Research Journal of Applied Sciences, Engineering and Technology 6(15): 2736-2739, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 27,2012 Accepted: February 01, 2013 Published: August 20, 2013

Data Quality Analysis on the Customer Churns in Telecom

Qin Hai-fei
Computer Science Department, ChuXiong Normal University, ChuXiong, YunNan, China

Abstract: With the development of information technology, data quality has become a focus of attention. This study studies the data source, the schema and the instance in telecommunication churn data, put forward different sector has different definition of data quality, as a basis work of data analysis for further study.

Keywords: Customer churn, data quality, instances, schema

INTRODUCTION

As a leading telecommunications information industry, own many of the data. How to guarantee the data is true and effective when the customer inquiries, How to let data become wealth, not burdensome. It becomes the goal of telecom operators pursue. To make the data better for telecommunications services, create more value, telecom enterprise must manage their data; ensure the data is true, accurate and reliable.

Data quality problem is the core issue of business intelligence, data quality direct impact on the data results, at present, because of the data from diverse sources, let data quality became complex .Because the object, purpose, use different, the definition of data quality is different.

DATA QUALITY DEFINITION

So far, the definition of data quality is diversity. Some time, many customer see data quality as information quality, that is not exactly. Since the 1950 s, people have started definition data quality from the angel (Center for Innovation in Engineering Education at Vanderbilt University, 2006; Song and Qin, 2007). We can find five categories of definitions, consumer-based, manufacturing-based, product-based, value-based, priori-based (Song and Qin, 2007).

'Data quality refers to the degree of excellence exhibited by the data in relation to the portrayal of the actual scenario' (http://en.wikipedia.org/wiki/Data-quality).

'Data quality is completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.' by Government of British Columbia defined.

'The totality of features and characteristics of data that bears on their ability to satisfy a given purpose, the sum of the degrees of excellence for factors related to data' by Glossary of Quality Assurance Terms.

'Data quality is the processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria.' 'Complete, standards based, consistent, accurate and time stamped by Glossary of data quality terms published' by IAIDQ (http://en.wikipedia.org/wiki/Data-quality).

From the above, Data quality standards the data get satisfaction in consistency, correctness, completeness and minimality (Aebi and Perrochon, 1993; Guo and Zhou, 2002; Han *et al*., 2008).

'Data Engineering demand, The Data quality requirements analysis and modeling think that there are many choose in quality metrics, the user should select the part of the application requirements' (Wang *et al*., 1993; Guo and Zhou, 2002).

From the above, this study analyzes the quality of data as to interesting topic based on the consumer, manufacturing, products, integrated multi-dimensional value of the data, According to the characteristics of the telecom data that both statistical data and real-time data, but also has web data.

DATA SOURCE ANALYSIS ON CHURN CUSTOMER

Different data sources, different subject-oriented and different requirements, the data quality is different. This study as to the churn customer data quality analysis. And the data of customer churn are mainly from the information service department in real-time database and statistical databases, the customer service department in history databases and reports, the network-centric in network services database, the market expansion department in budget databases (Fig. 1).

From Fig. 1,we can see telecom customers of the loss of data is multi-data source, the data quality analysis of telecommunications customer churn is to analysis the database of churn customers. From the point of view of the database, the quality of the data analysis is mainly the database schema and instance. The data quality problem in database can be seen as the problem of store data in three-level schema two mapping. In the

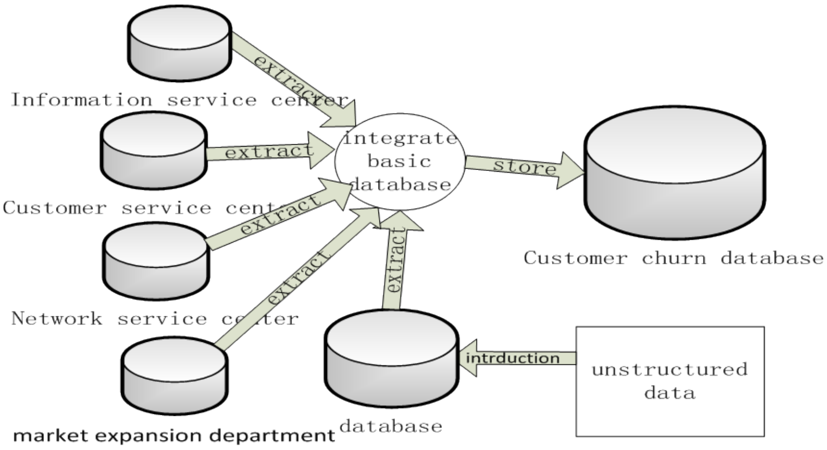

Fig. 1: Customer churn data source

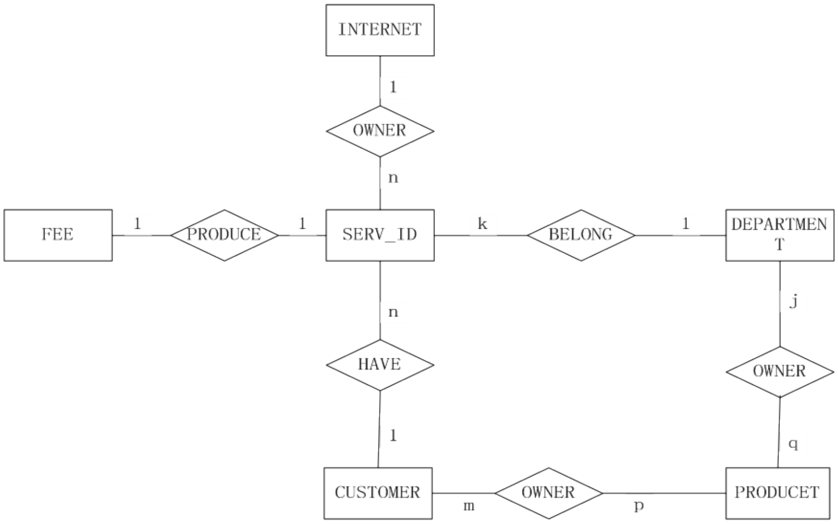

Fig. 2: Customer churn E-R diagram

three-level schema, the storage schema is defined by the database management system. This text mainly discusses external schema and schema. Figure 1 integrated analysis data, solved the problem of the external schema and schema. The schema is called relation schema, include single-schema and multi-relational schema. For multi-relational schema, need to check the relationship between schema and schema. In order to facilitate analysis, integrated all analytical data and stored in the customer churn database.

The analysis data from multi-source, the data existing redundancy, so need to analysis the data schema and instance.

DATA SCHEMA ANALYSIS

Data source is integrated with multiple data sources, involved in a number of different external schemas, so needs to analyze the entire schema.

First, according to the relationship of each data in the customer churn database, analyzed their contact and designed the different entity-relation diagram, combined with the entity- relation diagram. Solve the problem of different name synonymous and the same name meaning different. Solve the problem of object abstraction (the same object have different abstract in different applications), the entity properties and relation between the entities; solve the problem of the different of the attribute value type, range, set and so on. Make the merged entity relation map does not a naming conflict, structure conflict, properties conflict. It is Fig. 2 as follows.

From the Fig. 2, we can see from the SERV-ID is centre, because of the SERV-ID is unique in telecom, is centre in all work.

Checking schema, in a non-redundant mode conditions also need to check all external schema.

According to the control mechanism of data integrity provides the mechanism defined integrity constraints, integrity checking methods and default processing mode to check the external schema integrity constraints, to improve data quality.

Although the external schema definition primary and foreign keys, to improve the quality, but in the Large and Super large database, it will cause a greater impact on performance of system, so some information industry give up integrity constraints. Because of it leads to the data problems of kinds. Generally, for the smaller database, to improve data quality of external schema, it has strictly integrity constraint, for large systems, the integrity constraints does not strictly. These database performance-based considerations many table canceled integrity constraints and also to causing the problem of data quality, it will adopt instance analysis to solve quality problem.

EXAMPLE ANALYSIS

Example analysis is mainly check the content of external schema, check the data what issues? Mainly check data the content of table.

Missing values analysis: In addition to defining the primary key and foreign key in the database, the other attributes are likely to existence the missing values. It is because of what is input and what is right-wrong when input information. For this part of the data, it must determine whether the data existence. If it must, It can re-investigation, man-made or mathematical method (regression, Bayesian, decision tree, the mean or global constants, etc.) (Luai *et al*., 2006) to fill, Otherwise can ignore, when taking data can be limited conditions and did not take ignore data. In the analysis, fill the missing values is a complex and difficult work, so the database will adopt invalid value to solve it.

The same data have a different because of misspelling and inconsistent understand. Such as some people's names, place name and professional vocabulary abbreviations spelling have different, which is considered to be different data. Due to the problem, it can be established the specialized vocabulary correlation data dictionary table.

Vacancy values analysis mainly view data types, if the data type is consistent, we must check the credibility of the content to judge analysis.

Duplicated records analysis: Approximately duplicated records examining method, for the small data usually adopt a similar analysis after the first order. For the large data, you can define and edit distance function, or to establish common rules for the abbreviation dictionary table to be resolved.

Other example analysis: Except above analysis, it will divide data type into numeric and text to discuss other problem.

- **Numerical data analysis methods:**
    - **Probabilistic method:** Using probabilistic method test white-noise, adopt binning, clustering, regression and other methods to remove noise.
    - **Clustering method:** Using clustering method to collect data and aggregate
    - **Generalization method:** Using the high-rise data in data cube instead of the low-rise raw data
    - **Normalized method:** Using some method (min-max normalization, z-score normalization and normalization by decimal scaling) normal data
- **Text data analysis methods:**
    - **Matching method:** Using Dictionary table look-up and fuzzy Matching
    - **Contrast method:** Extracted from the text structure, separate parts, compared with the dictionary tables for verification
    - **Verification method:** According to the industry and domain knowledge analysis fields, verification dependencies between fields
    - **Unified Form method:** The same type of data used to represent a uniform format, such as date, phone number and gender and so on

From above, analysis the relationship instance, it is solve the problem of data content.

CONCLUSION

This study based on the definition of data quality, analysis the data of telecom customer churn. From the type of data quality point of view, telecommunications data is statistical, real-time, but also web data. From the definition of data quality point of view, data quality is consumer-based, manufacturing-based, product-based, value-based, etc. multi-dimensional data. We need to analyze the data in various departments in telecommunication and gives the precise definition of data quality, such as follow:

- The data from information services department is a true portrayal of customer consumption in telecommunications. Data quality requires data should have authenticity, accuracy, timeliness, effectiveness.
- The data from customer service center is customer own such as contact attribute, background properties is by the customer in their data. Data quality requires data should have interpretability, easy to understand, concise, consistent.
- The data from customer service center is telecom given, example for account ID attribute, is filled by

a telecommunications service. Data quality requires data should have accuracy, interoperability, flexibility, cohesion, consistency, substitutability, interpretability, easy understanding.

- The data from network service center is record by computer or record by telecom staff, is a key of telecommunications service whether successful. Data quality requires data should have Accuracy, completeness, consistency, real-time and uniqueness.
- The data from the market expansion department, such as budget and new product development data, is new data recorded by telecommunication. Data quality requires data should have applicability, accuracy, timeliness, accessibility, comparability and cohesion.

From above analysis, the definition of data quality is different because of the industry, field and use different. In future, we must make concrete analysis of concrete conditions to ensure data quality.

REFERENCES

Aebi, D. and L. Perrochon, 1993. Towards improving data quality. Proceedings of the International Conference on Information Systems and Management of Data. Delhi, 1993: 273-281.

Center for Innovation in Engineering Education at Vanderbilt University, 2006. What is Quality: Definitions and Contrasts? Retrieved from: http:/mot.Vuae.vanderbilt.edu/mt322/What is.htm.

Guo, Z.M. and A.Y. Zhou, 2002. Data quality and data cleansing: A survey. Res. J. Softw., 13(11): 2076-2082.

Han, J.Y., L.Z. Xu and Y.S. Hang, 2008. Review of data quality research. Comput. Sci., 2: 1-5.

Luai, A.S., S. Zyad and K. Basel, 2006. Data mining: A preprocessing engine. J. Comput. Sci., 2(9): 735-739.

Song, M. and Z. Qin, 2007. Reviews of foreign studies on data quality management. J. Inform., 2: 7-9.

Wang, R.Y., H.B. Kon and S.E. Madnick, 1993. Data quality requirements analysis and modeling. Proceedings of the 9th International Conference on Data Engineering. IEEE Computer Society, Vienna, 1993: 670-677.