INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

How to Improve Total Data Quality Management (TDQM) Using the Semantic Technology

Dr. Ahmed Said El Rawas
Head, Department of Management, College of Management and Technology,
AASTMT, Egypt
Dr. Haytham T. Alfeel
Associate Professor, Department of Information Systems,
College of Computer Science and Information System, Al Faium University, Egypt

Abstract:
Total data quality management is considered one of the most popular research areas in the last decade, where improving the quality plays the main role in the targeted achievement for any enterprise. Also, there is a lot of emerging information technologies which aid in improving the quality in different aspects. Through this research we will focus on how to improve the quality of data using the most recent technology in IT fields which is named Semantic Web technology by conducting set of hypotheses which aim to prove whether there is a the relation between the improving of web data and the usage of semantic web techniques or not. Also, we will focus on how to improve the performance of data query processing specially when having huge amount of dataset using the big data technologies

Keywords: *Total data quality management, semantic technology, SPARQL, big data, hive, SPARK*

1. Introduction

Nowadays, there is a huge amount of data found in different web platforms worldwide, one of the most important data sources is the Wikipedia [1] which includes information about everything in our life. Semantic Web technology [2] became one of the most important techniques which used to increase the quality of data through the web be transforming the data format from unstructured data to structured data. Through this research we will focus on clarifying the rule of semantic web [3] in improving the quality of data using different mechanisms and we will show how to improve the performance of data by reducing the response time of querying this information through semantic web query language aided by big data tools. The rest of research will is organized as follow: Section 2 focuses on the literature review which may be similar to our methodology. Section 3 discusses our main proposed methodology for this research. Section 4 presents the analysis of the proposed architecture according to its implementation and the conducted results. Finally, section 5 concludes our paper and discusses the possible directions for future work.

2. Research Problem

One of the main problems that faced Wikipedia is that it represents its information in un-structured format, which leads to quality problem according to the mismanagement of this information such as the difficulty of processing this information or query it. Also, through this research we will focus on another problem which relates to the huge size of Wikipedia data, however the sematic web technology help us to improve the quality of data but the problem still found when we talk about large amount of data which leads to another quality problem such as the performance and the response time of querying.

3. Research Objectives

The main objective is how to get use of this huge amount of data and put it in a simple form to produce an accurate results using quality technique semantic technology.

4. Research Hypothesis

The research wills discus certain hypotheses such as it,

- There is a significant relationship between improving the Total data quality management using the semantic technology.
- There is a significant relationship between improving the performance of data query using the semantic technology in big data environment.

5. Literature Review

The field of information quality (IQ) has experienced significant advances during its relatively brief history. Today, researchers and practitioners alike have moved beyond establishing information quality as an important field to resolving IQ problems—problems ranging from IQ definition, measurement, analysis, and improvement to tools, methods, and processes**.** The purpose of this TDQM methodology is to deliver high quality information products (IP) to information consumers [14].

Total Data Quality Management (TDQM) as proposed by Wang [13] is one of the most prominent methodologies for managing data quality. Based on the principle that the quality of data needs to be managed similar to the management of product quality, it describes an adjusted lifecycle of quality management suitable for data.

Semantic data is widely available and the development and application of ontologies have been gaining big momentum in a range of application domains, such as government organizations, healthcare or media [16], several semantic groups have been building or contributing to the development of ontologies

6. Classification of Data Quality Principles in the Semantic Web

The main goal behind using Linked Data is to easily enable knowledge sharing and publishing. The basic assumption is that the usefulness of Linked data will increase if it is more interlinked with other data; Tim Berners-Lee defined 4 keys principles for publishing [17]:

- Make the data available on the web: assign URIs to identify things.
- Make the data machine readable: use HTTP URIs so that looking up these names is easy.
- Use publishing standards: when the lookup is done provide useful information using standards like RDF.
- Link your data: include links to other resources to enable users to discover more things.

By following these guidelines, a certain level of uniformity is achieved, which increases the usability of data. To fully leverage all the benefits of the Semantic Web, data quality principles in Semantic Web should embrace and adopt the guidelines for Linked Open Data. Building on these principles and based on our experience with powerful data integration software to extract, transform, and load data from applications, databases and other data sources, we have derived five principles for data quality in the Semantic Web. These principles are:

- Quality of data source: This principle is related to the availability of the data and the credibility of the data source.
- Quality of raw data: This principle is mainly related to the absence of duplicates, entry mistakes, and noise in the data.
- Quality of the semantic conversion: This principle is related to the transformation of raw data into rich data by using vocabularies.
- Quality of the linking process: This principle is related to the quality of links between two datasets.
- Global quality: This principle is cross-cutting the other principles and covers the source, raw data, semantic conversion, reasoning and links quality.

7. Methodology

Improving the quality of a huge amount of data which is represented in Wikipedia content according to our predefined hypotheses is considered the main contribution of our work. The research methodology is based on semantic technologies that help us to enrich the content of web and increase the efficiency and usability of this data by converting the unstructured data to structured form; the thing which converts the view to web content from just represented data to an intelligent one.

Through semantic web technologies users can use an intelligent query language called SPARQL [4] which allows them to perform more difficult and complicated queries about the sematic data. Hence the semantic projects such as DBpedia and Linked Open Data [5] began to improve the quality of web data by moving the unstructured data of Wikipedia to its accurate and structured form. Also, Linked open data (LOD) enhance the quality of web content by linking different resources together to represent the same entity the thing which increase the availability and querying time performance.
Also one of our data quality improvement contribution is launching the Arabic Chapter of DBpedia [6,7] at the beginning of 2017 which aims to improve the quality of Arabic content through the web against its English one as clarified in this special case where there are two different keywords have the same meaning through the same predicate: “birth Name” and “birth name”. Although they have the same meaning but in English chapter of DB pedia [8], the retrieved results are different results as shown in Figure.1 and Figure.2. But by using the new Arabic chapter we improved the quality of data by using the Arabic property "**تاريخالميلاد**"to retrieve the same results in anyway.

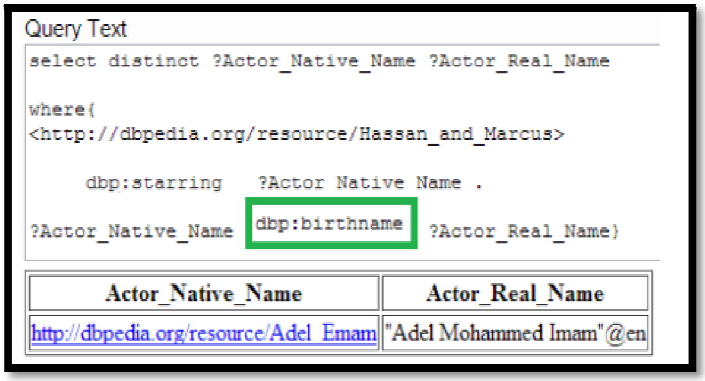

Figure 1: English SPARQL Query and Its Results Using Predicate "Birthname"

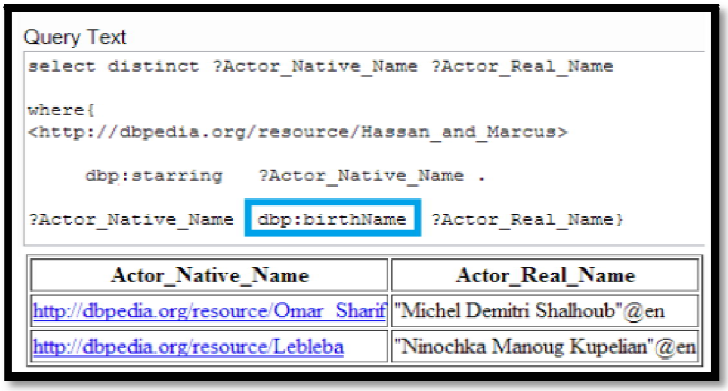

Figure 2: English SPARQL Query and Its Results Using Predicate "Birthname"

Also we can improve the quality of data in term of accuracy by perform a sophisticated queries using SPARQL which is not easy to be performed by any traditional query languages as shown in figure 3.

Figure 3: Sample of Sophisticated Arabic SPARQL Query about the Wife of MICROSOFT Owner

Secondly, we focus through this research of how to deal with the large amount of structured data using the new big data ecosystems platforms. So that our contribution here is to usea framework through which we can use big data [9] query languages over Hadoop [10] such as HiveQL [11]or SPARK SQL [12] instead of using a traditional semantic query language (SPARQL) to improve the quality of data by increasing the performance and query processing time aided by accurate results.

- In this framework, DBpedia datasets are converted into Comma-Separated Values (CSV) files instead of using its traditional format either RDF or N3 format.
- After that the converted CSV datasets are Loaded it in HDFS in Hadoop that we used here as a distributed file system. After that, Map-Reduce model [13] divides the massive datasets and performs parallel processing tasks on them.
- Then the hive ecosystem will be enabled over Hadoop platform to process the loaded data from HDFS by using its specific query language HiveQL.

Finally instead of using SPARQL as query language which will be able after this enhancement to perform any sophisticated query regarding to its huge size.Also we can improve the quality of retrieved data by using SPARK SQL as

recent query language of Big Data technologies instead of using HiveQL to prove that the quality of SPARK SQL query language Shark provide a better performance than Hive QL and hence SPARQL. So that it became the first choice to query a huge amount of semantic content as shown in Figure 4.

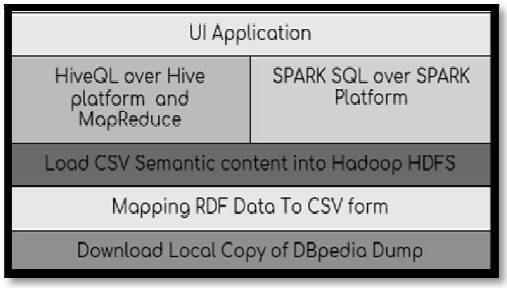

Figure 4: Proposed Framework

8. Implementation and Conducted Results

The proposed framework was delivered as a desktop application using Scala programming language to allow users to query any row of data stored in DB pedia datasets which are sized as 2 GBs in form of 4 million records. Physically, our practical cases needed to be evaluated through Intel Core I5 with 4 GBs memory and 160 GBs hard disks with Ubuntu 12.04 Linux operating system.

After performing our case studies we can prove our proposed hypothesis such as"to what extent the semantic web technology help in improving the quality of data?" and "How the using of Hive QL query language instead of SPARQL query language improve the quality of data in form of performance and accuracy in case of large amount of data?" We have set of test cases varying from simple to sophisticated queries that run using both SPARQL and HIVE-QL measuring the retrieval time of each query language for each test case as shown in figure 6.

Query Language	Qu 1	Qu 2	Qu 3	Qu 4	Qu 5
SPARQL	370 s	450 s	490 s	522 s	560 s
HiveQL	111 s	139 s	129 s	152 s	165 s
SPARK SQL	65 s	61 s	47 s	66 s	74 s

Figure 5: Query Execution Time for Different Queries Using Three Different Query Languages

Through this section we perform an evaluation study on our proposed framework to test the quality improvement by using the big data technologies and semantic web techniques to prove the truth of our hypotheses and vice versa. Our evaluation mythology depend on set of quality assurance factors such as availability, usability, stability, efficiency and performance of querying execution time as shown in figure 6.

Criterion	Availability	Stability	Efficiency	Usability	Performance
SPARQL	Full	Partial	Poor	Full	Poor
HiveQL	Full	Partial	Partial	Full	Partial
SPARK SQL	Full	Full	Full	Full	Full

Figure 6: Quality Evaluation Factors of Proposed Framework

Due to that, this methodology prove the truth of our hypotheses which lead us to that the semantic web has significant impact on improving the quality of web data which are suffered from non-usability due to set of problems such as un-structuring or heterogenetic problems.

9. Conclusions and Future work

Through this research we illustrated the impact of using information technologies such as semantic web technologies in improving the quality of data a converting it from unstructured format to structured one and hence increase the usability of this data in different aspect. Also, this research shows how to use semantic query language SPARQL to perform sophisticated queries which enable users to ask about more advanced information in accurate form. After that we clarified the important role of using big data techniques such as Hive and SPARK in improving the performance of data especially when users need to process huge amount of data as another aspect of quality management factors. Our future work will focus on how to improve the quality of data using artificial intellenegence and machine learning techniques to improve the data prediction and recommend the best scenarios for data usage.

10. References

i. Wikipedia, [Online]. Available: http://wikipedia.com [Accessed: 10- Oct- 2018].
ii. Open Link Virtuoso - Semantic Web Standards, 2015 [Online]. Available:https://www.w3.org/2001/sw/wiki/ Open Link Virtuoso [Accessed: 10- Dec- 2015].
iii. N. Shadbolt, T. Berners-Lee and W. Hall, 'The Semantic Web Revisited', IEEE Intel. Syst., vol.21, no. 3, pp. 96-101, 2006.
iv. N.Rakhmawati, J.Umbrich, M.Karnstedt, A. Hasnain, M.Hausenblas, 'Querying over Federated SPARQL Endpoints'A State of the Art Survey, pp.1306-1723, 2013.
v. L.Sikos, 'Linked Open Data.' In Mastering Structured Data on the Semantic Web, pp. 59-77, 2015
vi. H. Al-Feel, 'A Step towards the Arabic DBpedia', International Journal of ComputerApplications, vol. 80, no. 3, pp. 27-33, 2013.
vii. H. AL Feel, 'The Roadmap for the Arabic chapter of DBpedia', Wseas; 14th International Conference on Telecommunications and Informatics (TELE-INFO '15), in press, 2015
viii. C. Bizer, G. Kobilarov, S. Auer, C. Becker, J. Lehmann, R. Cyganiak and S. Hellmann,'DBpedia - A crystallization point for the Web of Data', Web Semantics: Science, Services andAgents on the World Wide Web, vol. 7, no. 3, pp. 154-165, 2009.
ix. S. LaValle. E. Lesser, R. Shockley, H. Rebecca, 'Big data, analytics and the path from insights to value', MIT, loan management review, vol.21, pp. 20–32, 2013
x. Apache Hadoop, 2018[Online]. Available: http://hadoop.apache.org/ [Accessed: 1-Oct- 2018]
xi. Apache Hive [Online]. Available: http://www.apache.org/hadoop/hive. [Accessed: 1-Oct- 2018]
xii. A. Lupher, 'Shark: SQL and Analytics with Cost-Based Query Optimization on Coarse-grained Distributed Memory', pp. 1-18, 2014.
xiii. J. Dean and S. Ghemawat, 'MapReduce', Communications of the ACM, vol. 51, no. 1, p. 107, 2008.
xiv. Wang, R.Y., Storey, V.C., and Firth, C.P. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering 7, 4 (1995), 623–640
xv. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM 41 (1998) 58-65
xvi. Towards an Ontology for e-Document Management in Public Administration – the Case of Schleswig-Holstein. Klischewski, R. 2012. s.l.: Proceedings HICSS-36, IEEE, 2012