Data and Information Quality: Research Themes and Evolving Trends

Emergent Research Forum

G. Shankaranarayanan
Babson College
gshankar@babson.edu

Roger Blake
University of Massachusetts Boston
roger.blake@umb.edu

Abstract

Along with data being increasingly viewed as a critical asset and the rise of "Big Data" with its potential uses, research of data and information quality has grown in importance to become a distinct area within information systems. Particularly with the ever more rapid changes in technology and adoption of data-driven decision-making, at this juncture it is important to take stock of data and information quality research by identifying the core topics and themes that distinguish this area. Next, it is important to understand its ongoing trends and patterns, which in turn will lead us to recognizing new and emerging research opportunities for researchers. This paper is a work in progress report on our study and the results for the first of these objectives, determining the core topics and themes that define the area of data and information quality research.

Keywords

Data quality, information quality, research frameworks, Latent Semantic Analysis (LSA)

Introduction

After beginning as one spread across many different disciplines, data and information quality research it has become a unified body of knowledge (Madnick, Wang, Lee, and Zhou, 2009). It was founded with the identification of quality dimensions (Wang and Strong, 1996), and subsequently progressed in multiple directions. Among the earliest was the concept of managing data as a product (Ballou et al. 1998), and data quality research has addressed the management of data quality, total data quality management (TDQM) (Wang et al., 1998), modeling the manufacture of data products (Shankaranarayanan et al., 2003), measurement of quality dimensions (e.g. Ballou and Pazer, 1996 and Shankaranayanan and Cai, 2006), using quality metadata for decision support (Fisher et al., 2003), economics of data quality (Even et al., 2007), and managing the quality of social media data (Shankaranarayanan et al., 2011). The above list is by no means exhaustive. We presented a preliminary version of our research at the AMCIS Conference in 2010, and since then a lot has changed over five years. With the growing importance of social media data and the explosion in the popularity of "data centric" analytics, we believe that this research area is at the threshold of a significant metamorphosis and explosion, especially given its potential applicability for the emerging concept of "Big Data" and the extensive use of social media data in decision-making today. Given this, we believe that it is critical to understand the current state of data quality research to assess trends and emerging opportunities. In this paper, we use the term "data quality", or "DQ" to include both data quality and information quality, as have other researchers whose work precedes ours.

The specific objectives of our research are: (1) to identify a clear set of research topics and themes to define the body of literature of data quality, (2) To identify associations (if any) between dimensions, one of the core concepts in data quality research, as well as the specific topics for which dimensions have been most rigorously studied and those where it has not, (3) To identify trends and additional analyses that can help us (4) recognize new and emerging opportunities for data quality research. This paper describes our work and results for the first.

The remainder of this paper is organized as follows. We first review prior work that summarizes and/or classifies data quality research to define the scope of this paper. We also describe the methodology we adopt for our analysis, Latent Semantic Analysis (LSA). We then describe the methodology and present the results, and lastly offer our conclusions together with directions for further research.

Relevant Literature

There have been numerous attempts to summarize the research on data quality. One of the first is the seminal work by Wang, Storey, and Firth (1995) that introduced data quality as a research topic. Since then, there have been studies to summarize, classify, and develop frameworks for DQ research such as those by Lima, Macada and Vargas (2006), Ge and Helfert (2007), Neely and Cook (2008), and Madnick, Wang, Lee, and Zhu (2009).

Wang, Storey, and Firth (1995) proposed a frameworks of data quality from a comprehensive analysis of publications through 1994 (1995). The authors compared data and data quality to a manufactured physical product and its quality and drew connections to managing data quality from established concepts for managing quality of physical products. Lima, Macada, and Vargas (2006) presented a summary of data quality research using articles published between 1995 and 2005. These derived relationships based on their judgment and intuition of the researchers to present a conceptual map of data quality.

Using a different perspective to classify data quality research, Ge and Helfert (2007) categorized the literature into that focusing on the assessment, management, and contextual aspects of data quality. A novel framework that combined the factors of "fitness for use" defined by Juran and Godfrey (2000) with the management elements defined by Wang et al. (1995) is offered by Neely and Cook (2008). Madnick, Wang, Lee, and Zhu (2009) used topics and methods to categorize data quality research and to develop a framework that allow researchers to characterize their own research. The authors defined research methods at different levels of hierarchy (treating some methods as subsets of others) and defined a hierarchy of topics and subtopics.

Each have these have defined a summary and categorization from their own point-of-view, and proposes its own different taxonomy. Although these offer useful insights into this area, each requires subjective judgements on the part of the researchers. We posit that there is a more interesting point-of-view that comes, not from the *researchers*, but from the *research* itself. Can literature tell us the core topics and the key themes within the research area? Can we understand what themes have risen to the forefront and the ones that are ebbing? Can we understand the evolution of research themes? Can we associate research topics with data quality dimensions? The summaries, classifications and frameworks proposed do not answer such questions.

As stated earlier, we believe that the research area is at a critical juncture. Up until 2010, data quality was considered somewhat anecdotal and esoteric. Today, because of the importance of data analytics and "big-data", the research area has witnessed an extraordinary growth in the last five years. This is a key motivating factor behind the work presented here.

Research Methodology

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique used to develop a semantic structure from a corpus of text documents. Based on Singular Value Decomposition (SVD) and bearing similarities to factor analysis, it is a "bag-of-words" approach in that it does not consider the order in which words appear, but rather their frequency. In many cases it is known to match human judgement quite well (Landauer et al. 1998). Details of the technique are provided by Deerwester et al. (1990).

LSA has been used to analyze literature for research topics in a range of fields, including the sciences (Stotesbury, 2003). In information systems, LSA has been used to analyze the field as a whole, notably in studies by Sidorova et al. (2008) and Evangelopoulos et al. (2012). Both provide the details of their approaches, and are what we follow for our work. LSA been also been used to analyze specific areas within information systems, such as for case-based research (Gordon et al. 2013). In the area of data quality, we earlier used LSA to develop our preliminary work (Blake, 2010) and a more detailed framework (Blake and Shankaranarayanan, 2010). This research is a follow-up and continuation of these two.

LSA has been compared with an alternate technique, Latent Dirichlet Allocation (LDA), by Zhang, Wu, and Huang (2014) who aimed at comparing the performance of LDA with LSA. The authors concluded that while the two algorithms were comparable at a high level, LDA was superior in identifying more detailed topics than LSA. We also plan to investigate LDA in future research.

Data Collection

Abstracts of articles on data quality were chosen from publications between 2000 and 2014. Abstracts were also added from conference proceedings. The two researchers each independently read the abstracts and determine relevance to data quality research and reach a consensus on the corpus through discussions. A complete list is provided in table 1. This list is growing as more work is published.

Journal/Conference Name	Count
International Conference on Information Quality	307
Americas Conference on Information Systems	76
International Journal on Information Quality	54
Journal of Data and Information Quality	30
European Conference on Information Systems	22
Information Quality in Information Systems	18
Decision Support Systems	15
Communications of ACM	9
Hawaii International Conference on Systems Sciences	9
Journal of Management Information Systems	7
Information and Management	6
Management Science	5
MIS Quarterly	5
Information Systems Research	4
Communications of AIS	4
General Category – Others	277
Total	848

Table 1a: Article count by Journal/Conference

Pre-2000	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
29	28	35	28	36	44	62	51	82	61	100	75	81	43	44	45	4

Table 1b: Article count by Year (00 = 2000)

Data Preparation

We applied several routine pre-processing steps to the texts of the abstracts in our corpus prior to applying LSA. First, we removed punctuation marks, special characters, and numeric values. In the second step, we removed stop words. After examining the remaining words, we removed additional words with little relevance to data quality such as “during”, “largely”, and “itself”. Finally, we excluded words consisting of less than three characters and words appearing less than five times throughout all abstracts. In the third step, we standardized frequently occurring phrases. The fourth step was to stem all words in the corpus. Stemming standardizes words having multiple variations with semantically equivalent meanings. Often these are words with the same root but with multiple suffixes.

Data Analyses

We analyze the prepared corpus of abstracts using LSA. LSA uses the context in which terms appear to measure term-to-term and document-to-document semantic similarities. We present our preliminary findings in the next section.

Preliminary Results and Conclusions

Results of the LSA were obtained for 5, 8, 10, 15, and 20 factor solutions. We have only presented our preliminary results for the 20, 10, and 5 factor solutions. For each factor solution, we examined the highest loading terms and the highest loading documents to interpret and label the factor. The labels were assigned by one researcher and confirmed/rejected by the other. In case of discord, the results were discussed and labeled in a manner that was satisfactory to both researchers.

The factors reported in Table 2 appear to have face validity. All of the key topics in data quality research appear to have been identified in the 20-factor solution. As we look at the reduced set of factors in the 10-factor solution, there is a clear merging of related topics that go into creating the factors in the 10-factor solutions. An approximate merging of factors from the 20-factor solution to the 10-factor solution is shown in Figure 1. Please note that the arrows are approximate for some factors in the 20-factor solution may have split when merging into 2 or more different factors in the 10-factor solution. For instance, 20F9 (Perceptual Assessment of DQ in Online Media) may have actually split when it merged with 10F9 and 10F5. We have only shown the one that we are able determine with confidence based on our preliminary analysis. We are still investigating these results in more depth and will be able to report better by the time the conference arrives. For instance, factor 10F5 (DQM in Online Data Exchanges) in the 10-factors solution emerges from the merge of 20F15 (DQM for Data Interchange), 20F9 (Perceptual Assessment of DQ in Online Media), 20F17 (DQM in Enterprise Systems), 20F11 (DQM in Supply Chains), and 20F13 (DQM for E-Commerce). This is not surprising considering that most of the enterprise systems are online and function across organizational and system boundaries. Other factors such as 10F3 (DQM for Healthcare) and 10F8 (Economics of DQ) appear to have stayed pure and relatively indivisible. Looking into the dimensions that are used in 10F9 may help us divide it further based on the dimensions examined in the articles.

20 Factor Solution	**10 Factor Solution**	**5 Factor Solution**
20F1 Quantitative Assessment Models for DQ	10F1 Strategies for DQM	5F1 Assessment of DQ
20F2 DQM in Data Mining	10F2 DQM in Databases	5F2 DQM in Databases
20F3 DQM for the Web	10F3 DQM for Healthcare	5F3 DQM for Web and Social Media
20F4 DQM for Decision Support	10F4 Representation Models for DQM	5F4 DQM for Decision Support
20F5 Trust and Risk Perceptions of DQ Online	10F5 DQM in Online Data Exchanges	5F5 DQM for Enterprise (and Systems)
20F6 DQM for Databases and Large Datasets	10F6 DQM for Decision Support	
20F7 DQM for Sensor Networks	10F7 DQM in Sensor Networks	
20F8 Economics of Data Quality	10F8 Economics of DQ	
20F9 Perceptual Assessment of DQ in Online and Social Media	10F9 Assessment of DQ	
20F10 DQ Metadata and Master Data Management	10F10 DQM for Web and Social Media	
20F11 DQM in Supply Chains		
20F12 Entity Matching solutions for DQM		
20F13 DQM for E-Commerce		
20F14 Representation Models for DQM		
20F15 DQM for Data Interchanges (Project Management)		
20F16 DQM in Healthcare		
20F17 DQM in Enterprise Systems (ERP)		
20F18 Process Oriented DQM		
20F19 DQM in Data Warehouses		
20F20 Architectural and Web-Service Models for DQM		

Table 2: Factor Labels for the 20, 10, and 5-Factor Solutions

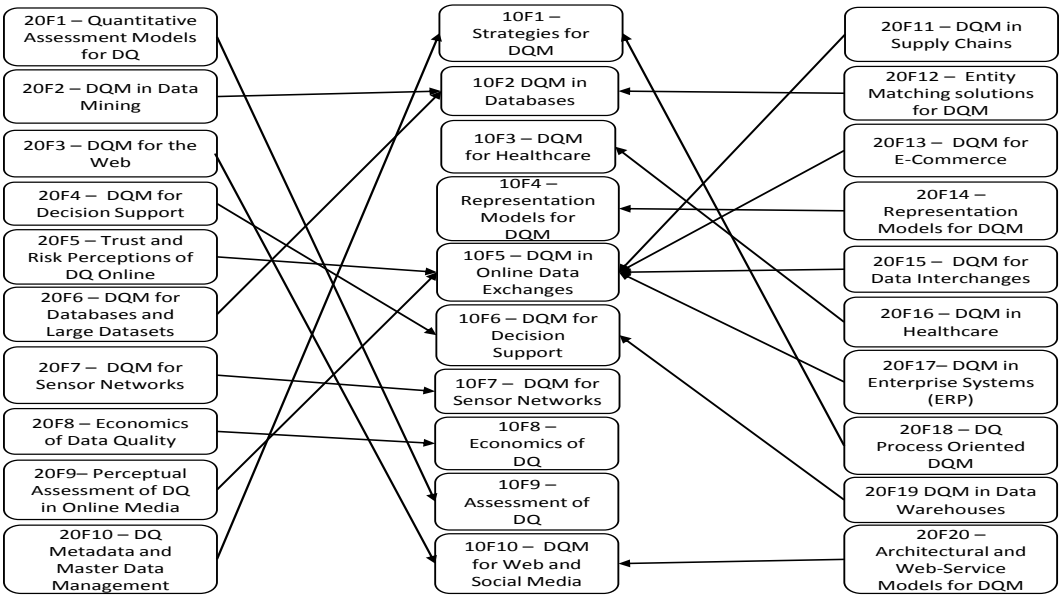

Figure 1: Approximate Mapping between the 20 and 10 Factor Solutions

Figure 2 describes how the 10-factor solution merges to create the 5-factor solution based on our preliminary analysis. The emergent 5 factors appear, reasonably consistent with the most recent research (Zhang et al. 2014). The authors identify "Data Quality Assessment", "Management of Data Quality", "Impact of Data Quality at the Organizational Level", "Data Quality and Databases", "Impact of Data Quality on Decision Making", and "Data Quality Application Areas" as the six core topics. Looking at our 5-factor solution, four of our factors are very similar to their core topics. Our "DQM for Enterprise and Systems" appears to encompass both DQM in organizations and applications of DQM.

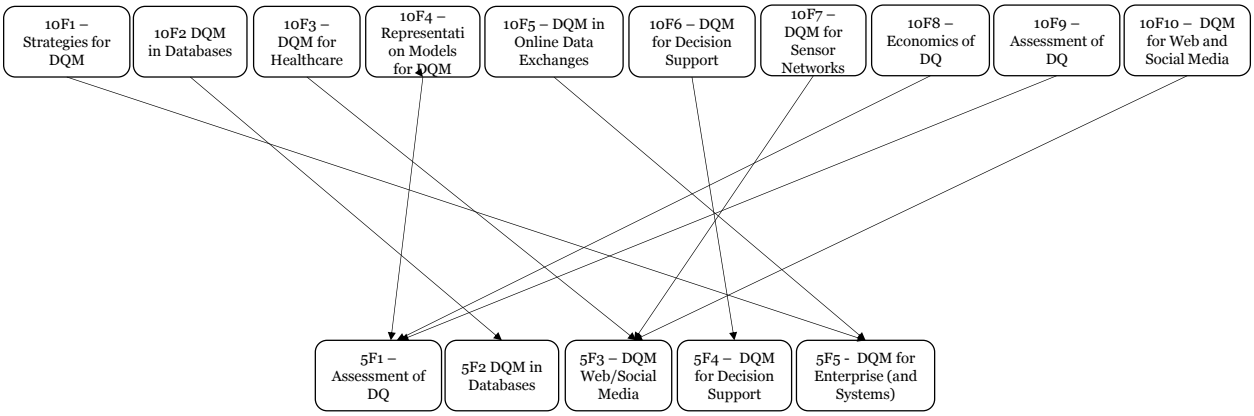

Figure 2: Approximate Mapping between the 10 and 5 Factor Solutions

In this paper we have presented a preliminary study to identify core topics and themes of data quality research. We identified the topics and themes by analyzing the texts of abstracts from almost 850 journal and conference articles published over the past 15+ years. We used Latent Semantic Analysis to measure term-to-term semantic similarity, and then used these similarity measures to load terms onto factors. We identified five core topics and twenty (20-factor solution) themes within based on the terms that loaded heavily on factors. We briefly compared the framework derived in this research with the results proposed by Zhang et al. (2014) for consistency of alignment. We are working on a mapping between data quality dimensions and themes. This will offer very interesting insights into the dimensions that are most well-researched, and those that have been relatively ignored. We are also extending this work to address the trends of research themes over the past 15+ years that will lead us to understanding new directions and opportunities in data quality research.

REFERENCES

Albert, S. and Whetten, D. 1985. "Organization identity," *Research on Organizational Behavior*, (7), pp. 263-295.

Ballou, D. P. and Pazer, H. L. 1985. "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff," *Information Systems Research*, (6).

Blake, R. 2010. "Identifying the core topics and themes in information quality research", in *Proceedings of the Americas Conference on Information Systems*, Lima, Peru.

Blake, R. and Shankaranarayanan, G. 2010. "Framing Data Quality Research: A Semantic Analysis Approach", in *Proceedings of the International Conference on Information Quality*, Little Rock, AR.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. 1990. "Indexing by latent semantic analysis," *Journal of the Society for Information Science* (41), pp. 391-407.

Evangelopoulos, N., Zhang, X., and Prybutok, V. 2012. "Latent Semantic Analysis: Five methodological recommendations, " *European Journal of Information Systems*, (21:1), pp. 70-86.

Even, A., Shankaranarayanan, G. and Berger, P. D. 2007. "Inequality in Utility of Data and its Implications for Data Management", in *Proceedings of the Seventeenth Annual Workshop on Information Technology and Systems (WITS 2007)*.

Fisher, C. W., Chengalur-Smith, I., & Ballou, D. P. 2003. "The impact of experience and time on the use of data quality information in decision making", *Information Systems Research* (14:2), pp. 170-188.

Ge, M. and Helfert, M. 2007. "A Review of Information Quality Research," in *Proceedings of the International Conference on Information Quality*, Cambridge, MA.

Gordon, S. R., Blake, R., & Shankaranarayanan, G. 2013. "Case-based research in information systems: Gaps and trends," *Journal of Information Technology Theory and Application (JITTA)*, 14(2).

Juran, J. M. and Godfrey, A. B. 2000. *Juran's Quality Handbook,* McGraw Hill International Editions: Industrial Engineering Series, 5th Edition.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. "Introduction to Latent Semantic Analysis," *Discourse Processes* (25), pp. 259-284.

Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y. 2006. *Journey to Data Quality*, The MIT Press, Cambridge, MA.

Lima, L., Maçada, G., and Vargas, L.M. 2006. "Research into information quality: A study of the state-of-the-art in IQ and its consolidation," in *Proceedings of the International Conference on Information Quality*, Cambridge, MA.

Madnick, S., Wang, R. Y., and Lee, Y. W. 2009. "Overview and Framework for Data and Information Quality Research," *ACM Journal of Information and Data Quality*, (1), pp. 1-22.

Neely, M. P. and Cook, J. 2008. "A Framework for Classification of the Data and Information Quality Literature and Preliminary Results (1996-2007)," in *Americas Conference on Information Systems (AMCIS)*, Toronto, CA.

Shankaranarayanan, G., Wang, R. Y. and Ziad, M. 2000. "IPMAP: Representing the Manufacture of an Information Product," in *Proceedings of the Information Quality Conference (IQ2000)*, Boston, MA.

Shankaranarayanan, G. and Cai, Y. 2006. "Supporting Data Quality Management in Decision Making," *Decision Support Systems* (42), pp. 302-317.

Shankaranarayanan, G., Iyer, B. and Stoddard, D. 2012. "Quality of Social Media Data and Implications of Social Media for Data Quality," in *Proceedings of the 17th International Conference on Information Quality (ICIQ – 2012)*, Paris, France, pp. 311-325.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T., "Uncovering the intellectual core of the information systems discipline," *MIS Quarterly*, vol. 32, 2008, pp. 467-482.

Stotesbury, H. 2003. "Evaluation in research article abstracts in the narrative and hard sciences," *Journal of English for Academic Purposes* (2), pp. 327-341.

Wang, R. Y., Storey, V. C., and Firth, P. 1995. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering* (7), pp. 623-640.

Wang, R. Y. and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Consumers," *Journal of Management Information Systems* (12), 1996, pp. 5-34.

Zhang, T., Wu, Y., and Huang, W. 2014. "Comparison of LSA and LDA in Data Quality/Information Quality Research," in *Proceedings of the 19th International Conference on Information Quality (ICIQ – 2014)*, Xi'an, China, pp. 99-112.