2011

A Systemic Method for Measuring Quality of Information on Weblog

Mohammad Javad Kargar

Australian Journal of Basic and Applied Sciences, 5(3): 44-50, 2011
ISSN 1991-8178

A Systemic Method for Measuring Quality of Information on Weblog

[1]Mohammad Javad Kargar, [2]Fatemeh Azimzadeh

[1]Department of Computer Engineering University of Science and Culture Tehran, Iran
[2]Department of Computer Engineering University Putra Malaysia, 43400 Serdang, Malaysia

Abstract: The vast amount of information on the World Wide Web is created and published by many different types of providers. Unlike books and journals, most of this information is not subject to editing or peer review by experts. This lack of quality control and the explosion of web sites make the task of finding quality information on the web especially critical. Meanwhile new facilities for producing web pages such as Blogs make this issue more significant because Blogs have simple content management tools enabling non-experts to build easily updatable web diaries or online journals. On the other hand despite a decade of active research in information quality (IQ) there is no framework for measuring information quality on the Blogs yet. This paper presents a framework for calculating and ranking quality of information on the Weblog. The current framework includes qualitative and quantitative criteria. The quantitative criteria were calculated automatically and the qualitative criteria were obtained by voting in the framework. The results of data analysis collected by system log files revealed high correlations between the criteria.

Key words: Information Quality; Web; Weblogs; Ranking.

INTRODUCTION

The amount of information on the web is growing rapidly. Measuring quality of information is one of the most important dimensions of information quality contexts. Despite the sizeable body of literature available on information quality, relatively few researchers have tackled the difficult task of quantifying some of the conceptual IQ definitions. The issue is more crucial where there is a focus on social networks.

Social networks refers to a range of web-based applications that allows users to interact and share information with one another Green and Pearson, (2005). The distinctive feature of such systems is the development of new ideas and concepts rather than technological innovation: Internet users are increasingly evolving from being an audience to forming a community that actively participates in the creation of content O'Reilly, (2007). With the emergence of a large number of wikis, weblogs, and social networking platforms like MySpace, Wikipedia, and Facebook, social networks has become very popular in the personal context.

In the current research, we focus on Weblogs. Weblogs are websites in which an author or a group of authors publishes articles sporadically or at regular intervals Raeth and S. Smolnik, (2009). The dynamic index page of a weblog lists the articles or extracts from them counter-chronologically so that the most recent item is listed first. Visitors can use this function to read the complete article, and they also have an opportunity to comments on it.

The author and other visitors can, in turn, respond to these comments, creating vivid discussions Ip and Wagner, (2008). Weblogs are often created by individuals or small groups of individuals, but the number of corporate weblogs is also steadily increasing Du and Wagner, (2006). The application areas of corporate weblogs are very diverse. Some corporate weblogs are only for internal use, but companies also apply this technology to market communications and public relation tasks Efimova and J. Grudin, (2007).

This paper presents a framework for measuring and ranking Weblogs. The presented framework has been tested on Persian (Farsi) Weblogs along with a Pesrian interface. The major reason for selecting Persian Weblogs as a test bed is that Persian, the official language of Iran, is the newcomer to the top 10 blogging languages Sifry, (2007).

Related Works:

Information quality research that started two decades ago has entered a new era where a growing number of researchers actively enhance the understanding of information quality problems and develop solutions to

Corresponding Author: Mohammad Javad Kargar, Department of Computer Engineering University of Science and Culture Tehran, Iran
Email: Kargar@usc.ac.i

emerging data quality issues. Recently, Madnick *et al* (2009) introduced a framework for characterizing information/data quality research along the dimensions of topic and method. Also, In our earlier work Kargar, *et al.,* (2009), we classified IQ research into four broad categories, which are: a) Many of resources have attempted to propose some IQ criteria for their respondents. For instance, Collins Memorial Library and Virtual Case Tyburski, (2007) have listed some IQ criteria.

b) The existing Literature has proposed different information quality models. These models include general purpose and special purpose models. General purpose models such as TDQM Wang and Strong, (1996), Naumann's model Naumann and Rolker, (2000) and AIMQ (2004) are the most popular general purpose models. special purpose models such as Data Warehouse Quality (DWQ) (1997), IQIP for information retrieval purposes Knight and J. Burn, (2005) intranet application Leung, (2001) and quality of information in Wikipedia (Stvilia, *et al*., 2005; 2005).

c) There are many researches which have tackled a few of criteria and have attempted to find methods for computing and measuring the criteria such as measuring timeliness (Zhang, *et al*., 2005; 2002) are examples of these works.

d) Some studies propose frameworks for evaluating the quality of conceptual models. . For instance, Moody *et al*., (2003) conducted an empirical analysis of the conceptual model quality framework proposed by Lindl and *et al*., (1994).

IQ criteria on the Weblog:

One of the most important issues in evaluating quality of information is to select the quality criteria because determining what to measure a difficult decision is. After studying the information quality literature, we encountered with many of the criteria with different classifications and interpretations. Actually, every study has interpreted and classified IQ criteria which are conformed to its context. In order to accurately define and measure the concept of information quality, it is not enough to identify the common elements of IQ frameworks as individual entities in their own right Knight and J. Burn, (2005). In fact, information quality needs to be assessed within the context of its generation Shanks and B. Corbitt, (1999) and intended use Katerattanakul and Siau, (1999). This is because the attributes of data quality can vary depending on the context in which the data is to be used Shankar and S. Watts, (2003).

Table I shows 18 selected IQ criteria and related sub-criteria for the Weblog. Of the 18 criteria, 9 criteria were obtained quantitatively by considered sub-criteria and the remaining 9 criteria were considered as qualitative criteria.

Weblog Content Management System:

Weblog management system is the heart of the framework. To implement the Weblog management system, it was decided to design a Content Management System (CMS). A content management system is a computer software system for organizing and facilitating collaborative creation of documents and other content. A content management system is a system used to manage the content of a Web site, Wikipedia, (2008). Content management systems are deployed primarily for interactive use by a potentially large number of contributors.

Many organizations have turned to CMS to publish data with the speed and freedom provided by the Web Rainville-Pitt and D'Amour, (2007). Many of modern applications have been developed by CMS. For example, the software for the website Wikipedia is based on a wiki, which is a particular type of content management system Wikipedia, (2008). Wiki systems are a form of content management system which enables a repository of information that may be updated easily by their users. Wiki systems such as wikipedia.org are similar to blogs in principle as they are based on user participation to add content Jazayeri, (2007). Harrison, (2006) believed that the CMS produces a Web page on the fly that looks just like ones handcrafted by humans.

Figure 1 shows the general structure of the CMS for ranking IQ on Weblog. For designing the Weblog management system, several technologies and tools were used. The current Weblog management system uses PHP, MySQL, HTML, CSS, JavaScript, and Ajax.

Administrator control panel is an interface that was designed for administrator to control, manage, and monitor the Weblog management system. User control panel is an interface designed for users to produce and manage contents of their Weblogs. A user is able to publish and manage his/her Weblog by logging to the user control panel. This panel provide features which user needs for managing a Weblog. Links management, adding new article, observing posted article, template selection and edition, general configuration of Weblog, comments configuration, friends' management, and sending files are the principle features of the user control system. IQ Management Modules calculate IQ scores for the Weblogs.

Table 1: IQ Criteria, Sub-criteria and Assessment Methods for the Weblog System

Criteria	Sub-criteria	Assessment Method
Understandability	-	Voting (user rating)
Informativeness	-	Voting
Representation	-	Voting
Accuracy	-	Voting
Completeness	-	Voting
Timeliness	-Last update -Last login	HTML Parsing
Believability	-	Voting
Concise	-	Voting
Cohesiveness	-	Voting
Maintainability	-Meta information checking	Parsing
Availability	-Ratio of visited links to failed links	Traffic Analyzer, HTML Parsing
Authority	-Number of written comments -Weblog Age -Number of posted articles -Number of external links -Number of internal links	Traffic Analyzer, counter
Latency	-Initial load time - Full load time	HTML Parsing
Popularity	-Number of received comments -Average of received comments -Number of visitors -Number of referred links	Traffic Analyzer, counter
Customer support	Customer support link	HTML Parsing
Amount of Data	-Weblog size	HTML Parsing
Objectivity	-	Voting
Redundancy	Ratio of multimedia elements to the overall information	HTML Parsing

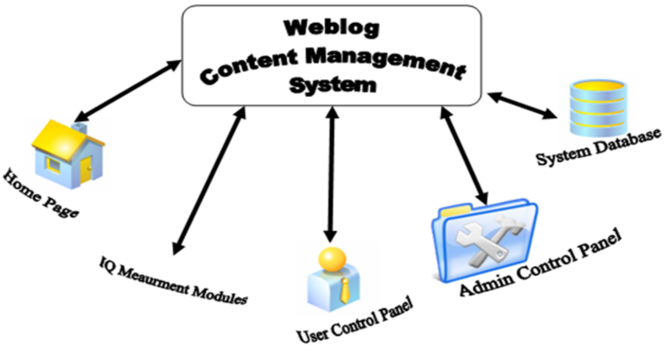

Fig 1: General Structure of the CMS for Ranking IQ on Weblog

Assessment Methods:

As seen in the Table I, the criteria, sub-criteria and the method for evaluation of each criterion are displayed. In general, The methods for evaluation of the criteria are classified into four groups:

1. Voting: This method is based on voting and is related to the qualitative criteria.
2. Parsing HTML: In a general sense Parsing is the process of analyzing an input sequence in order to determine its grammatical structure with respect to a given formal grammar.
3. Traffic Analyzer: Web Site Traffic Analyzer is a log analysis tool that monitors traffic to the Web site. The program lets us track traffic patterns to and from the Web site. Traffic Analyzer tracks how visitors flow through the site, including which pages visitors enter and what links they followed to find the site.

Implementation of Quantitative Criteria:

As mentioned earlier, 9 qualitative criteria were considered to measure IQ in Weblog. Each of the criteria was calculated by software module. We developed related source codes for the criteria. The source codes were developed by PHP in server side or by JavaScript in the client side. Some of the criteria were derived by previous research, while some of the criteria such as number of written comments, average number of received comments for each article were developed for the first time in the Weblog system. As comments are one of the most important attributes of Weblogs, number of written comments by Weblog's members was taken into account as a sub-criterion for authority. It is clear that written comments were directly influenced by the authors and can be considered as a sub-criterion for authority. For calculating the number of written comments, a counter module was written. Average number of received comments for each article is gained by:

$$Average\ number\ of\ received\ comment\ per\ entry = \frac{Total\ number\ of\ recived\ comments}{Total\ number\ of\ enties} \quad (1)$$

The rest of the sub-criteria were calculated and implemented by PHP code by related formula.which because of page limitation we pointed out only two sub-criteria.

Qualitative Criteria:

When measuring attributes of entities, we strive to keep our measurements objective. Although no measurement is truly objective, because there is always some degree of subjectivity about the entities and attributes, some measures are clearly more subjective than others. Subjective measures depend on the environment in which they are made Fenton and Pfleeger, (1997). On the other hand, quality is a matter of perception, and is often difficult to measure objectively. Like all other quality measures, it should be judged by the receiver.

Sizable portion of information quality criteria are often of subjective nature and can therefore not be assessed automatically. Some of the criteria are totally subjective such as understandability, believability, concise, and objectivity. Some of the criteria such as informativeness, representation, accuracy, completeness and cohesiveness mainly have subjective nature but some scholars have attempted to find automatic methods for measuring the criteria. Often the solutions require rigorous artificial intelligence techniques.

As noted earlier, the current framework comprised of 18 criteria, 9 of which were measured automatically and the remaining 9 criteria were obtained by voting. For this aim, a voting module was implemented as a division of Weblog's comment.

When users intended to leave a comment for a Weblog's post, in addition to writing comments could participate in the voting. There were 9 statements in voting division according to 9 criteria. Users could select scores between 1, as the lowest score, to 9 as the highest score. The results of voting were stored automatically in the system database.

Data Analysis and Correlations:

Correlation is primarily concerned with determining whether a relationship exists, to what magnitude and towards what direction it exists. A statistic analysis by SPSS shows there are significant correlations between many of IQ criteria. Here we point some the highest correlation (the correlation table is too big). From 153 pair correlations, 101 cases were significant. In other words, 66% of the correlations were significant. More importantly, of the correlations, 80% are significant at the .01 level.

Correlation between friends and referred is .623. This means that Weblogs which had more friends and were referred by other Weblogs. Correlation between last login and last update is .697 at the .01 level. This means that the last login has a strong linear relation with last update. In other words, the correlation states that when a Blogger logs into the system, he then usually updates his Weblog. Correlation between written comments and received comments is .639. This high correlation shows that the more comments a Blogger writes, the more comments he usually receives from other Bloggers. Correlation between referred and age is .668. This shows older Weblogs have been referred more by other Weblogs. Correlation between links and visited links is.9. This means that Weblogs which have more links could attract more visitors for the links. Correlation between first load time and full load time is .907. This is completely logical as first load time, which is related to the template of Weblog, influences full load time. Thus, Weblogs which had big templates encountered high first and full load times.

Also an analysis on qualitative data collected by voting shows there are very strong correlations between subjective variables. All the correlations are significant at the .01 level with a coefficient more than .96. The statistics show that Bloggers did not make significant distinction between subjective criteria for evaluating quality of information on Weblogs. Because of this, a voting-averages variable, which was as average of the nine voted criteria, was added. The use of the voting-averages as representative of the nine criteria is justifiable because correlation between them was more than 98%.

Table II shows correlation between voting-averages and sub-criteria which were obtained automatically by machine in our Weblog management system. From 18 sub-criteria, 11 cases were significant. Of these cases, seven sub-criteria are significant at the .01 level. On the other hand, the data shows that values of visited links, links, load time, comment per entry, Meta tag and multimedia rate have no correlation with voting-averages. This means that the variables did not have a relationship with the Bloggers' viewpoints about subjective information quality criteria.

Overall Quality of Information Score:

Given that all the variables, including the 18 sub-criteria calculated automatically and the nine variables obtained by voting were already saved in the system database, the overall quality of information for each Weblog could then be calculated. Thus, 27 variables were considered for measuring quality of information on the Weblogs. Since the scores for the criteria were measured on different measurement scales, standardization to a common dimension unit before calculating overall score must be done.

Min-max normalization is one of the standardization methods used to perform a linear transformation of the original data. Min-max normalization is calculated using the following formula Han and Kamber, (2006).

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \tag{2}$$

Table 2: correlation between Voting-Averages and 18 sub-criteria

Criteria	Voting average	
LastLogin	Pearson Correlation	.231(**)
	Sig. (2-tailed)	.000
LastUpdate	Pearson Correlation	.182(**)
	Sig. (2-tailed)	.001
Visitors	Pearson Correlation	.128(*)
	Sig. (2-tailed)	.023
Friends	Pearson Correlation	.141(*)
	Sig. (2-tailed)	.013
Referred	Pearson Correlation	.194(**)
	Sig. (2-tailed)	.001
Links	Pearson Correlation	.106
	Sig. (2-tailed)	.061
VisitedLinks	Pearson Correlation	.075
	Sig. (2-tailed)	.184
WrittenComments	Pearson Correlation	.178(**)
	Sig. (2-tailed)	.002
ReceivedComments	Pearson Correlation	.215(**)
	Sig. (2-tailed)	.000
Entries	Pearson Correlation	.185(**)
	Sig. (2-tailed)	.001
CommentPerEntry	Pearson Correlation	.035
	Sig. (2-tailed)	.535
FirstLoadTime	Pearson Correlation	.017
	Sig. (2-tailed)	.768
FullLoadTime	Pearson Correlation	.043
	Sig. (2-tailed)	.451
MultimediaRate	Pearson Correlation	-.093
	Sig. (2-tailed)	.101
METATag	Pearson Correlation	-.110
	Sig. (2-tailed)	.051
Age	Pearson Correlation	.142(*)
	Sig. (2-tailed)	.012
Availability	Pearson Correlation	.219(**)
	Sig. (2-tailed)	.000
WeblogSize	Pearson Correlation	.121(*)
	Sig. (2-tailed)	.032
Voting_average	Pearson Correlation	1
	Sig. (2-tailed)	

wherein the range was selected as 1 and 0 respectively for normalization of the criteria.

After normalization of the data, summated scores for each Weblog using the sum of the 27 variables were calculated. The summated score shows the score of each Weblog as a unique number. The following section discusses factor analysis as another method for calculating IQ scores in Weblogs.

Conclusions:

The aim of this research was to develop a framework for ranking information quality on Weblogs. In order to develop the framework, appropriate information quality criteria for Weblogs were first identified. Then the Weblog content management system was developed. The Weblog content management system contained all the facilities for content production on Weblog. Moreover, all the activities carried out by participants, as well as their information quality scores were recorded in the system database. The presented framework was ranked Weblogs based on selected IQ criteria.

The results of a statistical analysis collected by the system log files showed that there were significant correlations between many of the criteria. Moreover, because of the special nature of Weblogs, three special variables were considered and measured. These variables had not been considered in previous information quality research; namely the number of written comments, number of received comments and comment per entry, all of which were calculated automatically.

In future research, the authors plan to continue analyzing and data mining on the collected data related to IQ criteria and sub-criteria. It is clear that the current framework does not cover all the criteria for the Weblog. For future works, we plan to extend the framework to include more measures for IQ assessment.

REFERENCES

Du, H.S. and C. Wagner, 2006. "Weblog Success: Exploring the Role of Technology," *International Journal of Human Computer Studies,* 64: 789-798.

Efimova, L. and J. Grudin, 2007. "Crossing Boundaries: A Case Study of Employee Blogging," in *40th Hawaii International Conference on System Sciences (HICSS-40),* Big Island, Hawaii, USA.

Fenton, N.E. and S.L. Pfleeger, 1997. *Software Metrics: A Rigorous & Practical Approach*, second ed.: International Thomson Computer Press,

Green, D.T. and J.M. Pearson, 2005. "Social Software and Cyber Networks: Ties That Bind or Weak Associations within the Political Organization," in 38th Hawaii International Conference on System Sciences (HICSS-38), Big Island, Hawaii,

Han, J. and M. Kamber, 2006. *Data Mining: Concepts and Techniques*, 2 ed.: Elsevier Science & Technology,

Harrison, W., 2006. "Content Mismanagement Systems," *Software, IEEE,* 23: 5-8.

Jazayeri, M., 2007."Some Trends in Web Application Development," in *Future of Software Engineering (FOSE '07)*pp: 199-213.

Jarke, M. and Y. Vassiliou, 1997. "Data warehouse quality design: A review of the DWQ project," in *Proceeding of the International Conference on Information Quality (IQ),* Cambridge, MA.

Kargar, M.J., A.A. Ramli, H. Ibrahim and S.B. Noor, 2007. "Assessing Quality of Information on the Web Towards a Comprehensive Framework," in *14th IEEE International conference on Internet Communication Technology (ICT/MICC)* Malaysia: IEEE.

Katerattanakul, P. and K. Siau, 1999. "Measuring information quality of web sites: Development of an instrument," in *Proceedings of the 20th international conference on Information Systems*, Charlotte, North Carolina, United States, pp: 279-285.

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

O'Reilly, T., 2007. "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," Communications & Strategies, 65: 17-37.

Raeth, P. and S. Smolnik, 2009. "Towards Assessing the Success of Social Software in Corporate Environments," in Fifteenth *Americas Conference on Information Systems*, California, pp: 37-45.

Ip, K.F.R. and C. Wagner, 2008. "Weblogging:A study of social computing and its impact on organizations," *Decision Support Systems,* 45: 242-250.

Lee, Y.W., D.M. Strong, B.K. Kahn and R.Y. Wang, 2002. "AIMQ: a methodology for information quality assessment," *Information & Management,* 40: 133–146.

Leung, H.K.N., 2001."Quality metrics for intranet applications," *Information & Management*, 38: 137-152.

Lindland, O.I., G. Sindre and A. S¢lvberg., 1994. "Understanding Quality in Conceptual Modeling," *IEEE Software,* pp: 3.

Madnick, S.E., R.Y. Wang, Y.W. Lee and H. Zhu, 2009. "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality,* 1: 1-22.

Moody, D.L., G. Sindre, T. Brasethvik and A. Solvberg. 2003. "Evaluating the Quality of Information Models: Empirical Testing of a Conceptual Model Quality Framework," in *presented at 25th IEEE International Conference on Software Engineering (ICSE'2003)* Portland, Oregon,

Naumann, F. and C. Rolker, 2000. "Assessment methods for information quality criteria," in *Proceedings of 5th International Conference on Information Quality*, pp: 148-162.

Rainville-Pitt, S. and J.-M. D'Amour, 2007. "Using a CMS to create fully accessible websites," in *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, Banff, Canada, pp: 130-131.

Ricigliano, L., 2007."Criteria for Evaluating Information on the Web," http://library.ups.edu/research/handouts/author.htm Retrieved G. Tyburski, "Criteria for Quality in Information," http://virtualchase.com/quality/criteria.html,

Shanks, G. and B. Corbitt, 1999. "Understanding data quality: Social and cultural aspects," in *Proceedings of the 10th Australasian Conference on Information Systems*,

Shankar, G. and S. Watts, 2003. "A relevant, believable approach for data quality assessment," in *Proceedings of 8th International Conference on Information Quality*, pp: 178-189.

Sifry, D., 2007."The State of the Live Web," http://technorati.com/weblog/2007/04/328.html,

Stvilia, B., M.B. Twidale, L. Gasser and L.C. Smith, 2005. "Information quality discussions in Wikipedia," in *Proceeding of the International Conference on Knowledge Management (ICKM05)*, pp: 1-20.

Stvilia, B., M.B. Twidale, L.C. Smith, and L. Gasser, 2005. "Assessing information quality of a community-based encyclopedia," in *Proceedings of the International Conference on Information Quality (ICIQ)*, Cambridge, pp: 442-454.

Wang, R.Y. and D.M. Strong, 1996. "Beyond accuracy: what data quality means to data consumers.," *Journal of Management Information Systems,* 12: 5-34.

Wikipedia, 2008. "Content Management System," http://en.wikipedia.org/wiki/Content_management_system.

Zhang, Y., H. Zhu and S. Greenwood, 2005. "Empirical Validation of Website Timeliness Measures," in *Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC'05),* pp: 313-318.

Zhang, Y., H. Zhu, Q. Huo and S. Greenwood, 2002. "Measurement of Timeliness of Web-based Information Systems," in *Proceedings of the 6th World Multi-Conference on Systemic, Cybernetics and Informatics (SCI 2002).*