Information Quality on the World Wide Web: A Framework to Measure and its validation

Jaikrit S. KANDARI

Erick C. JONES

Fiona F. NAH
&
Ram R. BISHU
Industrial and Management Systems Engineering, University of Nebraska- Lincoln
Lincoln, NE 68503, USA

ABSTRACT

The scale and reach of published information on the World Wide Web dwarfs the printed paper world. Users are getting information from the web at the click of a button however they must dodge bad and sub-quality information before they can access and use quality content. Researchers have aimed to address this problem by suggesting various information quality frameworks. This article contends that these models though varied in their approach and application, share a greater commonality. It seeks to identify the common attributes that exist across these frameworks. A new framework for the measurement of information quality is developed and twenty two dimensions are identified for measuring information quality in context of the web from a user perspective. An online survey instrument is used for data collection. The research argues that WWW is not a homogeneous entity and should be understood from individual aspects and their interactions. It uses three independent variables of web domain, type of website and nationality to arrive at its conclusion. Results highlight nine IQ dimensions which are important across the whole web environment, while thirteen dimensions are dependent on the main effects or their interactions. The results also are in-line with Hofstede's cultural dimensions.

Keywords: Information Quality (IQ), IQ Frameworks, Hofstede's Cultural Dimensions

1. INTRODUCTION

The World Wide Web is arguably the largest available repository of data with the largest number of visitors searching for information (Herrera-Viedma et al. 2006). The scale and reach of published information on the web dwarfs the printed paper world. In many cases it happens without efficient information quality control (Herrera-Viedma et al. 2006). There are neither rules nor standards governing the type and quality of information that a writer can put on the web (Diligenti, Gori, & Maggine, 2004). One consequence of this oversight presents itself in the form of bad information.

The problem of information quality (IQ) has not escaped researchers' attention. Following general quality literature, Wang and Strong (1996) described information quality (IQ) as data that is 'fit-for use' by data consumers. They propose that assessing information quality (IQ) involves understanding it from the user's point of view. This article adopts this view and contends that data cannot be assessed independent of the people who use it. Literature search has revealed twenty major IQ frameworks which currently exist. Many of the subsequent researchers have adapted, expanded or validated the work of Wang and Strong (1996) at some levels while others have proposed different IQ perspectives. The main aim of this research is to identify the various IQ dimensions which are relevant in the web environment from a user perspective.

Further scope of the work which is not yet been researched in IQ literature is to investigate the behavior of these IQ dimensions across three independent variables of web domain, individual websites within a domain and national- culture differences. The research findings aim to build better understanding of IQ behavior on the World Wide Web from a user perspective.

2. INFORMTATION QUALITY AND IQ FRAMEWORKS

"Data" usually refers to information at its early stages of processing and "information", the product at a later stage (Strong et al. 1997b). In the context of this research the term "information" refers to both data and information and has been used interchangeably. Information quality (IQ) is commonly described in the literature as a multi-dimensional concept (Ballou et al. 1998; Klein, 2001; Aladwani et al. 2002; Gendrone et al., 2004). According to Strong et al. (1997a) high quality data is one that is fit for use by the data consumers. Information quality however is relative, as information considered useful for one person may not be 'fit' for another person's use (Tayi and Ballou, 1998). The next section traces the various IQ frameworks available in literature.

The authors Wang and Strong (1996) in their seminal work postulated the contextual IQ paradigm. They stated four information quality areas. The first- 'intrinsic data quality'- indicates that information has quality in its own right. It includes: accuracy, objectivity, believability and reputation. The second- 'contextual data quality'- requires that information should be provided on time and in appropriate amounts. It includes: relevancy, value-added, timeliness, completeness and appropriate amount of data. The third- 'representational data quality' comprises aspects related to the format of the information and its meaning. It includes: interpretability, ease of understanding, representational consistency and concise representation. Finally the fourth - 'accessibility data quality' emphasizes that information on the web must be easily accessible but secure. It includes: accessibility and access security.

Zeist and Hendricks (1996) presented the 'Extended ISO Model' which identified information quality characteristics and sub-characteristics. Their work was adapted by Leung (2001) to introduce the Adapted Extended ISO Model for Intranets. The period also saw application of IQ guidelines to build user-resources and 'how to' frameworks for the searchers of information (Knight, 2008). This was specifically directed to users of the World Wide Web. Notable frameworks in these were "CARS Checklist for Information Quality" (Harris, 1997), Web Evaluation Criteria (Beck 1997) and Web Wisdom (Alexander and Tate 1999). Some of criteria which kept

showing up and re-enforcing their importance were accuracy, objectivity and currency.

In 1999 Katerattanakul and Siau adapted the work of Wang and Strong (1996) to describe four IQ categories of individual websites while Shanks and Corbitt (1999) looked at the quality of data from a cultural aspect. Dedeke (2000), identified quality characteristics in terms of how they might be manifest in an electronic systems environment and Naumann and Rolker (2000) included subject, object and process criteria to define IQ criteria.

In context to the web, Zhu and Gauch (2000) suggested a quality metrics for information retrieval on the World Wide Web and Kahn et al. (2002) introduced the mapping of IQ dimensions into the PSP/IQ Model. Eppler & Muenzenmayer (2002) subdivided their suggested framework into content and media quality. The content quality is concerned about the quality of the information presented on the web. Media quality on the other hand is concerned about the quality of the medium used to deliver the web content. Adaption of Wang and Strong's (1996) work appeared again in 2002 when Liu and Chi (2002) introduced the "Evolutional Data Quality" framework primarily building on the foundation of four category IQ model. Klein (2002) too adopted the user-driven, consumption model of Wang and Strong (1996).

In recent literature Shankar and Watts (2003), suggested that accuracy, completeness, timeliness, believability and relevance are the core factors while discussing a theoretical model for data quality assessment. Sturges and Griffin (2003), Stvilia et al. (2005) and Song and Zahedi (2006) have contributed contextual models in the fields of archeological website quality, Wikipedia content and web-based health infomediaries respectively. Tombros et al. (2004) suggested five dimensions for judging quality in web pages and included IQ as one of the aspects of their identified web features. Liu and Huang (2005) in their work made mention of key dimensions like source, content, format and presentation, currency, accuracy and speed.

3. RESEARCH OBJECTIVES

Twenty major IQ frameworks were reviewed and some of the IQ dimensions kept re-enforcing their importance even across different contexts and environments, while some did not show up as frequently or not at all. This research is specifically aimed at the World Wide Web and hence the first research question:

"Which of the identified dimensions of IQ/DQ are relevant in the context of World Wide Web from the user perspective?"

The article is also interested in understanding the behavior of these IQ dimensions across differing web-domains and different websites within a web-domain. Hence- "How does IQ perception of the user change between different web domains and websites?"

This article argues that any website can potentially be visited by people from many different countries. Hall (1959, 1976) and Hofstede's (1980, 1980, 1983, 1984, 1985) cultural theories, suggest that different cultures have differing attitudes and views about web usability. It is an important area which cannot be overlooked and hence raised the third research question: "Can different national cultures lead to varying IQ dimensions?"

4. DEVELOPMENT OF THE FRAMEWORK

Knight (2008) states that despite the varied research contexts of the IQ frameworks and models, an analysis of the constructs column reveal a remarkable commonality amongst the eventual elements identified by various researchers as being important 'dimensions' of IQ. This study identified twenty three most frequently occurring dimensions. Table 1 provides a summary of the most common dimensions and the frequency with which they have appeared in the twenty IQ frameworks. It is interesting to note that all 15 dimensions proposed by Wang and Strong make the list. Timeliness ranked in 18 out of 20 frameworks. Accuracy, Accessibility, Amount of Data, Believability, Consistent Representation, Completeness, Objectivity and Relevancy showed their presence in 10 out of 20 frameworks. Usability and Usefulness with their presence 3 and 2 times in the list were at the bottom in the frequency table.

From a perspective of the web environment 'layout' and 'advertisement' were proposed as new dimensions which should be added to the list of 23. Thus a total of twenty five dimensions were proposed. These dimensions were defined in line with definitions understood in available literature. Up to four constructs were framed around each dimension and a focus group of five students was used to understand their perspective of the dimensions. Feedback from the focus group eliminated 'useful', 'usability' and 'layout'. The final list of 22 dimensions was used in subsequent research.

5. VALIDATION

The study uses a 2^3 or a 2*2*2 complete factorial between-subject research design. The model is shown in Table 2.

- Web Domain was tested at two levels namely the domains of NEWS and e-commerce.
- Nationality was tested at two levels, USA and INDIA for national culture.
- For 'Type of website within a domain' the study looked at two websites which were at opposite end of the spectrum within the same web domain. Website ranking portals www.alexa.com and www.compete.com were used to select the websites. The NEWS websites were of a foreign country by design, to control for learning effect/ bias if participants were given a news website of their host country. The websites selected for each domain are shown in Table 3.

Self-administered online survey questionnaire was selected as a means of data collection. The survey instrument was finalized after an iterative round of cognitive interviews. It consisted of 79 questions, 5 questions measured the demographics, 73 questions with up to 4 constructs measured each IQ dimension while 1 question was open ended to get a qualitative feedback from the users. A pilot study found Cronbach's Alpha values ranging from .563 to .899 thus suggesting that the instrument had a high reliability.

Participants were above 19 years of age. US respondents were from a large university in the mid-west region of United States while international respondents were from a research university in the western region of India. The final survey was hosted at www.surveymonkey.com. One of the requirements for users before the start of the survey was to complete a set of tasks designed to help them get a better picture of the construct definitions. The survey took around 30-45 minutes to complete.

DIMENSIONS	Wang and Strong	Zeist & Hendricks	Beck	Harris	Alexander &Tate	Katerattanakul & Shau	Shanks & Corbitt	Dedeke	Naumann & Rolker	Zhu & Gauch	Leung	Kahn et al	Liu & Chi	Eppler & Muenzenmayer	Klein	Shankar &Watts	Sturges & Griffin	Tombroset	Stvilia et el	Song & Zahedi	**TOTAL**
Accuracy	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X			X	17
Accessibility	X	X			X	X	X	X	X	X	X	X		X						X	12
Amount of Data	X			X		X		X	X			X	X	X	X		X	X			11
Availability		X						X	X	X				X					X	X	7
Believability	X			X		X	X		X	X		X	X	X		X			X	X	12
Consistent Representation	X					X	X	X	X		X	X	X	X	X		X		X		12
Completeness	X		X	X		X	X	X	X			X	X	X	X	X	X	X	X	X	16
Concise Representation	X					X		X	X		X	X		X	X		X				9
Efficiency		X						X		X	X	X	X	X			X			X	9
Navigation					X	X		X					X				X				5
Objectivity	X		X	X	X		X		X	X		X	X	X	X					X	12
Reputation	X			X		X	X		X			X							X		7
Relevancy	X					X		X	X	X	X	X	X	X	X	X				X	12
Reliability		X			X				X		X		X	X					X	X	8
Security	X	X						X	X		X	X	X	X					X	X	10
Timeliness	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		18
Understandability	X	X						X	X		X	X		X					X	X	9
Value added	X							X	X		X	X					X	X	X	X	9
Usability		X					X				X										3
Useful							X													X	2
Interpretability	X							X	X		X	X	X								6
Ease of Operation		X						X			X	X	X				X			X	7
Authority			X	X	X					X	X	X			X		X	X	X		10

Table 1: Tracing IQ Dimensional commonality in existing frameworks.

Factors	**Alias Used**	**Level 1**	**Level 2**
Web Domain	Domain	News	e-commerce
Website within each Domain	Type(Domain)	High Ranked (HR)	Low Ranked (LR)
Nationality	Nationality	USA	INDIA

Table 2: The between-subject research model

DOMAIN	**TYPE**	
	High Ranked Website	Low Ranked Website
NEWS	www.bbc.uk	www.star.co.uk
e-commerce	www.amazon.com	www.planetonline.com

Table 3: Website selection within each Domain

6. RESULTS

The research has three independent variables: Domain, Type of Website nested under Domain and Nationality. There were 22 dependent variables and ANOVA was run on SAS for each of the dependent variables the results are summarized in Table 4.

Summary Results						
Dimension	**Mean**	**Significance**				
		Domain	**Nation**	**Type(Domain)**	**Domain*Nation**	**Nation*Type(Domain)**
Accuracy	**4.17**	0.4123	0.281	0.1179	0.3212	0.0947
Accessibility	**4.14**	**0.0057**	0.2832	0.3022	0.6673	0.2954
Advertising	**3.85**	0.7442	0.8066	0.9967	0.6245	0.838
Amount of Data	**3.90**	0.7105	0.0647	0.461	1	0.5288
Authourity	**3.74**	0.0778	**0.0002**	0.101	0.6055	0.2013
Availaibility	3.41	**0.0145**	**<.0001**	0.9565	0.9688	0.9051
Believability	**4.13**	0.0698	0.2214	0.0692	0.9209	0.1704
Consistent Representation	**3.74**	0.1722	0.194	0.9094	0.09	0.1849
Completeness	**3.87**	0.304	0.4928	0.5371	0.3459	0.6894
Concise Representation	**3.79**	0.7582	0.4038	0.2715	0.2359	0.4826
Ease of Operation	3.43	**0.002**	**0.0037**	0.347	0.1575	0.593
Efficiency	**3.59**	0.4061	0.1154	0.9138	**0.0106**	0.2003
Interpretability	**3.60**	0.3493	**0.0107**	**0.0022**	0.1854	0.8531
Navigation	**4.05**	0.5858	0.4365	0.9158	0.8153	0.1106
Objectivity	**3.91**	0.5023	0.7655	0.6661	1	**0.0417**
Reputation	**3.86**	**0.0009**	0.107	0.2985	0.8064	0.3645
Relevancy	**3.70**	**<.0001**	0.6861	0.5321	0.165	0.0868
Reliability	**3.89**	**0.0474**	0.0798	0.5866	0.845	**0.02**
Security	**4.20**	**0.0033**	0.3498	0.79	**0.028**	0.4765
Timeliness	**3.76**	0.1862	**0.0042**	0.7999	0.5078	0.1244
Understand-ability	**4.19**	0.8793	0.5778	0.2036	0.8002	0.3178
ValueAdded	**3.97**	**0.0083**	**0.0181**	0.2523	0.8489	0.3954

Table 4 Summary results for 22 IQ dimensions against independent measures and their interactions

Figure 1 shows a four category result quadrant. High value of mean and no significant independent factors suggest that the IQ dimension is important across all nationalities, domain and web site types. High value of importance mean and significant main effect and/or interaction effect mean that IQ dimension though important depends on one or more of the independent factors. For results with low value of mean and no independent factor significant, it can be said that the dimensions are not important to IQ on the World Wide Web from a user perspective. Low mean and significant independent factors would mean that even though the dimension is not important from a user perspective in information quality it is still significantly impacted by the independent factors considered in the study.

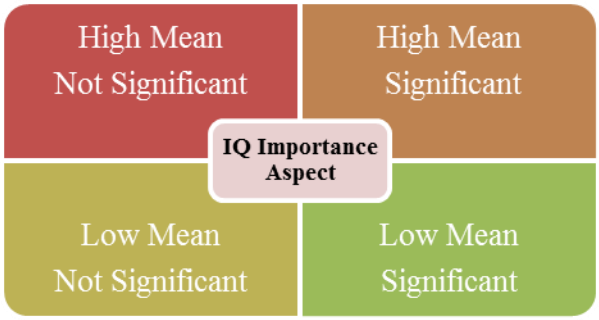

Figure 1: 2X2, four result quadrant

Summary results of Table 4 have been divided into the quadrants in Figure 2. Values of mean above 3.5 are considered higher and below 3.0 are considered low. Eleven dimensions fall in Ist quadrant, nine in the IInd quadrant, while none in quadrant IIIrd and IVth

Results in Table 4 indicate that nine dimensions are not impacted significantly by any of these independent factors. Dimensions included in the list which found no significant main effect or interaction effect are Accuracy, Advertising, Amount of Data, Believability, Consistent Representation, Completeness, Concise Representation, Navigation and Understandability.

Domain had a significant effect on three dependent dimensions, namely, Accessibility, Reputation and Relevancy. Main effect Nation has significant impact on two dependent IQ dimensions of Authority and Timeliness.

Domain and Nation are both seen as main effects for the IQ dimensions Value added, Availability and Ease of Operation. However Availability and Ease of Operation are not included in Figure 2 as they fall below the 3.5 mark but lie above 3.0 and hence are likely overlapping in the 1st and IVth quadrant. Interpretability is significantly impacted by Nation and Type nested under domain.

IQ dimensions of Security is dependent on the Domain as well as a Domain*Nation interaction. Graph of the main and interaction effects are plotted in Figure 4.1a and 4.1b respectively.

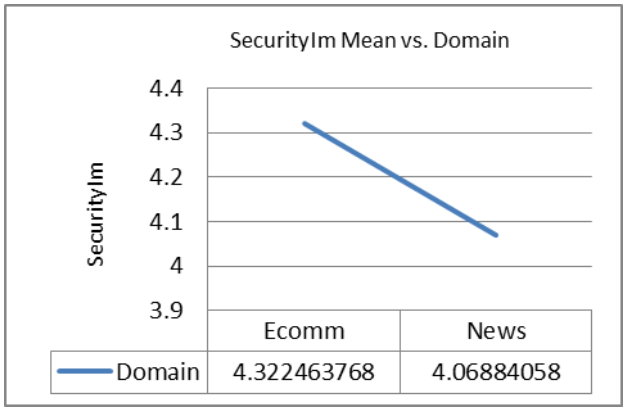

Figure 4.1a
Graph for Interaction Effect of Domain on Security

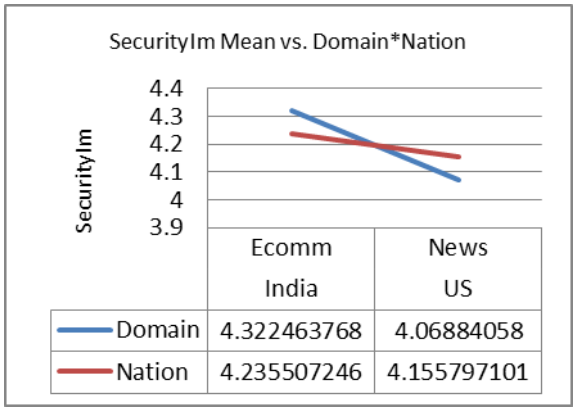

Figure 4.1 b
Graph for Interaction Effect of Domain and Nation on Security

Reliability is dependent on Domain and Nation*Type (Domain) interaction, while Efficiency and Objectivity show significant impact of Domain*Nation and Nation*Type (Domain) respectively.

Understandability	4.19	
Accuracy	4.17	
Believability	4.13	
Navigation	4.05	
Amount of Data	3.90	
Completeness	3.87	
Advertising	3.85	
Concise Representation	3.79	
Consistent Representation	3.74	
Accessibility	4.14	Domain
Reputation	3.86	Domain
Relevance	3.7	Domain
Authority	3.79	Nation
Timeliness	3.76	Nation
Value Added	3.97	Domain, Nation
Interpretability	3.6	Nation, Type (Domain)
Objectivity	3.91	Nation*Type (Domain)
Efficiency	3.59	Domain* Nation
Security	4.2	Domain, Domain* Nation
Reliability	3.89	Domain, Nation* Type (Domain)
NA		
NA		

Figure 2: Summary results divided into the result quadrant

7. DISCUSSION

Nine IQ dimensions did not show any variation across web domain, type of website and nationality, which indicated that they are important across the World Wide Web. Hence from the perspective of a web designer it is imperative that they give attention to the accuracy of the information on the website. They need to present it in a concise and consistent format at the same time ensuring that the information is complete and believable. Understandability is another dimension which ranked the highest in terms of importance across the web environment. In the web context end users want information which they can understand, with less advertisement and better navigation.

Accessibility, Reputation and Relevance are Domain dependent with high mean scores which indicates that user perception of these IQ dimensions varies across the domains but is still important. Nationality is significant for Authority and Timeliness which can find explanation in Hofstede's cultural dimensions of power distance and uncertainty avoidance index. Value added ease of operation and availability vary across Domain and Nationalities. This could be explained by the difference in perceived expectation users would have about the values, operating structure and availability from a NEWS domain versus an e-commerce domain. Mean values of security are expectedly very high for e-commerce as compared to NEWS. However there is a clear interaction of Domain and Nation in security this could potentially be explained by the culture dimensions in Hofstede's theory.

Qualitative analysis of the data from the research could result in more IQ dimensions relevant in context of the World Wide Web which have not yet been considered important in IQ literature. The two websites which have been selected based on the web-ranking portals could be validated using the feedback from the user on their perceived IQ ranks for the websites. These points could result in potential future work. However this research has succeeded in showcasing that the World Wide Web cannot be considered as one homogenous entity but is made up of a complex set of individual entities each of which may have its own set of IQ dimensions and requirements. However it also indicated that some of the IQ dimensions will stay important across the whole WWW while others may shift focus or change in importance relative to the domain, nationality or type of websites. It would be interesting to conduct this study in a wider range of web domains and across nationalities to contribute better generalizability of the results.

8. REFERENCES

1. Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. Information & Management, 39(6), 467-476.
2. Alexander, J. E., & Tate, M. A. (1999). Web wisdom: how to evaluate and create information quality on the web. Mahwah, NJ: Erlbaum.
3. Beck, S. (1997). Evaluation criteria: the good, the bad & the ugly; or why it's a good idea to evaluate Web sources. http://lib.nmsu.edu/instruction/evalcrit.html
4. Dedeke, A. (2000). A Conceptual Framework for Developing Quality Measures for Information Systems. Proc. of 5th International Conference on information quality, 126-128.
5. Diligenti, M., Gori, M., & Maggine, M. (2004). A unified probabilistic framework for Web page scoring systems. IEEE Transactions on Knowledge and Data Engineering, 16(1), 4–16.
6. Eppler, M., & Muenzenmayer, P. (2002). Measuring information quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. Proc. of 7th International Conference on information quality, 187- 196.
7. Gendron, M., & D'Onofrio, M. (2002). Formulation of a Decision Support Model Using Quality Attributes. Proc. of the 7th International Conference on Information Quality, MIT IQ Conference, 2002, 305-316.
8. Hall, E. T. (1959). The silent language. NY: Doubleday.
9. Hall, E. T. (1976). Beyond culture. NY: Anchor Press.
10. Harris, R. (1997). Evaluating Internet Research Sources. Virtual Salt http://www.virtualsalt.com/evalu8it.htm
11. Herrera-Viedma, E., Pasi. G., Lopez-Herrera, A., & Porcel, C. (2006). Evaluating the Information Quality of Web Sites: A Methodology Based on Fuzzy Computing With Words. Journal of the American Society for Information Science and Technology, 57(4), 538–549.
12. Hofstede, G. (1980). Culture's Consequences: International Differences in Work-Related Values. Beverly Hills, CA: Sage Publications.
13. Hofstede, G. (1983). The cultural relativity of organizational practices and theories.Journal of International Business Studies, 76-88.
14. Hofstede, G. (1984). Cultural dimensions in management and planning. Asia Pacific Journal of Management, 1(2), 81-99.
15. Hofstede, G. (1985). The interaction between national and organizational value system. Journal of Management Studies, 22(4), 347-357.
16. Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. Communications of the ACM, 45(4), 184 - 192.
17. Katerattanakul, P., & Siau, K. (1999). Measuring information quality of web sites: Development of an Instrument. Proc. of the 20th International Conference on Information Systems (ICIS'99), Charlotte, North Carolina, United States, 279-285
18. Klein, B. D. (2002). Internet data quality: Perceptions of graduate and undergraduate business students. Journal of Business and Management, 8(4).
19. Knight, S. (2008). User Perceptions of Information Quality in World Wide Web Information Retrieval Behavior. Dissertation: Edith Cowan University
20. Leung, H. K. N. (2001). Quality metrics for intranet applications. Information & Management, 38(3), 137-152.
21. Liu, L., & Chi, L. (2002). Evolutional Data Quality: A Theory-Specific View. Proc. Of the 7th International Conference on Information Quality, (MIT IQ Conference, 292-304.
22. Liu, Z., & Huang, X. (2005). Evaluating the credibility of scholarly information on the web: A cross cultural study. ScienceDirect, 37(2), 99-106.
23. Naumann, F., & Rolker, C. (2000). Assessment Methods for information quality Criteria. Proc. of 5th International Conference on information quality, 148-162.
24. Shankar, G., & Watts, S. (2003). A Relevant, Believable Approach for Data Quality Assessment. Proc. of 8th International Conference on information quality, 178-189.
25. Shanks, G., & Corbitt, B. (1999). Understanding Data Quality: Social and Cultural Aspects. Proc. 10th Australasian Conference on Information Systems, 785-797.
26. Song, J., & Zahedi, F. M. (2006). Trust in health infomediaries. Decision Support Systems, 43(2), 390-407
27. Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997a). Data quality in context. Communications of the ACM, 40(5), p.103 - 110.
28. Strong,D.M., Lee, Y.W., & Yang, R. (1997b). 10 Potholes in the Road to Information Quality. Computer, IEEE Computer society Press, 30(8), 38-46.
29. Sturges, P., & Griffin, A. (2003). The Archaeologist Undeceived: Selecting Quality
30. Archaeological Information from the Internet. Informing Science Journal, 6, pp.221- 232.
31. Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing Information Quality of a Community-Based Encyclopedia. 10th International Conference on Information Quality
32. Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. Communications of the ACM, 41(2), 54 - 57.
33. Tombros, A., Ruthven, I., & Jose, J. M. (2003). Searchers' Criteria for Assessing Web Pages. Proc. 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, July 28 - Aug 01, 2003, 385-386.
34. Wang, R.Y., & Strong, D.M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5-34.
35. Zeist, R. H. J., & Hendricks, P. R. H. (1996). Specifying software quality with the extended ISO model. Software Quality Management, 5(4), 273-284.
36. Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proc. 23rd annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, Athens, Greece, 288-295.