http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (8): 1523-1531, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1523.1531

An AHP-based Approach for Banking Data Quality Evaluation

[1]Keting Yin, [1]Yufei Pu, [2]Zirui Liu, [2]Qi Yu and [3]Bo Zhou
[1]School of Software Technology, Zhejiang University, Ningbo, 315048, China
[2]Bank of Dalian, Dalian, 116001, China
[3]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China

Abstract: With the growth of business in modern enterprises, data volume in enterprise is increasing significantly. The quantity of the data has entered the age of TB in most enterprises. As a result, data quality has become an important and attention-needed part in the construction of information systems in enterprises, especially in the banking industry, where data quality is even more crucial. Data quality issues in banks often lie in aspects such as incompleteness, inconsistency, invalidity, untimeliness, which has not only resulted in the bank's huge direct economic loss but has also influenced its strategic decisions to some extent. As an important part in data quality management, data quality assessment plays an indispensable role in data quality assurance. This study proposes an evaluation index system for banking data quality and an AHP-based evaluation method used to calculate weighting coefficients of each index in the data quality evaluation index system. The method proposed in this study is based on a real project of data quality assessment in a commercial bank. With a purpose to establish a whole set of data quality evaluation system and coefficient weighting method for the bank, the project produces accurate, systematic and scientific data quality assessment results, which is instructive to the implementation of data quality assessment in other banks.

Key words: Analytic hierarchy process, banking data quality, data quality assessment

INTRODUCTION

Data is a kind of product, so quality is its necessity. There are different definitions for Data Quality (DQ). Simply put, data quality is the extent to which the user's need is satisfied by the data. In many papers, DQ is the same as Information Quality (IQ), or is just defined as a set of attributes including Accuracy, Completeness, Timeliness and Consistency, etc. Data Quality Management (DQM) is a series of management activities such as recognizing, weighing, monitoring, warning of various data quality issues happening in each phase of data's life cycle, from planning, obtaining, storing, sharing to maintaining, applying, dying and to improve DQ through upgrading the management of the organizations. DQM is a cyclic process and the ultimate purpose is to win economic benefits for enterprises by increasing the value of data in use with the usage of reliable data.

A current survey on enterprise data management shows that more than a half of enterprises have already enjoyed a TB level of data quantity and some of them have even surpassed 10 TB. In a word, data management system in enterprise has entered the age of TB. Though most enterprises have paid enough attention to data quality, how to get started still remains a problem to enterprises. Financial and banking industries stand out in numerous industries because of their enormous data volumes, which have made their problems in DQM more severe.

With the diversification of banking business, the quantity of information system data continues to increase. Meanwhile, the changes in customer and business data as well as the integration of application systems have also contributed to the increasingly prominent banking data quality problems, such as incompleteness, inconsistency, invalidity, untimeliness in data (Xie, 2009). The occurrence of data quality issues is closely related to data source, for it is likely to be caused by the quality of a single data source or the inconsistency among several data sources. In the former situation, data quality of a single data source is mainly decided by its model's control over data completeness, while in the latter situation, the problem occurs in the former one will appear in different forms.

As a bank's basic data platform is often correlated with tens of important businesses and management systems and has plenty of application data sources as well as different kinds of databases which have their own extensions and discrepancies, a lot of data quality problems will occur. Currently, most of the banks do not have a set of complete and systematic data quality

Corresponding Author: Bo Zhou, College of Computer Science and Technology, Zhejiang University, 310027, Hangzhou, China Tel: +86-0574-27830788 Fax: +86-0574-27830606

evaluation index system, so the data quality of the banks could hardly be objectively and rationally evaluated. The data with quality problems will not only cause enormous direct economic loss of the bank, but will also influence its strategic decisions. Therefore, the evaluation of banking data quality has become an indispensable part in banking data quality management.

The selection of dimensions plays an important role in the process of data quality evaluation. Weighting on an individual dimension will easily result in the neglect of influences and relations between the evaluation indexes in an evaluation system consisted of multiple indexes, which is often inaccurate and partial. Thus, a multi-dimensional data quality evaluation system needs to be established.

Meanwhile, in order to indicate the relative importance of the dimensions, weight coefficient is always used to evaluate the importance of a dimension. Presently, there are tens of established methods for weight coefficient of evaluation index around the world. According to data source and computing process, those methods can be roughly classified into Subjective Weighting Method, Objective Weighting Method and Comprehensive Weighting Method and can be applied in different situations.

In consideration of banking data's characteristics such as multiple sources and influenced by different dimensions as well as the analysis of banking data platforms at the present stage, the contributions of this paper are as follows:

- Propose an evaluation index system suitable to data quality evaluation of the banks. This system is consisted of six main dimensions including Completeness, Validity, Timeliness, Accuracy, Consistency and Uniqueness and their sub-indexes
- Propose an AHP-based comprehensive weighting method which will weight according to the relative importance of the dimensions in the proposed system and quantize data quality in the basic banking data platform through weighted average
- Apply the proposed data quality evaluation system and AHP-based weight coefficient approach to a real data quality evaluation project of a commercial bank and conduct a detailed case analysis

RELATED WORK

Data quality: Data quality is a multi-dimensional definition (Pipino *et al*., 2002). And there are different definitions for it from different researchers. Huang *et al*. (1998) defines data quality as the extent to which the data satisfies the users' need. Kahn and Strong (1998) describes it as the degree to which data satisfies the expectation of particular users. Wang and Strong (1996) defines data quality as "Fitness for usage", which is based on the widely accepted conception of quality in total quality management so this definition is also widely accepted. Orr (1998) describes it as "the distance between data view expressed in the information system and real data".

As defined by Wang and Strong (1996), the evaluation of data quality relies on individual users. "Fitness for usage" by different users under different circumstances varies, so data quality is relative and cannot evaluate the quality of data independent from the users (Strong *et al*., 1997). So the identification of data quality evaluation dimensions has become a valuable study. Data quality dimensions are a set of attributes representing data quality structure or a single aspect of data.

Felix and Claudia (2000) proposed 22 evaluation indexes in three groups. Subjective indexes: Believability, Concise representation, Interpretability, Relevancy, Reputation, Understandability and Value-added. Objective indexes: Completeness, Customer support, Document, Objectivity, Price, Reliability, Security, Timeliness, Verifiability. Processing indexes: Accuracy, Amount of data, Availability, Consistent representation, Latency and Response time.

Different measurement tools, technologies and processes are needed for different data quality dimensions, which leads to the discrepancies in time, cost and human resource needed for the evaluation. Choose the suitable dimensions after a clear understanding of the work needed in each dimension evaluation, so a project can be better defined. The primary evaluation result of data quality dimensions is to determine the baseline and the rest are for continuous detection and information improvement, or business operation process.

As previously discussed, data quality is a relative conception, which has different definitions and evaluation criteria in different periods and fields. However, up to now, there is no systematic evaluation index system for data quality in banks in the literature. And as banking data has complicated types as well as sources, a single quality dimension cannot be used to evaluate and reflect the overall data quality precisely, an evaluation index system fit for data quality in banks is needed to assess data quality in banks' platforms, so that data quality management can be better achieved and improved.

Data quality evaluation: According to different criteria, there are different ways for the classification of data quality evaluation.

Subjective evaluation method and objective evaluation method: The evaluation methods can be classified into subjective method and objective method by whether the priority is defined subjectively or is calculated by mathematics (Pipino *et al*., 2002). Subjective evaluation method reflects the needs and experience of the stakeholders (including data collector, data maintainer and data consumer); objective evaluation method applies advanced mathematics in the evaluation process to calculate the relevant factors which have influenced data quality and give a digital evaluation of the overall data quality. The objective evaluation method reduces the influence of human factors in subjective method, which makes the result more objective and repeatable.

The present data quality measurement and evaluation takes the method which combines both subjective and objective approaches because of the intimate correlation between data quality/background and its users. Liu and Huang (2006) describes a combined method and the steps to improve data quality in a real application.

Technology-oriented method and question-oriented method: The evaluation methods can also be classified into technology-oriented methods and question-oriented methods according to data quality evaluation scenario selection.

Question-oriented method: Bobrowski *et al*. (1999) proposes a data quality evaluation method inside an organization. Firstly, a data quality criterion list is established with direct evaluation criteria and indirect ones. Traditional software evaluation method is adopted in the direct criteria evaluation, which is questionnaire and the scores used in indirect evaluation criteria are computed by the direct criteria.

Technology-oriented: The pivotal step in data quality evaluation is to select precisely the dimensions that can reflect the data quality in the enterprises. However, the establishment of indexes measuring data quality is decided by specific business scenarios, which proves to be the most efficient in practice. After the selection of dimensions, the next step is to evaluate the importance of each dimension with advanced mathematics. Generally, we assign a weight coefficient to each dimension and decide its relative importance through the calculation of the weight coefficient, so that data quality can be evaluated.

Evaluation methods in detail which is showed in Table 1: Simple-ratio method means the ratio of expected result (E) to total (T), reflecting data quality in some aspects. Maximum-Minimum method is suitable for a dimension in which multiple indexes need to be totaled (Liu and Huang, 2006). The core for evaluation is to find the minimum value and the maximum value. The minimum value method is a conservative evaluation while the maximum is a radical one. Weighted-average method is to assign a weight between [0-1] to each index and make their sum 1. The evaluation index of data quality is calculated through weighted average. Those methods above can evaluate data quality roughly but cannot reflect the influence of each dimension on overall quality comprehensively and systematically, let alone the interaction between the dimensions which can result in the change of data quality evaluation result.

The Analysis hierarchy process is used to decide weighting coefficient in a multi-purpose decision making problem or a comprehensive evaluation problem with a hierarchy structure. Dataset has different data quality evaluation requirements in different applications, so each of it should correspond to multiple data quality evaluation models. In a data quality evaluation model, a dataset can correspond to multiple evaluation dimensions and an evaluation dimension can correspond to multiple indexes. Yu (2011) applies the AHP-based approach to a complicated credit data quality evaluation and makes it a standard system. A hierarchy structure is established after analyzing the factors in the system and the weighting coefficients of factors' relative importance are calculated and ranked after an objective judgment of the factors in the hierarchy structure.

Data quality management: After establishing the connection between data and material product because of the similarity between information system environment and production environment, the principle, method, guidance and technology of Total Quality Management (TQM) are applied in data quality management progressively. Data quality management has formed its own maturity model, data quality inspection principles and data quality detect standards in the process of upgrading (Wei *et al*., 2013).

Table 1: Mathematic methods in data quality evaluation

Method	Subjective or objective	Oriented
Simple-ratio method	Objective	NA
Maximum-minimum method	Objective	Technology
Weighted-average method	Objective	Technology
Analysis hierarchy process	Combined	Technology

Ryu *et al*. (2006) imports a data quality management maturity model to evaluate the organization's data quality management ability. The model is similar to the software capability maturity model and can be divided into four levels: (1) Primary data management (2) Defined data management (3) Managed data management (4) Advanced data management.

Ryu *et al*. (2006) proposed five DQM principles according to the construction of a real data warehouse in a commercial bank: (1) Data quality detection should occupy as little data system resource as possible. (2) Data quality should be detected at regular intervals to assure its timeliness and accuracy. (3) Data quality detection should cover the whole process of data warehouse construction if possible. (4) Feedback as soon as possible. (5) Security level limit is needed in the data quality detection.

Though data quality detection can help the developers to find out erroneous data and rules efficiently and improve data quality, it lacks a scientific quantitative index to indicate the quality of data warehouse on the whole, so it has to be used together with data quality detection standards. The data quality standards mainly includes two parts: (1) Technical indexes (completeness, relevance, uniqueness, effectiveness, etc.), (2) Business indexes (authenticity, accuracy, consistency, intelligibility, etc). The standards are the major references in the investigation of data quality situation.

EVALUATION SYSTEM FOR BANKING DATA QUALITY

Analysis of banking data and platform: In the huge banking data system, a small problem happened in one dimension of data quality will leads to a huge financial loss. For example, the accuracy of cardholder's information is one of the most significant dimensions in credit card platform data. Each operation like card application, consumption and cancellation is correlated with it. Therefore, every single message record should be accurate to improve the quality of service and provide scientific basis for strategic decisions to increase profits and expand market share. As in the dimension of completeness, for example, suppose there is a file with 1000000 client records in the bank and the bank prepares to send ads mails to all the addresses recorded in the file. If the rate of completeness of the address information in this file is 98% and every ads mail costs 0.2 yuan for its production and delivery. If 10 mails are sent to per client each year, then the direct loss caused by incompleteness of the data can amount to 1000000*0.02*0.2*10 = 40000(yuan). If other dimensions (e.g., accuracy) of the data recorded are also taken in consideration, then the direct loss will increase.

The example above has proved how a simple problem in data quality can lead to a direct financial loss of the bank. In fact, the situation can be even more complicated. Currently, the banks have tens, or even over one hundred of data platforms and systems, such as core business system, credit management system, ECIF system and each system has different data source, data logic structure, data exchange and storage method, so the data quality dimensions influencing each data system are different, which further complicates data quality management inside a bank's platform and between platforms.

For example, in core business system, the operations that influence data quality mainly include data collection, data exchange and storage, data quality and decision supports, so validity, completeness, accuracy and consistency should be considered in evaluating data quality of core business system. As for credit management system, the main operations are data collection and storage, so validity, uniqueness, completeness are the main dimensions in the evaluation.

Banking data quality is the basis of client relationship management. A good client data quality management is crucial in the maintenance of old clients and obtainment of the new clients and necessary for the maintenance of a leading position in the market. Good data quality could provide a solid foundation for accurate information needed by decision support. What's more, good data quality is also required by risk management. Therefore, banking data quality management and improvement become a crucial part of bank business. The evaluation system for banking data quality and the AHP-based approach proposed in this study can serve as the basis for banking data platform judgment and good data quality establishment.

Classification of banking data quality dimensions: As previously mentioned, according to data quality management involved in banking data platform at present and the analysis of each dimension's influence on data quality as well as the mutual influence between the dimensions, this study proposes a new set of data quality evaluation index system. In light of the attributes of banking data quality, the system selects six dimensions which are most relevant with banking data quality including Completeness, Validity, Timeliness, Accuracy, Consistency and Uniqueness. The details are listed in Table 2.

Table 2: Dimensions

Dimension	Specification
Completeness	Including entity lost, attribute lost, record lost and field value lost
Timeliness	Including the timeliness and rapidity of data obtainment, transmission, procession, loading and representation
Validity	Including the efficiency of pattern, type, range and business rule
Consistency	Including the consistency of data difference and mutual contradiction between different systems
Accuracy	Including the consistent extent between a real data value and a standard data value and or a real data value and an acceptable data value
Uniqueness	Including the uniqueness of primary key and candidate key

Table 3: Detailed indexes classification

Dimension	Evaluation indexes	Sub-indexes	Specification
Completeness	Technical index completeness	Interface files	Completeness of interface files in transmission
		Data records	Record counts
	Business index completeness	Client information	Customer information loss
		Account information	Key information loss
Validity	Technical validity	Field classes validity	Field classes validity
		Value field classes validity	Field code validity
	Business validity	Business relation classes	Business relation check
		Business development trend classes	Business index trend
Timeliness	Technical timeliness	Data transmission	Timeliness of data transmission
		Data processing	Timeliness of data ETL processing
Consistency	Business consistency	Business statistic index consistency	Consistency of business index between current system and the old one
Accuracy	Business accuracy	Business index range	Accuracy of business index in a given range
Uniqueness	Uniqueness	Uniqueness in table	Uniqueness of primary key in table

Taking the attributes of banking platform data into consideration, the dimensions decided for evaluation can be divided into technical indexes and data indexes.

Technical indexes refer to indexes that have to take every aspect of the market activity into consideration, establish a mathematical model and give a computational formula, so that index value, a number represents the substance of some aspect of the market can be get. In the classification of banking data platform, technical indexes include interface count, interface completeness ratio, interface timeliness ratio, ETL timeliness, code validity, field value validity.

Business index refers to the indexes that correlate with a bank business and in banking data platform, business indexes mainly include vacant ratio, invalid ratio, error ratio, clients and accounts, etc.

The six dimensions can be further subdivided upon technical indexes and data indexes. Completeness and Validity can be subdivided into technical index completeness, business index completeness, technical index validity and business index validity. Timeliness, Consistency, Accuracy contain only one of the two, which are technical index timeliness, business index uniqueness, business index accuracy. Uniqueness can't be subdivided by this standard and correspond to uniqueness in the table.

Take Completeness as an example, after a subdivision into technical index completeness and business index completeness, it can still be subdivided according to the contents of technical index and business index mentioned above. Technical index completeness can be subdivided into interface file index and data file index, with the former representing the completeness of the interface file in transmission and the latter the completeness of record counts. Business index completeness can be subdivided into client information sub-index and account information sub-index, representing client information loss and key information respectively. After a second subdivision, the structure of the system is listed in Table 3.

The interface file index contains multiple interface count indexes, corresponding to the interface counts between the basic data platform and other data platforms. Each data platform corresponds to an interface count sub-index, like the interface counts between core business system and basic data platform and between credit card system and basic data platform. Meanwhile, these two indexes generate two derived indexes, indicating the rate of the completeness respectively.

The data record index mainly records the completeness rate of data interface records between the basic data platform and other platforms, such as the completeness rate of the records between the core business system and the basic data platform, or between the credit card system and basic data platform, etc., which are all derived indexes.

In business index completeness, data platform only contains client information sub-index, which evaluates the loss of the client information. It can be subdivided into basic index and its derived index. For example, the derived index for the increment of individual clients with vacant ID number that the core system deals with on that day should be the vacant rate of increment of individual ID types that the core system deals with on that day and the

derived index for corporate clients with a vacant core license code should be the vacant rate of corporate license code that the core system deals with.

AHP-BASED EVALUATION METHOD

The Analytic Hierarchy Process (AHP), proposed by Thomas L. Saaty in the 1970s, is a systematic and hierarchical analytical method with both qualitative and quantitative ways combined, which has effectively simplified and solved the originally complicated decisions (Qin and Zhang, 1999). Apply this technique to the data quality evaluation system proposed in this paper, we can come up with the following steps: (1) Structure the evaluation hierarchy, (2) Construct pairwise comparison matrices, (3) Obtain the priorities vector, (4) Test consistency, (5) Weigh data quality weight and evaluate. This section will apply AHP-based approach to the data quality evaluation system proposed and will demonstrate its systematicness, practicability and simplicity through a case study of a commercial bank's platform (Feng *et al*., 2012).

Structure the evaluation hierarchy: The first step in the application of AHP technique is to layer the data quality evaluation and construct a hierarchical data model. In the model, the complicated data quality evaluation is decomposed into indexes for evaluation. The indexes can form several levels according to their attributes and relations: (1) Goal level: Contains only one element-data quality evaluation, which is the intended goal of this paper. (2) Criteria level: Includes the dimensions used in the data quality evaluation. (3) Sub-criteria level: Includes the sub-indexes respect to the dimensions. (4) Alternative level: The alternatives needed to be assessed, refers to the data platform in evaluation in this study.

As the evaluation system is described in the previous section, we organize the system with the AHP-based evaluation method model and arrange the dimensions and indexes into the right place, establishing a data evaluation structure model like Fig. 1.

Construct pairwise comparison matrices: the hierarchy structure represents the relations between the dimensions and sub-indexes. As different dimensions and sub-indexes have different weight coefficient in evaluation, the coefficient of each dimension or sub-index in the hierarchy is given and the result is represented by a reciprocal matrix $A = (a_{ij})_{n\times n}$, A is called judgment matrix, in short of pairwise comparison judgment matrix indexes:

$$A=\begin{bmatrix} a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad (1)$$

In order to determine the value of a_{ij}, the 1~9 weight method is recommended to determine the values of the factors in the matrix and finish the formation of reciprocal matrices. The 1~9 weight method is displayed in Table 4 (Saaty, 1980).

According to the above matrix construct method, we estimate the coefficient of the relative importance of the dimensions and sub-indexes by using expert scoring method and paired comparison method and construct the matrix for the instance used in this section. To make it simple, the six dimensions are defined as: Accuracy C_1, Completeness C_2, Consistency C_3, Timeliness C_4, Uniqueness C_5, Validity C_6. The dimension judgment matrix A is showed in Table 5, with a corresponding weight priorities vector $\omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.

After the establishment of the dimension judgment matrix, it is time to determine the weight coefficient of the corresponding sub-indexes and establish a judgment matrix with respect to each dimension. Accuracy, consistency and uniqueness all have only one sub-index, so the relative weight coefficient of their sub-indexes is 1. Timeliness has two sub-indexes C_{41}, C_{42} so its relative weight coefficient is set to be 3 by the method mentioned previously. Completeness and validity each have four sub-indexes and the relative weight coefficients are listed in Table 6 and 7.

Table 4: Weight specification

Weight	Specification
1	Represents that one element is as important as the other one
3	Represents that one element is a little more important than the other one
5	Represents that one element is obviously more important than the other one
7	Represents that one element is far more important than the other one
9	Represents that one element is extremely more important than the other one
2, 4, 6, 8	The mid-value between of the above values
Reciprocals	i/j is defined as a_{ij}, so j/i is its reciprocal

Table 5: Dimensions judgment matrix A = {C_1, C_2, C_3, C_4, C_5, C_6}

Dimensions	C_1	C_2	C_3	C_4	C_5	C_6
C_1	1	1/6	1/2	1/3	2	1/4
C_2	6	1	5	3	9	2
C_3	2	1/5	1	1/2	4	1/4
C_4	3	1/3	2	1	6	1/2
C_5	1/2	1/9	1/4	1/6	1	1/8
C_6	4	1/2	4	2	8	1

Table 6: Completeness $A_2 \leftarrow \{C_{21}, C_{22}, C_{23}, C_{24}\}$

Completeness (C_2)	C_{21}	C_{22}	C_{23}	C_{24}
C_{21}	1	3	7	8
C_{22}	1/3	1	5	5
C_{23}	1/7	1/5	1	3
C_{24}	1/8	1/5	1/3	1

Table 7: Validity $A_6 \leftarrow \{C_{61}, C_{62}, C_{63}, C_{64}\}$

Validity (C_6)	C_{61}	C_{62}	C_{63}	C_{64}
C_{61}	1	2	9	6
C_{62}	1/2	1	3	3
C_{63}	1/9	1/3	1	1/2
C_{64}	1/6	1/3	2	1

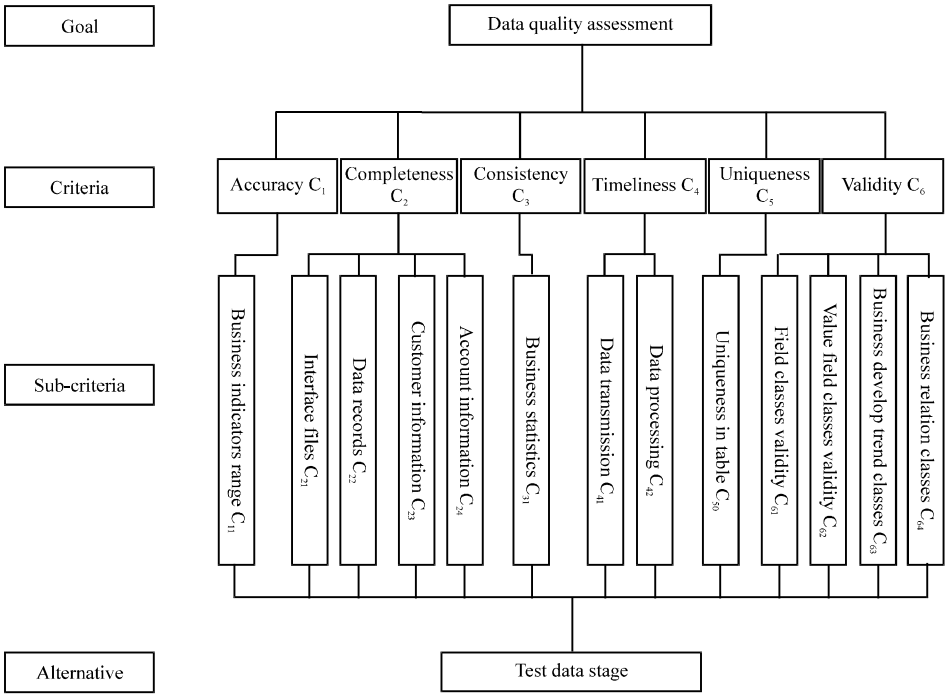

Fig. 1: AHP-based evaluation model of data quality evaluation

Obtain the priorities vector, Test the consistency: Next step is to calculate the eigenvector for each relative weight coefficient judgment matrix. Then, normalize the eigenvector (A normalized eigenvector can be used as priorities vector to present each dimension's or sub-index's relative importance to data quality (Saaty, 1998) ω_i (i = 1, 2, 3, ..., n):

$$\omega_i = \Sigma_j a_{ij} \Big/ \Sigma_{i,j} a_{ij} \quad (2)$$

Though the method of constructing paired comparison judgment matrix can reduce the interference of other factors and represent the difference in the influence of the factors in a pair, on the whole, there will be inconsistency to some extent inevitably:

$$a_{ij}\ a_{jk} = a_{ik},\ i, j, k = 1, 2 \quad (3)$$

A reciprocal matrix which satisfies Eq. 3 is called a consistent matrix. A consistency test is needed to determine whether a judgment matrix constructed should be accepted or not. The steps of consistency test are listed below.

First, compute the maximum eigenvalue λ_{max}, the equation is shown in 4:

$$\lambda_{max} = \frac{1}{n} \sum_{i=1}^{n} (A\omega)_i \Big/ (\omega)_i \quad (4)$$

Next, compute the consistency index (CI) of judgment matrix:

$$CI = (\lambda_{max}(A) - n) \Big/ (n-1) \quad (5)$$

Then search the average random consistency index (RI). The RI value is listed in Table 8 while n = 1, 2, ..., 9.

The final step is to compute the test value of consistency (CR):

$$CR = CI \Big/ RI \quad (6)$$

The closer the consistency test value (CR) is to 0, the more consistent the matrix is. Suppose when the CR <0.10 (Saaty, 1980) or λ_{max} = n, CI = 0, we can regard the consistency of the matrix acceptable and the matrix efficient, otherwise we will need to adjust and revise matrix A.

With the above steps we will calculate the consistency of matrix constructed previously. First calculate the dimension-level matrix and get a normalized eigenvector ω:

$$\omega = (0.0569, 0.4032, 0.0900, 0.1563, 0.0303, 0.2632)$$

Then test the consistency of this matrix by Eq. 4:

$$A\omega = \begin{bmatrix} (A\omega)_1 \\ (A\omega)_2 \\ (A\omega)_3 \\ (A\omega)_4 \\ (A\omega)_5 \\ (A\omega)_6 \end{bmatrix} = \begin{bmatrix} 1 & 1/6 & 1/2 & 1/3 & 2 & 1/4 \\ 6 & 1 & 5 & 3 & 9 & 2 \\ 2 & 1/5 & 1 & 1/2 & 4 & 1/4 \\ 3 & 1/3 & 2 & 1 & 6 & 1/2 \\ 1/2 & 1/9 & 1/4 & 1/6 & 1 & 1/8 \\ 4 & 1/2 & 4 & 2 & 8 & 1 \end{bmatrix} \begin{bmatrix} 0.0569 \\ 0.4032 \\ 0.0900 \\ 0.1563 \\ 0.0303 \\ 0.2632 \end{bmatrix}$$

Substitute 4 with Aω, work out that $\lambda_{max} = 6.1074$, CI = 0.0215, CR = 0.0173<0.1, which satisfies the demand of consistency.

With the same steps, we calculate the normalized eigenvector of each matrix we formed previously and test the consistency of them. The result is listed in Table 9.

Weight data quality and evaluate: The previous four steps have worked out the priority of each dimension respect to the data quality or each sub-index respect to the relative dimension. So the final step is to calculate the value which can be used to evaluate the banking data quality of the bank's basic data platform by the weighted average method.

Suppose there are three random samples of the same quantity collected from the same data platform under the same condition in January, February and March respectively. After a check of the bank's background program, a number of errors can be found in Table 10.

Finally, calculate the data quality quantized value of each month, January for example: Completeness (0.4302) = 0.5820×2+0.2786×3+0.0899×2+0.0495×1 = 2.2291, with the same formula, validity (0.2632) = 3.0078, Timeliness (0.1563) = 0.875, Consistency (0.0900) = 2, Accuracy (0.0569) = 1, Uniqueness (0.0303) = 2, So the assessment value R_1 = 0.4302×2.2291+0.2632×3.0078 +0.1563×0.875+0.0900×2+0.0569×1+0.0303×2 = 2.71, with the same calculate formula, R_2 = 2.22, R_3 = 1.81 representing January, February and March respectively. The smaller the number is, the better data quality a platform has. Judging from the variation trend of the result, the data quality quantized value of the platform is decreasing progressively, which means the overall data quality is improving. In this method, the data quality of the platform can be generally evaluated and its data quality management can be judged, which provides basic data for the bank's decisions.

Table 8: Random consistency value in average

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Table 9: Normalized eigenvector of sub-index matrix and consistency test result

k	1	2	3	4	5	6
ω_{k1}	1	0.5820	1	0.7500	1	0.5701
ω_{k2}	-	0.2786	-	0.2500	-	0.2605
ω_{k3}	-	0.0899	-	-	-	0.0662
ω_{k4}	-	0.0495	-	-	-	0.1032
λ_{kmax}	1.0000	4.1983	1.0000	2.0000	1.0000	4.0458
CI_k	0.0000	0.0661	0.0000	0.0000	0.0000	0.0153
CR_k	0*	0.0734*	0*	0*	0*	0.0170*

*Representing the consistency value (CR) <0.10, which satisfies the demand of the consistency

Table 10: Data quality evaluation

Dimension (weight)	Sub-index (weight in dimension)	January	February	March
Completeness (0.4302)	Interface files (0.5820)	2	4	3
	Data records (0.2786)	3	2	2
	Client information (0.0899)	2	1	1
	Account information (0.0495)	1	1	3
Validity (0.2632)	Field classes validity (0.5701)	4	2	2
	Value field classes validity (0.2605)	2	4	1
	Business relation classes (0.1032)	2	2	0
	Business development trend (0.0662)	0	1	3
Timeliness (0.1563)	Data transmission (0.7500)	5	0	2
	Data processing (0.2500)	2	4	1
Consistency (0.0900)	Business statistics (1)	2	1	1
Accuracy (0.0569)	Business index range (1)	1	0	2
Uniqueness (0.0303)	Uniqueness in table (1)	2	1	1

CONCLUSION

Based on a data quality evaluation project for a commercial bank, this study proposes an evaluation index system for banking data quality evaluation and an AHP-based approach for weighting coefficient of the criteria, which is applied in the weighting of the dimensions proposed in the evaluation system. Through the case study of a bank's data quality evaluation, the efficiency of quality evaluation and dimension weighting method is demonstrated, which is instructive to data quality evaluation in banks to some extent.

In the evaluation system proposed in this study, the meaning of the indexes/sub-indexes of different platforms in the banks is different and so is the weight of each sub-index. Therefore, this method cannot be used in the comparison of data quality between different platforms. The next study is to work out a solution to how to weight to make the data quality between different platforms comparable.

ACKNOWLEDGMENT

This research was supported by National Key Technology R&D Program of the Ministry of Science and Technology of China (No. 2013BAH01B03).

REFERENCES

Bobrowski, M., M. Marre and D. Yankelevich, 1999. A homogeneous framework to measure data quality. Proceedings of the 4th International Conference on Information Quality, October 22-24, 1999, MIT, Cambridge, MA., pp: 115-124.

Felix, N. and R, Claudia, 2000. Assessment methods for information quality criteria. Proceeding of the 5th International Conference on Information Quality, October 20-22, 2000, MIT, Cambridge, MA., pp: 148-162.

Feng, X., X. Qian and Q. Wu, 2012. A DS-AHP approach for Multi-attribute decision making problem with intuitionistic fuzzy information. Inform. Technol. J., 11: 1764-1769.

Huang, K.T., Y.W. Lee and R.Y. Wang, 1998. Quality Information and Knowledge Management. 1st Edn., Prentice Hall, New Jersey, ISBN-13: 978-0130101419.

Kahn, B.K. and D.M. Strong, 1998. Product and service performance model for information quality: An update. Proceedings of the 3rd International Conference on Information Quality, October 23-25, 1998, MIT, Cambridge, MA., pp: 102-115.

Liu, H. and Y. Huang, 2006. Data quality statistics and assessment methods. Stat. Decision, Vol. 3. 10.3969/j.issn.1002-6487.2006.03.014

Orr, K., 1998. Data quality and system theory. Commun. ACM, 41: 66-71.

Pipino, L.L., Y.W. Lee and R.Y. Wang, 2002. Data quality assessment. Commun. ACM., 45: 211-218.

Qin, J. and Y.P. Zhang, 1999. Application of contemporary statistical information analysis method in safety engineering-the principle of AHP. Ind. Saf. Dust Control, 4: 44-48.

Ryu, K.S., J.S. Park and J.H. Park, 2006. A data quality management maturity model. ETRI J., 28: 191-204.

Saaty, T.L., 1980. The Analytic Hierarchy Process. McGraw-Hill, New Yark.

Saaty, T.L., 1998. Ranking by eigenvector versus other methods in the analytic hierarchy process. Applied Math. Lett., 11: 121-125.

Strong, D.M., Y.W. Lee and R.Y. Wang, 1997. Data quality in context. Commun. ACM, 40: 103-110.

Wang, R.Y. and D.M. Strong, 1996. Beyond accuracy: What data quality means to data consumers. J. Manage. Inform. Syst., 12: 5-34.

Wei, S., W. Wei, J. Zhang, W. Wang and J.W. Zhao *et al*., 2013. Research of data quality assurance about ETL of telecom data warehouse. Inform. Technol. J., 12: 1839-1844.

Xie, F.C., 2009. Financial data warehouse data quality management oriented research and implementation. Xiamen University, Xiamen.

Yu, W., 2011. The empirical analysis of commercial bank's credit data quality evaluation based on fussy AHP. Shanghai Finance, 3: 102-105.