

A Novel Integrated Classifier for Handling Data Warehouse Anomalies

Peter Darcy, Bela Stantic, and Abdul Sattar

Institute for Integrated and Intelligent Information Systems
Griffith University

{P.Darcy, B.Stantic, A.Sattar}@griffith.edu.au

Abstract. Within databases employed in various commercial sectors, anomalies continue to persist and hinder the overall integrity of data. Typically, *Duplicate*, *Wrong* and *Missed* observations of spatial-temporal data causes the user to be not able to accurately utilise recorded information. In literature, different methods have been mentioned to clean data which fall into the category of either deterministic and probabilistic approaches. However, we believe that to ensure the maximum integrity, a data cleaning methodology must have properties of both of these categories to effectively eliminate the anomalies. To realise this, we have proposed a method which relies both on integrated deterministic and probabilistic classifiers using fusion techniques. We have empirically evaluated the proposed concept with state-of-the-art techniques and found that our approach improves the integrity of the resulting data set.

1 Introduction

Duplicate, *Wrong* and *Missing* data anomalies have continually hindered commercial sectors in the past resulting in error-prone data collection which seriously influence business processes. When the information is entered manually, based on additional information already recorded, the erroneous data can quite easily be discarded or approximated within a data set. Unfortunately, this is not the case within automatically recorded Spatial-Temporal data sets as the anomalies that persist are harder to rectify. These anomalies may include False-Positive readings such as *Duplicate* and *Wrong* observations or False-Negative readings such as *Missed* observations. This is especially found within automated data collection technology, such as Radio Frequency Identification (RFID), in which the anomalies can represent faults in sensors, intrusions, or missing objects.

Previous approaches have attempted to correct these anomalies at a deferred stage using deterministic or probabilistic approaches to identify and remove anomalies. However, in cases in which the persistent anomalies are particularly ambiguous, there is a need for a more intelligent methodology to clean the data. Noticing that there was a need for a fusion of both deterministic and probabilistic approaches to handle the most ambiguous anomalies, we have presented an integrated classifier that combines Non-Monotonic Reasoning, Bayesian Networks and Neural Networks to intelligently clean the error-prone observations. Through experimental evaluation, we have found that the Non-Monotonic Reasoning and Bayesian Network fusion methods resulted in

the highest achieving integrated classifier for both the False-Positive and False-Negative anomalies respectively, and provided a higher cleaning rate when compared to other state-of-the-art techniques.

The remainder of this paper is organised as follows: Section ?? will contain background information including RFID, Non-Monotonic Reasoning, Bayesian Networks and Neural Networks. We will introduce our methodology within Section ?? and highlight the motivation, architecture, intended scenario and assumptions. The Experimental Evaluation we performed are presented in Section ??, followed by our Conclusions found in Section ??.

2 Background

To conduct our experimentation and to determine if our approach of integrating classifiers together would significantly improve spatial-temporal databases, we have decided to test it on RFID data. We have chosen to integrate the Non-Monotonic Reasoning, Bayesian Network and Neural Network classifiers to create a novel and intelligent means of cleaning the anomalies. In the following section, we will provide a brief introduction to RFID, Non-Monotonic Reasoning, Bayesian Networks and Neural Networks.

Fig. 1: An example of how in an enclosed environment, there is the possibility of Duplicate (T2), Wrong (T3) or Missing (T4) readings.

2.1 RFID

Radio Frequency Identification (RFID) is a convenient technology employed in a wide array of commercial sectors which uses radio waves to allow communication between tagged items and readers [?]. RFID technology has already been employed to be used in various commercial sectors such as air package tracking, airport luggage monitoring and automatic pet identification. There are three different types of tags that may be utilised, the active, semi-active and passive. However, of the three, the passive tag is the easiest and most cost effective to implement due to its reasonable price and no battery being required [?].

Unfortunately, due to various anomalies found predominantly within the passive architecture, the cost effective passive RFID systems are only applied to a fraction of its potential utilisation. These anomalies include: Duplicate Readings, which occur when a tag is scanned twice where it should have only been recorded once; Wrong Readings, where data is found where it should not have been; and Missed Readings in which data is not recorded where it should be [?], [?]. If the problems were able to be eliminated, applications such as automatic scanning of items in a trolley in a supermarket may be implemented allowing saving in both cost and effort. Within Figure ??, R1-R4 represents the four readers of a small enclosed area such as a livestock area with mounted

scanners and T1-T5 represent tagged objects within the area such as cattle. Both T1 and T5 are read correctly as they are intended, however T2's reading is duplicated, T3 has a wrong reading and T4 is missed completely. The observational data found from this example, along with the readings that were supposed to be read may be found in Table ??.

In this example, the various anomalies which occur can produce hazardous conclusions for the users of the RFID system. The duplicate anomaly in which T2 is discovered in both R1 and R2 would cause the owner to not know the area which the cattle would be present in as it could be in either. This would be most problematic if the livestock area has designated zones for various animals. If the exact location is not known, the user would have to manually check the locations of the cattle. Similarly, if T3 appears in the quadrant where R2 is situated but is accidentally read by R2, the owner will assume the animal is in R4's zone and not R2 which could result in unnecessary work needed to check if the animal with the T3 tag attached is actually in the area where R4 is located. Finally, the most problematic anomaly would be T4 in which the observation is missed completely in the database, resulting in the owner believing that the animal may have escaped.

2.2 Non-Monotonic Reasoning

Non-Monotonic Reasoning (NMR) is a type of logic specifically designed to commence with many conclusions, and, as new information is presented, derive the correct solution. Clausal Defeasible Logic (CDL) is a type of NMR which was designed to be specifically run on a computer. It allows the option to use one of five proof algorithms, each with various amounts of ambiguity permitted. The different formulae allowed include the μ formula which will only allow factual information; the π formula which allows ambiguity to propagate; the β which blocks ambiguity; the α which allows allow the conjunction of π and β ; and the δ which only allows the disjunction of π and β [?], [?].

2.3 Bayesian Networks

A Bayesian Network is a means of probabilistically finding the most correct solution when given several pieces of information. Each of the probabilities of certain conclusion will be the product of all the information given up until that point. After each of the probabilities have been found, the conclusion achieving the highest probability is found to be the most accurate solution [?].

2.4 Neural Networks

An Artificial Neural Network refers to an intelligent classification technique which has been designed to emulate the processes of the human brain. Information is processed into a feature set which is then fed into the input nodes, passed through various amounts of hidden nodes and hidden layers which each have different weight calculations and is then passed into the output layer. It uses various training techniques such Genetic Algorithms to train the weights of the neurons to accurately classifier the information as the correct output [?].

Table 1: The recordings that took place from the example in Figure ?? and the observations that should have been recorded.

What is Recorded		
Tag EPC	Timestamp	Reader ID
T1	22/12/2010 10:32:43	R1
T2	22/12/2010 10:32:43	R1
T2	22/12/2010 10:32:43	R2
T3	22/12/2010 10:32:43	R4
T5	22/12/2010 10:32:43	R4
What is supposed to be Recorded		
Tag EPC	Timestamp	Reader ID
T1	22/12/2010 10:32:43	R1
T2	22/12/2010 10:32:43	R1
T3	22/12/2010 10:32:43	R2
T4	22/12/2010 10:32:43	R2
T5	22/12/2010 10:32:43	R4

Fig. 2: A high level diagram describing the information flow in our methodology and the steps that takes place from when the data extracted to when it is loaded back into the database.

3 Proposed Methodology

To correct the missed, wrong and duplicate readings found in various spatial-temporal databases including automatic capture technology such as RFID, we have chosen to employ the use of an integrated classifier architecture. Within this system, we take conclusions drawn from the three classifiers and use fusion techniques, such as a Non-Monotonic Reasoning algorithm, a Bayesian Network or taking the Majority answer, to derive a highly intelligent output. In this section, we identify the motivation behind developing our methodology, the architecture we created, the intended scenario of our approach and the assumptions needed for our concept run as intended.

3.1 Motivation

Radio Frequency Identification has been found to have limited functionality due to problems in the system such as data anomalies [?]. If these anomalies were eliminated, the applications that may benefit from RFID would be increased to various other commercial sectors thereby saving cost and effort. Previous approaches have been utilised to eliminate easily found anomalies, such as middleware algorithm used to determine a duplicate observation recorded in the same location in under a second, however these methodologies lack the intelligence needed to properly correct the stored observations to its maximum integrity. Additional past literature has individually stated that there is a need to use both deterministic and probabilistic methodologies to adequately clean the data [?], [?], [?]. With this in mind, we propose an approach that took advantage of both probabilistic and deterministic approaches to bring RFID data cleaning to a

higher level of integrity. We did this because we fundamentally believed that missing data require a level of probability to find the absent information. In contrast, we believe that both wrong and duplicated data will need to have a deterministic approach due to having the information already present and there is less need to rely on probability. We specifically chose two probabilistic approaches (the Bayesian Network and Neural Network) and one deterministic approach (Non-Monotonic Reasoning) to give a probabilistic advantage to the former methods. This is also the reason as to why we chose the global fusion of the classifiers as opposed to pairwise combinations. To counter this, we chose a the novel deterministic Non-Monotonic Reasoning as a fusion technique which permits additional bias to the the Non-Monotonic Reasoning conclusion in its logical rules.

3.2 Architecture

We have divided our methodology into four core components, the *Feature Set Definition*, *Classification*, *Classifier Integration* and *Loader*. Due to the vast differences between the false-positive and false-negative anomalies, we have different classifier integrations for both the duplicate/wrong data, and the missing data. As seen in Figure ??, the Original Data containing the RFID observations, along with the Geographical Data, is passed into the Feature Set Definition where crucial analytical features of the data are identified. This analytical information is then passed into the Classification component where the Non-Monotonic Reasoning, Bayesian Network and Neural Networks are used to determine if a reading is valid or not. The results of the Classifiers are then passed into the Classifier Integration which uses three fusion methods to intelligently determine the validity of each suspicious reading. Finally, after all the information is gathered, the methodology will finally either delete, keep or insert the correct values into the data set within the loader component.

Fig. 3: The visual representation of how the tag's data is broken up into streams and analysed for both the false-positive and false-negative anomalies with the crucial readings (A-D) around the suspicious data found.

Feature Set Definition The first action that the Feature Set Definition takes is to divide up the data into streams that follow the geographical path of each tag using the geographical data passed into the system. Once this is done, suspicious readings are found based on the geographical data supplied at the beginning. Only suspicious readings will be flagged by the system and all other observations will be ignored by our system. For example, if a reading occurs in two locations not within proximity concurrently, it will be flagged as suspicious as it may be a duplicate anomaly. Figure ?? describes the data found for both the false-positive and false-negative anomalies, and the data recorded for each. A major difference between both is that, due to the possibility of a duplicate observation, the spatial and temporal locations of A-D are needed for false-positive

anomalies whereas only the spatial locations are needed for the false-negative analysis. After these values have been calculated, various binary (true or false) analyses are performed on the values obtained from the Tag Streams. These mathematical operations may be found in Table ??.

Table 2: Each of the operations performed on the Tag Stream data to pass to the classifiers.

False-Positive Tag Stream	False-Negative Tag Stream
$b.loc \leftrightarrow x.loc$	$a == b$
$c.loc \leftrightarrow x.loc$	$b \leftrightarrow c$
$b.time == x.time$	$b == c$
$c.time == x.time$	$d == c$
$b.loc == x.loc$	$n == (s - 2)$
$c.loc == x.loc$	$n > (s - 2)$
$a.loc \leftrightarrow x.loc$	$n > (s - 2)$
$d.loc \leftrightarrow x.loc$	
$b.time \leftrightarrow x.time$	
$c.time \leftrightarrow x.time$	

Within Table ??, there are three main comparisons. The first is the equivalence comparison $==$ which will check if the left value is equal to the right. The second comparison is if the left value is within proximity \leftrightarrow to the right value in both spatial and temporal natures. The third comparison we make is to determine if the left value is greater than $>$ the right value. When determining if a value is within proximity, spatially this will mean if the geographical location of the two readers are physically close to each other whereas in the temporal sense, it will check if the time values are within a user-defined threshold. As a default, we have set it to 30 seconds. With regards to the False-Negative Tag Stream operations, n is the number of missing anomalies and s is the shortest path between both b and c . The reason we compare n to $s - 2$ is that the shortest path also contains b and c which may not be needed within the flagged values.

Classification After the crucial analytical data has been found, it will be passed on as a feature set to the classification component of the methodology to be determined if the suspicious observations should be deleted, kept, or inserted into the database. For the false-positive anomalies which contain either a suspected duplicate or wrong reading, there are two conclusions that may be drawn from the classifiers: either delete or keep the values. With regards to the false-negative anomalies, there are five possible conclusions that may be made each with different reader values combinations. When handling missing data in various data sets, there is a need to impute the data back into the database (i.e. generate possible answers to be used when lacking factual information). We have named permutations to be imputed back into the database as seen in Figure ??. The first two permutations consist of substituting the values of readers b and c for all missing data. The third includes finding the shortest path between readers b and c which will be s , and inserting it into the middle of the missing values with b and c values added around it if the shortest path does not cover all missing recorders. Finally,

the fourth and fifth permutation places the shortest path s to either the earliest or latest missing observations and substituting values c or b respectively.

Fig. 4: The five possible Permutations that may be chosen to be imputed for the missing RFID reader values. Please note that s in this figure refers to the shortest path between reader values b and c , and may occupy more than one observational record.

With regard to the classifiers used throughout our experimentation, we have followed the configurations mentioned in literature for correcting anomalies in databases [?]:

- As it has been shown to provide the highest cleaning rate, we have utilised only the α Non-Monotonic Reasoning formula rules for both the false-positive and false-negative data anomalies.
- We trained the Neural Network used for false-positive anomalies with a genetic algorithm that had a large amount of both chromosomes and generations.
- For the Bayesian Network with the false-positive anomalies, as well as both the Bayesian and Neural Networks used for the false-negative anomalies, we implemented a Genetic Algorithm which had a low amount of chromosomes which trained for a large amount of generations [?].
- All training consisted of the information described in the feature set definition with its respective correct output being processed by the various classifiers and modifying the networks to enhance the conclusions being drawn.

Classifier Integration In this work, we have proposed various methods of combining the classifiers together to develop a new intelligent means increasing the integrity of the conclusions being made. To this end, we have introduced three main fusion techniques to integrate the classifiers, the Non-Monotonic Reasoning Fusion (NMR Fusion), Bayesian Network Fusion (BN Fusion) and Majority Rules Fusion (MR Fusion). For the False-Positive anomalies, since the returned determination is either to keep or delete a value, we only need to integrate the classifiers once. However, due to the False-Negative anomalies finding five varied conclusions, we need to run the fusion algorithms five different times for each permutation.

Fig. 5: A visual representation of the rules we implemented when we created the logic engine to deterministically integrate the classifiers in the Non-Monotonic Reasoning Fusion.

With regard to the Non-Monotonic Reasoning Fusion, we took all the conclusions made in the classification component and put them into the logic engine depicted in Figure ???. Each of the values represent either the Bayesian Network (BN), Neural Network (NN) or Non-Monotonic Reasoning (NMR), and the arc between the antecedents

Table 3: A Table depicting the configuration of the values found within the Bayesian Network Fusion technique.

Conclusion	BN		NN		NMR	
	T	F	T	F	T	F
Positive	60%	40%	60%	40%	40%	60%
Negative	40%	60%	40%	60%	60%	40%

gives higher weighting to the value anti-clockwise (i.e. $\sim NMR$ will be have its conclusion overwritten if $BN \wedge NN$ is also proven). We also have made sure that the Non-Monotonic Reasoning will have a slightly higher bias than the probabilistic techniques as this fusion method is deterministic in nature. Table ?? contains the values and weighting we have given to each of the classifiers to be used in the Bayesian Network Fusion. Similarly to the NMR Fusion, we wish to give a slightly higher bias to the Bayesian and Neural Networks as the Bayesian Network Fusion is itself a probabilistic technique. The final technique we have employed, the Majority Rules Fusion, is an unbiased approach which will use the conclusion voted most by the three classifiers as its determination. We hope that by having deterministic, probaiblistic, and non biased fusion methods, we will observe varying results among the different anomalies. In the unlikely event that none of the permutations have been found to be chosen more than once for the false-negative anomalies, a weighting system is employed based on the scale of most unbiased to biased ($3>1>2>4>5$).

Loader After the decision has been made by the integrated classifier, our methodology will then proceed to either delete, keep, or insert the correct values in the data set. We have made the option to either modify the Original Data set if the user is comfortable with the enhanced data sets or to create a new data set keeping the original data set separate for added integrity. Being that this entire process is at a deferred stage of the capture cycle where all the data has been stored, in this work we did not consider the cost of cleaning. However, in the future, we would like to implement a version of this concept that will run in real-time at the stage of data capture.

3.3 Intended Scenario

We have intended to create our methodology for a scenario in which many readers are mounted around a known environment and tags are passing through the area to be scanned. Applications in which this is already conducted include a hospital in which surgical patients are monitored, airports which track luggage and the transportation of various items in a supply chain. It is crucial that a known environment is used in the scenario as the geographical locations of each of the readers and their proximity to one another must be recorded in the system.

3.4 Assumptions

There are two assumptions we have identified for this scenario relating both to the identification of the false positive and false negative anomalies. The first is that, as

stated in the intended scenario, the geographical locations of the readers must be known to the readers so that a tag which is recorded at abnormal locations may be flagged as a suspicious reading. The second assumption we make is that the time used to flag a missed reading is less than the time it takes for the tagged object to move from one readers scan range to another. Both of these assumptions are crucial as they provide the rules that our system follows to identify a suspicious set of observations to be corrected.

4 Experimental Evaluation

To properly test our integrated classifier, we have devised four experiments to determine the overall effectiveness and advantages it has over existing techniques. The first two experiments will be carried out to test the effectiveness of each of the fusion techniques for solving both false positive anomalies (duplicate and wrong data), and false negative anomalies (missing data). The second two experiments will take the best performing integrated classifier to compare it to state-of-the-art techniques currently used to enhance the integrity of RFID data. These experiments will be performed on multiple test beds to determine its effectiveness on varying amounts of anomalies.

4.1 Environment

To properly evaluate the effectiveness of our methodology, we have used simulated test cases of the information obtained from readers. We have created five test beds with 500, 1,000, 1,500, 2,000 and 2,500 test cases each to observe the performance of the approaches where there are various amounts of anomalies. Each of the test cases present within the test bed represent a found anomaly within the data sets. All code used in our methodology was written in the C++ language and executed in Microsoft Visual C++ 6.0. The computer used for this experimentation was a Microsoft Windows XP machine with Service Pack 3 Intel (R) Core 2 Duo CPU E8400 @ 3 GHz 2.99 GHz with 4 GB of RAM.

4.2 Results

Fig. 6: The False-Positive results of Bayesian Network (BN), Neural Network (NN), Non-Monotonic Reasoning (NMR), Fused Non-Monotonic Reasoning (FNMR), Fused Bayesian Network (FBN) and Fused Majority Rules (FMR) when tested against, 500, 1,000, 1,500, 2,000, 2,500 test cases and the average.

The first experiment we ran included testing the percentage of clean data for the Bayesian Network, Neural Network, Non-Montonic Reasoning, Fused Non-Monotonic Reasoning, Fused Bayesian Network and Fused Majority Rules classifiers when attempting to clean 500, 1,000, 1,500, 2,000 and 2,500 False-Positive test cases. From the False-Positive results found in Figure ??, the highest performing classifier average

has been found to be the Fused Non-Monotonic Reasoning classifier. The absolute highest performing classifier was also the Fused Non-Monotonic Reasoning classifier when attempting to clean 500 test cases. The least performing classifier for the False-Positive experiment was the Bayesian Network when attempting to clean 500 test cases. We believe that the advantage the Fused Non-Monotonic Reasoning classifier had was due to its deterministic architecture and the nature of False-Positive anomalies.

Fig. 7: The False-Negative results of Bayesian Network (BN), Neural Network (NN), Non-Monotonic Reasoning (NMR), Fused Non-Monotonic Reasoning (FNMR), Fused Bayesian Network (FBN) and Fused Majority Rules (FMR) when tested against, 500, 1,000, 1,500, 2,000, 2,500 test cases and the average.

In the second experimentation we conducted, we took the same classifiers and test case amounts, however we used False-Negative anomalies rather than False-Positive. The results, which may be viewed in Figure ??, has shown that the highest performing classifier average has been found to be the Fused Majority Rules classifier. The highest performing clean on the data sets has been found to be the Fused Majority Rules classifier as well when attempting to clean 1,500 test cases. The least achieving classifier for the False-Negative anomalies has been found to be the Bayesian Network classifier when attempting to clean 500 test cases. The results have shown that the unbiased nature of the Fused Majority Rules classifier has given it a clear advantage when cleaning False-Negative anomalies. It is also important to observe that as highlighted in the above results, of the two types of anomalies, it is harder to correct the False-Negative anomalies.

The Non-Monotonic Reasoning fused classifier was able to achieve the highest false-positive clean as its deterministic nature makes it ideal to clean wrong and duplicate data whereas probabilistic techniques would introduce an additional level of ambiguity. With regards to the false-negative anomalies, the Majority Rules fused classifier gained the highest cleaning rate due to it being able to accept all three classifiers without any bias. Additionally, we believe we may have obtained a higher result if we introduced a dynamically trained fused Bayesian Network or created a Fused Neural Network classifier. As this methodology is designed to be applied at a deferred stage of the RFID capture cycle, our experimentation was not concerned with the runtime performance. However, we would like to extend and modify our approach in the future to allow real-time processing in which case we will be taking the processing time into consideration for each classifier and focusing on the amount of time needed to generate rules or train the networks.

5 Conclusion

In this paper, we presented a methodology to clean anomalous Spatial-Temporal data using an integrated classifier. For this study, we investigated RFID technology as our case study as it continues to generate anomalies within the recorded data sets and has

a need to be rectified before it can be employed in various other commercial sectors. Through experimental evaluation, we have found that the highest performing fusion type for wrong and duplicate data was the Fused Non-Monotonic Reasoning classifier, while the Fused Majority Rules approach was the most effective for cleaning missing readings. We then compared each of the highest achieving integrated classifiers against state-of-the-art and currently utilised approaches and found that our technique provides superior integrity. With regard to future work, we would like to investigate other fusion approaches of additional classifiers such as the Support Vector Machine and other classifier training techniques. We would also like to apply our technique to various databases as we believe that our methodology is not limited to merely cleaning RFID data and may be applied to other spatial-temporal data collections as well. Also, as mentioned earlier, we would like to employ a real time implementation of this concept.

References

1. Yang, Q.: Activity recognition: linking low-level sensors to high-level intelligence. In: Proceedings of the 21st International Joint conference on Artificial intelligence (IJCAI). (2009) 20–25
2. Chawathe, S.S., Krishnamurthy, V., Ramachandran, S., Sarma, S.E.: Managing RFID Data. In: VLDB. (2004) 1189–1195
3. Jeffery, S.R., Garofalakis, M.N., Franklin, M.J.: Adaptive Cleaning for RFID Data Streams. In: VLDB. (2006) 163–174
4. Darcy, P., Stantic, B., Mitrokotsa, A., Sattar, A.: Detecting Intrusions within RFID Systems through Non-Monotonic Reasoning Cleaning. In: Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2010). (2010) 257–262
5. Billington, D.: Propositional Clausal Defeasible Logic. In: European Conference on Logics in Artificial Intelligence (JELIA). (2008) 34–47
6. Billington, D.: An Introduction to Clausal Defeasible Logic [online]. David Billington's Home Page (Aug 2007) Available from: <<http://www.cit.gu.edu.au/~db/research.pdf>>.
7. Zio, M.D., Scanu, M., Coppola, L., Luzi, O., Ponti, A.: Bayesian Networks for Imputation. Journal Of The Royal Statistical Society Series A **167**(2) (2004) 309–322
8. Blumenstein, M., Verma, B.: A Neural Based Segmentation and Recognition Technique for Handwritten Words. In: The 1998 IEEE International Joint Conference on Neural Networks Proceedings. Volume 3. (May 1998) 1738–1742
9. Darcy, P., Stantic, B., Derakhshan, R.: Correcting Stored RFID Data with Non-Monotonic Reasoning. Principles and Applications in Information Systems and Technology (PAIST) **1**(1) (2007) 65–77
10. Rao, J., Doraiswamy, S., Thakkar, H., Colby, L.S.: A Deferred Cleansing Method for RFID Data Analytics. In: VLDB. (2006) 175–186
11. Khoussainova, N., Balazinska, M., Suci, D.: Probabilistic Event Extraction from RFID Data. In: International Conference on Data Engineering. (2008) 1480–1482
12. Darcy, P., Stantic, B., Sattar, A.: Correcting Missing Data Anomalies with Clausal Defeasible Logic. In: Advances in Databases and Information Systems (ADBIS 2010). (2010) 149–163