CERTIFICATE

This certificate is awarded to

Agung Hadhiatma

for the contribution as Presenter of

Improving Data Quality in the Linked Open Data: A Survey

International Conference on Computing and Applied Informatics (ICCAI) 2017
"Empowering the Society through Information Technology, Computational Science, and Engineering Research"

Organized by
FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITAS SUMATERA UTARA

Medan (Indonesia), 29 - 30 November 2017

Drs. Mahyuddin K.M. Nasution, M.IT.,Ph.D.
Vice Rector for Research, Community Service, and Cooperation
Universitas Sumatera Utara

PROCEEDING PUBLISHER :
IOP Conference Series conferenceseries.iop.org IOP Publishing

Dr. Erna Budhiarti Nababan, M.IT.
Chairman ICCAI 2017

Improving data quality in the linked open data: a survey

IOP Conf. Series: Journal of Physics: Conf. Series **978** (2018) 012026 doi:10.1088/1742-6596/978/1/012026

Improving data quality in the linked open data: a survey

A Hadhiatma[1]

[1]Department of Informatics, Sanata Dharma University, Yogyakarta, Indonesia

Email: agunghad@usd.ac.id.

Abstract. The Linked Open Data (LOD) is "web of data", a different paradigm from "web of document" commonly used today. However, the huge LOD still suffers from data quality problems such as completeness, consistency, and accuracy. Data quality problems relate to designing effective methods both to manage and to retrieve information at various data quality levels. Based on review from papers and journals, addressing data quality requires some standards functioning to (1) identification of data quality problems, (2) assessment of data quality for a given context, and (3) correction of data quality problems. However, mostly the methods and strategies dealing with the LOD data quality were not as an integrative approach. Hence, based on those standards and an integrative approach, there are opportunities to improve the LOD data quality in the term of incompleteness, inaccuracy and inconsistency, considering to its schema and ontology, namely ontology refinement. Moreover, the term of the ontology refinement means that it copes not only to improve data quality but also to enrich the LOD. Therefore, it needs (1) a standard for data quality assessment and evaluation which is more appropriate to the LOD; (2) a framework of methods based on statistical relational learning that can improve the correction of data quality problems as well as enrich the LOD.

1. Introduction

The Linked Open Data (LOD) project is the implementation of the semantic web idea [1], shifting representation from openly "web of document" to "web of data". The LOD can be derived from either structured [2], semi-structured [3] or unstructured data format [4]. The LOD, which is developed, linked and integrated from diverse data sources, has been evolving to be open, large distributed knowledge representation. As a consequence, it results several major issues, e.g. data quality, data interoperability, data management and scalability. One of these challenging issues in the LOD is about data quality. In general, data quality is defined as the fitness for use [5]. Several researches, e.g. [6, 7, 8], indicate that the LOD still inherits with data quality problems especially in terms of completeness, consistency, and accuracy.

Data quality in the semantic web can be classified into dataset quality [9], link quality [10] and schema quality [11].

- Dataset quality issues show that firstly, dataset or instances may be incorrect (noisy data), and incomplete (missing data) and secondly, regarding to their relations, datasets may be incorrect, incomplete, and inconsistent
- Link quality refers to the measured level that links between instances are not missing and weather every link is useful, and appropriate or not. Interlinking in the LOD means that dataset triples in Resource Description Framework (RDF), consisting a various of Uniform Resource Identifiers (URIs), form links from an entity as a subject to an entity as an object.

- Schema quality consists of completeness and consistency. Completeness of schema is the level that entities, attributes, classes, links are not missing regarding to their schema. Consistency of schema is the degree that entities, attributes, classes and links are coherent and free of logical conflict or contradiction with respect to their schema. In the LOD, a schema is represented by its ontology.

The challenges of data quality in the LOD are how to efficiently manage and to effectively analyse data and knowledge in big volume, since managing big data becomes more complex (that data is expanding massively, is mostly unstructured format as distributed graph, is controlled autonomously and has various levels of quality).

Dealing data quality means to address of some issues e.g. assessing information in various data quality levels, detecting errors, finding missing data, and improving data quality. Hence, the issues of data quality correspond to establishing methods and models to manage and to retrieve information. The poorer the quality of data, the worse the information can be retrieved. In the other side, according to [12],the growing volume of distributed semantic web leads to the more complex distributed graph and consequently becomes the greater potential information (hidden information or implicit knowledge) to be extracted but the harder to be derived.

2. Improving data quality in semantic web (the LOD): strategies and methods.

As the use of semantic web has been growing fast, it is necessary and challenging to deal with data quality on the semantic web. Consistency, accuracy and completeness of the semantic web have been recognized as an important problem in the recent decades. Searching and retrieving information in various data quality levels in semantic web may produce inaccurate or irrelevant results.

There have been many efforts to deal with consistency, accuracy and completeness in the Semantic Web. The author in [13] developed a model to address imprecise or uncertain data in order to improve recall and to achieve a better performance matrix in information retrieval. Several authors used logical reasoning model to predict missing data. Based on ontological background knowledge and a dataset of statements, logical reasoning can find inconsistency and infer new statements as predictions of missing data. Semantic web has framework including both syntax and model-theoretic semantics and inherits Description Logics, a knowledge representation supporting for logical reasoning. Logical reasoning plays an important role on semantic web. But logical reasoning methods have two main drawbacks. Firstly, for computation, it is hard to follow the growing of the web size [14]. Secondly, it is difficult to infer information on noisy, probabilistic and uncomplete data, identified in [15]. To deal with uncertainty in semantic web, many researchers have proposed OWL (Ontology Web Language, a family of knowledge representation languages) combined with special mathematical approaches, such as fuzzy extensions [16]. With extended OWL, the new approach of reasoning methods and algorithms have been specifically developed by[17] to derive relevant and unambiguous information from uncertain and unorganized data, but they still have some limitations.

Meanwhile, in many applications of data mining researches, successful approaches to predict missing data can be achieved by inductive learning from the huge data. Then, for the web search engines and applications, there have been increasing researches combining two research areas: web and data mining, well known as web mining. The author [18] successfully used inductive learning methods for data exploration on semantic web, namely semantic web mining. Subsequently, several approaches followed to use inductive learning methods for predicting data and finding missing data in the area of semantic web; there are listed as follow: Paulheim [4] proposed specifically association rule algorithm to detect similar co-occurrence patterns between types on DBpedia dataset. The authors in [19] employed k-nearest neighbours for predicting types, based on feature input of class and related resources. Sleeman and Finin [20] exploited a labelled training set to predict the type of instances. For inductive learning methods, it can be stated that several past studies tried to detect patterns and find missing information using the concepts including machine learning, statistical methods and exploitation of external knowledge.

Traditional data mining explores hidden knowledge by utilizing vector data input with independent variables. All data in the form of a vector input look like a single table, which feds to a machine learning for deriving hidden information or finding missing data. However, if the such machine learning try to search patterns that it need relational tables or graph as inputs (for instance web and linked data), the accuracy of learning would reduce and even be unreliable. These causes may be that in the traditional data mining, the machine learning does not consider about information of linkages and relationship, and it is hard to utilize rich information of a graph structure or schema. Unlike traditional data mining algorithms which find for patterns in a single table, relational data mining algorithms search for patterns among multiple relational tables. Considering the graph structure of multiple relational tables, relational learning algorithms will extract more rich information than learning on a single table. The most extensively applicable method and research for relational learning is Inductive Logic Programming (ILP) [21]. ILP is a relational learning approach which constitutes intersection of inductive learning and logic programming, employing concepts and methods from both machine learning and logic programming. However, ILP works well with deterministic structured data, but not with probabilistic structured data.

To handle probabilistic data, the ILP and the statistical machine learning concept are combined, involving the concept of machine learning, probability theory, and logic, which trigger to the emerging research area, namely Statistical Relational Learning (SRL). SRL methods and techniques evolve to represent, reason, and learn on data or knowledge with complex relational and probabilistic structure [22]. Indeed, SRL functionally works on interlinked, sparse, noisy, incomplete, and conflicting data. Researches in [23,24] resumed that SRL can operate in semantic web for extracting hidden information and finding missing data. Several SRL representations, frameworks and models have been proposed, for examples, Probabilistic Relational Models, Relational Dependency Network, Bayesian Logic Programming, and Markov Logic Networks. Among SRL models, we focus on Markov Logic Networks. Markov logic networks (MLNs) are a statistical relational model that generalizes weight first-order logic and Markov Networks [46]. MLN leaning consists of 3 main components: structural learning, weight learning, and Maximum a Posteriori (MAP) inference. MLN structure learning for huge data is hard and challenging. Structure learning should repeatedly learn the structure of graph or schema and weight learning process computes the weight of the structure with expensive statistical inference (NP-Hard MAP inference). Both structure and weight learning work to evaluate its relational structure. Several researches have tried to improve MLN computation performances of weight and structure learning. The author [25] conducted efforts to reduce computation load of the pseudo log likelihood and its gradient in the Weighted Pseudo Log Likelihood (WPLL) evaluation function. The authors [26] proposed Markov Logic Structure Learner (MSL) model for learning data structure by optimizing a likelihood-type measure.

Meanwhile, the LOD is incorporated with its schemas and ontologies, aiming both to function machine data interpretation and to enhance data accessibility. Ontology learning can supports for semi or automatic support ontology construction [11] and can partially replace the hard effort of manual ontology construction [27]. Those researches show that recent ontology learning framework have attempted to establish ontology by learning from various types of data sets for examples: database, text, html, and xml, but they seldom learning and exploring from the LOD.

Research by [28] proposed Statistical Relational Learning (SRL) methods to inductively learn ontologies from a repository containing noisy data. Similarly, SRL can learn Ontology Web Language (OWL) axioms from the Linked Data which has probabilistic unstructured graph representation. For an example, Zhu [29] suggested ontology learning approach that Description Logic and Bayesian Network are incorporated to deal with incomplete data in semantic web for constructing schema axioms from data axioms. Exploring the LOD by relational learning with especially focusing to its ontology is challenging and necessary because of some reasons. According to Buhmann [30], in the LOD, there are still less datasets which are structured and referred to their ontologies. The LOD research community still put less attention to the ontological layer and the correspondences between

datasets and their ontology. Moreover, Ontology expressiveness can still be enhanced for more useful and powerful applications.

Only few researches used some guidelines as data quality assessment methodologies for improving data quality in the LOD, although several data quality assessment methodologies have been proposed. Research in [31] presented a systematic review of data quality assessment methodologies especially to the LOD. The review has resumed several assessment approaches, extracted a number of data quality dimensions and investigated relevant methods and tools. This research also showed that the flexibility of the tools with regard to enhance the level of automation still needs improvement. Several researchers also proposed data quality assessment and evaluation particularly in Semantic Web [32][33] and methods for detecting error presented by [34,35,36]. Furber and Hepp[32] presented Semantic Web Information Quality Assessment (SWIQA), a methodology that adopts rule templates to represents quality requirements. The rule templates are automatically used to detect error as well as to measure quality scores for five dimensions.

3. Discussion

From the literature review above, we have identified some potential issues as a background to define objectives of our next research, as below:

Schema and ontology quality is a degree to which entities, attributes, classes, links are complete, accurate, and consistent or coherent in the integration context. On the other hand, according to Zhu [35 29], the LOD is still lack of expressive ontologies, and instance datasets based on their ontologies. The LOD itself consists of mainly instance data, but still less representing data to ontological layer. Based on these facts, we intend to investigate some main research questions. Firstly, can ontology expressiveness be enhanced from data underlying on it? Secondly, does ontology expressiveness become a significance factor to improve data quality? Finally, how efficiently detect and curate incomplete, inaccurate, inconsistent data and links relating to other data, their schema and ontology in a context of integrated data? Hence, it may be significant that dealing with data quality should be corresponded to schema and ontology of that LOD.

For detecting and curating data quality, Zaveri [31] presented a comprehensive systematic review of data quality assessment methodologies and tools (methods) applied to the LOD. He showed that enhancing flexibility of the data quality tools still requires much effort and open discussion. The flexibility regards subsequent works to increase the level of automation and minimize user involvement. In addition, some tools are hard to configure and require a considerable amount of setting. Although the tools are simple to use, but they still work with limited scope or need much human interpretation. Also, in general, those other previous papers presents that the LOD quality is an essential but efforts to standardize data quality tracking and assurance are still less and how the data quality dimensions and metrics should be defined is still open discussion.

We suggest that firstly, to effectively and efficiently improve data quality needs the assessment methodologies as a guide and secondly, our next research should focus on dealing data quality in the LOD, considering relational data, schema and ontology in integrative context. Therefore, the main purposes of the future research are to achieve:

- A selected or modified standard for data quality assessment and evaluation which is more appropriate to facilitate the identification of the data quality problems and the assessment of data quality for an integration context
- A framework of methods based on statistical relational learning that can improve the correction of data quality problems in the LOD through ontology learning. The measured performances for handling data quality in huge LOD by statistical relational learning are learning time, level of learning automation (flexibility), ontology expressiveness, scalability, and number of data and links which can be repaired.

Using existing data quality assessment methodologies, we would identify to which data quality problems in the LOD (in the terms of completeness, consistency, and accuracy) correspond to their

own relational data, schema and ontology in integration context. Also, we would study which data quality assessment methodology is appropriate to asses these problems or whether the methodology requires to be updated.

For improving data quality, we are interested to focus on the concept of ontology enrichment and refinement. Ontology enrichment techniques are a process to generate and add data and link extensions from and to the linked data which is bounded with its ontology. The extensions can include new entities, concepts and relations. Meanwhile, ontology refinement relates to improving and correcting an existing ontology. Ontology refinement process consist of several tasks such as updating, adding, deleting axioms and links in existing ontology, aiming to make semantic web to be more detailed and relevant. Ontology enrichment and refinement can be established through ontology learning.

Then, we would propose several strategies to address data quality problems which have been identified by a suitable data quality assessment. Firstly, improving data quality in the LOD, considering in its schema and ontology, can be approached by ontology learning namely ontology refinement. Secondly, enhancing ontology expressiveness in the LOD may be addressed by ontology learning namely ontology enrichment. Finally, by referring to [23, 24], it is believed that relational learning such as Markov Logic Network (MLN) may be adjusted as a tool to handle ontology refinement and enrichment. The existing RDF data and ontology in the LOD can be enriched and refined into a new one, which aims to improve data quality as well as to enhance ontology expressiveness.

Based on the strategies, we would propose a new framework of methods to improve data quality as well as learning performances. In the framework, the detected data quality problems are then repaired by ontology learning, adopting some techniques/methods of relational machine learning. Subsequently, these learning performances in some variables are measured and compared. Then, based on evaluation of some chosen techniques/methods in relational machine learning, we would propose two main steps to develop a new framework, as presented below:

- Studying and developing suitable techniques of the relational learning to address data quality problems in integrative context: Relational learning consists of structure learning and weight learning. To deal with relational learning in the LOD is a challenging task because structure learning should search in a huge search space and a weight learning process requires computationally expensive statistical inference[37][26].
- Establishing certain domain knowledge and rules as a background knowledge represented by ontology: a background knowledge is incorporated into a relational machine learning aims to enhance a level of automation, control learning process, improve effective learning, and constrain learning spaces. In an addition, background knowledge represented as ontology would support relational machine learning in logical reasoning. Background knowledge could be captured from LOD itself or others.

The result of Framework would be evaluated by a selected assessment and evaluation methodology. Weight learning and structural learning performance can be measured using average conditional log-likelihood of test atoms (CLL) and average area under the precision-recall curve (AUC).

4. Conclusion

Most previous researches have conducted research to improve data quality not as considering an integrative view but as a partial approach. Considering integrative view, it is opportunity to improve data quality in the LOD in the term of incompleteness, inaccuracy and inconsistency referring to its schema and ontology through ontology learning, approached with relational leaning. Relational learning utilizes the combination of machine learning, probability theory, and logic. In addition, relational learning in big data should consider huge searching space for its structural learning and computationally expensive statistical inference for its weight learning.

Investigating the characteristic of data quality in the LOD is by adopting appropriate assessment methodology, purposed to develop mechanism and knowledge to detect and evaluate the quality dimension/criteria. The mechanism and the knowledge are used formally to curate and enrich dataset and links. Alternative methods to curate and enrich dataset and links could be established by modified relational learning. Existing RDF data and underlying ontology could be enriched and refined into a new one which aims to improve data quality as well as to enhance ontology expressiveness. The performances of a method in relational learning for improving the LOD quality are measured by a number of data and links which can be improved as well as enriched, learning time, a level of learning automation, ontology expressiveness, and scalability.

We highlight that the quality of the LOD includes a number of novel aspect such as coherence and consistency with regard to the ontology and the schema in the context of integrated data. However, there are still research opportunities to formulize how the data quality dimensions to be determined and how data quality tracking, assessment, and assurance to be implemented.

References

[1] Augenstein I, Pado S and Rudolph S 2012 LODifier: Generating linked data from unstructured text. *In Proc.Conf on Extended Semantic Web Conference*(Greece).
[2] Bizer C and Cyganiak R 2006 D2R server-publishing relational databases on the semantic web. *Proc.Conf on The 5thInternational Semantic Web*.
[3] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P N, Hellmann S, Morsey M, Kleef P V, Auer S and Bizer C 2014 DBpedia-A large-scale multilingual knowledge base extracted from wikipedia *J. Sem. Web*.
[4] Paulheim H 2012 Browsing linked open data with auto complete *Proc Semantic Web Challenge* (Sydney).
[5] Knight S and Burn J 2005 Developing a framework for assessing information quality on the world wide web *J. Inf. Sci* **8** pp 159-172.
[6] Furber C and Hepp M 2010 Using semantic web resources for data quality management *Proc Int.Conf on The 17 the Knowledge Engineering and Knowledge Management* vol 6317 (Lisbon : Springer) pp 211-225.
[7] Furber C and Hepp M 2010 Using SPARQL and SPIN for data quality management on the semantic web *Proc.Int.Conf on The 13th Business Information Systems* vol LNBIP 47 (Berlin:Springer) pp 35-46.
[8] Hogan A, Harth A, Passant A, Decker S and Polleres A 2010 Weaving the pedantic web *Proc.Int.Works on The 3rd Linked Data on the Web*.
[9] Lei Y, Nikolov A, Uren V and Motta E 2007 Detecting quality problems in Semantic Metadata without the presence of a gold standard *Proc.Works on Evaluation of Ontologies for the Web at the WWW EON* pp 51-60.
[10] Guert C, Groth P, Stadler C and Lehmann L 2012 Assessing linked data mappings using network measures *Proc on The 9th Extended Semantic Web*(Crete).
[11] Furber C and Hepp M 2011 SWIGA: A semantic web information quality assessment framework. *Proc.Int on European Information Systems* (Helsinki).
[12] Singh M P and Huhns M N 2005 *Service-Oriented Computing: Semantics, Processes, Agents*. Wiley, 1st edition.
[13] Paulheim H and Pan J Z 2012 Why the semantic web should become more imprecise *What will the Semantic Web look like 10 years from now*.
[14] The Large Knowledge Collider. EU FP 7 Large-Scale Integrating Project 2008 (online: http://larkc.eu/ LarKC)
[15] Ji Q, Gao Z and Huang Z 2011 Reasoning with Noisy Semantic *Data Proc.Int on the 8th Extended Semantic Web ESWC* (Crete).
[16] Straccia U 2005 Towards a fuzzy description logic for the semantic web *Proc .Conf on The 2nd European Semantic Web: The Semantic Web: Research and Applications* vol LNCS 3532

pp. 167–181 (Springer).
[17] Lukasiewicz T. Probabilistic default reasoning with conditional constraints 2002 *J. Annals of Math. and Artifi. Intell* 34(1-3) pp 35-88
[18] d'Amato C and Fanizzi N, Esposito F 2010 Inductive learning for the semantic web: What does it buy? *Proc.Int on the 8th Extended Semantic Web ESWC* (Crete) vol1 (Springer) pp 53-59.
[19] Giovanni A, Gangemi A, Presutti V and Ciancarini P 2012 Type Inference through the Analysis of Wikipedia Links *Proc.Works on Linked Data on the Web LDOW*(Lyon).
[20] Sleeman J and Finin T 2013 Type prediction for efficient co reference resolution in heterogeneous semantic graphs *Proc.Int on The 7th IEEE Semantic Computing* (California).
[21] Lavrac N and Dzeroski S 1994 *Inductive logic programming techniques and applications* (Ellis Horwood).
[22] Getoor L and Taskar B 2007 *Introduction to statistical relational learning* (The MIT Press).
[23] Stumme G, Hotho A and Berendt B 2006 Semantic web mining -state of the art and future directions *J. Web. Sem* vol 4 no 2 pp 124–143.
[24] Tresp V, Bundschus M, Rettingerand A and Huang Y 2008 Towards machine learning on the semantic web *Proc.Int.Works Uncertainty Reasoning for the Semantic Web URSW 2005-2007, Revised Selected and Invited Papers* (Springer) vol 5327 pp 282–314.
[25] Richardson M and Domingos P 2006 Markov logic network *J. Mach. Learn*.
[26] Kok S and Domingos P 2010 Learning markov logic network structure using structural motifs. *Proc.Int.Conf on the 27th Machine Learning* (Haifa : Omnipress).
[27] Maedche A and Staab S. 2001 Ontology learning for the semantic web 2001 J. IEEE Intell. Syst vol 16 no 2 pp 72-79.
[28] Zhu M, Gao Z, Pan J Z, Zhao Y, Xu Y and Quan Z 2013 Ontology learning from incomplete semantic web data by BelNet *Proc Int.Conf on The 25th IEEE Tools with Artificial Intelligence* (Herndon,USA).
[29] Zhu M 2011 DC proposal: Ontology learning from noisy linked data *Proc.Int.Conf on The 10th The Semantic Web* (Berlin: Springer-Verlag) part II pp 373-380.
[30] Buhmann L, Fleischacher D, Lehmann J, Melo A and Volker J 2014 Inductive lexical learning of class expressions *J. Knowl. Eng and Knowl. Man* **8676** 42-53.
[31] Zaveri A, Rulab A, Maurino A, Pietrobon R, Lehmann J, Auer S 2014 Quality Assessment for Linked Data: A Survey *J.Sem. Web*.
[32] Furber C and Hepp M 2011 SWIQA A semantic WEB information quality assessment framework. *Proc.Euro.Conf on Information Systems ECIS*.
[33] Bizer C and Cyganiak R 2009 Quality driven information filtering using the WIQA policy framework. *J. Web Semantics*.
[34] Tropper G, Knuth M and Sack H 2012 DBpedia ontology enrichment for inconsistency detection *Proc.Int on the 8th on Semantic Systems I-SEMANTICS*.
[35] Lehmann J and Buhmann L 2010 ORE- A tool for repairing and enriching knowledge bases. *Proc.Conf on the 9th International Semantic Web ISWC*(Shanghai).
[36] Paulheim H 2014 Identyfing wrong links between datasets by multi-dimensional outlier detection *Proc. Int. Works on Debugging Ontologies and Ontology Mappings* (USA).
[37] Kok S and Domingos P 2009 Learning markov logic network structure via hypergraph lifting *Proc.Int.Conf on the 26th Machine Learning* (Montreal :Omnipress) pp 505-512.