A STUDY OF DATA QUALITY ISSUES IN MOBILE TELECOM OPERATORS

Naiem Khodabandhloo Yeganeh and Shazia Sadiq
School of Information Technology and Electrical Engineering
The University of Queensland, St Lucia, QLD 4072, Brisbane, Australia

Keywords: Data Quality, Telecommunication Operators, Case Study.

Abstract: Telecommunication operators currently servicing mobile users world-wide have dramatically increased in the last few years. Although most of the operators use similar technologies and equipment provided by world leaders in the field such as Ericsson, Nokia-Siemens, Motorola, etc, it can be observed that many vendors utilize propriety methods and processes for maintaining network status and collecting statistical data for detailed monitoring of network elements. This data forms their competitive differentiation and hence is extremely valuable for the organization. However, in this paper we will demonstrate through a case study based on a GSM operator in Iran, how this mission critical data can be fraught with serious data quality problems, leading to diminished capacity to take appropriate action and ultimately achieve customer satisfaction. We will further present a taxonomy of data quality problems derived from the case study. A breif survey of reported literature on data quality is presented in the context of the taxonomy, which can not only be utilized as a framework to classify and understand data quality problems in the telecommunication domain but can also be used for other domains with similar information systems landscapes.

1 INTRODUCTION

A mobile telecom operator (MTO) is a company that provides mobile telecommunication services for its end customers and it usually utilizes a variety of technologies, vendors, and sub-contractors to provide such services. Although these different systems and organization are not designed by MTO, MTO is ultimately responsible for customer satisfaction and has to monitor its network precisely and efficiently. This variety inevitably leads to data quality problems.

In this paper we present a study of data quality problems derived from the telecom domain. We firstly present background literature on data quality and an overview on problems and solutions in section 2. In section 3, a taxonomy of data quality problems for MTO is presented and different real-world data quality issues in a MTO are studied. In section 4, an analysis of MTO data quality problems and data quality solutions is presented. Finally, in section 5, the outlook of future works with a conclusion is presented.

2 BACKGROUND LITERATURE

Consequents of poor quality of data have been experienced in almost all domains. From the research perspective, data quality has been addressed in different contexts, including statistics, management science, and computer science (Scannapieco, 2005). To understand the concept, various research works have defined a number of quality dimensions (Scannapieco, 2005), (Wang, 1995), (Fox, 1994).

Data quality dimensions characterize data properties. Many dimensions are defined for assessment of quality of data that give us the means to measure the quality of data as a value. These dimensions can be grouped in four categories (Fox, 1994), (Scannapieco, 2005): accuracy, currency, completeness, and consistency.

A variety of solutions are proposed for data quality problem. These solutions can be categorized into groups regardless of the domain they are applied on:

Semantic integrity constraints are used for maintaining data integrity. Integrity constraints can be used for preventing database from loosing quality as well as repairing quality problems.

Record linkage solutions are used to detect duplicate record which are not always identical in

Khodabandhloo Yeganeh N. and Sadiq S. (2008).
A STUDY OF DATA QUALITY ISSUES IN MOBILE TELECOM OPERATORS.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - DISI*, pages 441-444
DOI: 10.5220/0001689904410444

value, but represent the same entity. For example, same name can be written in two different formats, but might identify the same person or a SMS can be modified during the distribution from one to one, but conveying the same meaning.

Data lineage or provenance solutions are classified as annotation and non-annotation based approaches where back tracing is suggested to address auditory or reliability problems.

Schema mapping solutions are designed to map schemas between different data sources or views. See (Rahm, 2000) where issues and approaches of database integration are studied taking schema mapping requirements into consideration.

Data uncertainty and completeness are important unsolved problems in databases and it is important to efficiently integrate models of uncertainty in database systems (Silberschatz, 1991).

3 TAXONOMY OF DATA QUALITY ISSUES FOR MTO

MTO data has specific requirements due to its nature, but many data quality dimensions as described in the previous section apply to this field generally. Below, we provide some examples of how various data quality dimensions can be found in MTO data.

Figure 1 shows the weekly average of the daily peak hours of the difference in the number of traffic channel assignments reported by the two major components of a GSM network: radio and core, at the same time from the same vendor. We see an average 2% to 5% difference. This difference will decrease call setup success rate and answer seizure rate which will have bad effects on customer satisfaction and the operator's income.

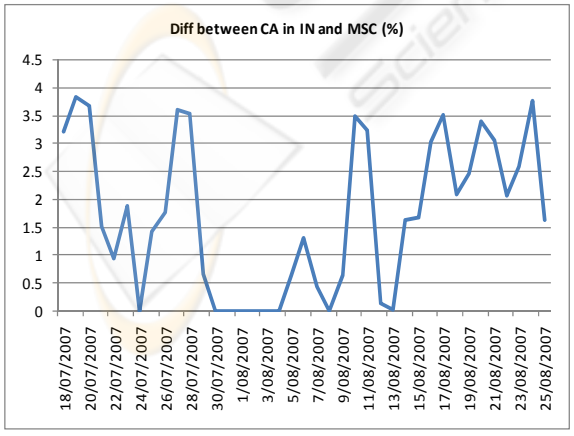

Figure 1: Difference between call attempts reported by IN versus MSC.

Different parts of a telecom network are independently designed and managed, but they co-operate as a single system. This independence causes inconsistency problems in database. For example, it is likely to have a name mismatch problem between different elements of the network. For example, you need to know that Trunk Group with name TGRP_BSC28 is going to the Base Station Controller named BSC_QAZVIN for the city of Qazvin and these names should be linked together.

Data lineage is also an important problem in such databases; for example, analysis of statistics may show a sharp change in some indicators at a past time where engineers or managers may need to know what was changed, who changed it, and why it was changed.

Below, we categorize MTO data quality problems into four major categories to constitute taxonomy for MTO specific data quality issues:

Lack of Consistency. Inconsistency emerges when same measurements from different parts of the network don't mach. This difference can have several reasons including; different vendors, unknown errors, inexact time synchronization between different elements or other technical reasons.

Lack of Sufficient Error Recognition. Sometimes all of the errors are not calculated and thus, the values do not match. For example, there is a rule that the number of incoming calls to a switch should be equal to the number of outgoing calls plus the number of rejected, congested or overloaded calls. However our study shows marked differences due to uncounted rejected calls. This difference is an effect of data incompleteness.

Lack of Standardization. There is still no global standard for data interaction by different networks and it is not always possible to map counters and measurements from different vendors, e.g. to calculate the number of Failed Random Access Channels in a network consisting of both Alcatel and Siemens radio equipments).

Lack of Synchronization. Time related issues frequently occur in MTO data, e.g. when the microwave link is disconnected temporarily, the number of active traffic channels might report a fake traffic. This issue is a direct effect of using out-of-date data in the system.

4 RELATIONSHIP WITH REPORTED LITERATURE

Although, all data quality solutions provided may not be directly or practically applicable for industry

use, in the discussion below we extract relevant techniques from literature and establish their applicability for MTO data quality requirements as developed through our case study.

Table 1 and Table 2 show relations between data quality problems in a MTO and traditional data quality domain in brief. Table 2 infers that data quality issues discussed for MTOs are present in other domains and solutions proposed therein can be applied in the context of MTOs. The numbers in cells correspond to references as below:

1. (Bohannon, 2007), (Cong, 2007)
2. (Elmagarmid, 2007), (Gravano, 2001), (Tejada, 2001) (Scanappieco, 2007), (Dasu, 2002), (Rahm, 2000), (Miller, 2000)
3. (Buneman, 2002), (Bhagwat, 2004) (Srivastava, 2007)
4. (Scanappieco, 2007), (Dasu, 2002), (Batani, 1986), (Rahm, 2000), (Miller, 2000)
5. (Silberschatz, 1991), (Thone, 1995)
6. (Ballou, 2003), (Chomicki, 2005) ,(Churces, 2004), (Al-lawati, 2005), (Freedman, 2004)

Table1: Relation between traditional DQ problems and MTO data quality problems.

MTO / DQ Problem / Dimension	(i)	(ii)	(iii)	(iv)
Currency/ Timeliness				2
Accuracy/ Reliability	1	3	2	
Completeness		5		
Consistency	6		4	

Table 2: Relation between DQ solutions and MTO data quality problems.

MTO / DQ Problem / Solution	(i)	(ii)	(iii)	(iv)
Semantic Integrity	1	1		1
Record Linkage	6		2	
Data Lineage		3		
Schema mapping			4	
Certainty Completeness		5		

(i) Lack of Consistency; (ii) Lack of Error Recognition; (iii) Lack of Standardization; (iv) Lack of Synchronization.

Lack of Consistency. (Ballou, 2003) provides a model to calculate trade off between consistency and completeness of data. Works such as join estimation are also useful for improving the consistency of data. (Chomicki, 2005) (Elmagarmid, 2007) , (Churces, 2004) , (Al-lawati, 2005) and (Freedman, 2004) proposes some methods for record-linkage. For a complete survey of record linkage see (Elmgarmid, 2007). These techniques can be used to find inconsistencies in statistical data. (Bohannon, 2007), propose and utilize a concept called conditional functional dependencies that can be useful to maintain MTO database consistency.

Lack of Sufficient Error Recognition. A closely associated aspect of error recognition is error tracking. Backtracking for data lineage discussed in (Buneman, 2002), (Bhagwat, 2004), and (Srivastava, 2007) can be applied to determine if a changed record is truly dirty or the change is deliberate.

Lack of Standardization. Some automatic standardization can be defined within the database system, e.g. methodology presented in (Dasu, 2002) which automatically collects schema model of the database and extracts dependencies can be useful. Techniques for similarity analysis (Elmagarmid, 2007), approximate joins (Gravano, 2001) and mapping tables (Tejada, 2001) can assist here.

Lack of Synchronization. Currency, age and timeliness of data are key characteristics of the data. (Bullou, 1985) provides a model for data quality which supports data expiration. (Pernici, 2002) presents a methodological framework for information systems where currentness of data is important.

5 OUTLOOK

The number and diversity of mobile telecommunication operators world-wide has led to a highly competitive industry. A major contributing factor towards achieving the above goal is the quality of the data upon which both operations as well as the business decisions rely on.

We have conducted a case study of a GSM operator and studied in detail how data quality issues emerge in the various activities of the organization. Through this study we have derived a taxonomy of data quality problems for MTOs. The taxonomy provides a forum to analyse and relate the domain specific problem for MTOs to the wider data quality problem and solution space.

We would like to remark that the data quality problems mentioned in the previous sections do not represent the entire spectrum of data quality issues for MTOs. The nature of these organizations and their sensitivity to the quality problems was

demonstrated by examples which usually occur in the radio, core and statistics databases. The operations of MTOs have several other aspects which were not considered in our study.

Through this study, we hope that data quality problems for MTOs in specific and for the telecommunication industry in general can be better understood. We also envisage that this study can provide the roadmap for industry relevant research on data quality.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the help of the unnamed MTO through which this case study was developed

REFERENCES

Al-Lawati, A., Lee, D., McDaniel, P. 2005. *.locking-aware Private Record Linkage,* IQIS

Ballou, D. P., Pazer, H. 2003. *Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts.* IEEE Transactions on Knowledge and Data Engineering , vol. 15, no. 1.

Ballou, D.P., Pazer, H.L. 1985. *Modeling data and process quality in multi-input, multioutput information systems*. Management Science, vol.31, no.2, 1985.

Bhagwat, D., Chiticariu, L., Vijayvargiya, T. G. 2004. *An Annotation Management System for Relational Databases.* VLDB.

Bohannon, P., Fan, W., Geerts, F., Jia, X., Kementsietsidis, A. 2007. *Conditional functional dependencies for data cleaning*. In ICDE.

Buneman, P., Khanna, S. 2002. *On Propagation of Deletions and Annotations through Views.* PODS

Chomicki J., Marcinkowski. J. 2005. *Minimal-change integrity maintenance using tuple deletions.* Inf. Comput., 197:90–121.

Churces, T., Christen, P. 2004. *Some Methods for Blindfolded Record Linkage*. BMC Medical Informatics and Decision Making 4, no. 9.

Cong, G., Fan, W., Geerts, F., Jia, X., Ma S. 2007. *Improving Data Quality: Consistency and Accuracy*. VLDB: 315-326.

Dasu, T., Johnson, T., Muthukrishnan, S., Shkapenyuk, V. 2002. *Mining database structure; or, how to build a data quality browser.* SIGMOD Conference : 240-251.

Elmagarmid, A. K., Panagiotis, G.I., Verykios, S.V. 2007. *Duplicate Record Detection: A Survey.* IEEE TKDE 19, no. 1.

Fox J. C., Levitin, A., Redman, T. 1994. *The Notion of Data and Its Quality Dimensions.* Inf. Process. Manage. 30(1): 9-20.

Freedman, M. J., Nissim, K., Pinkas, B. 2004. *Efficient Private Matching and Set Intersection*. EUROCRYPT.

Gravano, L., Ipeirotis, P.G., Jagadish, H. V., Koudas, N., Muthukrishnan, S., Srivastava, D. 2001. *Approximate string joins in a database (almost) for free.* In Proceedings of the 27th International Conference on Very Large Databases (VLDB), pages 491–500.

GSM Association. 2007. http://www.gsmworld.com/technology/glossary.sht8

Miller, R., Haas, L., Hernandez. M. 2000. *Schema mapping as query discovery.* In Proc. 26th VLDB Conf., pages 77-88.

Pernici, B., Scannapieco M. 2002. *Data Quality in Web Information Systems.* ER : 397-413.

Rahm E., Do. H. 2000. *Data cleaning: Problems and current approaches.* IEEE Data Engineering Bulletin, 23(4):1-11.

Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A. K. 2007. *Privacy preserving schema and data matching.* SIGMOD Conference : 653-664.

Scannapieco, M., Missier, P., Batini, C. 2005. *Data Quality at a Glance*. Datenbank-Spektrum 14: 6-14

Silbershutz, A., Stonebreaker, M., AND Ullman, J. D. 1991. *Database systems: Achievements and opportunities.* In Communication with ACM 34, 10, 110–119.

Srivastava, D., Velegrakis, Y. 2007. *Intensional Associations Between Data and Metadata.* SIGMOD.

Tejada, .S, Craig, A., Knoblocka, A., Minton, S. 2001. *Learning object identification rules for information integration*. Information Systems Volume 26, Issue 8, Pages 607-633.

Thone, H., Kiessling, W., Guntzer, U. 1995. *On cautious probabilistic inference and default detachment.* Ann. Oper. Res. 55, 195–224.

Wang, R. Y., Storey V. C., Firth, C. P. 1995. *A Framework for Analysis of Data Quality Research.* IEEE Trans. Knowl. Data Eng. 7(4): 623-640