UNIVERSITI PUTRA MALAYSIA

EA FRAMEWORK FOR EVALUATING INFORMATION QUALITY OF PERSIAN WEBLOGS

MOHAMMAD JAVAD KARGAR BIDEH

A FRAMEWORK FOR EVALUATING INFORMATION QUALITY OF PERSIAN WEBLOGS

By

MOHAMMAD JAVAD KARGAR BIDEH

Thesis Submitted to the School of Graduate Studies, University Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

September 2008

To my parents, my wife and my son.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

A FRAMEWORK FOR EVALUATING INFORMATION QUALITY OF PERSIAN WEBLOGS

By

MOHAMMAD JAVAD KARGAR BIDEH

April 2008

Chair: Associate Professor Abd Rahman Ramli, PhD

Faculty: Engineering

The World Wide Web is a great tool for exploring all kinds of information. Unlike books and journals, most of this information is unfiltered, i.e. not subject to editing or peer review by experts. This lack of quality control and the large increase in number of web sites make the task of finding quality information on the web especially critical. Meanwhile, new facilities for producing web pages such as weblogs make this issue more significant because Blogs are simple content management tools that enable non-experts to build easily updatable web diaries or online journals. Despite a decade of active research, a comprehensive methodology for the assessment of Information Quality (IQ) is lacking. Specifically, no framework for measuring information quality on the weblogs is currently available.

After identifying and prioritizing IQ criteria on Weblogs, a Weblog management system that automatically calculates and collects IQ scores for created Weblogs is developed. The system is implemented on Persian Weblogs. Results of analysis of data collected by the Weblog management system show that there are significant correlations between many of the information quality variables. In addition, an analysis of the data revealed seven IQ dimensions on the Weblogs. Each of the dimensions was comprised of related IQ variables. Coefficients are identified for each variable in order to facilitate IQ measurement on the Weblogs. Moreover, statistical analysis shows that three specific sub-criteria for Weblogs; namely the number of written comments, number of received comments and comment per entry influence information quality on the Weblogs and interestingly fall into same dimension.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

SATU RANGKA KERJA UNTUK MENILAI KUALITI MAKLUMAT WEBLOGS PERSIAN

Oleh

MOHAMMAD JAVAD KARGAR BIDEH

April 2008

Pengerusi: Profesor Madya Abd Rahman Ramli, PhD

Fakulti: Kejuruteraan

Web Lebar Sedunia adalah satu perkakas yang hebat untuk menjelajahi semua jenis maklumat. Tidak seperti buku-buku dan jurnal-jurnal, kebanyakan maklumat ini tidak ditapis., i.e. bukan subjek suntingan atau pemeriksaan semula oleh pakar-pakar. Kekurangan kawalan kualiti dan peningkatan yang besar dalam bilangan laman-laman web menyebabkan tugasan pencarian maklumat yang kualiti ke atas web agak kritikal. Sementara itu, kemudahan-kemudahan yang baru untuk menghasilkan laman-laman web seperti Weblog-Weblog menyebabkan isu ini menjadi lebih signifikan kerana blog-blog ialah perkakas pengurusan muatan mudah yang membolehkan orang-orang bukan pakar membina secara mudah catatan-catatan harian web atau jurnal-jurnal dalam talian yang boleh dikemaskinikan. Walaupun satu dekad penyelidikan aktif, satu kaedah komprehensif untuk penilaian Kualiti Maklumat (KM) agak kurang. Khususnya, sehingga kini masih tiada rangka kerja untuk menilai kualiti maklumat ke atas Weblog-Weblog.

Setelah pengidentitian dan pengutamaan kriteria KM ke atas Weblog-Weblog, satu sistem pengurusan Weblog yang mengira dan mengumpul perhitungan KM secara automatik untuk binaan Weblog-Weblog dibangunkan. Sistem ini dibangunkan ke atas Weblog Persian. Hasil daripada data analisis yang dikumpul oleh sistem pengurusan Weblog menunjuk bahawa ada beberapa korelasi yang signifikan antara kebanyakan pembolehubah-pembolehubah KM. Tambahan pula, satu data analisis menunjukkan ada tujuh dimensi KM ke atas Weblog-Weblog. Setiap dimensi mengandungi pembolehubah-pembolehubah KM yang berkaitan. Pemalar-pemalar ditentukan untuk setiap pembolehubah dengan tujuan untuk memudahkan penilaian KM ke atas Weblog-Weblog. Di samping itu, analisis statistik menunjukkan ada tiga sub-kriteria yang spesifik untuk Weblog-Weblog, iaitu bilangan komen yang disampaikan, bilangan komen yang diperolehi, dan komen per masukan mempengaruhi KM ke atas Weblog-Weblog dan kepentingan jatuh ke dalam dimensi yang sama.

ACKNOWLEDGEMENTS

Acknowledgement is not a play of words, but an attitude of mind. If words are considered as the symbol of approval and tokens of appreciation, then let the words play the heralding role to expressing my gratitude.

First and foremost of all, I pay my obeisance and gratitude to the Allah for giving me the ability to carry out the research work and completing it.

I would like to express my sincere and deep gratitude to my supervisor, Associate Prof. Dr. Abd Rahman Ramli. His unwavering support and advice throughout my two years of PhD study enabled me to focus on what I needed to learn and complete my studies on time.

Special thanks to my co-supervisor, Associate Prof. Dr. Hamidah Ibrahim for her helpful comments and suggestions in completing this thesis. Also thanks to Dr. Samsul. B. Noor for his support in my research committee.

I would like to thank my friends and colleagues for their motivation, support and help accorded throughout my study in University Putra Malaysia. This is also extended to everyone who helped me directly or indirectly in making my graduate studies smooth journey.

Last but not the least, I would like to express my gratitude and appreciation to my family for their guidance, encouragements, moral support and their patience in tolerating my idiosyncrasies throughout my course of study and research work.

This thesis was submitted to the Senate of University Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Abd Rahman Ramli, Phd
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Hamidah Ibrahim, Phd
Associate Professor
Faculty of Computer Science
Universiti Putra Malaysia
(Member)

Samsul Bahari B. Mohd Noor, Phd
Lecturer
Faculty of Engineering
Universiti Putra Malaysia
(Member)

AINI IDERIS, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Pura Malaysia

Date: 13 November 2008

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

MOHAMMAD JAVAD KARGAR BIDEH

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREAVIATION

AIMQ	A Methodology for Information Quality
CMS	Content Management System
DoI	Denial of Information
DoS	Denial of Service
DWQ	Data Warehouse Quality
HUF	Homepage Update Frequency
IQ	Information Quality
IQIP	Identify, Quantify, Implement, and Perfect
IS	Information System
LAMP	Linux, Apache, MySQL, PHP/Perl/Python
MAMP	Mac, Apache, MySQL, PHP/Perl/Python
MTDTP	Mean Time Delay To Publish
PHP	Hypertext Preprocessor
PSP	Product and Service Performance
QoI	Quality of Information
QoS	Quality of Service
SES	Site Evolution Speed
TDTP	Time Delay To Publish
WAMP	Windows, Apache, MySQL, PHP/Perl/Python

CHAPTER 1

INTRODUCTION

The vast amount of information on the World Wide Web is created and published by many different types of providers, including businesses, organizations, governments, and individuals. Unlike books and journals, most of this information is unfiltered, i.e. not subject to editing or peer review by experts. So it's important to evaluate the Web sources one uses. Any source one finds is written for specific reasons that may or may not be useful for everybody purposes. The University of California, Berkeley study on how much information is created each year clearly illustrates the problem [1]:

- In 2002, about 5 Exabytes of new information was created in print, film, magnetic and optical formats. Five Exabytes is equivalent to 37,000 times the size of the United States Library of Congress book collection or 800 megabytes per person based on the world population.

- From 1999 to 2002, information in these formats grew at a rate of 30% per year. Ninety-two percent of this information was stored on magnetic media [2].

- Ninety-two percent of this information was stored on magnetic media.

While it is useful to have access to so much diverse and uncensored material, it is important to remember that internet browsers and search engines do not discern between valid, useful information and the inaccurate, useless stuff. Unlike most print publications which have editors and editorial boards to screen and select content, any individual or group can publish on the World Wide Web. This lack of quality control and the explosion of web sites make the task of finding quality information on the web especially critical.

Another aspect of this issue is role of information in decision. How much is the information worth? In the context of decisions, the value of information is the expected increase in utility of the decision as a result of having the information. This issue is more significant when variety of information sources, distributed, unknown locations and different forms of information presentations are considered. Moreover, users who vary in their preferences and background knowledge which is required to interpret the information and motivation for accessing it, gather information to perform many different tasks [3].

At present, content is considered to be the most important element of websites [4] and is seen to be directly related to website success [5]. To encourage repeat visits, visitors need to be provided with appropriate, complete and clear information [6].

Many Internet applications, e.g., digital libraries and electronic commerce, are built around information flows. Their main goal is to transport the right information to the right user at the right time. From school children to experts who manage critical national scale systems, an increasing number of information consumers depend on information content that is relevant, accurate and satisfactory in serving the request.

Ahamad et.al. [7] believed that providing Quality of Information (QoI) in large networked information flow applications is a research challenge that immediately follows the Quality of Service (QoS) research. In analogy to the many dimensions of QoS, there are also many dimensions of QoI, such as the consistency, timeliness, reliability, trustworthiness, and density/richness of information [7].

On the other hand in the early days of the Web, the technology was new and therefore only webmasters as specialists could make web pages. As the Web continues to develop, new technologies facilitate environments for producing web pages. Weblogs or blogs are the latest ways by which students, businessmen, and many others publish their mentality. Blogs are simple content management tools enabling non-experts to build easily updatable web diaries or online journals. They are published chronologically, with links and commentary on various issues of interest. Weblog tools enable the author to describe and edit the small contents via a web browser and transform the contents form text format to HTML files.

Blog became a popular media for publishing information on the internet [8] and has come into the spotlight in the World Wide Web [9]. Ohmukai et.al. [9] called these frequently-posted contents as small contents. A vast number of the small contents and citations among Weblog communities are increasing day by day. Some efforts such as topic discovery, trend analysis and content ranking are applied to these large amounts of information. In May 2007, Blog search engine Technorati tracked more than 70 million blogs. Every day 120,000 new blogs are created and 1.5 million posts are made [10]. Figure 1.1 shows the number and growth of Weblogs from March 2003 until March 2007.

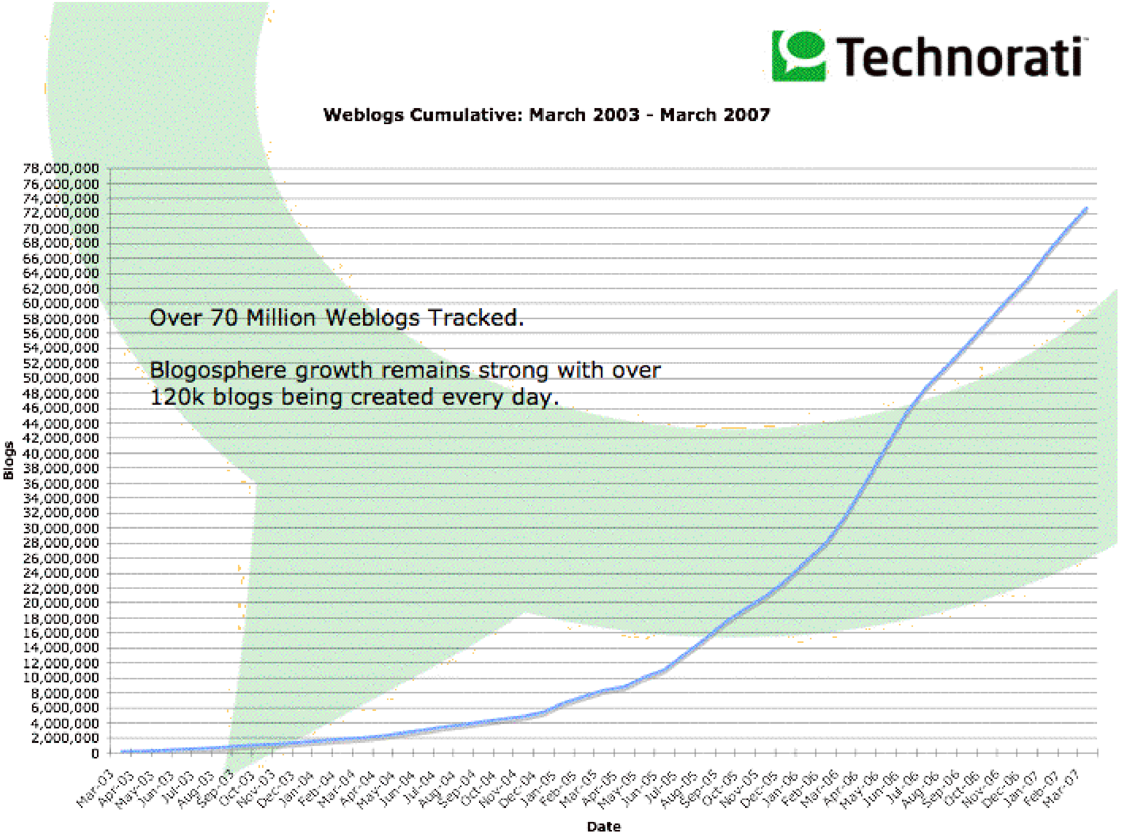

Figure 1.1: Number and Growth of Weblogs from March 2003 until March 2007 [10]

Ranking information source on web and Weblog can help users to better select their information sources. A fundamental assumption made by many information rich applications is the ability of the system to find and deliver information with satisfactory Quality of Information when such information is needed [7]. Despite a decade of research and practice, only piece meal, ad hoc techniques are available for measuring, analyzing and improving Information Quality (IQ) on the Web.

Undoubtedly a practical IQ model can be used to rank information sources by quality of information metrics.

1.1 Motivation and Problem Statements

The World Wide Web offers information and data from all over the world. Because so much information is available, and because that information can appear to be fairly "anonymous", it is necessary to develop skills to evaluate what one finds. There is no filtering process for the web. Because anyone can create a web page, fraudulent web pages can appear equally with articles from peer-reviewed journals.

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. Evaluating web sites quality requires appropriate evaluation criteria. Many of existing criteria are not easy to measure and require methods such as heuristic evaluations, or/and empirical usability tests.

Determining what to measure is a difficult decision: often is focused on attributes that are convenient or easy to measure rather than those that are needed.

Generally quality evaluation approaches suffer from several limitations:

- There is a general aim to define very general criteria, not addressing the specific type of site or page. There are differences among e-government, information target specific and large public sites. These differences must be taken into account when measuring the characteristics of the sites, which should be appropriately weighted. For example, a link rich page can be considered a positive element for informative parts of a site, while could disturb in a service specific section/page, where the user should be driven to accomplish his/her task in a linear manner.

- Criteria are not orthogonal. Same characteristics are often considered more than once, so contributing to a higher or lower score, depending on they have been fulfilled or not. However, this is unavoidable.

- IQ criteria are often of subjective nature and can therefore not be assessed automatically, i.e., independent of the user [11]. In the other hand the perception of the quality changes from different user perspectives: the final user is interested in external quality related to the usability and functionality of the site, while the developer is more interested to the internal quality related to software maintainability and portability.

- Information sources usually are autonomous and often do not publish useful (and possibly compromising) quality metadata. Many sources even take measures to hinder IQ assessment.

• The enormous amount of data to be assessed impedes assessment of the entire information set. Thus sampling techniques are often necessary which decrease the precision of the assessed scores.

• Information from autonomous sources is subject to sometimes surprising changes in content and quality [11].

• Finally, to define a metrics, we need measurable characteristics and a rigorous approach [12].

Despite the sizeable body of literature available on Information Quality, relatively few researchers have tackled the difficult task of quantifying some of the conceptual definitions IQ. In fact, a general criticism within the IQ research field is that most approaches lack methods or even suggestions [11]. Particularly there is not any framework for measuring IQ in Weblogs.

Apart from the quality of information on the Web issue, there are relatively rich sets of quality assessment frameworks and tools for evaluating web pages but there is not any practical framework for evaluating a Weblog as special case of web. Even page rank as a service which has been developed by Google does not cover many of Weblogs. Google page rank shows page rank for just a few of popular Weblogs which are visited frequently.

Developing a model for evaluating quality of information on the Weblogs provide a bed for ranking Weblogs. We believe that quality of content of a Weblog can evaluate quality of Weblog because Weblogs have same structures and similar templates. What makes Weblogs different from each other is the content. Therefore quality of information on Weblog can be declared as quality of Weblog considerably.

Ranking information quality of Weblogs provides a context for controlling quality of Weblogs. The quality control helps to standardize criteria and models for Weblog quality and information quality on Weblog. Moreover ranking Weblogs based on information quality criteria encourage Weblog owners to produce more valuable contents. The ranking system constructs a competitive environment for gaining higher score between Weblog owners. The motion will improve quality of whole Weblog system ultimately.

An important aspect of developing information quality model is that can be employed by search engine. It is clear that a search engine based on information quality criteria can find quality information on the web more efficiently in comparison with a search engine which does not employ information quality factors. Therefore evaluation of Weblogs can be lead to improvement of search engines and crawlers performance. Improvement of search engines results customer satisfactory and finding useful information for user application and meets consumer expectations.