20th International Conference on Information Quality (ICIQ 2015)

Making High Quality the Norm

Cambridge, Massachusetts, USA
24 July 2015

ISBN: 978-1-5108-2017-3

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Printed by Curran Associates, Inc. (2016)

MIT Information Quality Program
Attn: Dr. Richard Wang
225 Crafts Road
Chestnut Hill MA 02467
USA

Phone: (617) 304-3120

rwang@mit.edu

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

THE 15TH INTERNATIONAL CONFERENCE ON INFORMATION QUALITY

Table of Contents

Abstracts

Parallel Session 3-A: Entity Resolution Part I **Friday: July 24, 2015 – 12:30 to 1:30 pm**	**E51-145** **Session Chair: Ismael Caballero Munoz-Reja**

Entity Resolution Research and Opportunities

John R. Talburt

Abstract: Entity Resolution (ER) is the process of determining whether two references to real-world objects are referring to the same object or to different objects. This presentation will focus on a discussion of the new challenges in ER and Master Data Management (MDM) brought on by Big Data. Already straining to attain acceptable performance with "normal size" data, matching across large data is even more daunting. The talk will include an overview of some of the most important areas of research that seek to adapt traditional methods of blocking and transitive closure for use in the big data tools such as Hadoop map/reduce.

A Framework for Collecting, Extracting and Managing Event Identity Information from Twitter

Debanjan Mahata and John R. Talburt pg. 18

Abstract: With the popularity of Twitter, there has been voluminous growth in the digital footprints of real-life events in the Internet. The references to different types of events in Twitter have the potential to provide extremely valuable information to researchers and organizations, which could be mined and analyzed for making major decisions. There are tremendous applications in the areas of real-life event analysis, opinion mining, reference tracking, online advertising, recommendation engines, cyber security, event management, enterprise data integration, among others. Thus, there is a need of a generic framework that can collect different event references, extract identity information of the events from them and maintain the information persistently for resolving new references to the events and provide updated analytics. The presented research establishes the design and implementation of such a framework from the perspective of Event Identity Information Management (EIIM) in the domain of Twitter. The paper introduces the problem of EIIM in Twitter, discusses the prevalent challenges and proposes the design of a framework capable of managing persistent identity information of pre-specified set of events. We explore the applications of the research, validate the different components of the framework and conclude with our comments on various criteria showing high efficacy and practical utility of our proposed framework.

Abstracts

Parallel Session 3-B: Assessing the Value of IQ **Friday: July 24, 2015 – 12:30 to 1:30 pm**	**E51-149** **Session Chair: Wei Zhang**

Information Product: How Information Consumers' Perception of 'Fitness for Use' Can Be Affected

Farjam Eshraghian and Stephen A. Harwood pg. 34

Abstract: Emerging paradigms such as big data, business data analytics and business data science increased the importance of focusing on the notion of data and information quality more than before. Although more mechanized processes are being used to prepare pieces of information to be fit-for-use for information consumers, how information consumers perceive the quality of information and its difference with the actual quality of information gets more important as information consumers are at the position to judge the quality. Motivated by the notion of Perceived Information Quality (PIQ), this study explains the role of adoption and implementation of Information Technology (IT) in the process of perceiving information quality by information consumers. For the purpose of this research, seven Iranian organizations were studied. Due to lack of proper communication and institutional link between these organizations and the technology developers, it has been challenging for the studied organizations to implement the adopted technologies in the context of use. This study finds that the solution which is provided for information quality issues after the implementation of a type of information technology has direct impact on perceived information quality. This solution can be the combination of technical solutions by the technology developers and contextual-situated solutions by local experts and the balance of these solutions can influence the information consumers' perception. This research unfolds how the mechanism of technology implementation may reform the perceived information quality by drawing upon Information Product theory from the literature of information quality and the theory of Social Learning from the literature of social studies of technology. This mechanism of influence can be conceptualized as follows: a generic type of information technology is perceived to be useful and adopted by an organization. The organization faces challenges in implementing the new technology including dealing with information quality issues due to specific institutional and organizational context. The organization tries to provide a combination of solutions to overcome this challenge. The type and balance of solutions can form the perceived information quality by information consumers within the organization.

Measuring Sociocultural Factors of Success in Information Quality Projects

Therese L. Williams, David K. Becker, Carmen Robinson, and John R. Talburt pg. 48

Abstract: Information and data quality practitioners are in general agreement that social, cultural, and organizational factors are the most important in determining the success or failure of an organization's data quality programs. This paper presents some of the first research undertaken to substantiate these anecdotal claims. The paper describes a survey of recent graduates from the University of Arkansas at Little Rock Information Quality Graduate Program. In the survey the graduates rate how much influence these sociocultural factors had on the outcomes of their data quality projects. The results of the survey support the practitioners' claims.

Abstracts

Parallel Session 3-C: IQ in CRM Systems **Friday: July 24, 2015 – 12:30 to 1:30 pm**	**E51-151** **Session Chair: Beverly Kahn**

Self-Healing Customer Data Quality Issues Through Interpretation of Unstructured Data – Text Mining and Machine Learning Applied (Project Irene)

Kannan Chandrasekaran and Delphine Clement pg. 70

Abstract: In traditional customer data management, it is rare that unstructured data, such as those stored in notes, is automatically and systematically interpreted to drive proactive data cleansing. And yet, it is a fact that sellers forming teams to manage their accounts in CRM keep exchanging verbatim through notes during the entire lifecycle of a customer relationship. The notes aim for instance at updating the players around the detection of new leads but also to warn about a key contact changing job or leaving the company or about an upcoming modification of company coordinates, in fact various comments that, if mined in an unsupervised way through machine learning, would allow an Enterprise data management group to operationalize a service of self-healing data quality, thereby giving both time back and up to date structured data to sellers. This practice oriented paper will discuss the results and method of a proof of concept developed and tested for mining Microsoft CRM notes, in its first phase looking after the cluster of contacts leaving their company. This paper will detail the business context and the technical realization of the project and will demonstrate why machine learning is instrumental to go from prototype to operationalization. This paper will also illustrate the value of the process developed, for the business on the one hand for whom data quality becomes ambient, and for the data management services group on the other hand for whom this automated approach allows more relevancy and pro-activeness in the data correction.

Statistical Quality Control Framework for Crowd-Worker in ER-In-house Crowdsourcing System

Morteza Saberi, Omar Khadeer Hussain, and Elizabeth Chang pg. 88

Abstract: These days, poor data quality is prevalent in organizations. This poor quality negatively effect on accuracy of organization decision making. The problem of dirty data is more severe for organization's customer relation-ship management (CRM) and prevents it from effective performance. One type of dirty data is duplicate records that correspond to the same entities. Presence of duplicate profiles in an organization's database prevents an organization to have a clear picture of customers' profile. Thus, developing efficient Entity resolution (ER) technique in a given organization is essential. Recently, crowdsourcing technique has been used to improve the accuracy of entity resolution process that make use of human intelligence to label the data and make it ready for further processing by entity resolution (ER) algorithms. However, labelling of data by humans is an error prone process that affects the process of entity resolution and eventually overall performance of crowd. Thus controlling the quality of labeling task is an essential for crowdsourcing systems. However, this task becomes more challenging due to unavailability of ground data. In this study, we focus on contact centers and employ Customer Service Representatives (CSRs) as crowd-worker for ER-Crowdsourcing system. A statistical quality control (SQC) framework is proposed to control the quality of CSRs labeling. The proposed SQC framework should be able to estimate the true error of CSRs in order to monitor their labeling accuracy and performance. To this end, a Hybrid Gold- plurality (HGP) Algorithm is proposed that estimate CSR's true error. The proposed HGP algorithm is capable of an appropriate accuracy in error estimation as it is composed of both Masking and Detection crowd-worker quality control mechanisms. Synthetic dataset is used to demonstrate the applicability of the SQC framework.

Abstracts

Parallel Session 5-A: Entity Resolution Part II **Friday: July 24, 2015 – 2:50 to 3:50 pm**	**E51-145** **Session Chair: Isaac Osesina**

Confidence Rating for Interactive Identity Resolution

Fumiko Kobayashi and John R. Talburt pg. 105

Abstract: Entity identity information management (EIIM) systems provide the information technology support for master data management (MDM) systems. One of the most important configurations of an EIIM system is identity resolution. In an identity resolution configuration, the EIIM system accepts entity identity information and returns the corresponding entity identifier. In the original EIIM model, identity resolution was only a batch operation. After the model was extended with an Identity Management Service (IMS) to decouple identity resolution from batch EIIM processing, identity resolution moved into the interactive realm with additional functionality. This paper discusses the design for one of these additional functions called an identity resolution confidence rating. The confidence rating is a measure of the likelihood that the returned identifier is accurate. This paper proposes a model for generating confidence ratings based on an assessment of the differences between the match scores for competing entities. The paper also includes results from experiments performed on an IMS system supporting a Student Identity Master Data Management System.

A Visualization System to Support Clerical Review, Correction, and Confirmation Assertions in Entity Identity Information Management

Cheng Chen, Daniel Pullen, and John R. Talburt pg. 120

Abstract: Clerical review of Entity Resolution (ER) is crucial to maintain entity identity integrity in Entity Identity Information Management (EIIM). This paper proposes the design of a new visualization system that reduces the effort and increases the accuracy of the clerical review process. The system also supports correction and confirmation assertions in EIIM for saving clerical review results to entity identity structures. The paper also discusses three modes of the visualization tool to support different clerical review purposes. Finally the paper describes how the design has been implemented to interact with the OYSTER open source EIIM System.

Abstracts

Parallel Session 5-B: IQ and Data Science Research **Friday: July 24, 2015 – 2:50 to 3:50 pm**	**E51-149** **Session Chair: Arun Sundararaman**

Reference Model for Transparent Data Supply

Sami Laine and Carol Lee pg. 135

Abstract: Human action can be interpreted from two alternative perspectives: prescriptive planning and emerging situations. Traditionally, Data and Information Quality Research has emphasized quality controls and standardization at data entry situations to improve the accuracy of data. However, data standards are only one factor that might describe the actual meaning of data instances. Likewise, quality controls are just one of the many factors that can influence the actual quality of data. In reality, the recorded data might not conform to specified data standards and data quality controls might have been circumvented. The paradigm of Situated Action emphasizes that the actual state of data can only be understood in relation to the actual circumstances and available resources at each data entry situation. Therefore, we constructed the Reference Model for Transparent Data Supply. The purpose of the reference model is to point out factors that should be explicitly recognized, considered and monitored while interpreting the actual meaning and quality of situated data. It consists of general requirements and general components as required by Explanatory Design Theories and it is founded on widely used frameworks from three disciplines: Human Computer-Interaction, Data and Information Quality, and Software Engineering. The reference model implies that variations in and between any of the modelled context factors will lead to distinctive data quality error profiles and can cause significant semantic heterogeneity. Unless these context factors and their variations are recognized, the validity of data-driven decision making might be compromised due to hidden accuracy errors and latent semantic heterogeneity. In the future, the reference model can be used to guide the development of transparent provenance across information production processes.

Timestamp Accuracy in Healthcare Business Process Improvement

Sami Laine, Juha Soikkeli, Toni Ruohonen, and Marko Nieminen pg. 150

Abstract: Business process improvement is a challenge in a complex environment, such as healthcare. Currently, it can involve a lot of manual data gathering, modelling and analyses. The results of data-driven analyses can be questioned due to inaccurate data affected by subtle contextual and human factors. Originally, we conducted two individual research projects that approached this same problem from qualitative and quantitative perspectives. According to our qualitative data quality research, current ambiguous and erroneous administrative timestamps should be more precise. A software development project indicates that the automatic process mining, discovery, modelling, and simulation could support healthcare process improvements by automating analytics and predictive simulations. However, contextual heterogeneity and data quality must be assessed carefully, and currently requires a lot of manual efforts. Our qualitative analyses identified and explained timestamp errors and semantic heterogeneity that cannot be currently identified from the data layer: data standards, data models and data sets. The synthesis of our findings indicates that software systems should collect metadata about user interactions and user interface structures. Additional metadata could then be used to discover and explain the actual meaning and contextual quality of recorded transaction data. We also noted that the automatic process visualization software, that discovers actual processes, could be more widely used to identify organizational data quality problems and opportunities for process improvement.

Abstracts

Parallel Session 5-C: Success Factors for IQ Projects **Friday: July 24, 2015 – 2:50 to 3:50 pm**	**E51-151** **Session Chair: Elizabeth Pierce**

An Objective Measurement of Information Value Using Application Traces in Infomediary: A Case Study of a Credit Reporting System in China

Mei Song and Jin Wang pg. 165

Abstract: A new objective method to measure the information value (IV) in infomediary is presented in the paper using application traces. It can be used to the hierarchical management of information quality in infomediary. As the credit infomediary in China, credit reporting system collects data from banks and public departments according to unified interface specification, and integrates and computes data based on certain rules to provide better credit services. Information value can be differentiated into three levels by analysis in credit reporting system. The basic level is the score more than 60 (level I), which is focus and should adopt most strict management; The middle level is between 40 and 60 (level II), which is secondary and should adopt moderate management. The last level is less than 40 (level III), which is bottom and with loosen and going management. The finding also provides valuable insight to the information quality management of other infomediaries.

Data Quality Analytics: Key Problems Arising from the Repurposing of Manufacturing Data

Philip Woodall and Anthony Wainman pg. 174

Abstract: Repurposing data is when data is used for a completely different decision/task to what it was originally intended to be used for. This is often the case in data analytics when data, which has been captured by the business as part of its normal operations, is used by data scientists to derive business insight. However, when data is collected for its primary purpose some consideration is given to ensure that the level of data quality is "fit for purpose". Data repurposing, by definition, is using data for a different purpose, and therefore the original quality levels may not be suitable for the secondary purpose. Using interviews with various manufacturers, this paper describes examples of repurposing in manufacturing, how manufacturing organizations repurpose data, data quality problems that arise specifically from this, and how the problems are currently addressed. From these results we present a framework which manufacturers can use to help identify and mitigate the issues caused when attempting to repurpose data.

Note: These authors were unable to attend ICIQ to present their papers in person; however, they granted us permission to include their papers in the Conference Proceedings.