ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

Data Representation in web portals

Dr.A.Muthu Kumaravel[1]

MCA Department, Bharath Institute of Science and Technology, Bharath University, Chennai – 73[1]

ABSTRACT: Data mining is a process of inferring knowledge from huge volume of data. Data mining can be carried out on data represented in quantitative, textual, or multimedia forms. Online information systems are characterized by the presentation of a large amount of data to a wide audience, the quality of which can be very heterogeneous. The information systems on the web need to publish good quality of information in the shortest possible time after it is available from information sources. Quality data have to be perfect, accurate, complete, consistent, timely, and flexible enough to meet user needs. Data Quality is a very important aspect in web services. To perk up data quality, it is now and then necessary to dirt free the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points, removing unnecessary data fields, identifying anomalous data, and standardizing data formats. The Quality of the data presented by the web portals has to be analyzed. Intrinsic and representational categories of data quality are very important in the web portal to give the data in most effective manner. This paper presents the study about the attributes of quality representation of data and a case study about how effective, the data representation has been made with “Science & Technology” column of ‘The Hindu’ daily news paper web portal.

KEY WORDS: Data Quality, Intrinsic DQ, Representational DQ, Web Portal.

I. INTRODUCTION

Due to the advancement and growth of Information and Communications Technology, all the information of every universal activity like News, Health, Entertainment, Education, etc., are available in websites through internet. The World Wide Web is a repository of various data. But there is a question of quality of data published in the websites. Data quality is a new research area that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data, also measured by the structure and consistency that is, how the data has been represented in the web portal. A web portal or public portal is a web site that has lot of information from multiple sources on the web. It organizes the information in an easy user-friendly manner. In worldwide numerous users use web portals to obtain information for their work and to help with decision making. The users and data consumers need to ensure that the data obtained are right for their needs. Thus the organizations that provide Web portals need to offer data that meet user requirements. Data quality represents a common interest between data consumers and portal providers. Data quality plays an important role in the efficiency and effectiveness of organizations and businesses.

II. CLASSIFICATION OF DATA QUALITY

Data Quality is classified into four categories, Intrinsic DQ, Accessibility DQ, Contextual DQ and Representational DQ. Each category has many dimensions like Accuracy, Completeness, Consistency, Timeliness, etc. from literature survey [2] in Table1. Accuracy of data is the degree to which data correctly reflects the real world object or an event being described. An example of data Accuracy is the bank balance in the customer's account is the real value customer deserves from the Bank. Completeness of data is the extent to which the expected attributes of data are provided. For example, a customer data is considered as complete if all customer addresses, contact details and other information are available and also the data of all customers is available. Consistency of data means that data across the enterprise should be in synchronized with each other or the absence of data conflicts. An example of data in-consistency is a credit card is cancelled, and inactive, but the card billing status shows due. The timeliness of data is extremely important which depends on user expectation. Quality of data in the web portals can be analyzed using the survey method. The survey has been made with the web users who are regular to use the online “The Hindu” web portal.

TABLE 1 DQ CATEGORIES AND DIMENSIONS

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Timeliness/Currency.
Accessibility DQ	Accessibility, Access Security
Contextual DQ	Relevancy,Value-added, Completeness
Representational DQ	Content coverage/Amount of data, Consistent Representation/Writing Style, Interactivity, Layout, Multimedia Presentation, Navigation Quality,Organization, Achieves/Documentation.

The scope of the study in this paper includes only the intrinsic and representational data quality categories of "Science & Technology" column of 'The Hindu' web portal. Table 2 shows the Data quality, its dimensions and its definitions.

TABLE 2 DEFINITIONS FOR THE DQ DIMENSIONS

Category	Dimensions	Definitions
Intrinsic	Accuracy	Ensure data are the correct and valid values.
	Timeliness or Currency	The news is up to date. Information in the articles is useful to our work or life.
Representational	Content coverage	The website includes appropriate information and features.
	Consistent Representation or Writing style	The pages of the portal should be consistent in style. Choose a style and apply it to all the pages in the portal. Alternatively, try not to use more than two or three styles.
	Interactivity	Easy to effectively retrieve specific information on the site.
	Layout	The art of the overall design of a page, such as arrangement of graphics and text
	Multimedia presentation	Use of audio and video content.
	Navigation	Links to other websites

	quality	or between pages
	Organization	The information presented on the pages of the portal should be organized by combining various visual characteristics such as size of letters, images, colours, data grouping etc.
	Archives	Storage and provision of past articles or past newspapers.

III. QUALITY ANALYSIS

Data Quality (DQ) is often defined as “fitness for use”, i.e., the ability of a collection of data to meet user requirements [3, 14].

This definition and the current view of assessing DQ, involve understanding DQ from the users point of view [15]. Newspapers can provide online versions, that are not mirror images of print versions, instead offer something extra such as interactive features or information that could not fit in print version [1]. There are number of newspapers available on internet some with general information and some papers are complete with archives. The Hindu newspaper is one among the complete newspaper available on the internet via the web portal http://www.thehindu.com/.The online web portal of this paper consists of many columns which covers various information every day. But the case study in this paper has analyzed the data qualities like Intrinsic DQ, and Representational DQ in the ‘S&T’ (Science & Technology) column alone.

The “S&T” Column of the portal includes several sub columns like Agriculture, Energy & Environment, Gadgets, Internet, Science and Technology. The survey has been done by feedback analysis using statistical tool. A questionnaire has been framed and the feedback has been collected from the undergraduate and postgraduate Students, Research scholars, Academicians of various disciplines and web users who go through this portal in a regular basis.

The questionnaire has been framed with 5 to 6 questions for each dimension. The web user has to enter their rating percentage values in the specified columns.

TABLE 3 ATTRIBUTE QUESTIONNAIRES

1. Accuracy: Ensure data are the correct and valid values.				
Questions:	Low	Medium	High	Very High
A. The presence of advice for agriculture growth and advancement from concerned experts.				
B. The information about Energy &Environment.				
C. The arrival of new gadgets.				
D. The information of new web addresses and contact details.				
E. The innovative research projects explained clearly.				
2. Timeliness or Currency: The news is up to date. Information in the articles is useful to our work or life.				
Questions:	Low	Medium	High	Very High
A. The information about agriculture are up to date.				
B. The information about current energy & Environment.				
C. The issues about various gadgets are given in right time.				
D. The new research opportunities reach the user in time.				
E. The latest news in science and technology are provided in time.				
F. The writing style in S&T column changes periodically.				

Likewise more than 80 feedback forms collected and calculated the average of each dimension. Table 3 shows the part of the attribute questionnaire.

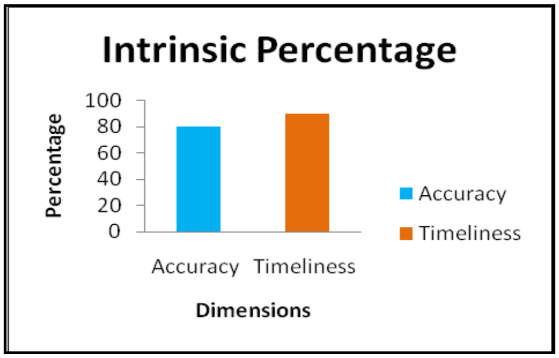
IV. INTRINSIC QUALITY

The Intrinsic DQ specifies the basic qualities of data like accuracy and timeliness. Accuracy ensure data are correct and valid values, Timeliness refers to the information is up to date and the articles are useful to our work or life. Chart1 represents the Intrinsic DQ in which the accuracy is 80% and the timeliness is 90%. On an average, the intrinsic quality of data, that's accuracy and timeliness is measured as 85% from the feedback collected.

TABLE 4 INTRINSIC DQ PERCENTAGES

Accuracy %	Timeliness %
80	90

Chart 1: Intrinsic Data Quality



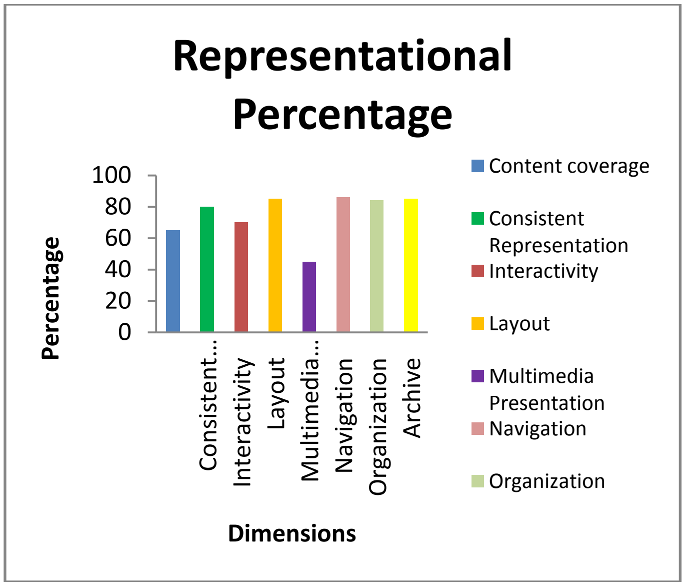
V. REPRESENTATIONAL DATA QUALITY

The Representational DQ specifies the way in which the data are presented or made available in the web portal.

The representational DQ includes content coverage, writing style, interactivity, layout, multimedia presentation, navigation, organization and archive. These factors help the online web portal to present their information in a most effective manner to the wide user. Chart2 represents the Representational DQ in which the data representational quality has been observed through various factors.

From the chart2 it is observed that the navigation of data is very high as 86%, and the Layout, organization and archive of the presentation of data are high and found to be 85%, 84 % and 85% with a very small difference of 1% among them from the feedback collected.

Chart 2: Representational DQ



Content coverage and interactivity are found to be 65% and 70%. Multimedia presentation is found to be a medium value of 45%.

VI. CONCLUSION

Understanding content and consumer preferences is unique, rather than asking consumers to describe what kind of news and information they want and how they should be covered, this study measured online newspaper content and measured consumer reaction. The study on the “S&T” column of “The Hindu” web portal have shown just the amount of presence of Intrinsic and Representational Data qualities which is quantified by their Data quality dimensions as previously mentioned in the data classifications section. Through quantifying the data quality dimensions, the study has been made with the exact presence of intrinsic and representational data qualities. This paper has made a sample study to quantify the Data qualities through their dimensions, so that importance can be given to areas in which a poor quantifying measure is shown. Future study can lead to all the columns of the paper, identification of lacking data quality in the portal, suggestions to improve the data quality can also be included .

TABLE 5 REPRESENTATIONAL DQ PERCENTAGES

Content coverage %	Consistent Representation%	Interactivity%	Layout%	Multimedia Presentation%	Navigation%	Organization%	Archive%
65	80	70	85	45	86	84	85

REFERENCES

[1] Chyi, H.I. & Lasorsa D., Access, Use and Preferences for Online Newspapers. Newspaper Research Journal, 1999, 20(4), 2-13.

[2] M. Angelica Caro,Coral Calero, Ismael Caballero, Mario Piattini., Data Quality In Web Applications: A State Of The Art ,IADIS International Conference on WWW/Internet 2005, pp 364-368.

[3] C. Cappiello, C. Francalanci, and B. Pernici., Data quality assessment from the user´s perspective in International Workshop on Information Quality in Information Systems, (IQIS2004). 2004. Paris, Francia: ACM. p. 68-73.

[4] Caro, C. Calero, H. Sahraoui, and M. Piattini, A Bayesian Network to Represent a Data Quality Model. International Journal on Information Quality, 2007. Accepted for publication in the inaugural issue 2007.

[5] InduShobha N. Chengalur-Smith, Donald P. Ballou, Harold L. Pazer, The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. IEEE Transactions on Knowledge and Data Engineering 11(6): 853-864, 1999.

[6] Monica Bobrowski, Martina Marr, Daniel Yankelevich: A Homogeneous Framework to Measure Data Quality. In MIT Conference on Information Quality (IQ), 115-124, 1999.

[7] Cappiello, C., et al., 2004. Data quality assessment from the user´s perspective. Proc. IQIS2004, pp: 68-73.

[8] Eppler, M. and Muenzenmayer, P.,2002. Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Aplication Methodology. Proc. of ICIQ2002, pp: 187-196.

[9] Pernici, B. and Scannapieco, M.,2002. Data Quality in Web Information Systems. Proceeding of the 21st International Conference on Conceptual Modeling, pp: 397-413.

[10] Chen, K & Yen, DC 2004, 'Improving the quality of online presence through interactivity', Information & Management, vol. 42, No. 1, p. 217.

[11] M. Gertz, T. Ozsu, G. Saake, and K.-U. Sattler, Report on the Dagstuhl Seminar "Data Quality on the Web". SIGMOD Record, 2004. vol. 33, No. 1: p. 127-132.

[12] P. Katerattanakul and K. Siau. Measuring Information Quality of Web Sites: Development of an Instrument. in Proceeding of the 20th International Conference on Information System. 1999. p. 279-285.

[13] Caro, C. Calero, I. Caballero, and M. Piattini. Defining a Data Quality Model for Web Portals. in WISE2006, The 7th International Conference on Web Information Systems Engineering. 2006. Wuhan, China: Springer LNCS 4255. p. 363-374.

[14] D. Strong, Y. Lee, and R. Wang, Data Quality in Context. Communications of the ACM, 1997. Vol. 40, Nº 5: p. 103 -110.

[15] S.A. Knight and J.M. Burn, Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 2005. 8: p. 159-172.

[16] Mohamed Haneefa K and Shyma Nellikka, Content Analysis of Online English Newspapers in India , DESIDOC Journal of Library & Information Technology, Vol. 30, No. 4, July2010