Identification of Data Patterns using Data Quality Techniques

Naveenkumar SR[1], Dr. G Mahadevan[2] and R.Vanitha[3]

[1]Research Scholar of Ph.D., Department of Computer Science, Rayalaseema University, Kurnool, India

[2]Principal, Annai College of Engineering, Kumbakonam, India

[3]Assistant Professor, Annai College of Engineering, Kumbakonam, India

Abstract: *The aim of this paper is to spotlight problems in classification of Address data and customer details in unstructured data files, quality analysis and to debate potential research chance to attain high data quality inside a corporation. The review adopted systematic literature review methodology supported analysis articles printed in journals and conference proceedings. we tend to developed a review strategy supported specific themes like current analysis space in knowledge quality, essential dimensions in knowledge quality, knowledge quality management model and methodologies and data quality assessment strategies. Supported the review strategy, we tend to choose relevant analysis articles, extract and synthesis the knowledge to answer our analysis queries. The review highlights the advancement of knowledge quality analysis to correspond its real world application and discuss the accessible gap for future analysis.*

Research gap like organizations management, data quality impact towards the organization and information connected technical solutions for data quality dominated the first years of knowledge quality analysis. However, since the web is currently going down because the new data supply the rising of latest analysis areas like knowledge quality assessment for internet and massive data is inevitable. This review conjointly identifies and discusses essential knowledge quality dimensions in organization like knowledge completeness, consistency, accuracy and timeliness. We tend to conjointly compare and highlight gaps in knowledge quality management model and methodologies. Existing model and methodologies capabilities are restricted to the structured knowledge kind and limit its ability to assess data quality in internet and massive data. Finally, we tend to uncover accessible strategies in knowledge quality assessment and highlight its limitation for future analysis. This review is very important to spotlight and analyze limitation of existing knowledge quality analysis associated with the recent wants in data quality like unstructured data kind and massive data

Keywords: *Data Quality, Information Quality Management Model, Assessment strategies, Database*

1. Introduction

Achieving high information quality has become a very important part in managing data among a corporation. Possessing high information quality might facilitate a corporation to formulate higher business strategy and unveil business pattern for higher cognitive process. Failure in providing high information quality to the organization has brought varied problems like false call thanks to incorrect data, high value of operation and lack of client satisfaction. Moreover, the increasing numbers of information out there nowadays with unknown quality levels more challenges to optimally analyze and create use of data that are relevant to the organization. High data quality has been outlined as a knowledge that's appropriate use and ready to meet the purposed set by data user. This definition clearly instructed that quality of information extremely dependent to the context of data usage and synergies to the client desires, ability to use and skill to access knowledge. Thus, in knowledge quality assessment and improvement method, participation of information users and alternative data stakeholders that are involve throughout data entry, processing and knowledge analysis is very important. Numerous strategies are projected to assess data quality from the context of information users and alternative data stakeholders as well as exploitation survey, and form.

Researchers adopted surveys and form to gather needs and expectations from information user, information entry personnel and different data neutral. Adoption of surveys and questionnaires is very important so as to outline information quality attributes and data quality dimensions to attain high data quality among its context. Applied math ways like correlation analysis is then wont to determine correlation between attributes and classify

https://doi.org/10.17758/ERPUB5.ER09191005

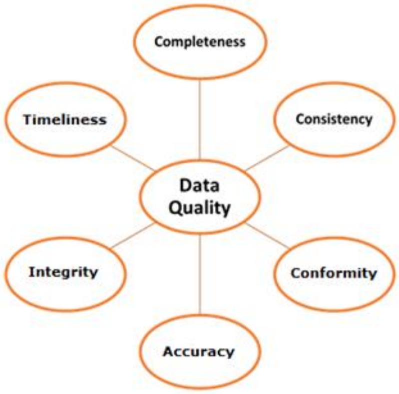
the attributes into information quality dimensions as an example data completeness, information consistency, information accuracy and data timeliness. On the opposite hand, findings of the analysis facilitate research worker to spot explanation for information quality issues and later, suggestion on information quality improvement is created. A lot of progress has been established in information quality analysis that isn’t solely restricted to the adoption of surveys and questionnaires as mentioned before. Thus, we have a tendency to urge to answer queries relating to progress in data quality analysis through a review of information quality research articles that has been revealed before this. Our main intention in doing this review is to spotlight potential problems in information quality analysis and to debate potential empty research gap in data quality research particularly in managing data quality among the organization. moreover, this review is meant to facilitate information quality implementation among the organization by discussing the strengths and weaknesses of existing data quality management model and data quality assessment ways.

The discussion provided during this review is restricted solely to the present proposals in information quality analysis which will be directly enforced in any organization with specific functions like business. In doing this study, we have a tendency to excluded analysis articles that projected activity metrics for information quality and data quality framework while not correct assessment technique. By doing this, we have a tendency to slender our discussion solely on analysis articles that projected information quality resolution that's appropriate for direct implementation in organization.

2. Research Methodology

This paper involves systematic literature review of information quality analysis in organizations. Supported the analysis queries, we have a tendency to known specific themes that are associated with knowledge quality analysis space, knowledge quality dimensions, knowledge quality management and data quality assessment strategies. Then, we have a tendency to construct our review strategy as a suggestion throughout analysis articles choice, information extraction and knowledge synthesis so as to answer the analysis queries. We have a tendency to limit our review to the analysis articles revealed in journals and conference proceedings.

Critical Dimensions of Data Quality to achieve success in business, you wish to form selections quick and supported the correct data. One among the necessary functions of a knowledge warehouse and enterprise business intelligence resolution is to produce users with a snap-shot of their business at any given purpose of your time. This permits call manufacturers to realize higher insight into their business and market in order that they'll creates elections quicker and better. While a business intelligence system makes it abundant easier to research and report on the information loaded into a knowledge warehouse system, the existence of information alone doesn't make sure that executives create selections smoothly the quality of the information is equally as necessary. Consider a high-level meeting to review company performance: if you learn that 2 reports compiled from purportedly similar set of information mirror two completely different revenue figures, nobody will grasp that figures are correct, that may cause necessary selections to be delayed whereas the “truth” is investigated. One of the causes of information quality problems is in supply data that's housed during a patchwork of operational systems and enterprise applications. Every of those knowledge sources will have scattered or misplaced values, noncurrent and duplicate records, and inconsistent (or undefined) knowledge standards and formats across customers, products, transactions, financials and additional. Data quality issues may arise once an enterprise consolidates knowledge throughout a merger or acquisition. However may be the most important contributor to data quality problems is that the information are being entered, edited, maintained, manipulated and reported on by folks. To maintain the accuracy and worth of the business-critical operational data that impact strategic decision-making, businesses ought to implement a knowledge quality strategy that embeds data quality techniques into their business processes and into their enterprise applications and data integration. On the surface, it's obvious that knowledge quality is concerning improvement up dangerous data – data that are missing, incorrect or invalid in how. However so as to make sure knowledge are trustworthy, it's necessary to know the key dimensions of information quality to assess however the information are “bad” within the1st place.



Conformity:

Conformity means that the information is following the set of normal data definitions like data sort, size and format. Instance, date of birth of client is within the format “mm/dd/yyyy” Questions you'll be able to raise yourself: Do information values be fits the desired formats? If therefore, do all the information values benefit those formats? Maintaining conformity to specific formats is very important.

Accuracy:

Accuracy is that the degree to that information properly reflects the $64000 world object OR an occurrence being delineated. Examples: Sales of the business unit are the 4000worth.Address of associate worker within the employee information is that the real address. Questions you'll be able to raise yourself: Do information objects accurately represent the “real world” values they're expected to model? Are there incorrect spellings of product or person names, addresses, and even untimely or not current data? These problems will impact operational and advanced analytics applications

Consistency:

Consistency means that knowledge across all systems reflects a similar data and is in synch with one another across the enterprise. Examples: A business unit standing is closed however there are sales for that business unit. Employee standing is terminated however pay status is active.

Completeness:

Completeness is outlined obviously comprehensiveness. Knowledge is complete obligatory data is missing. As long because the data meets the expectations then the information taken into account complete.

Consistency:

Consistency means that knowledge across all systems reflects a similar data and is in synch with one another across the enterprise. Examples: A business unit standing is closed however there are sales for that business unit. Employee standing is terminated however pay status is active.

Systematic Management of Data Quality in Organization.

Managing information quality dimensions and raising these dimensions through a scientific method are necessary to make sure high data quality among the organization. For this reason, numerous researches are done to propose a model and methodologies for systematic information quality management. a complete information Quality Management (CIQM) [4] has been projected earlier to support the conception of ‘data as a product’. With this idea, high information quality may be achieved by replicating physical production of prime quality product. CIQM extended Total Quality Management (TQM) framework that employed in physical production. The methodologies begin with the definition of knowledge product (IP). At this stage, the science has its own characteristics and needs so as to realize prime quality state. Then, the data quality (IQ) metrics is developed and wont to live the science. The measuring results then analyzed victimization applied math method management, pattern recognition and Pareto chart. Lastly, improvement being created on the science victimization in for producing Analysis Matrix supported the analysis done before. The supply of varied tools for result analysis as mentioned before helps organization to implement CIQM. However, many arguments occur once comparison information production to physical production. These included the power of knowledge to be shared among user. Secondly, information might not arrive in time once required and it's troublesome to assign many quality dimensions like quality to physical production. CIQM was designed for managing information quality in databases and current technologies together with massive data could limit the usage. This can be because of the range of knowledge varieties out there in massive data. Future work is feasible to revamp the framework by together with alternative information sources in massive data.

Data Integrity Methodology (DIM) [30] has been planned later and printed the requirements to attain information integrity by addressing the inspiration of knowledge itself. Data integrity thought of information ability to fulfill organizations strategic objectives. However, to attain the quality of knowledge, a framework of data integrity ought to be fulfilled. The framework enclosed information policy, organization capabilities, information administration, design, process, validation, communication and framework compliance. On the opposite hand, the planned methodologies supplemental another innovate information quality management that is to reassure data quality once the development method being created.

AIM Quality (AIMQ) model includes of Product and repair Performance Model for data Quality (PSP/IQ), IQA instrument to live information quality and data quality gap analysis technique to boost the data quality. Form is employed during this model to assess data quality. Additional applied math analysis is then getting used to spot data quality drawback space. The used of PSP/IQ is because of the target of achieving top quality data guided by the size attributes particularly, intrinsic, objective and accessibility.

Another example of knowledge quality management model is information Quality Management Maturity Model (IQMMM). The inspiration of this model is to boost organization quality and because the consequence it might give top quality of knowledge. During this model, structure of integrated databases being managed by standardizing its information. Standardization of information is divided into many stages like logical, physical and mapping information data. Alternative information quality management model and methodologies mentioned before doesn't manage data quality throughout the mixing of varied databases across the organization. This model stressed the wants of information integration so as to reinforce data accuracy and consistency. Moreover, its ability to confirm top quality of information throughout information integration area in additional price.

Many researches tired information quality target structured data sort compared to different styles of data like semi-structured and unstructured data. However, during this review, we tend to found many models that are appropriate for either structured, semi structured or unstructured information sort.

One in every of them is that the Complete information Quality Management (CIQM). CIQM counsel theoretical, empirical and intuitive approach to assess information quality [26]. It comprised of 3 stages; state reconstruction, assessment and selection of best improvement method. The advantage of CDQM is that the flexibility of the methodology to support structured, unstructured and semi-structured variety of information. However, there's no outlined activity methodology or calculation to live information quality dimensions in CDQM. Thus, the implementation of CDQM within the organization is troublesome. We tend to summarize the strengths and weaknesses of the information quality model and methodology in Table a pair of supported characteristics found within the literature.

As represented in every of information quality management model, information quality levels have to be compelled to be live and assess before additional analysis is done. Following section can mentioned information quality assessment methodology found throughout this review.

References

[1] S. W. Tee, P.L. Bowen, P. Doyle, F.H. Rohde, "Factors influencing organizations to improve data quality in their information systems," Accounting & Finance, vol. 47, pp. 335-355, 2007.

[2] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, "Methodologiesfor data quality assessment and improvement,” ACM Computing Surveys (CSUR), vol. 41, p. 16, 2009.

[3] W. Eckerson, "Data Warehousing Special Report: Data quality andthe bottom line," Applications Development Trends May, 2002.

[4] Y.Y.R. Wang, R.Y. Wang, M. Ziad, Y.W. Lee, Data quality vol. 23: Springer, 2001.

[5] F. Casati, M.C. Shan, M. Sayal, "Investigating business processes,"ed: Google Patents, 2009.

[6] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, C.Batini, "Managing data quality in cooperative information systems," On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, pp. 486-502, 2002.

[7] C. Batini and M. Scannapieca, Data quality: Concepts, methodologiesand techniques: Springer-Verlag New York Inc, 2006.

[8] V. Peralta, "Data quality evaluation in data integration systems,"Université de Versailles (chair) Raúl RUGGIA Professor, Universidad de la República, Uruguay, 2008.

[9] F. G. Alizamini, M.M. Pedram, M. Alishahi, K. Badie, "Data qualityimprovement using fuzzy association rules," 2010, pp. V1-468-V1-472.

[10] Y. Man, L. Wei, H. Gang, G. Juntao, "A noval data qualitycontrolling and assessing model based on rules," 2010, pp. 29-32.

[11] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions inontological foundations," Communications of the ACM, vol. 39, pp.86-95, 1996.

[12] KQ. Wang, SR. Tong, L. Roucoules, B. Eynard, “Analysis of data quality and information quality problems in digital manufacturing," 2008, pp. 439-443.