Framework for Handling Data Veracity in Big Data

Muna Al-Jepoori[1] and Zuhoor Abdullah Al-Khanjari[2]

[1] Computing, Canterbury Christ Church University, Canterbury, UK, [2]Computer Science, Sultan Qaboos University,Muscat, Oman

[1]muna.al-jepoori@canterbury.ac.uk, [2]zuhoor@squ.edu.om

ABSTRACT

Big Data is the term used for massive amount of data collected by different means and in various formats. Data Veracity refers to the uncertainty of available data; this means that the quality of the collected data cannot be trusted. This paper reports on ongoing research based on using the Semantic Web technology to verify user entered data and increase dependability on Big Data. Validating, cleaning and reducing collected data are the major activities required to enhance the quality of the collected data.

Keywords: *Big Data; Veracity, User generated data, automatically collected data, Crowed validation of data, Fake news.*

1. INTRODUCTION

Data is the most important asset that can contribute to solving many of the current issues that are directly or indirectly affecting our lives. Our world has become very complicated, and the need to find solutions to certain problems has never been as urgent. Societies are facing a variety of problems that range from the need for security, finding ways to combat certain diseases, improving the performance of businesses, and providing for diverse areas of applications that contributes to the wellbeing of mankind.

Big Data is the term used for massive amount of digital data collected from various sources. Data Veracity refers to the uncertainty of available data; in such a case, quality and accuracy are difficult to control [1].

In the field of Data collection, processing and analysis, there has been great progress in collecting vast amounts of data as a results of many initiatives and projects. The Open Linked data [2] provide a huge amount of reasonably verified data that can be used with a high degree of confidence. Data collected using the traditional databases are normally accurate. However, the way of collecting and processing data is changing, and the need to control the integrity of the collected data is growing.

This paper is structured as follows: it begins by introducing the trends in data collections and the current technology that is used in Big Data projects. Section 3 presents the proposed framework for handling veracity in Big Data, and in section 4, the implementation plan and testing plan are presented. The final section provides a summery and directions for future work.

2. DATA COLLECTION AND NEW TECHNOLOGY

2.1 Trends in Collecting Data

In the early days of computing, data was collected using highly controlled processes that insured the integrity of data during data conversion, data collection and data entry; specially trained staff verified entered data before such data is adopted. With the emergence of web-based applications, people started entering their own data with a good degree of validation mechanism.

Most of data that is being produced nowadays comes from images and text. For such data to be useful in a reasonable amount of time, automated information extraction tools are used to collect, classify, identify and recognize objects. The big question here is “How correct is the output from these automatic tools?”

Currently, with the emergence of social media and the potential benefits of the collected data, people are entering their own data with a high degree of freedom. Thus, people can claim whatever they want; for example, if an 8 years child claims he is 18 years old, we have no way of validating such information. We cannot even tell whether any of the information provided by users is correct. Unfortunately, people are expressing opinions and reporting untrue information in their posts. Such a context presents major questions:

- How can we accept or reject untrusted information?
- What are the consequences of keeping wrong information, wrong ideas, wrong beliefs, and dangerous deceptions?

- Who is responsible?
- How legal is it to allow such information to be in the public domain?

In addition, big data became massive when we started collecting data with the help of machines. Servers are recording users' behavior in every online activity, machines are collecting and recording weather data, road traffic capturing images and videos that are automatically recorded and analyzed. More than 2.5 quintillion bytes of data are generated every day [3]. IBM expected that 80% of the available data in 2015 are uncertain [4]. The two main sources of uncertainty come from social networks, as there is little control on data content and expressions, and the Internet of things, as network devices rely on sensors that might not provide accurate data for many reasons.

2.2 Current Technology and Big Data

Big Data acquisition, preprocessing, storage and management have become possible due to the progress in manufacturing integrated circuit technology, the increase in network capacity, and the advancement of parallel computing. This provided for the production of powerful technology and software such as Apache Hadoop and its different components [4]. The Semantic Web technology, standards, concepts and data sets have provided for the Big Data projects to become a reality.

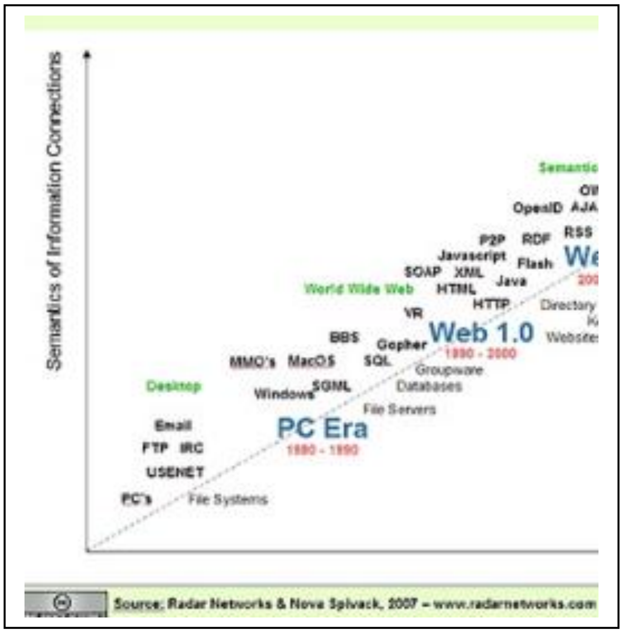
The term Semantic Web was coined by Tim Berners-Lee, now the director of the World Wide Web Consortium (W3C) which oversees the development of the Semantic Web standards. The Semantic Web is a 'web of data'; unlike the current "Web of Document", the Semantic Web enables software to "understand" the meaning of the information on the Web. Web pages in the Semantic Web are annotated with machine accessible data that enables automated agents to access the Web more intelligently and perform tasks on behalf of users [5, 6].

Big Data projects benefitted from the Semantic Web technology, the extensive effort of the W3C consortium, and the standard technology produced as a result of extensive research towards the full implementation of the Semantic Web.

Apache Hadoop is an open source software project that enables the distributed processing of large data, using clusters of servers. Big Data can be implemented on one server at the start and then scaled up to thousands of machines. Modules in Hadoop are designed to automatically detect and handle hardware failure using software with very a high degree of fault tolerance. This has eliminated the need to rely on high-end hardware [7]. Fig. 1 summarizes the timelines and web generations together with major technology used [8]; it is obvious that we are currently in the era of the Semantic Web where we have:

- Semantic Web technologies: RDF, OWL, SPARQL, Semantic application platforms, datastores, and open linked data (LOD) datasets [2]. In addition, we have microdata, schema.org and JSON-LD that have been successfully used in many of the Big Data projects, such as Facebook. Google knowledge graph and knowledge vault is using a variety of sources to enable better reasoning [9]; it uses LOD dataset, such as DBpedia and US government Food and Drugs Administration (FDA) datasets; it also uses the Central Intelligence Agency (CIA) world Factbook produced for US policymakers, Facebook, and many other specific sources that provide information used in the semantic search leading to transforming Google's search engine to a knowledge engine [10].
- Distributed databases - enabled by Semantic Web technologies and open source software.
- Intelligent applications- with natural language processing, machine learning, and machine reasoning [6].

In summary, we are currently witnessing the evolution of the web towards fulfilling the semantic web vision as anticipated by Tim-Berners Lee in 1989 [11]; Big Data is a natural product of all the scientific efforts and progresses in web technology.



Source: Radar Networks & Nova Spivack, 2007 – www.radarnetworks.com

Fig. 1. Timelines and Intelligence in connections

3. Framework for Reducing Veracity in Big Data

Tim-Berners Lee proposed the layered approach to the Semantic Web [12], shown in Fig 2. Trust is the most difficult layer to implement, many different solutions have been proposed, however, very little progress have been made.

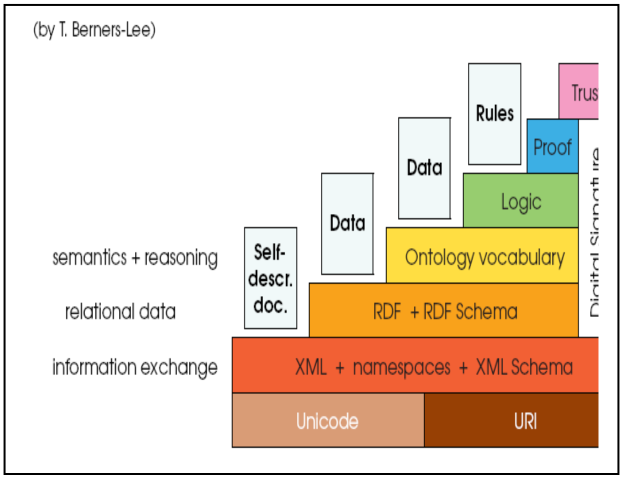


Fig. 2. The layered approach to the Semantic Web

Lack of Trust is also the main problem in Big Data. The framework suggested below is an attempt towards contributing to increasing trust by reducing data Veracity. The idea is based on the successful effort of the Sematic Web community, directed by W3C, and vast amount of knowledge already collected for the Semantic Web. Fig. 3 shows the general architecture suggested to use available Trusted data and to create and use Control Knowledge (CK). Trusted data is built using currently available datasets, together with the specialized database, and the databases available at collaborating governments and private institutions. The framework enables validating, clearing, user entered data, and it includes:

- Untrusted Data- this view or repository contains untrusted data that is collected from sources that cannot be validated at the point of collection, such as Facebook data and some other automatically collected data
- Trusted data- this is the repository of highly trusted semantically annotated data. Validation is based on LOD datasets and any other data converted from the currently used businesses and enterprises applications; this mainly includes any data in the public domain that can be accessed without restrictions. It may also include a restricted view of business private data or highly redistricted data.
- Crowd – one way to validate data is Crowd validation. The role of the crowd is to validate untrusted information; a dedicated web based system will be introduced. Crowd are to have access to selected information in the domain of interest; in particular, users will be asked to validate pieces of information that might include Fake News or wrong information. Users will be asked to agree or disagree with such pieces of information. The higher the number of responses supporting the validity of information, the higher validity score the information gets. The Crowd will be recruited to work on special critical tasks following well identified process that ensures adequate and unbiased answerers. Some topics are critical enough to deserve the effort and the cost of such Crowd validation process.
- The Control Knowledge part, created as a result of Crowd validation, includes references to untrusted information and the number of trust points it has scored.

Trusted data and the CK are used to validate untrusted data and promote the verified objects and relationships to trusted data, using algorithms that depend on the content and context of data.

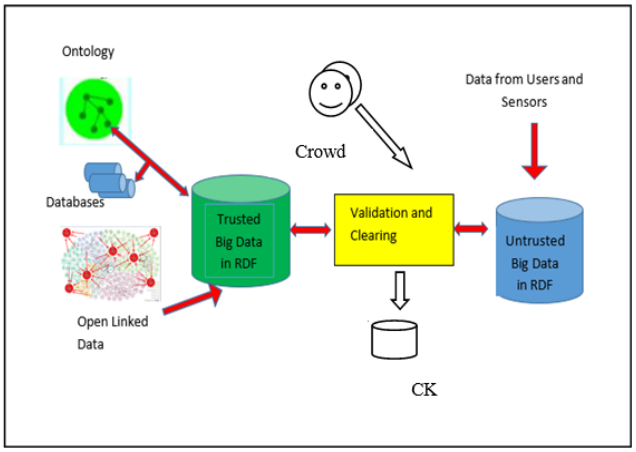


Fig. 3. Framework for Handling Veracity in Big Data

4. IMPLEMENTATION PLAN AND TESTING STRATEGY

The initial phase of implementation will be carried out by processing parts of the Facebook graph, using related LOD datasets of BBC news together with domain Ontology. The selected domain is Fake News on Facebook.

Trusted Big Data will be generated, while untrusted data will be subject to crowd validation.

5. SUMMARY AND CONCLUSIONS

The framework presented in this paper aims at providing processes that validate and clean the data that is currently part of untrusted Big Data projects. Using control knowledge to verify some of the uncertain data and to decide what to validate, neglect, delete, or keep for further checking. Such strategy is in fact the normal process followed by people to validate facts. The aim is to gradually increase trusted data to the extent that it provides for applications with a higher degree of trust. This work is still ongoing research that is benefiting from the vast amount of knowledge already collected for the Semantic Web [13, 14].

REFERENCES

[1] IBM, The Four V's of Big Data, available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data, last accessed on 20/12/2017

[2] Berners-Lee, Linked data, 2009, available from: http://www.w3.org/DesignIssues/LinkedData.html, last accessed on 20/12/2017

[3] IBM, Bringing big data to the enterprise, available from: http://www-01.ibm.com/software/data/bigdata/ data at the speed of business, last accessed on 20/12/2017

[4] I. Claverier-Berge, Solutions Big Data IBM, 2012 available from: http:/www-5.ibm.com/fr/events/netezzaDM_2012/Solutions_Big_Data.pdf, last accessed on 20/12/2017

[5] T. Berners-Lee , J. Hendler ,and O. Lassila , The Semantic Web, Scientific American, May 2001

[6] T. Berners-Lee, Business Model for the Semantic Web, 2001

[7] https://hadoop.apache.org/, last accessed on 20/12/2017

[8] N. Spivack Web 3.0: The Third Generation Web is Coming, available from: https://lifeboat.com/ex/web.3.0 last accessed on 20/12/2017

[9] Google, Google Knowledge Graph Search API, available from: https://developers.google.com/knowledge-graph/ accessed on 20/12/2017

[10] S. Leslie, Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data,APress, 2015

[11] T. Berners-Lee, The road to Semantic Web, 1989

[12] G. Antoniou and F. van Harmelen. A Semantic Web Primer. MIT Press, 2004. ISBN 0-262-01210-3.

[13] T. Berners-Lee, The Next Web, 2009, available from: https://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=en, last accessed on 20/12/2017

[14] C. Bizer, A Jentzsch, R. Cyganiak , State of the LOD Cloud,Available from: http://www4.wiwiss.fu-berlin.de/lodcloud/state/, last accessed on 20/12/2017