International Journal of Computer Engineering Science (IJCES)
Volume 2 Issue 11 (November 2012) ISSN : 2250:3439
https://sites.google.com/site/ijcesjournal http://www.ijces.com/

Analysis of Relative Importance of Data Quality Dimensions for Distributed Systems

Gopalkrishna Joshi[1], Narasimha H Ayachit[1], Kamakshi Prasad V[2]
[1] BVB College of Engineering & Technology, Hubli – 580031 (India)
{ghjoshi,ayachit}@bvb.edu
[2] Jawaharlal Nehru Technological University, Kukatpally, Hyderabad (India)
kamakshi.prasad@gmail.com

Abstract. The Increasing complexity of the processes and their distributed nature in enterprises is resulting in generation of data that is both huge and complex. And data quality is playing an important role as decision making in enterprises is dependent on the data. This data quality is a multidimensional concept. However, there does not exist a commonly accepted set of the dimensions and analysis of data quality in the literature by the concerned. Further, all the dimensions available in literature may not be of relevance in a particular context of information system and not all of these dimensions may enjoy the same importance in a context. Practitioners in the field choose dimensions of data quality based on intuitive understanding, industrial experience or literature review. There does not exist a rigorously defined mechanism of choosing appropriate dimensions for an information system under consideration in a particular context. In this paper, the authors propose a novel method of choosing appropriate dimensions of data quality for an information system bringing in the perspective of data consumer. This method is based on Analytic Hierarchic Process (AHP) popularly used in multi-criterion decision making and the demonstration of the same is done in the context of distributed information systems

Keywords: Data Quality, Data Quality Dimensions, Distributed Systems, TDQM, AHP.

1 Introduction

The data in the digital era is growing at an enormous rate. The dependence of organisations on data in decision making is increasing day by day. The problems related to data are also growing with this growth of data. Data problems are observed to result in lost revenues and market share, reduced profits, and customer dissatisfaction. Poor quality data is estimated to result in the increase of operational cost by at least 10% (and probably as much as 20%) of revenue [1]. Today, most organizations use data in two ways: transactional/operational use (running the business), and analytic use (improving the business). Both usage scenarios rely on high quality data, suggesting the need for processes to ensure that data is of sufficient quality to meet all the business needs. Therefore, data quality has assumed significant importance.

Data quality is a multidimensional concept. Though the terms like accuracy, correctness etc., are frequently used by the practitioners and researchers in the field as dimensions of data quality, there does not exist a rigorously defined set of data quality dimensions acceptable by all[2,3]. Further, all data quality dimensions may not be required to be considered while designing a particular information system and all of them may not assume the same importance while designing the information systems considering the data quality aspects. The choice of data quality dimensions for an information system under consideration is mostly done based on intuitive understanding, industrial experience or literature review [2] and there does not exist a defined mechanism to choose the appropriate set of dimensions of data quality for an information system.

A method has been proposed in this paper by the authors to choose appropriate data quality dimensions using a decision making method called Analytic Hierarchic Process (AHP). This has been demonstrated for distributed systems as one type of types of Information Systems.

2 Related Work

The In the following paragraphs in this section, are presented the information regarding earlier work related to data quality dimensions and Analytical Hierarchy Processing (AHP).

2.1 Background

There are several data quality frameworks available in the literature all of them focusing on a set of data quality dimensions suiting a particular context of use. The choice of data quality dimensions for an information system under consideration is mostly done based on intuitive understanding, industrial experience or literature review [8,9,10] . A few of the frameworks look beyond the data quality dimensions and propose processes, tools and techniques covering the life cycle management of data [11,12,13]. Total Data Quality Management (TDQM), the data quality management framework proposed at MIT is one such framework used by practitioners and researchers both. This framework makes use of the concept of Information Manufacturing Process which converts the data into information product used by the data consumer in decision making process[5,6,7]. This framework brings in the perspective of data consumer in deciding the quality of the information product he is using in his decision making process. An earlier work by Wong and Strong proposes data quality dimensions from data consumers' perspective [4]. It lists fifteen data quality dimensions that are of importance from data consumers' perspective. These fifteen data quality dimensions are classified in four categories viz. intrinsic, representational, and contextual and access. However, it requires a careful study about which of these fifteen dimensions of data qualities are of importance for an information system under consideration. Analytic Hierarchy Processing is the technique proposed to be in making the decision about the choice of data quality dimensions for an information system.

2.2 Analytic Hierarchy Processing (AHP)

Decision making has become mathematical science today [15]. Comparison is the natural way human beings make decisions. But, as the number of objects for comparison increases, decision making turns out to be a difficult task. Multi-criteria decision making finds its mention in several situations in management sciences. Analytic Hierarchy Process (AHP) is a widely used approach in multi criterion decision making [17,18,19,20,21].

2.3 Introduction to AHP

AHP is a theory of measurement through pair wise comparisons done between any two criteria a and b in a given relation and relies on the judgments of experts to derive priority scales. It is these scales that measure intangibles in relative terms [16]. The comparisons are made using a scale of absolute judgments that represents, how much more one element dominates another with respect to a given attribute. Table 1 shows the scale that may be used for this purpose of quantifying the pairwise comparison [14].

The judgments may be inconsistent, and how to measure inconsistency and improve the judgments, when possible to obtain better consistency is a concern of the AHP which is measured in the form of Consistency Index (CI) and Consistency Ratio (CR) that are explained in the subsequent part of the paper.

2.4 Working of AHP

Consider n elements to be compared, $C_1 \ldots C_n$ and denote the relative 'weight' (or priority or significance) of C_i with respect to C_j by a_{ij} and form a square matrix $A=(a_{ij})$ of order n with the constraints that $a_{ij} = 1/a_{ji}$, for $i \neq j$, and $a_{ii} = 1$, all i. Such a matrix is said to be a reciprocal matrix. However, this reciprocal relation may not remain valid in certain typical cases. The weights are consistent if they are transitive, that is $a_{ik} = a_{ij}a_{jk}$ for all i, j, and k. Such a matrix might exist if the a_{ij} are calculated from exactly measured data. Then find a vector ω of order n such that $A\omega = \lambda\omega$. For such a matrix, ω is said to be an eigenvector (of order n) and λ is an eigenvalue. For a consistent matrix, $\lambda = n$.

For matrices involving human judgment, the condition $a_{ik} = a_{ij}a_{jk}$ does not hold as human judgments are inconsistent to a greater or lesser degree. In such a case the ω vector satisfies the equation $A\omega= \lambda_{max}\omega$ and $\lambda_{max} \geq n$. The difference, if any, between λ_{max} and n is an indication of the inconsistency of the judgments. If $\lambda_{max} = n$ then the judgments have turned out to be consistent. Finally, a Consistency Index can be calculated from (λ_{max}-n)/(n-1). That needs to be assessed against judgments made completely at random and Saaty has calculated large samples of random matrices of increasing order and the Consistency Indices of those matrices. A true Consistency Ratio is calculated by dividing the Consistency Index for the set of judgments by the Index for the corresponding random matrix. Saaty suggests that if that ratio exceeds 0.1 the set of judgments may be too inconsistent to be reliable. In practice, CRs of more than 0.1 sometimes have to be accepted. Perfect consistency of judgments is indicated by a CR having value of 0.

Table 1. Saaty's table for pair-wise comparison

Intensity of Importance	**Definition**	**Explanation**
1	Equal importance	Two factors contribute equally to the objective
3	Somewhat more important	Experience and judgment slightly favor one over the other
5	Much more important	Experience and judgment strongly favor one over the other
7	Very much more important	Experience and judgment very strongly favor one over the other.
9	Absolutely more important	The evidence favoring one over the other is of the highest possible validity.

3 Methodology

In this section is demonstrated the process of computing data quality dimensions for a typical type of information system viz. Distributed system. A distributed system has the resources and application distributed among a network of systems that are geographically distributed. Even though some degree of heterogeneity may occur, the data design is done centrally. Even though the does not exist unanimity about the choice of data quality dimensions, the data quality dimensions proposed by Wang and Strong were considered as the basis for the study [4].

3.1 User Data Collection

A survey was conducted by the authors among the users of the distributed systems. The survey questionnaire was designed keeping the target respondents in mind. The respondents were asked to compare the importance of two dimensions of data quality at any given point of time. The first question contained only two dimensions of data quality for comparison. As the respondent progressed through the survey, a new dimension of data quality was introduced every time for comparison. The respondents were asked a total of 14 questions to compare the dimensions of data quality and were asked to rate the importance on the scale proposed by Saaty as mentioned in Table 2.1.

By taking the average rating of the respondents for every rating the reciprocal matrix is built. AHP is applied on the collected data to compute the importance of the data quality dimensions as given in the algorithms of section 3.2.

3.2 AHP Algorithms to Compute Priority of the Data Quality Dimensions, CI and CR

Algorithm ComputeDQPriority

```
// Input: The pair wise comparison values  for all n dimensions of the data quality
//Output: The Priority of Data Quality Dimensions, stored in W
1.Build Reciprocal Matrix A[ ] [ ]:
  Initialise all the elements of principal diagonal to 1
  for each of the of row i, up to n
    for each of the column j, j > I, up to n
      Read A[i][j], that indicates the preference of dimension i over dimension j
      A[j][i] = A[i][j]
    end for
  end for
2.Build Normalised Eigene Vector W:
for each  row i
    Compute geometric mean GM(i),  which is nth root of the product of all
    elements of row i.
 end for
Compute the average of the Geometric Means of all the rows, AvGM.
for each row,  i
    W(i ) = GM(i) / AvGM.
end for
```

end ComputeDQPriority

```
Algorithm  ComputeCInCR
    // Input : The Reciprocal Matrix A[n][n] and Normalised Eigene Vector W[n]
   //  Output: CI and CR
   1.Compute LamdaMax
      Compute Matrix A3 = A[ ][ ] X W.
      Compute LamdaMax = (Sum of all n elements of matrix A3) / n.
   2.Compute consistency Index CI using the formula,
                    CI = ( LamdaMax – n) / ( n – 1).
   3. Compute consistency ratio CR = CI / RI, where RI is random index, whose value
      for n to be picked from Random Index Table.
end ComputeCInCR
```

4 Results

A survey was conducted by the authors among the users of the distributed systems. The survey questionnaire was designed keeping the target respondents in mind. The respondents were asked to compare the importance of two dimensions of data quality at any given point of time. The respondents of the survey included professionals who have been using distributed information systems. The minimum experience of the respondents is about 06 years. By taking the average rating of the respondents for every rating the reciprocal matrix was built. AHP is applied on the collected data to compute the importance of the data quality dimensions as given in the algorithm of section 3.2. The results of application of AHP to this survey data is given in Table 2.

The quality of survey data is indicated by the factors consistency index and consistency ratio. The consistency index and consistency ratio computation is done by using the formulae described in section 3.5. The values of CI and CR for the data of the survey are found to be as follows:

Consistency Index = 0.0820 Consistency Ratio = 0.0516

The consistency ratio is expected be around 0.1. Hence the quality of data obtained through survey is considered consistent.

Prod(i) in Table 2 refers to the importance of data quality dimension i in comparison with all the remaining dimensions (done through pair wise comparison).

It speaks of the overall strength of the relationships that data quality dimension i has with the remaining data quality dimensions.

Aggregation of individual judgments in a group into a single representative judgment for the entire group should be done such that reciprocal of synthesized judgments is equal to the syntheses of the reciprocals of these judgments. It has been proved that geometric mean is the way to do that. Hence, geometric mean computation is done for the decisions of every data quality dimension's pair wise comparisons and is shown as GM(i).

Table 2. AHP Calculations with 15 DQ dimensions for Distributed Systems

PRODUCT(i)	GM(i)	W(i)	IP	A3	A3i/Wi
19.9936	1.2210	0.0763	0.8616	1.1640	15.2482
52.0833	1.3015	0.0814	0.9184	1.2411	15.2532
101.2500	1.3605	0.0851	0.9600	1.2931	15.2036
169.5740	1.4081	0.0880	0.9935	1.3537	15.3776
105.6563	1.3644	0.0853	0.9627	1.3014	15.2582
75.1339	1.3337	0.0834	0.9411	1.2898	15.4694
188.4155	1.4180	0.0886	1.0005	1.3703	15.4579
70.6860	1.3283	0.0830	0.9372	1.2991	15.6439
0.0001	0.5350	0.0334	0.3775	0.6048	18.0804
0.2217	0.9045	0.0565	0.6382	0.6048	10.6958
0.0391	0.8057	0.0504	0.5685	0.8631	17.1358
0.0000	0.4762	0.0298	0.3360	0.7732	25.9693
0.0000	0.4937	0.0309	0.3483	0.4846	15.7011
1.2500	1.0150	0.0635	0.7162	1.0077	15.8811
1.5625	1.0302	0.0644	0.7269	1.0203	15.8421
				LambdaMax	16.1478
				CI	**0.0820**
				RI	1.5900
				CR	**0.0516**

The vector W, is the Eigen vector which is the priority vector. W(i) speaks of the importance of a data quality dimension i in relation to the remaining data quality dimension. The computation of Eigen vector that indicates the priorities is shown in Figure 1.

Idealised Priority computation is done for these data quality dimensions by taking largest of W (i) as 1 and scaling the other ones accordingly. This vector IP speaks of the ranking of the data quality dimensions based on importance.

By observing vector W, it can be seen that the following data quality dimensions viz. Interpretability, Amount of Data, Ease of Understanding, Concise Representation and Consistent Representation are found to have values less than 0.06, which is the value of each dimension considering equal importance for the entire fifteen dimensions. Considering the equal probability of distribution for all the 15 data quality dimensions which comes to 0.06, the following data quality dimensions are found to be of importance in the context of distributed systems: Believability, Accuracy, Objectivity, Reputation, Relevance, Value Added, Timeliness, Completeness, Access and Security.

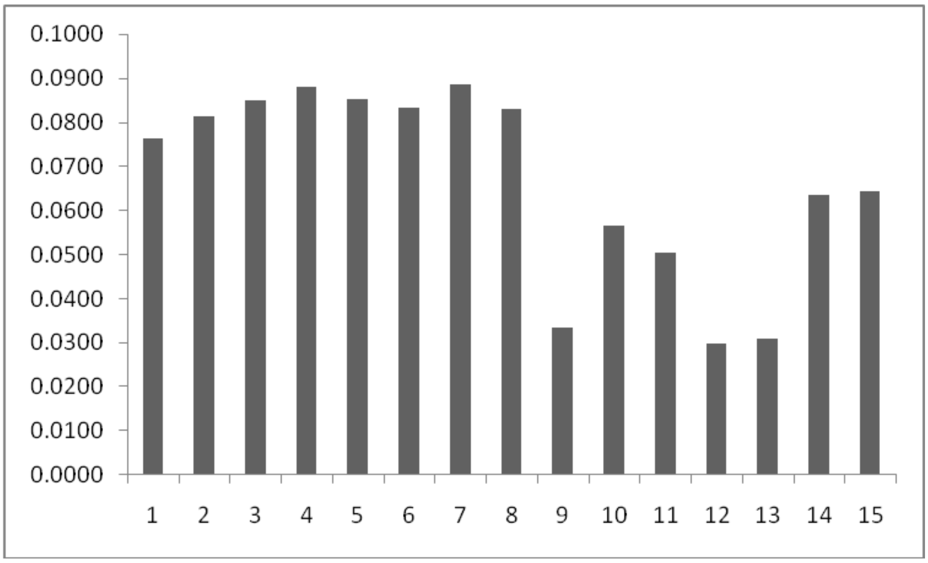

Figure 1. The Eigen Vector Values of Data Quality Dimensions

The quality of survey data is indicated by the factors consistency index and consistency ratio. The consistency index and consistency ratio computation is done by using the formulae described in section 3.5. These values are found to be as follows:

Consistency Index = 0.0820 Consistency Ratio = 0.0516

The consistency ratio is expected be around 0.1.

Having found that the five identified dimensions have lesser importance, their presence on the overall decision making might have had although not major, minor skew. This minor variation does affect the decision making where the importance is almost same, but their hierarchy may change. In view of this, AHP is once again applied to the data obtained by eliminating the dimensions already identified viz. Interpretability, Amount of Data, Ease of Understanding, Concise Representation and Consistent Representation. The results of application of AHP for this data are as shown Table 3.

The Eigen Vector values of these 10 quality dimensions are found to be in the range 0.0932and 0.1082, which means that the spread is less. Further the values of Consistency Index and Consistency Ratio are as follows:

Consistency Index = 0.0068 Consistency Ratio = 0.0046

This shows that the data quality survey data is improved after eliminating the least important dimensions of data quality.

Table 3. AHP Calculations with 10 DQ dimensions for Distributed Systems

PRODUCT(i)	**GM(i)**	**W(i)**	**IP**	**A3**	**A3i/Wi**
0.3492	0.9323	0.0932	0.9030	0.8613	9.6940
0.6944	0.9760	0.0975	0.9649	0.9017	9.8943
1.4400	1.0246	0.1024	1.0381	0.9466	10.1400
3.2813	1.0824	0.1082	1.1424	1.0000	10.5618
1.2727	1.0162	0.1015	1.0277	0.9388	10.1207
0.8571	0.9898	0.0989	1.0014	0.9144	10.1256
1.5625	1.0302	0.1029	1.0472	0.9517	10.1728
0.5120	0.9564	0.0956	0.9448	0.8835	9.8866
1.0000	1.0000	0.0999	1.0000	0.9238	10.0078
1.0000	1.0000	0.0999	1.0000	0.9238	10.0078
				LambdaMax	10.0611
				CI	**0.0068**
				RI	1.4900
				CR	**0.0046**

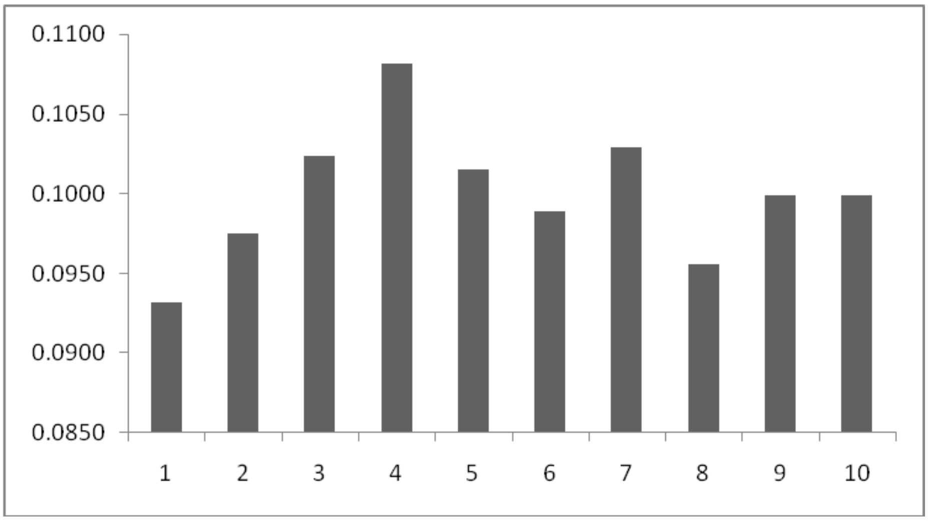

Figure 2. The Eigen Vector Values for 10 Data Quality Dimensions

5 Conclusion

The work in this paper shows a method of choosing appropriate dimensions of data quality in the context of distributed systems. In addition, the dimensions finally chosen are ranked in the order of their priority so that it also can act as input to the designer. However, the limitation of the survey could be in terms of the response of the respondents to the survey. The choice of respondents plays an important role on the outcome of such processes.

6.References

1. Redman, T.C : Data: An Unfolding Quality Disaster. *DM Review, August 2004*.

2.Wang, R.Y., Storey, V.C., and Firth, C.P.: A framework for Analysis of Data Quality Research, IEEE Trans. On Knowledge Data Engg. 7, 4, pp 623-640 (1995).

3.Wand, Yair., and Wang R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations, CACM, Vol.39, No.11, pp86-95(1996).

4. Wang, Richard., Strong Diane M. Beyond Accuracy: What Data Quality Means to Data Consumer, Journal of Management Information System: 12,4 (1996).

5.Shankaranarayan G, Wang R, and. Ziad M. Modelling the Manufacture of an Information Product with IP-MAP. In: Proceedings of the 6th International Conference on Information Quality (ICIQ 2000), Boston, MA, (2000)

6.Wang R. Y,Lee Y.L., Pipino L, and Strong D.M: "Manage Your Information as a Product," Sloan Management Review, vol. 39, pp. 95-105, (1998).

7.P. P. Chen: The Entity-Relationship Model - Toward a Unified View of Data, ACM Transactions on Database Systems, vol. 1, pp. 166-193, (1976).

8.Ballou, D.P, Pazer, H.L.: Modelling Data and process Quality in multi-input and multi-output information systems. Manage.Scie.31, 2, pp. 150-162, (1985).

9.Firth,C.P and Wang, R.Y. :Data Quality Systems: Evaluation and Implementation. Cambridge Market Intelligence, London (1996)

10.Kriebel C.H.: Evaluating the quality of information systems. Design and Implementation of Computer Based Information Systems. Sijthtoff & Noordhoff, Germantown (1979).

11. Ismael Caballero et al., IQM3: Information Quality Management Maturity Model, Journal of Universal Computer Science, vol. 14, no. 22, 3658-3685,(2008).

12.English, L :Improving Data Warehouse and Business Information Quality. Wiley & Sons: New York(1999).

13.Wang, R.Y. : A product perspective on Total Data Quality Management, Communications of the ACM, Vol.41, No.2, (1998).

14.Coyle Geoff: Practical Strategy. Open Access Material. AHP Pearson Education Limited, (2004).

15.Figuera, J., Greco, S. and Ehrgott, M. (Eds) :Multiple Criteria Decision Analysis, State of the Art Surveys, New York: Springer, (2005).

16.Saaty Thomas L., Decision making with the analytic hierarchy process, Int. J. Services Sciences, Vol. 1, No. 1, pp.83-98.(2008).

17.Saaty, T.L. :How to make a decision: the analytic hierarchy process, Interfaces, Vol. 24,No. 6, pp.19–43,(1994).

18.Saaty, T.L.: Theory and Applications of the Analytic Network Process, Pittsburgh, PA: RWS Publications,(2005).

19. Saaty, T.L. and Alexander, J. :Conflict Resolution: The Analytic Hierarchy Process, New York: Praeger,(1989).

20. Saaty, T.L. and Vargas, L.G. :Models, Methods, Concepts and Applications of the AnalyticHierarchy Process, Boston: Kluwer Academic Publishers (2000).

21. Saaty, T.L. and Vargas, L.G.: Decision Making with the Analytic Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks, New York: Springer (2006).