Lecture Notes in Computer Science

The LNCS series reports state of the art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI and LNBI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online.

Detailed information on LNCS can be found at www.springer.com/lncs

Proposals for publication should be sent to LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
E-mail: lncs@springer.com

ISSN 0302-9743

Lecture Notes in Computer Science
LNCS LNAI LNBI

springer.com

Aberer et al. (Eds.)

LNCS 4255

Information Systems – WISE 2006

LNCS 4255

Karl Aberer Zhiyong Peng
Elke A. Rundensteiner Yanchun Zhang
Xuhui Li (Eds.)

Web Information Systems – WISE 2006

7th International Conference on
Web Information Systems Engineering
Wuhan, China, October 2006, Proceedings

Springer

Lecture Notes in Computer Science 4255

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Karl Aberer Zhiyong Peng
Elke A. Rundensteiner Yanchun Zhang
Xuhui Li (Eds.)

Web Information Systems – WISE 2006

7th International Conference on
Web Information Systems Engineering
Wuhan, China, October 23-26, 2006
Proceedings

Springer

Volume Editors

Karl Aberer
EPFL
Switzerland
E-mail: karl.aberer@epfl.ch

Zhiyong Peng
Wuhan University
China
E-mail: peng@whu.edu.cn

Elke A. Rundensteiner
Worcester Polytechnic Institute
USA
E-mail: rundenst@cs.wpi.edu

Yanchun Zhang
Victoria University
Australia
E-mail: yzhang@csm.vu.edu.au

Xuhui Li
Wuhan University
China
E-mail: lixuhui@whu.edu.cn

Library of Congress Control Number: 2006934596

CR Subject Classification (1998): H.4, H.2, H.3, H.5, K.4.4, C.2.4, I.2

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-48105-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-48105-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11912873 06/3142 5 4 3 2 1 0

Preface

WISE 2006 was held in Wuhan, China, during October 23 -26, hosted by Wuhan University. As more and more diverse information becomes available on the Internet and in the Web, novel theoretical and technical approaches to building and improving Web information systems become of vital importance to the further development of information technology. Following the successful conferences from 2000 to 2005, WISE 2006 provided a premium forum for researchers, professionals, and industrial practitioners from around the world to share their rapidly evolving knowledge and to report on new advances in Web information systems.

WISE 2006 received 183 submissions from 20 countries worldwide. From these submissions, the Program Committee selected 37 regular papers and 17 short papers, corresponding to an acceptance ratio of 20% and 9%, respectively. This volume also includes invited keynote papers, given by three leading experts at WISE 2006: M. Tamer Özsu (University of Waterloo), Katsumi Tanaka (Kyoto University) and Lizhu Zhou (Tsinghua University).

The submission and review process worked as follows. Each submission was assigned to three or four Program Committee members for review. During the discussion phase PC members and PC chairs carefully analyzed in particular papers that received divergent evaluations. Based on the review scores and the results of the discussions, the Program Chairs made the final decision.

Three workshops were held in conjunction with WISE 2006. The workshop papers are published in a separate proceedings by Springer in its *Lecture Notes in Computer Science* series (LNCS 4256).

We are grateful to the Program Committee members who helped tremendously in reviewing the large number of submissions, to Yanchun Zhang, Xuhui Li and Qing Li for their great support in the conference organization, to Yurong Chen for setting up the Web site and processing the submissions and registrations. Finally, we would like to thank Wuhan University for their support in organizing the conference.

October 2006

Karl Aberer
Zhiyong Peng
Elke Rundensteiner

Organization

Organization Committee

General Co-chairs

Yanxiang He, Wuhan University, China
M. Tamer Özsu, University of Waterloo, Canada

Program Committee Co-chairs

Karl Aberer, EPFL, Switzerland
Zhiyong Peng, Wuhan University, China
Elke Rundensteiner, WPI, USA

Workshop Co-chairs

Ling Feng, University of Twente, Netherlands
Guoren Wang, Northeastern University, China

Publicity Co-chairs

Qing Li, City University of Hong Kong, China
Xiaoyong Du, Reming University, China

Publication Co-chairs

Yanchun Zhang, Victoria University, Australia
Xuhui Li, Wuhan University, China

WISE Society Representive

Xiaohua Jia, City University of Hong Kong, China

Local Organization Committee

Shi Ying, Wuhan University, China
Youyu Gong, Wuhan University, China
Chuanhe Huang, Wuhan University, China

Program Committee

Bernd Amann, France
Periklis Andritsos, Italy
Budak Arpinar, USA
Bharat K. Bhargava, USA

Alex Borgida, USA
Wojciech Cellary, Poland
Shermann S. M. Chan, Japan
Edward Chang, USA
Enhong Chen, China
Li Chen, USA
Songting Chen, USA
Vassilis Christophides, Greece
Kajal Claypool, USA
Alfredo Cuzzocrea, Italy
Ernesto Damiani, Italy
Alex Delis, USA
Oscar Díaz, Spain
Gill Dobbie, New Zealand
David Embley, USA
Piero Fraternali, Italy
Avi Gal, Israel
Dimitrios Georgakopoulos, USA
Dina Goldin, USA
Vivekanand Gopalkrishnan, Singapore
Peter Haase, Germany
Mohand-Said Hacid, France
Wook-Shin Han, Korea
Yanbo Han, China
Christian Huemer, Austria
Mizuho Iwaihara, Japan
H.V. Jagadish, USA
Qun Jin, Japan
Kamal Karlapalem, India
Hiroyuki Kitagawa, Japan
Nick Koudas, Canada
Harumi Kuno, USA
Herman Lam, USA
Wang-Chien Lee, USA
Shijun Li, China
XueMin Lin, Australia
Tok Wang Ling, Singapore
Chengfei Liu , Australia
Sanjay Madria, USA
Murali Mani, USA
Massimo Mecella, Italy
Carlo Meghini, Italy
Xiaofeng Meng, China
Mukesh K. Mohania, India
Wolfgang Nejdl, Germany
Anne Hee Hiong Ngu, USA
Zaiqing Nie, China
Moira C. Norrie, Switzerland
Tekin Ozsoyoglu, USA
Jignesh M. Patel, USA
Alexandra Poulovassilis, UK
Cartic Ramakrishnan, USA
Michael Rys, USA
Monica Scannapieco, Italy
Klaus-Dieter Schewe, New Zealand
Heiko Schuldt, Switzerland
Mark Segal, USA
Tony Shan, USA
Timothy K. Shih, Taiwan
Steffen Staab, Germany
Gerd Stumme, Germany
Hong Su, USA
Jianwen Su, USA
Aixin Sun, Singapore
Keishi Tajima, Japan
Hideyuki Takada, Japan
Olga de Troyer, Belgium
Michalis Vazirgiannis, Greece
Jari Veijalainen, Finland
Yannis Velegrakis, USA
Andreas Wombacher, Netherlands
Yuhong Xiong, China
Jian Yang, Australia
Masatoshi Yoshikawa, Japan
Ge Yu, China
Jeffrey Yu, China
Philip Yu, USA
Donghui Zhang, USA
Aoying Zhou, China
Xiaofang Zhou, Australia

Table of Contents

Keynote Papers

Session 1: Web Search

Session 2: Web Retrieval

Session 3: Web Workflows

Session 4: Web Services

Session 5: Web Mining

Session 6: Performant Web Systems

Session 7: Web Information Systems

Session 8: Web Document Analysis

Session 9: Quality, Security and Trust

Session 10: Semantic Web and Integration

Session 11: XML Query Processing

Session 12: Multimedia and User Interface

Internet-Scale Data Distribution: Some Research Problems

M. Tamer Özsu

David R. Cheriton School of Computer Science, University of Waterloo
tozsu@uwaterloo.ca

Abstract. The increasing growths of the Internet, the Web and mobile environments have had a push-pull effect. On the one hand, these developments have increased the variety and scale of distributed applications. The data management requirements of these applications are considerably different from those applications for which most of the existing distributed data management technology was developed. On the other hand, they represent significant changes in the underlying technologies on which these systems are built. There has been considerable attention to this issue in the last decade and there are important progress on a number of fronts. However, there are remaining issues that require attention. In this talk, I will review some of the developments, some areas in which there has been progress, and highlight some of the issues that still require attention.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, p. 1, 2006.

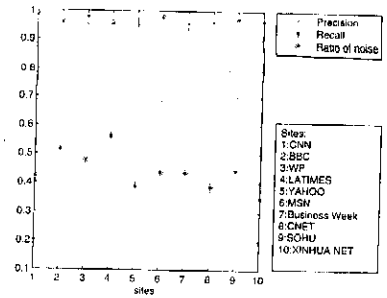

Fig. 2. Block level Precision, Recall values and Ratio of Noise

5 Conclusion

We devised a simple but powerful approach to identify primary portions within Web pages. The VIPS algorithm is adopted to partition a Web page into multiple semantic blocks. Shannon's information entropy is adopted to represent each feature's ability for differentiating. And the weighted Naïve Bayes classifier is presented to eliminate the redundant blocks.

The approach is tested with several important English and Chinese news sites and achieved precise results. Evidently, it can be applied to general Web IR systems to reduce the size of index and increase the precision of retrieval.

References

1. Lin, S.-H. and Ho, J.-M.: Discovering Informative Content Blocks from Web Documents, in the proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD'02).(2002)
2. Sandip Debnath, Prasenjit Mitra, C. Lee Giles.: Automatic Extraction of Informative Blocks from Webpages.SAC'05 Santa Fe, New Mexico, USA. March 1317 (2005)
3. Gupta,S., Kaiser,G., Neistadt,D. and Grimm,P.: DOM based Content Extraction of HTML Documents, in the proceedings of the 12th World Wide Web conference (WWW 2003), Budapest, Hungary. May (2003).
4. Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma.: Learning Block Importance Models for Web Pages. WWW 2004, New York, USA. May 17-22 (2004)
5. Zhang Zhigang,Chen Jing,Li Xiaoming.: An Approach to Reduce Noise in HTML Pages. JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION . April 23, (2004)
6. Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y., VIPS: a vision-based page segmentation algorithm, Microsoft Technical Report. MSR-TR-2003-79, (2003)
7. C. E. Shannon. A mathematical theory of communication.Bell System Technical Journal, 27:398-403.(1948)

Defining a Data Quality Model for Web Portals

Angélica Caro[1], Coral Calero[2], Ismael Caballero[2], and Mario Piattini[2]

[1] Department of Auditoria e Informática, University of Bio Bio
La Castilla s/n, Chillán, Chile
ancaro@inf-cr.uclm.es
[2] Alarcos Research Group, Information Systems and Technologies Department
UCLM-SOLUZIONA Research and Development Institute.
University of Castilla-La Mancha
{Coral.Calero, Ismael.Caballero, Mario.Piattini}@uclm.es

Abstract. Advances in technology and the use of the Internet have favoured the appearance of a great variety of Web applications, among them Web Portals. These applications are important information sources and/or means of accessing information. Many people need to obtain information by means of these applications and they need to ensure that this information is suitable for the use they want to give it. In other words, they need to assess the quality of the data.

In recent years, several research projects were conducted on topic of Web Data Quality. However, there is still a lack of specific proposals for the data quality of portals. In this paper we introduce a model for the data quality in Web portals (PDQM). PDQM has been built upon the foundation of three key aspects: a set of Web data quality attributes identified in the literature in this area, data quality expectations of data consumers on the Internet, and the functionalities that a Web portal may offer to its users.

Keywords: Data Quality, Information Quality, Web Portal, Quality Model.

1 Introduction

Over the past decade the number of organizations which owns Web portals grows dramatically. Their have established portals to complement, substitute or widen existing services to their clients. In general, portals provide users with access to different data sources (providers) [19], as well as to on-line information and information-related services [31]. Also they create a working environment that users can easily navigate in order to find the information they specifically need to quickly perform their operational or strategic functions and make decisions [7]. It will be up to the owners of a Web portal to obtain and maintain a high level in the quality of information [17].

In the literature, the concept of Information or Data Quality is often defined as "fitness for use", i.e., the ability of a data collection to meet user requirements [5],[27]. Besides, the terms "data" and "information" are often used as synonyms [28]. In this work we will use them as synonymous.

Research on data quality (DQ) began in the context of information systems [18],[27] and it has been extended to contexts such as cooperative systems [9],[20],[30], data warehouses [2],[32] or e-commerce [1],[13], amongst others. Due

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, pp. 363 – 374, 2006.

to the particular characteristics of Web applications and their differences from the traditional information systems [31], the research community started to deal with the subject of DQ on the Web [11].

However there are no works on DQ that address the particular context of Web portals [6], in spite of the fact that some works highlight the DQ as one of the relevant factors in the quality of a portal [22],[25]. Likewise, except for few works in the DQ area, like [4],[5],[29], most of them have looked at quality from the data producers or data custodians perspective and not from the data consumers perspective [4]. The last perspective differs from the two others in two important aspects [4]: (1) data consumer has no control over the quality of available data and (2) the aim of consumers is to find data that match their personal needs, rather than provide data that meet the needs of others.

In this paper, we present a portal data quality model (PDQM), focused on the data consumer's perspective. As key pieces in the construction of our model we have taken (1) a set of Web DQ attributes identified in the literature, (2) the DQ expectations of data consumers on the Internet, described by Redman in [26], and (3) the functionalities that a portal Web may offer its users [7].

To produce the PDQM model, we defined the process shown in Fig.1. During the first phase, we have recompiled Web DQ attributes from the literature, which we believe should therefore be applicable to Web portals. In the second phase we have built a matrix for the classification of the attributes obtained in previous phase. This matrix reflects two basic aspects considered in our model: the data consumer perspective and the basic functionalities which a data consumer uses to interact with a Web portal.

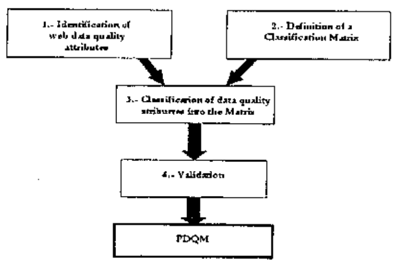

Fig. 1. Phases in the development of the PDQM

In the third phase we used the obtained matrix, to analyse the applicability of each Web DQ attribute in a Web portal. Finally, in the fourth phase, we must validate PDQM. The remainder of this paper is organized as follows. In section 2, we describe the phase of identification of attributes of Web DQ and show a summary of the results obtained (phase 1 of Fig.1). Section 3 is dedicated to the description of the phase of classification matrix construction (phase 2 of Fig.1). We set out the result of classification of attributes on the matrix in section 4 (phase 3 of Fig.1). Section 5 shows how we aim to validate our model (phase 4 of Fig.1). In section 6 we present the conclusions and the future work.

2 Identification of Attributes of Web Data Quality

The first stage in the development of our model consisted of a systematic review of the relevant literature [15]. From this task we aimed to identify data quality attributes which have been proposed for different domains in the Web context (Web sites [8], 14],[23], data integration [3],[24], e-commerce [13], Web information portals [25], cooperative e-services [9], decision making [11], organizational networks [21] and DQ on the Web [10]). The idea was to take advantage of work already carried out in the Web context and apply it to Web portals.

In this review we studied 55 works, published between 1995 and 2004. From the studied work we selected the projects in which DQ attributes applicable to the Web context were proposed. We thus obtained a total of 100 Web DQ attributes. We wanted to reduce this number, having also detected certain synonymous amongst the attributes identified. Those attributes were combined and also the ones which had similar name and meaning, obtaining a final set of 41 attributes. Table 1 shows these attributes, pointing out for each of these the work where they were put forward, as well as the total number of pieces of work where they can be found referred to. In addition, the symbols × and ⊗ were used to represent how they were combined (× indicates the same name and meaning and ⊗ marks the fact that only the meaning is the same).

Table 1. Web Data Quality Attributes 1-41

Author	Year	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise Representation	Consistent Representation	Cost Effectiveness	Customer Support	Currency	Documentation	Duplicates	Ease of operation	Expiration	Flexibility	Granularity	Interactive	Internal Consistency
Naumann and Rolker	2000		×	×			×	×	×	×	×		×		×							
Katerattanakul and Siau	1999, 2001	×	⊗			×																
Eppler	2001	×	×		×				×	⊗	⊗			×			⊗				×	⊗
Fugini et al	2002		⊗		×			⊗	×													×
Pernici and Scannapieco	2002		×						×									×				
Graefe	2003	⊗					⊗	×														
Bouzeghoub and Peralta	2004													×								
Gertz	2004								×					×		×				×		
Melkas	2004	×	×	⊗				×	×	×	×	×					×		×			
Moustakis	2004				⊗			⊗														
Yang et al.	2004		×					⊗	×					×								
Number of references		4	7	2	3	1	2	6	7	3	3	1	1	4	1	1	2	1	1	1	1	2

Author	Interpretability	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's Information	Timeliness	Traceability	Understandability	Validity	Value-added	Number of Attributes
Naumann and Rolker	×	×			×			×	×	×	×	×	×			×	⊗	×		×	22
Katerattanakul and Siau							×								×			⊗			6
Eppler			×									⊗	×			×	×	⊗			16
Fugini et al									⊗				⊗			×					8
Pernici and Scannapieco										⊗											4
Graefe	×			×					⊗										⊗	⊗	8
Bouzeghoub and Peralta																×					2
Gertz						×															5
Melkas	×				×				×		×	⊗	×			×	×	⊗		×	20
Moustakis									⊗					×							4
Yang et al.									⊗												5
Number of references	3	1	1	1	2	1	1	1	6	2	2	3	4	1	1	5	3	4	1	3	

3 Matrix Classification

In the second step we defined a matrix which would allow us to perform a preliminary analysis of how applicable these attributes are to the domain of Web portals. The matrix was defined based on the relationship that exists between: (1) The functionalities of a Web portal, identified in [7]: Data Points and Integration, Taxonomy, Search Capabilities, Help Features, Content Management, Processes and Actions, Communication and Collaboration, Personalization, Presentation, Administration and Security; and (2) The data quality expectations of Internet consumers as stated in [26]: Privacy, Content, Quality of Values, Presentation, Improvement and Commitment

On this matrix we carried out an analysis of what expectations were applicable to each of the different functionalities that a portal offers to a data consumer represented in Fig.2 with a "√" mark.

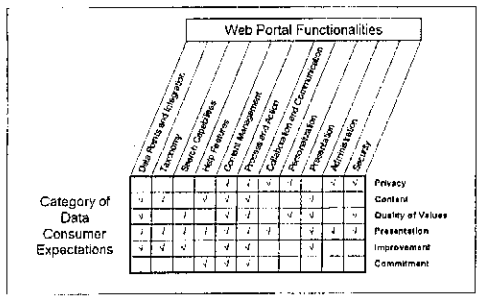

Fig. 2. Matrix for the classification of attributes of Web data quality

In the following paragraphs we explain each relationship (functionality, expectation):

- *Data Points and Integration.* They provide the ability to access information from a wide range of internal and external information sources and display the resulting information at the single point-of-access desktop. The expectations applied to this functionality are: *Content* (consumers need a description of portal areas covered, use of published data, etc.), *Quality of value* (data consumer should expect the result of searches to be correct, up-to-date and complete), *Presentation* (formats, language, and other aspects are very important for easy interpretation) and *Improvement* (users want to participate with their opinions in the portal improvements and to know what the results of applying them are).

- *Taxonomy.* It provides information context (including the organization-specific categories that reflect and support the organization's business). The expectations applied to this functionality are: *Content* (consumers need a description of which data are published and how they should be used, as well as easy-to-understand definitions of every important term, etc.), *Presentation* (formats and language in the taxonomy are very important for easy interpretation; users should expect to find instructions when reading the data), and *Improvement* (the user should expect to convey his/her comments on data in the taxonomy and know what the result of improvements are).

- *Search Capabilities.* This provides several services for Web portal users and needs to support searches across the company, the World Wide Web, and in search engine catalogs and indexes. The expectations applied to this functionality are: *Quality of values* (the data consumer should expect the result of searches to be correct, up-to-date and complete), *Presentation* (formats and language are important for consumers, both for searching and for easy interpretation of results) and *Improvement* (the consumer should expect to convey his/her comments on data in the taxonomy and be aware of the result of improvements).

- *Help Features.* These provide help when using the Web portal. The expectations applied to this functionality are: *Presentation* (formats, language, and other aspects are very important for easy interpretation of help texts) and *Commitment* (the consumer should be able to ask and obtain answers to any question regarding the proper use or meaning of data, update schedules, etc, easily.).

- *Content Management.* This function supports content creation, authorization, and inclusion in (or exclusion from) Web portal collections. The expectations applied to this functionality are: *Privacy* (a privacy policy for all consumers to manage, to access sources and to guarantee Web portals data should exist), *Content* (consumers need a description of data collections), *Quality of values* (a consumer should expect all data values to be correct, up-to-date and complete), *Presentation* (formats and language should be appropriate for easy interpretation), *Improvement* (the consumer should expect to convey his/her comments on contents and their management and be aware of the result of the improvements) and *Commitment* (the consumer should be able to ask any question regarding the proper use or meaning of data, etc.,).

- *Process and Action.* This function enables the Web portal user to initiate and participate in a business process of a portal owner. The expectations for this functionality are: *Privacy* (the data consumer should expect there to be a privacy policy to manage the data about the business on the portal), *Content* (consumers should expect to find descriptions about the data published for the processes and actions, their uses, etc.), *Quality of values* (that all data associated to this function are correct, up-to-date and complete), *Presentation* (formats and other aspects are very important in order to interpret data), *Improvement* (the consumer should expect to convey their comments on contents and their management and to know the improvements) and *Commitment* (the consumer should be able to ask and obtain an answer to any question).

- *Collaboration and Communication.* This function facilitates discussion, locating innovative ideas, and recognizing resourceful solutions. The expectations for this functionality are: *Privacy* (the consumer should expect a privacy policy for all consumers that participate in activities of this function), and *Commitment* (a consumer should be able to ask and have any question answered regarding the proper use or meaning of data for the collaboration and/or communication, etc.).

- *Personalization.* This is a critical component in creating a working environment that is organized and configured specifically for each user. The expectations applied to this functionality are: *Privacy* (the consumer should expect privacy and security as regards their personalized data, profile, etc.), and *Quality of values* (data about the user profile should be correct and up-to-date).

- *Presentation.* It provides both the knowledge desktop and the visual experience to the Web portal user that encapsulates all of the portal's functionality. The expectations for this functionality are: *Content* (the presentation of a Web portal should include data about areas covered, appropriate and inappropriate uses, definitions, etc.), *Quality of values* (the data of this function should be correct, up-to-date and complete.), *Presentation* (formats, language, and other aspects are very important for the easy interpretation and appropriate use of data.) and *Improvement* (the consumer should expect to convey their comments on contents and their management).

- *Administration*. This function provides a service for deploying maintenance activities or tasks associated with the Web portal system. The expectations for this functionality are: *Privacy* (Data consumers need security for data about the portal administration) and *Quality of values* (Data about tasks or activities of administration should be correct and complete).

- *Security*. This provides a description of the levels of access that each user (groups of users) are allowed for each portal application and software function included in the Web portal. The expectations for this functionality are: *Privacy* (the consumer needs a privacy policy regarding the data of the levels of access.), *Quality of values* (data about the levels of access should be correct and up-to-date) and *Presentation* (data about security should be in a format and language for easy interpretation).

4 Classification

The third phase of the development process in the PDQM model (see Figure 1), consisted in classifying the Web data quality attributes (shown in section 2) in each of the relationships (functionality, expectation) established on the classification matrix created in stage 2 (and presented in section 3). In this paper we do not show the attributes applicable to each relationship (functionality, expectation), we just set out a summary of the attributes applicable to each portal functionality (Table 2).

Table 2. Data quality attributes for functionality

Functionalities	Accessibility	Accuracy	Amount of data	Applicability	Attractiveness	Availability	Believability	Completeness	Concise Representation	Consistent Representation	Cost effectiveness	Customer support	Currency	Documentation	Duplicates	Ease of operation	Explication	Flexibility	Granularity	Interactive	Internal consistency	Integrity	Latency	Maintainable	Novelty	Objectivity	Ontology	Organization	Price	Relevancy	Reliability	Reputation	Response time	Security	Specialization	Source's information	Timeliness	Traceability	Understandability	Validity	Value-added	Total of Attributes
Data Points and Integration	✓	✓	✓				✓		✓			✓	✓				✓								✓					✓	✓	✓					✓	✓	✓		11
Taxonomy	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓													✓	✓	✓					✓	✓	✓		12
Search Capabilities	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓													✓	✓	✓					✓	✓	✓		12
Help Features	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Content Management	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Process and Action	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Collaboration and Communication	✓																✓	✓																			✓	✓	✓		6
Personalization	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Presentation	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Administration	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Security	✓	✓	✓	✓			✓		✓			✓	✓			✓	✓								✓					✓	✓	✓					✓	✓	✓		13
Number of References	7	4	9	2	1	3	6	5	9	1	0	8	6	1	1	8	4	1	0	0	0	5	0	0	3	2	0	1	0	7	7	2	0	5	3	1	0	7	1	1	8	

As can be seen in Table 2 there are some quality attributes which were not classified in the matrix. This is basically due to the fact that these are not able to be assessed by the data consumers, for example- Ontology and Latency.

5 Validation

Until this point, in the production of the PDQM model we have established relationships between the functionalities of a web portal and the expectations of the Internet users. On the basis of these relationships we have identified, intuitively, Web DQ attributes which could be applicable in a portal obtaining the first version of the PDQM model. The next phase consists in the validation of our model (see Fig.1).

In this process we decided to carry out a study by means of a survey. In the survey, consumers of web portal data would be asked to give their opinion about what aspects they think are important when assessing the quality of the data they get from a portal.

We decided to draw up independent questionnaires for each one of the functionalities because we thought that to use only a questionnaire for the whole model would be tiring for the subjects. In this paper we will describe the validation tasks carried out on a single portal functionality.

The first functionality we chose was Data Points and Integration. This decision was made on the basis that this functionality is the first one that the consumer comes across on entering a Web portal and also because this aspect will help the portal user to assess whether the data which the portal offers match their expectations or not.

5.1 Definition of the Objectives of the Survey

The first activity carried out corresponds to the definition of the objectives of the survey, since these are directly related to the information that will be gathered [16]. For the chosen functionality these are:

1. To obtain the opinion of the consumers of a Web portal with respect to how they characterizes the quality of the data obtained in a portal through the functionality of Data Points and Integration.
2. To ascertain the importance given by the data consumers to each one of the DQ attributes identified in each relationship (Data Points and Integration, expectation).
3. To identify other aspects, in particular, Data Points and Integration which are important for consumers but which have not been considered in our model.

5.2 Choosing of the Subjects

With the above objectives in mind, the target population of our survey was defined as the whole set of Web portal data consumers. As to work with the whole population was not feasible, we used just a representative sample of it. That is, people who usually use portals to get data. To obtain the sample we used the snowball sampling method.

5.3 Preparation of the Questionnaire

In this section, we explain the stages used for preparing the questionnaire (the instruments is in http://FreeOnlineSurveys.com/rendersurvey.asp?id=140254).

1. **Search in Literature.** In the pertinent literature we looked for studies in which surveys had taken place in the validation of data quality attributes. Amongst those found we could mention [8],[23],[29].
2. **Construction of the instrument (or use of an existing one).** As there are advantages in using an already-proven instrument [16], we have built a new questionnaire based on the one that has been created in [29]. The questionnaire was constructed with an automatic tool and was posted on the URL given by the tool's provider.

3. **Choice of questions.** Basically, to produce the questions, we have considered the purpose and goals of the survey [16]. We therefore chose 2 questions for the first objective, one (a compound one) for the second and another for the third. All this was done taking into account factors such as the level of understanding of those being interviewed, an appropriate number of questions and the standardization of reply formats. A set of demographic questions were included too. Apart from all this, we included a question where we asked the subjects their opinion about the survey itself. With this we pretended to pick up on any defect or problem as far as the questionnaire is concerned.
4. **Making the questions.** We tried to create questions which have evident sense and which are specific [16]. The language used is conventional, expressing simple ideas. Negative questions are not included.
5. **Type of questions.** Our instrument mixes open and closed questions. In the first question, an open question, we asked the respondents about the attributes they considered important for a data point (the functionality under study). In the second question, open too, we asked about the attributes which were important for the data obtained through a data point. In the third question, another open question, we showed the 15 attributes classified for the functionality and we asked the respondents about other attributes that they consider need to be included. In the fourth question, a closed question, the subjects had to give a value (among 1 and 7) to the importance given for each one of the 15 attributes classified for the functionality. After this, we asked, with 6 closed questions, demographic data. Finally, in the last question, an open question, we asked the subject's opinion about the survey.
6. **Format of the questionnaire.** Our questionnaire was self-administered (it will be put on the Web and the user would be able to access to it when and where he/she decides). In order to check the format and instructions of the questionnaire we have used a check list proposed, to this purpose, in [16].
7. **Evaluation of the survey instrument.** In spite of the fact that we use a previously -proven model for our survey, we have carried out a pilot study whose aim was to ensure that the survey was a reliable instrument. For the pilot study we asked a group of data consumers, via e-mail, to take part. Twenty participants were notified in this way, of whom 15 replied. We thus reached a response rate of 75%. Among those who replied, and after analysing the observations given, a new version of the questionnaire was produced. The changes consisted, basically, in adjusting the presentation format, so that in each one, key aspects would be clearly highlighted.

5.4 Application and Results of the survey

The sample on which we applied the survey was obtained by means of personal contacts. The questionnaire was posted on the Web and each person included in the sample (150 approximately) was sent an e-mail. The e-mail contained an invitation to take part in the survey (at this point we explained its purpose, as well as the importance of the cooperation of each person). We also asked them to send the e-mail to their contacts in order to obtain more answers to our survey. The Web address of the questionnaire was provided at the same time.

A total of 91 responses were obtained within two weeks. After data screening, we eliminated 32 incomplete and repeated questionnaires. As a result, the total effective sample was 69 subjects (46% of the subjects contacted).

With the demographic question we obtained the respondents' profile. 43% of the respondents were female and 57% were male; 58% were between 36 and 45 years old; 97% used the Web more than once a day; 94% used Web portals; and 97% had more than 3 years of experience using the Web.

For the Data Points and Integration functionality, in our classification we considered 15 data quality attributes (see Table 2). The results of the first open question showed that the most mentioned attributes were: Accessibility, Understandability, Currency and Consistent representation (all considered in our model for the functionality under study), Organization, Source's Information and Response time (all considered in our model but not for this functionality). All these attributes were mentioned by more than 10% of the participants. The results of the second open question showed the following attributes as the most mentioned: Accuracy, Relevance (considered in our model for this functionality), Organization, Source's Information and Response time (considered in our model but not for this functionality). All these attributes were mentioned by more than 10% of the participants.

In the third open question we showed all the attributes considered in our model and we asked the data consumer for other attributes that they consider necessary. As a result, the most-proposed attribute was Attractiveness with 22%, Security with 12% and Source's Information, Response time and Ease of Operation with 10%. All were considered in our model but not classified within this functionality.

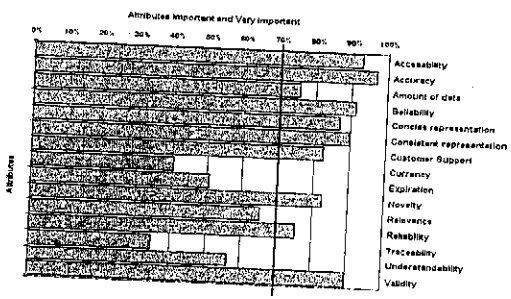

Fig. 3. Importance assigned by respondents to the data quality attributes proposed

Lastly, in the fourth question (the closed one), the participants had to assign a degree of importance between levels 1 and 7 (1 not important and 7 very important) to each attribute. The attributes that had at least 70% of preferences (adding the percentages of level 6 and 7) will be considered as important for the data consumer. Among the fifteen attributes, Accessibility, Currency, Amount of data, Reliability, Believability, Understandability, Accuracy, Relevance, Consistent representation and Validity appeared to be relevant (see Fig.3). This result coincided with our initial

classification of DQ attributes for "Data point and Integration" functionality, in at least 66 %.

With the results of this validation we are able to fine-tune our model. This means, for the functionality being studied, looking at only the ten attributes that have been judged most important by the respondents (adding some particular attribute proposed by them). We do believe, however, that it would be better to await the result of the survey on all the functionalities, and then fully adjust our model to obtain a definitive PDQM model.

Finally, although we have results just for one portal functionality considered in our model, we believe that allows us to draw some conclusions. One of them is that our model, in regards to the functionality Data points and Integration, is very close to the data consumer perspective. We affirm this because our assignation of DQ attributes to evaluate this functionality coincides to a very large degree with their responses. Furthermore, we consider that the results of this survey demonstrated that using the surveys to validate our model is appropriate. Thus we can also best-fit our model to the data consumer perspective.

6 Conclusions and the Future Work

The Web portals are applications that have been positioned like information sources and/or as means of accessing information over the last decade. The other side to this is that those who need information by means of these portals need to be sure, somehow, that this information is indeed suitable for the use they wish. In other words, they really need to assess the level of the quality of the data obtained.

In the literature we have studied, there are no specific proposals for data quality models for Web portals. In this article we have presented a preliminary version of a data quality model for Web portals (PDQM) that consider the consumers point of view. This has been built on three fundamental aspects: a set of Web data quality attributes set out in the relevant literature, data quality expectations of data consumers on the Internet, and the functionalities which a Web portal may offer its users.

As a first step in the validation of the model, we carried out a survey to validate the "Data point and Integration" functionality. We started with this functionality because we believe it is the most important one from the data consumer perspective. In order to achieve a complete validation of our PDQM model, we will apply the questionnaire in full, i.e. with the 11 questionnaires created (one per functionality). The population sample chosen for this phase corresponds to the users of the portal of Castilla-La Mancha (www.castillalamancha.es).

We are very aware that giving 11 questionnaires to a person might turn out to be really tiring and none- too- motivating, so just three of the 11 questionnaires will be given to each person undergoing this survey. The said three will be chosen at random. Besides that, a Web application will be installed in the portal we have selected and this will manage the free distribution of the survey to the users who log on to the portal. The application will be made up of three modules: an administrative one (by which we can administer the survey and the application), an analyzing module (which will show the results: statistics, diagrams, response rate, etc) and a data collection module (which creates the questionnaires for users from a random choice of 3 questionnaires from the possible 11 and which will be in charge of storing the answers. At the end of the time allowed for the application, when an acceptable level of response has been reached, we will analyse the results obtaining the final version of PDQM.

As future work, once we obtain a final version of PDQM, we will use it as basis to de-fine a framework for evaluating and improving the DQ for Web portals. Our aim is to offer the different consumers of Web portal data the possibility of evaluating, accor-ding to their needs and criteria, the DQ they receive from the Web portal.

Acknowledgments. This research is part of the following projects: CALIPO (TIC2003-07804-C05-03), CALIPSO (TIN20005-24055-E) supported by the Ministerio de Educación y Ciencia (Spain) and COMPETISOFT (506PI0287) financed by CYTED.

References

1. Aboelmeged, M., A Soft System Perspective on Information Quality in Electronic Commerce, In Proceedings of the Fifth Conference on Information Quality, 2000.
2. Bouzeghoub, M. and Kedad, Z., Quality in Data Warehousing, in Information and Database Quality, Piattini, M., Calero, C., and Genero, M., Eds.: Kluwer Academic Publishers, 2001.
3. Bouzeghoub, M. and Peralta, V., A Framework for Analysis of data Freshness, In International Workshop on Information Quality in Information Systems, (IQIS2004), Paris, France, 2004.
4. Burgess, M., Fiddian, N., and Gray, W., Quality Measures and The Information Consumer, In Proceedings of the 9th International Conference on Information Quality, 2004.
5. Cappiello, C., Francalanci, C., and Pernici, B., Data quality assessment from the user's perspective, In International Workshop on Information Quality in Information Systems, (IQIS2004), Paris, Francia, 2004.
6. Caro , A., Calero, C., Caballero, I., and Piattini, M., Data quality in web applications: A state of the art, In IADIS International Conference WWW/Internet 2005, Lisboa-Portugal, 2005.
7. Collins, H., Corporate Portal Definition and Features: AMACOM, 2001.
8. Eppler, M., Algesheimer, R., and Dimpfel, M., Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework, In Proceeding of the Eighth International Conference on Information Quality, 2003.
9. Fugini, M., Mecella, M., Plebani, P., Pernici, B., and Scannapieco, M., Data Quality in Cooperative Web Information Systems, 2002.
10. Gertz, M., Ozsu, T., Saake, G., and Sattler, K.-U., Report on the Dagstuhl Seminar "Data Quality on the Web", SIGMOD Record, vol. 33, Nº 1, pp. 127-132, 2004.
11. Graefe, G., Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality, In Proceeding of the Eighth International Conference on Information Quality, 2003.
12. Huang, K.-T., Lee, Y., and Wang, R., Quality information and knowledge: Prentice Hall PTR, 1999.

13. Katerattanakul, P. and Siau, K., Information quality in internet commerce desing. In Information and Database Quality, Piattini, M., Calero, C., and Genero, M., Eds.: Kluwer Academic Publishers, 2001.
14. Katerattanakul, P. and Siau, K., Measuring Information Quality of Web Sites: Development of an Instrument, In Proceeding of the 20th International Conference on Information System, 1999.
15. Kitchenham, B., Procedures for Performing Systematic Reviews, 0400011T.1, 2004.
16. Kitchenham, B. and Pfleeger , S. L., Principles of survey research: part 3: constructing a survey instrument, SIGSOFT Softw. Eng. Notes, ACM Press, 27, pp. 20-24, 2002.
17. Kopcso, D., Pipino, L., and Rybolt, W., The Assesment of Web Site Quality. In Proceeding of the Fifth International Conference on Information Quality, 2000.
18. Lee, Y., AIMQ: a methodology for information quality assessment, Information and Management. Elsevier Science, pp. 133-146, 2002.
19. Mahdavi, M., Shepherd, J., and Benatallah, B., A Collaborative Approach for Caching Dynamic Data in Portal Applications. In Proceedings of the fifteenth conference on Australian database, 2004.
20. Marchetti, C., Mecella, M., Scannapieco, M., and Virgillito, A., Enabling Data Quality Notification in Cooperative Information Systems through a Web-service based Architecture, In Proceeding of the Fourth International Conference on Web Information Systems Engineering, 2003.
21. Melkas, H., Analyzing Information Quality in Virtual service Networks with Qualitative Interview Data, In Proceeding of the Ninth International Conference on Information Quality, 2004.
22. Moraga, M. Á., Calero, C., and Piattini, M., Comparing different quality models for portals. (2006). To appear on Online Information Review.
23. Moustakis, V., Litos, C., Dalivigas, A., and Tsironis, L., Website Quality Assesment Criteria, In Proceeding of the Ninth International Conference on Information Quality, 2004.
24. Naumann, F. and Rolker, C., Assesment Methods for Information Quality Criteria, In Proceeding of the Fifth International Conference on Information Quality, 2000.
25. Pressman, R., Software Engineering: a Practitioner's Approach, Fifth ed: McGraw-Hill, 2001.
26. Redman, T., Data Quality: The field guide. Boston: Digital Press, 2000.
27. Strong, D., Lee, Y., and Wang, R., Data Quality in Context, Communications of the ACM, Vol. 40, N° 5, pp. 103 -110, 1997.
28. Wang, R., A Product Perspective on Total data Quality Management, Communications of the ACM, Vol. 41, N° 2, pp. 54-65, 1998.
29. Wang, R. and Strong, D., Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems: Armonk: Spring 1996, 12, pp. 5-33, 1996.
30. Winkler, W., Methods for evaluating and creating data quality, Information Systems, N° 29, pp. 531-550, 2004.
31. Yang, Z. a. C., S. and Zhou, Z. and Zhou, N., Development and validation of an instrument to measure user perceived service quality of information presenting Web portals, Information and Management, Elsevier Science, 42, pp. 575-589, 2004.
32. Zhu, Y. and Buchmann, A., Evaluating and Selecting Web Sources as external Information Resources of a Data Warehouse, In Proceeding of the 3rd International Conference on Web Information Systems Engineering, 2002.

Finding Reliable Recommendations for Trust Model

Weiwei Yuan, Donghai Guan, Sungyoung Lee, Youngkoo Lee*, and Andrey Gavrilov

Department of Computer Engineering, Kyung Hee University, Korea
{weiwei, donghai, sylee, avg}@oslab.khu.ac.kr, yklee@khu.ac.kr

Abstract. This paper presents a novel context-based approach to find reliable recommendations for trust model in ubiquitous environments. Context is used in our approach to analyze the user's activity, state and intention. Incremental learning based neural network is used to dispose the context in order to detect doubtful recommendations. This approach has distinct advantages when dealing with randomly given irresponsible recommendations, individual unfair recommendations as well as unfair recommendations flooding regardless of from recommenders who always give malicious recommendations or "inside job" (recommenders who acted honest previous suddenly give unfair recommendations), which is lack of consideration in the previous works. The incremental learning based neural network used in our approach also enables to filter out the unfair recommendations with limited information about the recommenders. Our simulation results show that our approach can effectively find reliable recommendations in different scenarios and a comparison is also given between previous works and our method.

1 Introduction

Computational models of trust have been proposed for ubiquitous environments because they are capable of deciding on the runtime whether to provide services to requesters who are either unfamiliar with service providers or do not have enough access rights to certain services. The basis for the trust model to make decision for unfamiliar service requesters are the recommendations given by recommenders who have past interaction history with the requesters. However, in the large-scale, open, dynamic and distributed ubiquitous environments, there may possibly exist numerous self-interested recommenders who give unfair recommendations to maximize their own gains (perhaps at the cost of others). Since recommendations given by recommenders are the key point for the trust model to make decision, finding ways to avoid or reduce the influence of unfair positive or negative recommendations from self-interested recommenders is a fundamental problem for trust model in ubiquitous environments. At the same time, because of the highly dynamic nature of ubiquitous environments, it is not always easy to get enough information about the recommenders. Hence the trust model is required to find the reliable recommendations with limited information about the recommenders.

* Corresponding author.

K. Aberer et al. (Eds.): WISE 2006, LNCS 4255, pp. 375 – 386, 2006.