With this fact alone, it is expected that big amount of data are sure to be captured by the system. Plus the fact that RFID readers are automatic scanners that accumulate billions of data depending on the reader's reading range and as well the number of items present. In fact, compared to typical barcode applications, RFID applications generate more data about ten to hundred times the data volume of typical bar code applications [1-4].

These characteristics would help us delve deeper into RFID system and be able to understand the problems arising due to these characteristics of RFID data. Next step is to identify different data reduction techniques that address problems created by these RFID data characteristics.

B. Data Reduction Techniques

Different reduction techniques are being proposed by numerous researchers to reduce terabytes of data captured and scanned by RFID readers. These techniques are often focused on dirty data, either missed data or redundant data, captured by RFID readers. Redundant data is one of the major problems in RFID Systems. Since RFID system wants to capture updated status of items in certain locations, readers have a preprogrammed scanning interval time. With these, issue on redundancy arises since there are multiple scans on a certain item that undergo no change at a long period of time.

For example, an item in the store stayed for 1 week in the grocery shelf. Scanners in the store are preprogrammed to scan every 2 minutes and the store opens 24/7. If we will compute the totality of the number of scans that these readers do, it can reach to almost 5040 scans per reader/week not to mention the number of tags scanned per second. With these, it involves thousands or millions of redundant data captured depending on how large the organization is.

With the above example, it is but evident that steps should be taken to reduce the number of redundant data in RFID systems. Several studies explore these issues and tried to come up with several solutions. Here are a few of them:

In [7] Mylyy discusses the need for data cleaning since there are many unreliable readings in RFID Systems. The article formalized these readings to three scenarios: false negative readings, false positive readings and duplicate readings. The article also discusses two aspects of redundancy in RFID and that is data redundancy and as well as RFID readers redundancy. With these in mind, smoothing algorithm was introduced which is based on the sliding window approach. This algorithm intends to reduce duplicates and perform merging operation within a predefined threshold. In [7] Mylyy algorithm, four operations were presented: first is the insertion of the incoming reading into the window, the second is the removal of the expired readings from the said window, the counting of the readings and the fourth is the setting of

output_flag for the readings that satisfy the threshold condition.

In [1] Wang & liu tell that one way to filter redundant data is by using a sliding window. A filter will be scanned within a sliding window from multiple readers and if there are duplicates in the readings, these will be deleted. As an example, if the two readers, Rx and Ry have the same EPC value within T (Time), then one of them is dropped.

OBSERVATION(Rx, e, Tx),
OBSERVATION(Ry, e, Ty), Rx <> Ry,
within(Tx, Ty, T)
-> DROP: OBSERVATION (Rx, e, Tx)

Another technique to reduce data captured by RFID Systems is generalization. According to [2], data are relatively dependent on the users' needs and generalization is a must to reduce to the minimum level of abstraction that is interesting to the users of the said RFID Systems. For example, Unilever, a company that manufactures shampoos, conditioners, toothpastes, etc., it is only interested on inventory of how many boxes of toothpastes are delivered to a certain Department store and not how many pieces of toothpastes are delivered. But in this case, certain internal rules should be structured to maintain clarity such as a box contains 100 pieces of toothpastes or a pallet contains 12 boxes of toothpastes.

Some data cleaning algorithm worked on the data streams of RFID system. One example of that is the SMURF algorithm which is a declarative and self – tuning smoothing filter that aimed to produce accurate data streams for individual tag – id readings and accurate aggregate estimate over large tag population. It also deals with a sliding – window processor for data smoothing and a mechanism that detects mobile tags. The algorithm uses techniques from sampling theory such as binomial sampling to make the cleaning more statistically based [10].

Numerous data reduction techniques are presented in hundreds of articles published in the IT community. This is a proof that indeed, the largeness of data being captured by RFID System needs to be addressed properly for RFID to ubiquitous in every organizations whether in Retail or Supply Chain Management or any organization that need automatic and fast tracking and inventory of data. With continuous evolution of RFID technology, more and more solutions to the problems are being introduced to solve this one major obstacle in RFID systems. In an effort to contribute something essential to the IT community, another data reduction technique is being proposed by this paper.

III. PROPOSED DATA REDUCTION ALGORITHM

A. Electronic Catalog

Electronic catalogs are electronic representations of information about products and services of a certain organization. It is usually hierarchical in form based on a certain human defined category or taxonomy [8]. Electronic catalogs are usually used in the field of e-commerce and have gained attention to many companies who wanted to innovate and explore new ways to attract customers especially in this internet era.

Most electronic catalogs use relational database to manage large amount of data especially if the organization is a big one. Usually, electronic catalog uses product's category to group many products into one single entity. This concept of electronic catalog can be integrated into the concept of data reduction of RFID. We believe that categorization is one way to reduce data in a very simple but effective way. As in the previous sections of this study, different data reduction techniques are being presented and that, this study attempts to bring in an algorithm that introduces a concept of Electronic Catalog integrated into RFID system.

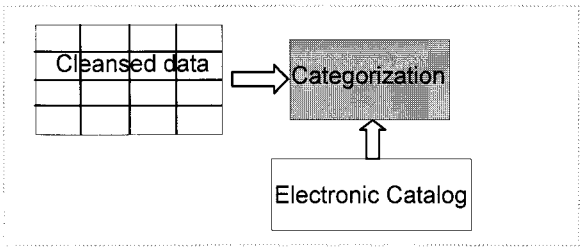

Fig. 1 RFID Data Reduction Using Electronic Catalog

For instance, RFID system captures data from the tags and that these data are cleansed based on several data cleaning algorithms presented at the later part of this paper. After the cleansing of data on one – reader and multiple - readers levels, these cleansed data undergo the categorization reduction process that will minimize the amount of data. An electronic catalog is referenced to do the categorization reduction process.

Since RFID data can be applied to numerous organizations and we are dealing with data reduction, this study is limited to retail organizations that are interested in order packages during delivery of stocks and are interested as well in individual items to be released to the consumers. The algorithms involved in this study focuses on cleaning irrelevant data and taking advantage of certain level of interests of a retail organization that will lead to data reduction.

B. Categorization Reduction

Ideas: The item will be scanned and be categorized based on its description and the location of the reader. Before categorization, the item should be cross-examined to the items scanned by the other readers to be sure that the other readers didn't scan the same item. After making sure that there are no duplications, the categorization happens. In the case of redundant readings of a single reader, this can be resolved by revising the tuple attributes of an RFID reading.

Examples: Soap will be in toiletry (type) and coke will be in the category of soda. With this, using the categorization and together with the location of the readers, we will be able to merge these items to one category thus taking advantage of groupings. This will cause the database to be smaller.

Assumptions: There are multiple readers in a certain location. Reader1 captures information with its EPC, location and time. EPC is the unique identification, location is the place where the reader is and the time is the time when the scanning is done.

Descriptions of the presented Algorithm

1) Readers scan the items with (EPC, Location, Time)
2) Checking should be done on the tuples scan by one reader. If there is duplication in the readings of a particular reader, all other readings will be deleted and the updated reading will be retained. This is possible by comparing the time scanned by the same reader.
3) After the individual cleaning the data scanned by individual readers, compare items scanned by one reader to the other readers through their EPC and time scanned. If a reader has the same information to the other readers, one reading should be deleted (least recently scanned).
4) After the update, the items should be categorized on a predetermined type together with its EPC and location such as Soap (toiletry), Coke(Soda), Shampoo(toiletry)
5) Using this information, items with the same category and location will be merged for a certain period of time until changes in the location is determined thus reducing the number of items saved in that period of time

Procedure

1) Individual Reader Cleansing
   1^{st} Tuple → $Reader_1$(EPC,Location,Time)
   2^{nd} Tuple → $Reader_1$(EPC,Location,Time)
   IF EPC of 1^{st} Tuple = EPC of the 2^{nd} Tuple
   Compare Time of 1^{st} Tuple > Time of the 2^{nd} Tuple
   Delete 2^{nd} Tuple
2) Between Readers Cleansing
   1^{st} Tuple → $Reader_1$(EPC,Location,Time)
   2^{nd} Tuple → $Reader_2$(EPC,Location,Time)
   IF EPC of $Reader_1$ = EPC of the $Reader_2$
   $Reader_1$(Time) < $Reader_2$(Time)
   Delete 1^{st} Tuple
3) Categorization Reduction
   Refer to Electronic Catalog with EPC
   Return Category from Electronic Catalog
   Add Number of items attribute
   (EPC,Location,Time,Category,NumberofItems)

In the below example, it clearly shows that there are 7 tuples in just two category. Instead of storing these into 7 separate tuples, it can be reduced into two tuples to save space. This is a design suited for retail stores that has a seemingly determined number of categories in the shelves. The algorithm is applicable to items that usually stay in one place and takes a lot of time to be moved into other places. Examples of these are stocks of items in the warehouses. This algorithm takes advantage of grouping concepts and time slicing concepts.

Table 1 Before Reduction Process

Table: ***Product***			
EPC	Location	Time	Category
0614141000734 203886	Shelf1	2008-11-04 11:20:00	00998
0614141000734 837464	Shelf1	2008-11-04 11:30:00	00998
0614141000734 938463	Shelf1	2008-11-04 11:30:00	00998
0614141000733 093648	Shelf1	2008-11-04 11:40:00	00976
0614141000733 103939	Shelf1	2008-11-04 11:50:00	00976
0614141000734 193837	Shelf1	2008-11-04 11:20:00	00998
0614141000733 877373	Shelf1	2008-11-04 11:30:00	00976

Table 2 After Reduction Process

Table: ***Product***				
EPC	Location	Time	Category	No
0614141000734 938463	Shelf1	2008-11-04 11:30:00	00998	4
0614141000733 877373	Shelf1	2008-11-04 11:30:00	00976	3

EPC in the Table2 will display only the highest among the values of the same category. EPC values are based on the values from the article [11]. The theory of this to use a certain level of abstraction that is of interest to the user and as well as taking advantage of groupings of certain objects. In the

aspect of the database structure used in this algorithm, it is shown in the figure below the relational database proposed for the said algorithm.

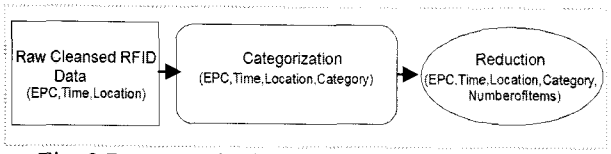

Fig. 2 Processes in the Categorization Reduction

C. Example of Query Processing after Reduction

Once the data are reduced using the process of categorization reduction, queries can be performed using the reduction table termed vacuumed table or alternative raw table. In the relational database structure without vacuuming, raw data can be still stored in the database in case these data are needed in the future. Assume an example of a query as following: When did the toothpaste with the EPC=061414100073493846 3 entered the store?

Categorization Table

Cat_ Code	Description
00998	Toothpaste

EPC = 06141410007349 38463

The cursor starts from the highest EPC that represents the whole category and go downward as it searches for the right EPC

EPC	Location	Time	Category
0614141 0007349 38463	Shelf1	2008-11-04 11:30:00	00998
0614141 0007348 37464	Shelf1	2008-11-04 11:30:00	00998
0614141 0007342 03886	Shelf1	2008-11-04 11:20:00	00998
0614141 0007341 93837	Shelf1	2008-11-04 11:20:00	00998
0614141 0007338 77373	Shelf1	2008-11-04 11:30:00	00976
0614141 0007331 03939	Shelf1	2008-11-04 11:50:00	00976
0614141 0007330 93648	Shelf1	2008-11-04 11:40:00	00976

Fig. 3 Example of Query Processing after Categorization Reduction

In the above query, a system with vacuuming should support its users in interpreting the results of queries using table of categorization reduction. A system requires the property of faithful history encoding, there is a need to go back to the lowest level of abstraction which is the raw data itself. We will be able to locate the data from the reduced table through knowing first the category of the said item. In the query above, the toothpaste is the category and we assume that category toothpaste is 00998. In the reduced table, the processor will compare the EPC of the questioned object with the EPC representing the whole category. If it doesn't match, then it searches from the highest and moves downward to locate the equivalent EPC within the said category. Fig. 3 shows the example of query processing after reduction.

IV. PERFORMANCE STUDY

Examining the results of the new data reduction algorithm is really important to know whether the said algorithm is an effective one. This paper compared two conditions, dirty data captured by RFID readers and the resulting cleansed data after using the reduction algorithm proposed by this paper.

RFID readers captured 25,000 tuples with 15 identified categories in a retail store. Categories used are predetermined already by the company and based on its own level of abstraction. Using categorization reduction, from 25,000 original tuples, are reduced to just 10,000 tuples reducing them by 40% on tuple size. In terms of actual data storage, parameters are set to 40 bytes per tuple and that a block size/page size of a disk is 16kb. The original amount of data to be stored in the disk can be 1,000,000kb equivalent to 62,500 blocks, but using the reduction process, it is reduced to only 400,000kb or 25,000 blocks. It is evident that the more categories involved, the bigger the reduction is. But since each algorithm has its weak point, same is true to this data reduction algorithm. One drawback of this algorithm is when categories are minimal; reduction on data size is also minimal.

V. CONCLUSION

RFID is a wireless technology that introduces new boundaries to the world. It offers better features and functionalities nonexistent to other technologies. Costs of infrastructures in RFID are slowly declining that it has an inverse effect on the technology's demands to organizations worldwide. However, despite the promise of RFID's technology, one cannot deny that there are downside effects of it. In this paper, we presented an overview of the RFID systems, identified different characteristics of RFID data and gathered different data reduction techniques. Our contribution is that we developed a new data reduction algorithm. Using the categorization reduction, data captured by RFID readers are cleansed and reduced. Query after categorization is also presented to show how data can be extracted after the reduction process. Performance

study is also discussed to analyze effects of the new data reduction algorithm.

REFERENCES

[1] F. Wang and P. Liu, "Temporal Management of RFID Data," *31st International Conference on Very Large Databases*, Norway, 2005, p. 1135

[2] H. Gonzalez, J. Han, X. Li and D. Klabjan, "Warehousing and Analyzing Massive RFID Data Sets," *22nd IEEE ICDE Conference*, April 2006, p. 4

[3] Y. Tu and S. Piramuthu, "Reducing False Reads in RFID-Embedded Supply Chains," *Journal of Theoretical and Applied Electronic Commerce Research*, Electronic Version Vol. 3, Issue 2, August 2008, pp. 60-70

[4] D. Lyle, "Understanding and Solving the RFID Data Dilemma," *Journal of Theoretical and Applied Electronic Commerce Research*, August, 2004

[5] LogicaCMG, "Making Waves: RFID Adoption in Returnable Packaging," *LogicaCMG*, Netherlands, 2004, pp. 22-24

[6] J. Han, X. Li, H. Gonzalez and J. Lee, "Mining Massive RFID, Trajectory, and Traffic Data Sets," *SIGKDD'08*, August, 2008

[7] O. Mylyy, "RFID Data Management, Aggregation and Filtering," 2007, Available: epic.hpi.uni-potsdam.de/pub/Home/Publications/RFID-Paper_SS2007_Oleksandr_Mylyy.pdf

[8] K. Kiryoong, K. Dongkyu, K. Jeuk, L. Ighoo and L. Sang-goo, "Issues and Trends of Information Technology Management in Contemporary Organizations," Mehdi Khosrow-Pour, 2002, p. 323

[9] Q.E.D Systems, "Active and Passive RFID: Two Distinct, But Complementary, Technologies for Real-Time Supply Chain Visibility," 2002 Available: www.autoid.org/2002_Documents/sc31_wg4/docs_501-520/520_18000-7_WhitePaper.pdf

[10] S. Jeffery, M. Garofalakis and M. Franklin, "Adaptive Cleaning for RFID Data Streams," *32nd International Conference on Very Large Databases*, September, 2006, pp.163-174

[11]EPC Global, "RFID & EPC Essentials," Available: www.epcglobalinc.org/what/cookbook/chapter1/002-- RFID_EPC_Essentials_v1.pdf

Gecynth Torre Namo is currently a Graduate Student of Silla University, Busan, South Korea. She completed her B.S. Degree Major in Computer Science from University of Saint La Salle, Philippines last March, 2001.

Hong-Won Yun
He received his B.S. and the Ph.D. degrees at the Department of Computer Science from Pusan National University, Korea, in 1986 and 1998, respectively. He is a professor at the Department of Information Technology, Silla University in Korea. His research interests include temporal database and semantic web.