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

ISBN 978-3-540-76992-7

9 783540 769927

Lecture Notes in Computer Science
LNCS LNAI LNBI

springer.com

Benatallah et al. (Eds.)

LNCS 4831

Web Information Systems Engineering – WISE 2007

WISE

LNCS 4831

Boualem Benatallah Fabio Casati
Dimitrios Georgakopoulos Claudio Bartolini
Wasim Sadiq Claude Godart (Eds.)

Web Information Systems Engineering – WISE 2007

8th International Conference on Web Information Systems Engineering Nancy, France, December 2007, Proceedings

Springer

45

Lecture Notes in Computer Science 4831

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

Boualem Benatallah Fabio Casati
Dimitrios Georgakopoulos
Claudio Bartolini Wasim Sadiq
Claude Godart (Eds.)

Web Information Systems Engineering – WISE 2007

8th International Conference on
Web Information Systems Engineering
Nancy, France, December 3-7, 2007
Proceedings

Springer

Volume Editors

Boualem Benatallah
The University of New South Wales, Sydney, NSW 2052, Australia
E-mail: boualem@cse.unsw.edu.au

Fabio Casati
University of Trento, 38050 Trento, Italy
E-mail: casati@dit.unitn.it

Dimitrios Georgakopoulos
Telcordia, Austin Research Center, Austin, TX 78701, USA
E-mail: dimitris@research.telcordia.com

Claudio Bartolini

HP Laboratories, Palo Alto, CA 94304, USA
E-mail: claudio.bartolini@hp.com

Wasim Sadiq
SAP Australia, Brisbane, QLD 4000, Australia
E-mail: wasim.sadiq@sap.com

Claude Godart
LORIA-ECOO, 54506 Vandœuvre-lès-Nancy, France
E-mail: claude.godart@loria.fr

Library of Congress Control Number: 2007939800

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 3-540-76992-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-76992-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12197309 06/3180 5 4 3 2 1 0

Preface

WISE 2007 was held in Nancy, France, during December 3–6, hosted by Nancy University and INRIA Grand-Est. The aim of this conference was to provide an international forum for researchers, professionals, and industrial practitioners to share their knowledge in the rapidly growing area of Web technologies, methodologies and applications. Previous WISE conferences were held in Hong Kong, China (2000), Kyoto, Japan (2001), Singapore (2002), Rome, Italy (2003), Brisbane, Australia (2004), New York, USA (2005), and Wuhan, China (2006).

The call for papers created a large interest. Around 200 paper submissions arrived from 41 different countries (Europe: 40%, Asia: 33%, Pacific: 10%, North America: 7%, South America: 7%, Africa: 2%). The international Program Committee selected 40 full-papers (acceptance rate of 20%) and 18 short papers (acceptance rate of 9%). As a result, the technical track of the WISE 2007 program offered 13 sessions of full-paper presentation including one industrial session and 5 sessions of short papers. The selected papers cover a wide and important variety of issues in Web information systems engineering such as querying; trust; caching and distribution; interfaces; events and information filtering; data extraction; transformation and matching; ontologies; rewriting, routing and personalization; agents and mining; quality of services and management and modelling. A few selected papers from WISE 2007 will be published in a special issue of the *World Wide Web Journal*, by Springer. In addition, $1000 value was awarded to the authors of the paper selected for the “Yahiko Kambayashi Best Paper.” We thank all authors who submitted their papers and the Program Committee members and external reviewers for their excellent work.

Finally, WISE 2007 included two prestigious keynotes given by Eric Billingsley, eBay research and Lutz Heuser, SAP research, one panel, one preconference tutorial, and six workshops.

We would also like to acknowledge the local organization team, in particular Anne-Lise Charbonnier and François Charoy. We also thank Mohand-Said Hacid and Mathias Weske as Workshop Chairs, Manfred Hauswirth as Panel Chair, Mike Papazoglou as Tutorial Chair, Olivier Perrin, Michael Sheng and Mingjun Xiao as Publicity Chairs, Qing Li, Marek Rusinkiewicz and Yanchun Zhang for the relationship with previous events and the WISE Society, and Ustun Yildiz for his work in the editing proceedings.

We hope that the present proceedings will contain enough food for thought to push the Web towards many exciting innovations for tomorrow's society.

September 2007

Boualem Benatallah
Fabio Casati
Dimitrios Georgakopoulos
Claudio Bartolini
Wasim Sadiq
Claude Godart

Organization

General Chairs	Claude Godart, France Qing Li, China
Program Chairs	Boualem Benatallah, Australia Fabio Casati, Italy Dimitrios Georgakopoulos, USA
Industrial Program Chairs	Claudio Bartolini, USA Wasim Sadiq, Australia
Workshop Chairs	Mohand-Said Hacid, France Mathias Weske, Germany
Tutorial Chair	Mike Papazoglou, The Netherlands
Panel Chair	Manfred Hauswirth, Ireland
Publicity Chairs	Olivier Perrin, France Michael Sheng, Australia Mingjun Xiao, China
Publication Chair	Claude Godart, France
Wise Society Representatives	Yanchun Zhang, Australia Marek Rusinkiewicz, USA
Local Organization Chair	François Charoy, France
Local Organization Committee	Anne-Lise Charbonnier, France Laurence Félicité, France Nawal Guermouche, France Olivier Perrin, France Mohsen Rouached, France Hala Skaf, France Ustun Yildiz, France

Program Committee

Karl Aberer, Switzerland
Marco Aiello, The Netherlands
Rafael Alonso, USA
Toshiyuki Amagasa, Japan

Yasuhito Asano, Japan
Karim Baina, Morocco
Luciano Baresi, Italy
Ladjel Bellatreche, France
Elisa Bertino, USA
Athman Bouguettaya, USA
Shawn Bowers, USA
Alex Buchamnn, Germany
Dave Buttler, USA
Rajkumar Buyya, Australia
Wojciech Cellary, Poland
François Charoy, France
Arbee L.P. Chen, Taiwan
Soon Ae Chun, USA
Andrzej Cichocki, USA
Christine Collet, France
Gill Dobbie, New Zealand
Marlon Dumas, Australia
Schahram Dustdar, Austria
Johann Eder, Austria
Fernand Feltz, Luxembourg
Ling Feng, The Netherlands
James Geller, USA
Paul Grefen, The Netherlands
Daniela Grigori, France
Mohand-Said Hacid, France
Wook-Shin Han, Korea
Jun Han, Australia
Christian Huemer, Austria
Patrick C.K. Hung, Canada
Yoshiharu Ishikawa, Japan
George Karabatis, USA
Hiroyuki Kitagawa, Japan
Masaru Kitsuregawa, Japan
Herman Lam, USA
Jianzhong Li, China
Xuemin Lin, Australia
Tok Wang Ling, Singapore
Chengfei Liu, Australia
Peng Liu, USA
Jianguo Lu, Canada
Pat Martin, Canada
Michael Maximilien, USA
Xiaofeng Meng, China
Paolo Misser, UK
Mukesh Mohania, India
Noureddine Mouaddib, France
Miyuki Nakano, Japan
Ekawit Nantajeewarawat, Thailand
Amedeo Napoli, France
Anne Hee Hiong Ngu, USA
Beng Chin Ooi, Singapore
Mourad Ouzzani, USA
Helen Paik, Australia
Thimios Panagos, USA
Zhiyong Peng, China
Olivier Perrin, France
Jean-Marc Petit, France
Dimitris Plexousakis, Greece
Rodion Podorozhny, USA
Fethi A. Rabhi, Australia
Thomas Risse, Germany
Norbert Ritter, Germany
Uwe Roehm, Australia
Colette Rolland, France
Keun Ho Ryu, Korea
Shazia Sadiq, Australia
Regis Saint-Paul, Australia
Yucel Saygin, Turkey
Monica Scannapieco, Italy
Klaus-Dieter Schewe, New Zealand
Ming-Chien Shan, USA
Tony Shan, USA
Michael Sheng, Australia
Amit Sheth, USA
Jianwen Su, USA
Aixin Sun, Singapore
Stefan Tai, USA
Kian-Lee Tan, Singapore
Samir Tata, France
Farouk Toumani, France
Aphrodite Tsalgatidou, Greece
Julien Vayssiere, Australia
Yannis Velegrakis, Italy
Wei Wang, Australia
Raymond Wong, Australia
Guotong Xie, China
Jian Yang, Australia
Ge Yu, China
LiangZhao Zeng, USA
Xiaofang Zhou, Australia
Aoying Zhou, China

Sponsoring Institutions

Nancy-Université

INRIA

Loria

I3

région lorraine

Grand Nancy
communauté urbaine

emisa.org

Table of Contents

Session 1: Querying

Session 2: Trust

Session 3: Caching and Distribution

Session 4: Interfaces

Session 5: Events and Information Filtering

Session 6: Data Extraction, Transformation, and Matching

Session 7: Ontologies

Session 8: Rewriting, Routing, and Personalisation

Session 9: Agents and Mining

Session 10: QoS and Management

Session 11: Modeling

Session 12: Topics

Industrial Session

Short Paper Session 1

Short Paper Session 2

Short Paper Session 3

Short Paper Session 4

Short Paper Session 5

Development Process of the Operational Version of PDQM

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

[1] Department of Computer Science and Information Technologies, University of Bio Bio
Chillán, Chile
mcaro@ubiobio.cl
[2] Alarcos Research Group. Information Systems and Technologies Department
UCLM-INDRA Research and Development Institute.
University of Castilla-La Mancha
{Coral.Calero, Mario.Piattini}@uclm.es

Abstract. PDQM is a web portal data quality model. This model is centered on the data consumer perspective and for its construction we have developed a process which is divided into two parts. In the first part we defined the theoretical version of PDQM and as a result a set of 33 data quality attributes that can be used to evaluate the data quality in portals were identified. The second part consisted of the conversion of PDQM into an operational model. For this, we adopted a probabilistic approach by using Bayesian networks. In this paper, we show the development of this second part, which was divided into four phases: (1) Definition of a criterion to organize the PDQM's attributes, (2) Generation of a Bayesian network to represent PDQM, (3) Definition of measures and the node probability tables for the Bayesian network and (4) The validation of PDQM.

Keywords: Data Quality, Information Quality, Web Portal, Data Quality Evaluation, Bayesian Network, Measures.

1 Introduction

In literature, the concept of Data or Information Quality (hereafter referred to as DQ) is often defined as "fitness for use", i.e., the ability of a collection of data to meet a user's requirements [3, 14]. This definition and the current view of assessing DQ, involve understanding DQ from the users' point of view [8].

Advances in technology and the use of the Internet have favoured the emergence of a large number of web applications, including web portals. A web portal (WP) is a site that aggregates information from multiple sources on the web and organizes this material in an easy user-friendly manner [16]. In the last years the number of organizations which own WPs has grown dramatically. They have established WPs with which to complement, substitute or widen existing services to their clients [17]. Many people use data obtained from WPs to develop their work and to make decisions. These users need to be sure that the data obtained from the WPs are appropriate for the use they need to make of it. Likewise, the WP owners need to deliver data that meet the user's requirements in order to achieve the user's preference. Therefore, DQ represents a common interest between data consumers and WP providers.

B. Benatallah et al. (Eds.): WISE 2007, LNCS 4831, pp. 436–448, 2007.

In recent years, the research community has started to look into the area of DQ on the web [7]. However, although some studies suggest that DQ is one of the relevant factors when measuring the quality of a WP [10, 17], few address the DQ in WPs. Along with this, another important factor to consider is the relevance of users (or data consumers) in DQ evaluation and the necessity of proposals dealing with this topic [2, 3, 7].

Consequently, our research aims is to create a DQ model for web portals, named PDQM, which focuses upon the data consumer's perspective. To this end, we have divided our work into two parts. The first consisted of the theoretical definition of PDQM [4], which resulted in the identification of 33 DQ attributes that can be used to assess a portal's DQ. The second, presented in this paper, is concerned with converting the theoretical model into an operational one. This conversion consists of specifying the DQ attributes of PDQM in an operational way. This means defining a structure with which to organize the DQ attributes, and to associate measures and criteria for them.

Considering the subjectivity of the data consumer's perspective and the uncertainty inherent in quality perception [6], we chose to use a probabilistic approach by means of Bayesian networks, such as that proposed in [9], to transform PDQM into an operational model. A Bayesian network (BN) is a directed acyclic graph where nodes represent variables (factors) and arcs represent dependence relationships between variables. Arcs in a BN connect parent to child nodes, where a child node's probability distribution is conditional to its parent node's distribution. Arcs, nodes and probabilities can be elicited from experts and/or empirical data, and probabilities are conveyed by using Node probability tables (NPTs) which are associated to nodes [13]. In our context, BNs offer an interesting framework with which it is possible to: (1) Represent the interrelations between DQ attributes in an intuitive and explicit way by connecting influencing factors to influenced ones, (2) Deal with subjectivity and uncertainty by using probabilities (3) Use the obtained network to predict/estimate the DQ of a portal and (4) Isolate responsible factors in the case of low data quality.

This paper focuses on the process of converting PDQM into an operational model. That is, the creation of the Bayesian network that represents PDQM, its preparation in order to use it in the DQ assessment and its validation.

The rest of the paper is organized as follows. Section 2 presents a summary of the theoretical definition of PDQM and describes the process used to generate the operational model. Section 3 presents the criterion used to organize the PDQM's attributes. Section 4 shows the generation of the structure for the BN that will support PDQM. The definition of measures and NPTs for the BN is described in Section 5. The validation of PDQM is explained in Section 6. Finally, conclusions are given in Section 7.

2 Defining a Data Quality Model

To produce the portal data quality model (PDQM), we have defined a process which we have divided into two parts. The first part corresponded to the theoretical definition of PDQM and was based on the key aspects that represent the data consumer's perspective and the main characteristics of WPs. As a result of this first part we obtained a set 33 of DQ attributes that can be used to assess DQ in WPs (see Table 1). All the details of the development of the theoretical version of PDQM can be found in [4].

Table 1. Data Quality Attributes of PDQM

Accessibility	Consistent Representation	Novelty	Timeliness
Accuracy	Customer Support	Objectivity	Traceability
Amount of Data	Documentation	Organization	Understandability
Applicability	Duplicates	Relevancy	Currency
Attractiveness	Ease of Operation	Reliability	Validity
Availability	Expiration	Reputation	Value added
Believability	Flexibility	Response Time	
Completeness	Interactivity	Security	
Concise Representation	Interpretability	Specialization	

The second part consists of the transformation of the theoretical model into an operational one. To do this, we decided to use a probabilistic approach by using BN. The second part is composed of four phases. During the first phase, we have defined the criteria with which to organize the DQ attributes of PDQM. In the second phase, we have generated the graphical structure of PDQM (BN graph). In the third phase, we have prepared PDQM to be used in an evaluation process. Finally, the fourth phase corresponds to the model validation (see, Fig. 1).

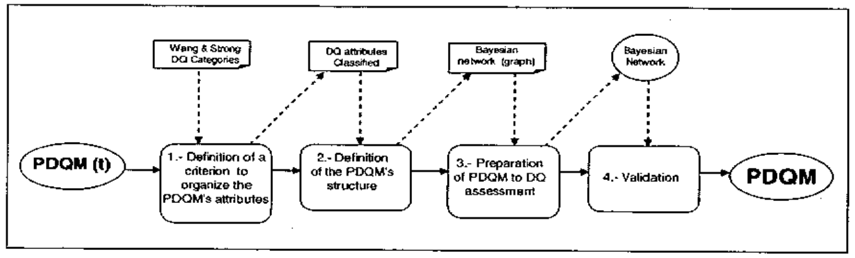

Wang & Strong DQ Categories
DQ attributes Classified
Bayesian network (graph)
Bayesian Network
PDQM (t)
1.- Definition of a criterion to organize the PDQM's attributes
2.- Definition of the PDQM's structure
3.- Preparation of PDQM to DQ assessment
4.- Validation
PDQM

Fig. 1. The development process of the operational version of PDQM

The following sections describe the conversion of PDQM into an operational model.

3 Phase 1: Definition of a Criterion to Organize the PDQM's Attributes

As explained in [11], a BN can be built by starting from semantically meaningful units called network fragments. A fragment is a set of related random variables that can be constructed and reasoned about separately from other fragments. Thus an initial phase when building the BN for PDQM, was to define a criterion that allowed us to organize the DQ attributes into a hierarchical structure, with the possibility of creating network fragments.

We used the conceptual DQ framework developed in [14] as a criterion for organizing the DQ attributes. However, in our work we have renamed and redefined the Accessibility category, calling it the Operational category. The idea was to emphasize the importance of the role of systems, not only with respect to accessibility and security, but also with respect to aspects such as personalization, collaboration, etc. Having done all this, and taking the definition of each DQ category into account, we have classified all the DQ attributes of PDQM into the categories seen below in Table 2. Thus, we have identified 4 network fragments based on this classification, one per category.

Table 2. Classification of DQ Attributes of PDQM into DQ Categories

DQ Category	DQ Attributes
Intrinsic: This denotes that data have quality in their own right.	Accuracy, Objectivity, Believability, Reputation, Currency, Duplicates, Expiration, Traceability
Operational: This emphasizes the importance of the role of systems; that is, the system must be accessible but secure in order to allow personalization and collaboration, amongst other aspects.	Accessibility, Security, Interactivity, Availability, Customer support, Ease of operation, Response time
Contextual: This highlights the requirement which states that DQ must be considered in the context of the task in hand.	Applicability, Completeness, Flexibility, Novelty, Reliability, Relevancy, Specialization, Timeliness, Validity, Value-Added
Representational: This denotes that the system must present data in such a way as to be interpretable and easy to understand, as well as concisely and consistently represented.	Interpretability, Understandability, Concise Representation, Consistent Representation, Amount of Data, Attractiveness, Documentation, Organization

4 Phase 2: Definition of the PDQM's Structure

In order to generate new levels in the BN, we established relationships of direct influences between the attributes in each category. These relationships were established by using the DQ categories and the DQ attributes definitions, together with our perceptions and experience. Thus, each relationship is supported by a premise that represents the direct influence between an attribute and its parent attribute. As an example of how this works, Table 3 shows the relationships established in the DQ Representational category.

Table 3. Relationships between DQ attributes in the DQ Representational category

	Relation of Direct Influence		**Premise that supports the direct influence relationships**
	Level 2	**Level 3**	
DQ Representational (Level 1)	Concise Representation	-	If data are compactly represented without superfluous elements then they will be better represented.
	Consistent Representation	-	If data are always presented in the same format, are compatible with previous data and consistent with other sources, then they will be better represented.
	Understandability	Interpretability	If data are appropriately presented in language and units for users' capability then they will be understood better.
		Amount of data	If the quantity or volume of data delivered by the WP is appropriate then they will be understood better.
		Documentation	If data have useful documents with meta information then they will be understood better.
		Organization	If data are organized with a consistent combination of visual settings then they will be understood better.
	Attractiveness	Organization	If data are organized with a consistent combination of visual settings then they will be more attractive to data consumers.

Taking these relationships as basis, we built the BN graph which represents PDQM, see Fig. 2.

In the BN which was created, four levels can be distinguished. Level 0, where PDQ is the node that represents DQ in the whole WP. Level 1, where nodes represent DQ in each DQ category in a WP (obviously, the PDQ node is defined in terms of the others 4). Level 2, where nodes represent the DQ attributes with a direct influence upon each of the DQ categories, and Level 3, where nodes represent the DQ attributes with a direct influence upon each of the DQ attributes in Level 2.

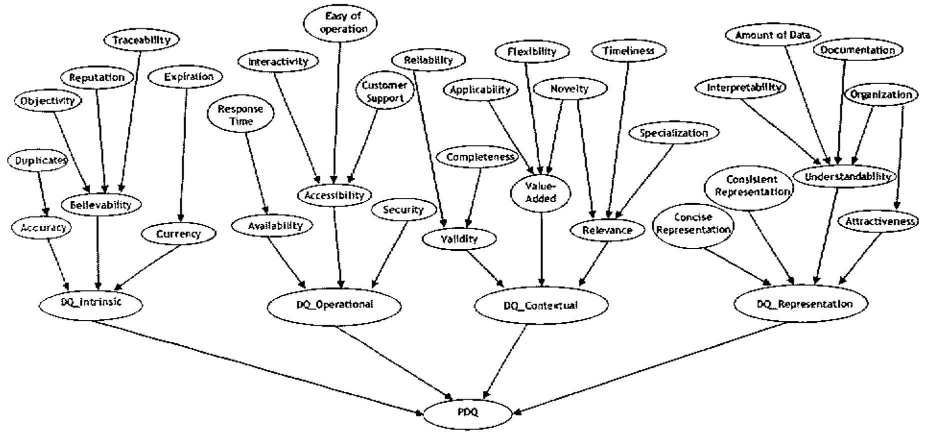


Fig. 2. BN graph to represent PDQM

In the following phase the BN must be prepared to be used in an evaluation process.

5 Phase 3: Preparation of PDQM for DQ Assessment

Having taken into consideration the size of the BN generated in the previous phase, and although our final objective is to create a comprehensive BN model for PDQM, we decided to develop this phase separately for each fragment network (DQ_Intrinsic, DQ_Operational, DQ_Contextual and DQ_Representational). In this paper we shall in particular work with the DQ_Representational fragment. To prepare the fragment network to be used in the DQ assessment, the following sub-phases will be developed:

a. If necessary, artificial nodes will be created to simplify the fragment network, i.e., to reduce the number of parents for each node.
b. Measures for the quantifiable variables (entry nodes) in the fragment network will be defined.
c. The NPTs for each intermediate node in the fragment network will be defined.

Phase a: Simplifying the fragment network. The original sub-network had two nodes with four parents (Understandability and DQ_Representational) so we decided to create two synthetic nodes (Representation and Volume of Data) in order to reduce the combinatory explosion in the following step during the preparation of the NPTs. In Fig. 3 we will show the original sub-network (graph 1) and the sub-network with the synthetic nodes created (graph 2).

Phase b: Defining qantifiable variables for fragment network. This sub-phase consists of the definition of measures for the quantifiable variables in the DQ_Representational fragment. We therefore defined an indicator for each entry node in the fragment (see Fig. 3, graph 3). In general, we selected and defined measures according to each attribute's (entry node) definition. To calculate each indicator we followed two methods: (1) base and derived measures are used when objective measures can be obtained (a measure is derived from another base or from derived measures [1]) or (2) data consumer valuations are used when the attribute is subjective. In both cases the indicators will take a numerical value of between 0 and 1. For each indicator, the labels that represent the fuzzy sets associated with that indicator were defined by using a fuzzy approach, and by considering the possible values that the indicator may take. Finally, a membership function was defined to determine the degree of membership of each indicator with respect to the fuzzy labels. In Table 4 we show a summary of the indicators defined for this fragment.

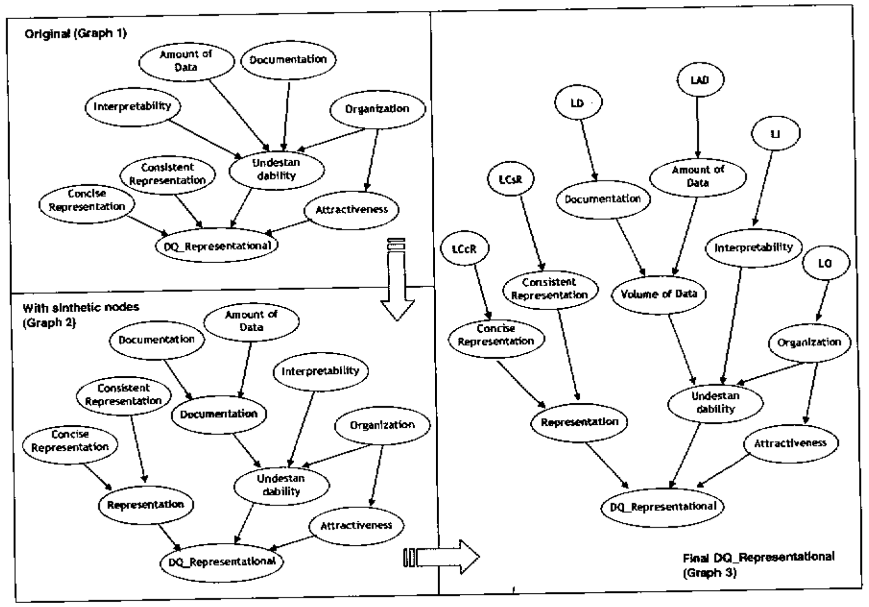


Fig. 3. Preparation of Sub-network DQ_Representational for assessment

As an example, we will explain the definition of the LAD indicator. To calculate the LAD indicator we have established an *Analysis Model* that includes a *formula* which gives us a numerical value and a *Decision Criteria* in the form of a membership function (see Table 5). This will later allow us to determine the degree of membership of each measure with respect to fuzzy labels. Note, as is explained by Thomas in [15], that the membership function degrees can be used as probabilities with the condition that both the fuzzy clustering algorithm and the approximation method preserve the condition that the sum of the membership degrees is always equal to 1.

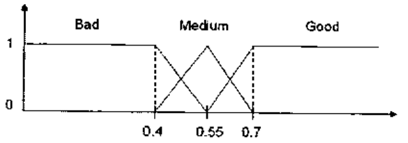
The analysis model (formula and decision criteria) attached to each of these indicators was determined from an analysis of literature, or from common-sense assumptions about the preferences of data consumers in WPs.

In the case of the LAD indicator, our intention is to use the formula to represent the fact that data consumers estimate the amount of data that exists, by assessing the amount and distribution of images, links and words that a WP delivers on each page. We assign more importance (0.4) to the amount of words because it has more impact on users: they do not feel comfortable if they have to read too much [12].

Table 4. Indicators defined for the fragment network

Name	Description
Level of Concise Representation (LCcR)	To measure ***Concise Representation*** (*The extent to which data are compactly represented without superfluous or not related elements*). To calculate *LCcR*, measures associated with the amount and size of paragraphs and the use of tables to represent data in a compact form were considered.
Level of Consistent Representation (LCsR)	To measure ***Consistent Representation*** (*The extent to which data are always presented in the same format, are compatible with previous data and are consistent with other sources*). The measures defined are centred on the consistency of the format and on compatibility with the other pages in the WP. For this indicator measures based on the use of Style in the pages of the WP and in the correspondence between a source page and the destination pages were defined.
Level of Documentation (LD)	To measure ***Documentation*** (*Quantity and utility of the documents with metadata*). To calculate LD measures related to the basic documentation that a WP presents to data consumers were defined. In particular, the simple documentation associated with the hyperlinks and images on the pages was considered.
Level of Amount of Data (LAD)	To measure ***Amount of Data*** (*The extent to which the quantity or volume of data delivered by the WP is appropriate*). Understanding the fact that, from the data consumer's perspective, the amount of data is concerned with the distribution of data within the pages in the WP. To calculate LAD the data in text form (words), in hyperlink form (links) and in visual form (images) were considered.
Level of Interpretabilit y (LI)	To measure ***Interpretability*** (*The extent to which data are expressed in language and units appropriate for the consumer's capability*). As we considered that the evaluation of this attribute was too subjective, a check list for its measurement was used. Each item in the check list will be valuated with a number from 1 to 10. These values are subsequently transformed into a [0, 1] range.
Organization (LO)	To measure ***Organization*** (*The organization, visual settings or typographical features (color, text, font, images, etc.) and the consistent combinations of these various components*). To calculate LO measures that verify the existence of groups of data in the pages (tables, frames, etc.), the use of colors, text with different fonts, titles, etc. were considered.

Table 5. Analysis model of LAD indicator

<table>
<tr><th colspan="2">LAD (Level of Amount of Data)</th></tr>
<tr><td>Formula to Calculate LAD</td><td>Decision Criteria</td></tr>
<tr><td>LAD = DWP * 0.4 + DLP * 0.3 + DIP * 0.3</td><td rowspan="3">
</td></tr>
<tr><td>Derived Measures</td></tr>
<tr><td>DWP: Distribution of Words per page
DLP: Distribution of Links per page
DIP: Distribution of Images per page</td></tr>
</table>

For the decision criteria (taking into consideration that several studies show that users prefer data presented in a concise form [12]), the membership function transforms the numerical value of the indicator into one of the following labels: Good, Medium and Bad.

Phase c: Defining the NPTs for fragment network. The intermediate nodes are nodes which are defined by their parents and are not directly measurable. Thus, their NPTs were made by expert judgment.

On the other hand, after having taken into account the importance of considering the task context of users and the processes by which users access and manipulate data to meet their task requirements [14] in a DQ evaluation process, we considered that the probability distribution may differ according to the WP context.

This implies that sub-phase c must be developed by considering a specific domain of WPs. In this work we have started to consider the educational context and we have defined the NPTs by considering that this fragment will be applied to university WPs. Table 6 shows the NPTs for the nodes in the third level of the fragment.

Thus, the DQ_Representational fragment is prepared to evaluate the DQ in university WPs. The last phase in generating the operational model for this fragment is that of its validation. In the next section we shall describe the validation experiment which was developed and the results which were obtained.

Table 6. Node probability tables for Level 2 in fragment

Level 2

Consistent Representation

LCsR	Low	Medium	High
Bad	0.9	0.05	0.01
Medium	0.09	0.9	0.09
Good	0.01	0.05	0.9

Volume of Data

Documentation	Bad			Medium			Good		
Amount of Data	Bad	Medium	Good	Bad	Medium	Good	Bad	Medium	Good
Bad	0.9	0.8	0.5	0.8	0.3	0.15	0.5	0.15	0.01
Medium	0.09	0.15	0.3	0.15	0.4	0.25	0.3	0.25	0.09
Good	0.01	0.05	0.2	0.05	0.3	0.6	0.2	0.6	0.9

Concise Representation

LCcR	Low	Medium	High
Bad	0.9	0.05	0.01
Medium	0.09	0.9	0.09
Good	0.01	0.05	0.9

Interpretability

LI	Low	Medium	High
Low	0.9	0.05	0.01
Medium	0.09	0.9	0.09
High	0.01	0.05	0.9

Organization

LO	Low	Medium	High
Bad	0.9	0.05	0.01
Medium	0.09	0.9	0.09
Good	0.01	0.05	0.9

6 Phase 4: Validation of PDQM

The method defined to validate PDQM consisted of using two different strategies to evaluate the representational DQ in a given WP. One of them evaluated the DQ with a group of subjects and the other evaluated it with PDQM. We next compared the results obtained to determine whether the evaluation made with PDQM was similar to that made with the subjects. That is, whether the model represented the data consumer's perspective.

Therefore, for the first strategy we developed an experiment by which to obtain the judgments of a group of subjects about a DQ representational in a university WP. In this experiment, the subjects were asked for their partial valuations of each DQ attribute in the fragment and for their valuation of the global representational DQ in the WP.

For the second assessment strategy, we built a tool that implements the fragment of the DQ_Representational. This tool allows us to automatically measure the quantifiable variables and, from the values obtained, to obtain the entry data for the BN that will give us the evaluation of PDQM. In the following subsections we will describe the experiment, the automatic evaluation and the comparison of both results in greater detail.

6.1 The Experiment

The subjects who took part in the experiment were a group of students from the University of Castilla-La Mancha in Spain. The group was composed of 79 students enrolled in the final year (third) of Computer Science (MSc). All of the subjects had previous experience in the use of WPs as data consumers. The experimental material was composed of one document including: the instructions and motivations, the URL of a university WP, three activities to be developed in the WP, and a set of 9 questions in which we requested their valuations for the DQ representational in the WP. The first 8 valuations were requested for each of the DQ attributes in the DQ_Representational fragment and the last question attempted to gauge the global DQ Representational in the WP. As a result of this experiment we obtained the valuations shown in Table 6.

Table 6. Valuations given by the subjects for the DQ Representational

Attribute Evaluated	Valuations		
	Low/Bad	Medium	High/Good
Attractiveness	30%	61%	9%
Organization	37%	44%	19%
Amount of Data	18%	49%	33%
Understandability	32%	47%	21%
Interpretability	6%	45%	48%
Documentation	16%	49%	34%
Consistent Representation	18%	53%	29%
Concise Representation	16%	52%	32%
Portal	**17%**	**68%**	**16%**

6.2 The Automatic Evaluation

The PoDQA tool [5] is the application that will support the PDQM model. Its aim is to give the user information about the DQ level in a given WP (at present it is just a prototype and only the Representational DQ is supported). The tool downloads and analyzes the pages of the WP, in order to calculate the defined measures using the public information in WPs.

Thus, for a given WP PoDQA will calculate the measures associated with the indicators: LCsR, LCcR, LD, LAD, LI, LO. Each indicator will take a value of between 0 and 1. This value will be transformed into a set of probabilities for the corresponding labels. Each of these values will be the input for the corresponding input node. With this value, and by using its probability table, each node generates a result that is propagated, via a causal link, to the child nodes for the whole network until the level of the DQ Representational is obtained.

We used PoDQA to evaluate the same university WP that was used in the experiment. As a result we obtained the values for each indicator (see Table 7) which were transformed into a valid entry for the BN. These values were entered in the BN which, finally, generated the level of DQ representational in the WP (like as in Fig. 4).

Table 7. Values obtained for the indicators of the DQ_Representational fragment

LCsR	LCcR	LD	LAD	LI	LO
0.12	0.99	0.46	0.99	0.5	0.44

6.3 Comparing the Results Obtained

When comparing the results obtained with the two evaluation strategies, we can observe that, in general, they are very different, see Table 8.

In effect, with regard to the final evaluation, Table 8 (last row) shows that while in the experiment the subjects evaluated the DQ at a *Medium* level (68%), with the automatic evaluation the DQ was evaluated at the same value for the *Medium* and *High* levels (in both cases 40%). With regard to the partial values, that is, for each DQ attribute, the results are also very different. The reason for this is, in our opinion, that the results given for the indicators are, in some cases, very extreme (see for example the values for LCcR and LCsR). Consequently, the nodes with most differences are the child nodes of the nodes that represent the indicators that take these extreme values.

Table 8. Valuations obtained from the experiment and valuations calculated automatically

Attribute Evaluated	Low/Bad		Medium		High/Good	
	Subj.	PDQA	Subj.	PDQA	Subj.	PDQA
Attractiveness	30%	34%	61%	44%	9%	22%
Organization	37%	26%	44%	66%	19%	8%
Amount of Data	18%	6%	49%	13%	33%	81%
Understandability	32%	52%	47%	23%	21%	25%
Interpretability	6%	43%	45%	49%	48%	7%
Documentation	16%	9%	49%	82%	34%	9%
Consistent Representation	18%	81%	53%	13%	29%	6%
Concise Representation	16%	6%	52%	13%	32%	81%
Portal	**17%**	**20%**	**68%**	**40%**	**16%**	**40%**

A preliminary interpretation of these results is that PDQM is more demanding than the subjects and needs to be adjusted. Thus, we attempted to reduce these differences by adjusting the NPTs and recalculating the representational DQ. The results obtained can be observed in Fig. 4, which shows the BN and the values calculated for it to each DQ attribute and the representational DQ level, and Table 9, which allow us to compare the values obtained.

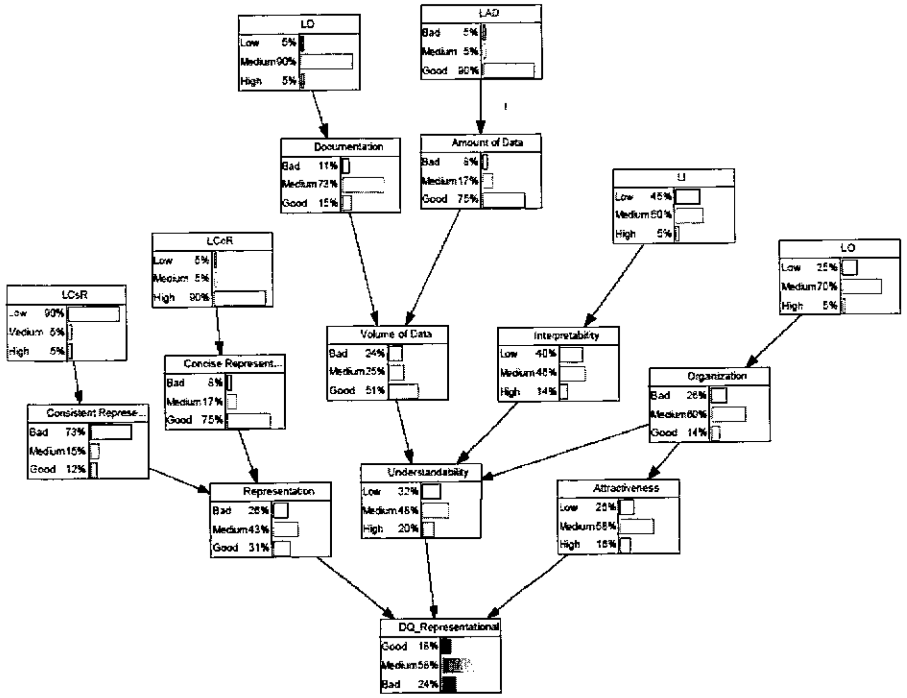


Fig. 4. New results adjusting the node probability tables in the BN of DQ Representational

As a result of this new configuration the general result of the automatic evaluation is closer to the subjects' evaluations. However, in spite of the fact that both evaluations gave their result as *Medium*, total coincidence between the values calculated does not exist (see last row in Table 9). Moreover, the partial values also have a better fit than in the first calculation, but do not totally coincide. See for example the differences between the Interpretability and Consistent Representation attributes for the valuations Low/Bad. We again believe that the main reason for this is the extreme values of the indicators. But this is not the only reason. Together with the former problem, we believe that the design of the WP evaluated may also influence this result. For example, to calculate the *Level of Amount of Data*, it is necessary to know the *distribution of words per page*. The measured WP presents values for this measure which can be considered as outliers (they take extreme values that do not follow a uniform distribution). Obviously, these values need to be removed from the calculation of the measure. Because of this we are now refining the calculations made by the tool by detecting and eliminating the outliers in our measures.

Table 9. New valuations obtained from PDQA with the new configuration of PDQM

Attribute	**Low/Bad**		**Medium**		**High/Good**	
Evaluated	**Subj.**	**PDQA**	**Subj.**	**PDQA**	**Subj.**	**PDQA**
Attractiveness	30%	26%	61%	58%	9%	16%
Organization	37%	26%	44%	60%	19%	14%
Amount of Data	18%	8%	49%	17%	33%	75%
Understandability	32%	32%	47%	48%	21%	20%
Interpretability	6%	40%	45%	46%	48%	14%
Documentation	16%	11%	49%	73%	34%	15%
Consistent Representation	18%	73%	53%	15%	29%	12%
Concise Representation	16%	8%	52%	17%	32%	75%
Portal	**17%**	**18%**	**68%**	**58%**	**16%**	**24%**

Of course we also need to repeat the experience carried out on just one WP in order to be sure that the BN accurately estimates the Representational DQ of any WP.

7 Conclusions and Future Work

In this paper, we have presented a work which consists of the development of PDQM, a DQ model for web portals. In the first part of our work, which is briefly mentioned in this paper, we have defined a theoretical version of PDQM composed of a set of 33 DQ attributes that can be used for DQ evaluation in WPs. In the second part, which is described in more detail, we have presented the process developed to convert the theoretical model into an operational model. For this purpose, we have chosen a probabilistic approach, by using a BN, due to the fact that many issues in quality assessment such as threshold value definition, measure combination, and uncertainty are circumvented.

We have thus defined a BN to support PDQM and we have built a tool that implements a sub-part of PDQM. The relevance of the approach used has been demonstrated in the first validation of our model. We believe that our proposal for DQ evaluation in WPs is a good alterative for data consumers. It may even be useful for WP developers who wish know whether their WPs have a good DQ level for data consumers.

We believe that one of the advantages of our model will be its flexibility. Indeed, the idea is to develop a model that can be adapted to both the goal and the context of evaluation. From the goal perspective, the user can choose the fragment that evaluates the characteristics he/she is interested in. From the context point of view, the parameters (NPTs) can be changed to consider the specific context of the WP evaluated.

As future work, we first plan to develop new validations which consider a greater number of WPs and which will allow us to refine PDQM. Another aspect to be considered is that of extending the definition of PDQM to other WP contexts. Lastly, we plan to extend the model to the other fragments and to include them in the PoDQA tool.

Acknowledgments. This research is part of the following projects: ESFINGE (TIC2006-15175-C05-05) granted by the Dirección General de Investigación del Ministerio de Ciencia y Tecnología (Spain), CALIPSO (TIN20005-24055-E) supported by the Minis-terio de Educación y Ciencia (Spain), DIMENSIONS (PBC-05-012-1) supported by FEDER and by the "Consejería de Educación y Ciencia, Junta de Comunidades de Castilla-La Mancha" (Spain) and COMPETISOFT (506AC0287) financed by CYTED.

References

1. Bertoa, M., García, F., Vallecillo, A.: An Ontology for Software Measument. In: Calero, C., Ruiz, F., Piattini, M. (eds.) Ontologies for Software Engineering and Software Technology (2006)
2. Burgess, M., Fiddian, N., Gray, W.: Quality Measures and The Information Consumer. In: Proceeding of the 9th International Conference on Information Quality (2004)
3. Cappiello, C., Francalanci, C., Pernici, B.: Data quality assessment from the user's perspective. In: International Workshop on Information Quality in Information Systems, Paris, Francia, ACM, New York (2004)
4. Caro, A., Calero, C., Caballero, I., Piattini, M.: Defining a Data Quality Model for Web Portals. In: Aberer, K., Peng, Z., Rundensteiner, E.A., Zhang, Y., Li, X. (eds.) WISE 2006. LNCS, vol. 4255, Springer, Heidelberg (2006)
5. Caro, A., Calero, C., de Salamanca, J.E., Piattini, M.: A prototype tool to measure the data quality in Web portals. In: The 4th Software Measurement European Forum, Roma, Italia (2007)
6. Eppler, M.: Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes. Springer, Heidelberg (2003)
7. Gertz, M., Ozsu, T., Saake, G., Sattler, K.-U.: Report on the Dagstuhl Seminar "Data Quality on the Web". SIGMOD Record 33(1), 127–132 (2004)
8. Knight, S.A., Burn, J.M.: Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal 8, 159–172 (2005)
9. Malak, G., Sahraoui, H., Badri, L., Badri, M.: Modeling Web-Based Applications Quality: A Probabilistic Approach. In: 7th International Conference on Web Information Systems Engineering, Wuhan, China (2006)
10. Moraga, M.Á., Calero, C., Piattini, M.: Comparing different quality models for portals. Online Information Review. 30(5), 555–568 (2006)
11. Neil, M., Fenton, N.E., Nielsen, L.: Building large-scale Bayesian Networks. The Knowledge Engineering Review 15(3), 257–284 (2000)

12. Nielsen, J.: Designing Web Usability: The Practice of Simplicity. New Riders Publishing, Indianapolis (2000)
13. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
14. Strong, D., Lee, Y., Wang, R.: Data Quality in Context. Communications of the ACM 40(5), 103–110 (1997)
15. Thomas, S.F.: Possibilistic uncertainty and statistical inference. In: ORSA/TIMS Meeting, Houston, Texas (1981)
16. Xiao, L., Dasgupta, S.: User Satisfaction with Web Portals: An empirical Study. In: Gao, Y. (ed.) Web Systems Design and Online Consumer Behavior, pp. 193–205. Idea Group Publishing, USA (2005)
17. Yang, Z., Cai, S., Zhou, Z., Zhou, N.: Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information and Management, vol. 42, pp. 575–589. Elsevier, Amsterdam (2004)

A New Reputation Mechanism Against Dishonest Recommendations in P2P Systems*

Junsheng Chang, Huaimin Wang, Gang Yin, and Yangbin Tang

School of Computer, National University of Defense Technology,
HuNan Changsha 410073, China
cjs7908@163.com

Abstract. In peer-to-peer (P2P) systems, peers often must interact with unknown or unfamiliar peers without the benefit of trusted third parties or authorities to mediate the interactions. Trust management through reputation mechanism to facilitate such interactions is recognized as an important element of P2P systems. However current P2P reputation mechanism can not process such strategic recommendations as correlative and collusive ratings. Furthermore in them there exists unfairness to blameless peers. This paper presents a new reputation mechanism for P2P systems. It has a unique feature: a recommender's credibility and level of confidence about the recommendation is considered in order to achieve a more accurate calculation of reputations and fair evaluation of recommendations. Theoretic analysis and simulation show that the reputation mechanism we proposed can help peers effectively detect dishonest recommendations in a variety of scenarios where more complex malicious strategies are introduced.

1 Introduction

P2P (Peer-to-Peer) technology has been widely used in file-sharing applications, distributed computing, e-market and information management [1]. The open and dynamic nature of the peer-to-peer networks is both beneficial and harmful to the working of the system. Problems such as free-riders and malicious users could lead to serious problems in the correct and useful functioning of the system. As shown by existing work, such as [2, 3, 4, 5, 6, 7], reputation-based trust management systems can successfully minimize the potential damages to a system by computing the trustworthiness of a certain peer from that peer's behavior history. However, there are some vulnerabilities of a reputation-based trust model. One of the detrimental vulnerabilities is that malicious peers submit dishonest recommendations and collude with each other to boost their own ratings or bad-mouth non-malicious peers [7]. The situation is made much worse when a group of malicious peers make collusive attempts to manipulate the ratings [8, 9].

* Supported by National Basic Research Program of china under Grand (No.2005CB321800), National Natural Science Foundation of China under Grant (No.90412011, 60625203) and National High-Tech Research and Development Plan of China under Grant (2005AA112030).

B. Benatallah et al. (Eds.): WISE 2007, LNCS 4831, pp. 449–460, 2007.