International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

An Enhanced Framework to Ensure Big Data Veracity in Social Business Intelligence

Mohamed OUBEZZA[1], Ali EL HORE[2], Jamal EL KAFI[3]
[1]LAROSERIE Laboratory Department of Computer Sciences, Faculty of Sciences, El-Jadida, Morocco
[2]DIS Laboratory Department of Computer Sciences, Faculty of Sciences, El-Jadida, Morocco
[3]LAROSERIE Laboratory Department of Computer Sciences, Faculty of Sciences, El-Jadida, Morocco
Email address: [1]Oubezza.m@ucd.ac.ma

***Abstract*— *Today, the Internet, characterized by its dynamics, is a fantastic source of data for companies. With the tools and techniques currently available, the processing of this data in real-time is no longer a concern; the main concern of decision-makers is the veracity of the data, especially in sensitive areas such as Business Intelligence. In this paper, we present a framework that tends to help decision-makers to ensure the veracity of Social Big Data to improve Social Business Intelligence. Our method checks the veracity of RDF triples based on domain ontology to control the structure of RDF triplets. Then to check the correspondences of the ontology properties, we use the WordNet semantic database, which allows us to check the subject and object classes of each triple RDF. Then the WikiData and DBpedia databases are used to ensure the veracity via a SPARQL query.*

***Keywords*— *Big Data, Ontology, Social Business Intelligence, Veracity*.

I. INTRODUCTION

Big Data is characterized by the 5V, which refers to five key elements to be taken into account and optimized as part of an approach to handle big data management. These 5V are Volume, Velocity, Variety, Visualization, and Veracity. The first four elements are well handled and the solutions that manage them are now mature. But veracity remains the most complicated element to deal with, especially in areas such as Business Intelligence, even with the introduction of new disciplines such as Social Business Intelligence.

Social business intelligence is a discipline that combines corporate data with unstructured data generated by users. This allows decision-makers to analyze their cases according to the trends in their environment.

The semantic web technologies and knowledge management systems (KMS) are based on knowledge presentation languages such as RDFS and OWL. And allows separating the metadata of the schema (Terminological BOX) from the instance data (Assertional BOX).

The information generated on the Internet is far from reliable, and manual verification of the veracity by experts is impossible. A first check makes it possible to ensure the logical consistency of the information, by comparing the ABOX instances and the TBOX scheme. The tools developed for this purpose are numerous and show good performance. But the effective verification of ABox, which presents the most important part of the data, remains less discussed, some previous work tries to compare two pieces of information, depending on their source. These approaches are not, in reality, exploitable because in reality the information comes alone and comes from a single source.

In this paper we propose a Framework to ensure the veracity of the information, trying to enrich the ontology with additional RDF. For this we will refer to the WordNet database to know the predicate class of RDF and then generate the corresponding standard RDF, then we determine the predicate sys names which allows building SPARQL queries to be sent to WikiData and DBpedia for additional evidence. The infrastructure of our Framework is based on Apache SPARK for parallel processing, which stores intermediate data in the RAM and shows good performance compared to Apache Hadoop.

The rest of this paper is structured as follows: in part 2 we present a study of the existing systems, part 3 will be dedicated to Social Business Intelligence approaches, then we will present our Framework in part 4 and the results of the evaluation in part 5. Finally, the conclusion and future directions will be presented in Part 6.

II. RELATED WORK

The amount of information available on the Internet is an important source of knowledge bases in all fields. But, since anyone can publish anything, about any topic. A verification of the veracity of this information is required. Several studies have been done on this subject.

Data Fusion tends to merge several data from different sources that share the same schema, several works in this framework try to identify the true values among the merged elements. A Survey of Data Fusion Techniques is presented by [1], The work of [2][3] focuses on the FreeBase knowledge base to get more information and then uses Classification and Machine Learning techniques to determine the level of truthfulness of the information. The work of [4] uses the same approach but adds a reliability value for each source to expand the knowledge base. The limit of this work is the amount of data needed to evaluate information.

The works of [5] process a single piece of information and verify its veracity on the Internet by calculating a confidential value according to several parameters. But this approach is unable to handle false propaganda information as shown in [6].

Finally, the work of [7] presents a Framework that offers processes that validate and clean up data that is part of unstructured big data. The weak point of this Framework is

Mohamed OUBEZZA, Ali EL HORE, and Jamal EL KAFI, "An Enhanced Framework to Ensure Big Data Veracity in Social Business Intelligence," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 4, pp. 252-256, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

that it is based on Crowd, which requires user interaction, and for us, this is a semi-automatic approach that is difficult to control in real situations.

III. Social Business Intelligence

A. Social Big Data

The social big data [8] comes from the combination of the efforts of the two domains social media and big data. Large social data will be based on the analysis of large amounts of data that could come from multiple distributed sources. Therefore, the social analysis of large data is inherently interdisciplinary and covers areas such as data mining, machine learning, information retrieval, statistics, natural language processing, semantic web, ontologies, and large data computing. Their applications can be extended to a wide range of areas such as health, political forecasting, e-commerce, cybercrime, counter-terrorism, public opinion analysis, and social network analysis.

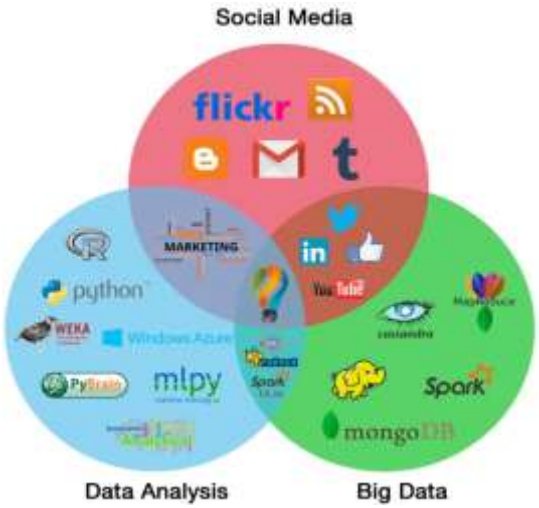

Fig. 1. The social Big Data [8].

Collecting, merging, processing and analyzing large amounts of social media data from unstructured (or semi-structured) sources to extract valuable knowledge is an extremely difficult task that has not been fully resolved. Traditional data management methods, algorithms, frameworks, and tools have become inadequate to handle a large amount of data. This issue has created a large number of open problems and challenges in the field of large social data related to different aspects such as knowledge representation, data management, data processing, data analysis, and data visualization. However, given the very large amount of heterogeneous data from social media, one of the main challenges is to identify valuable data and how to analyze it to discover useful knowledge that improves decision-making for individual users and businesses.

To properly analyze social media data, traditional analysis techniques and methods must be adapted and integrated into the new data paradigms that have emerged for large-scale data processing. Various large data frameworks such as Apache Hadoop and Apache Spark have been developed to enable the effective application of data mining methods and machine learning algorithms in different domains.

B. Social Business Intelligence

Social Business Intelligence (SBI) is the emerging discipline that aims at effectively and efficiently combining corporate data with User Generated Content to let decision-makers analyze and improve their business based on the trends and moods perceived from the environment [9].

C. Social Business Intelligence Architecture

The reference architecture we have chosen to support our approach to Social BI is described in Fig. 2, with a focus on the integration between sensitive and business data, achieved in a non-invasive way by extracting certain business flows from the company's data warehouse and integrating them with those that carry user-generated text content to provide users with business intelligence capabilities [10].

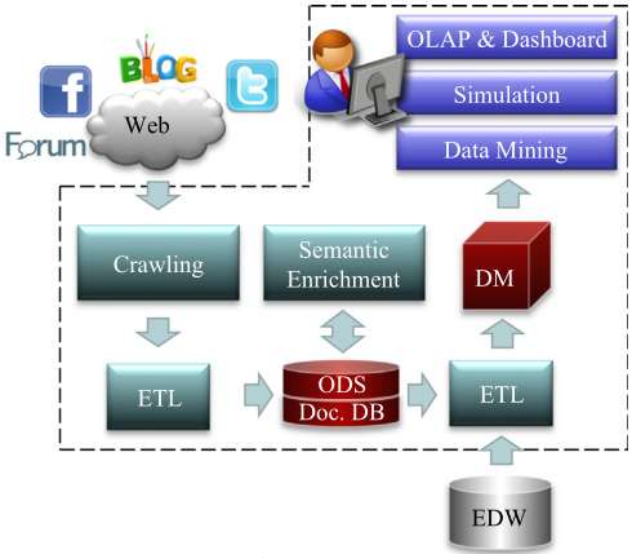

Fig. 2. Architecture of Social Business Intelligence [10].

The Crawling component performs a set of keyword-based queries to retrieve clips (and available metadata) that are within the subject's scope.

The target of crawler research can be the entire Web or a set of user-defined Web sources (e. g. blogs, forums, websites and social networks). The semi-structured output of the crawler is transformed into a structured form and loaded into the operational data memory (ODS), which stores all relevant data on clips, their authors and source channels; for this purpose, a relational ODS can be coupled with a document-oriented database that can store and efficiently search clip text. The ODS also represents all subjects in the field and their relationships.

The semantic enrichment component works on the ODS to extract semantic information hidden in clip texts. Depending on the technology adopted (e.g., supervised machine learning or lexical techniques), this information may include the clip's simple sentences, its subject(s), syntactic and semantic relationships between words or the feeling associated with an entire sentence or each subject it contains.

Mohamed OUBEZZA, Ali EL HORE, and Jamal EL KAFI, "An Enhanced Framework to Ensure Big Data Veracity in Social Business Intelligence," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 4, pp. 252-256, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

The ETL component periodically extracts data on clips and topics from the ODS, integrates them with business data extracted from the EDW of the enterprise data warehouse and loads them into the Data Mart (DM). The DM stores integrated data as a set of multidimensional cubes, using metastars for subject hierarchies; these cubes support the decision-making process in three complementary ways:

1. OLAP and dashboard: users can explore the UGC from different angles and effectively control overall social sentiment. The use of OLAP tools for the multidimensional analysis of the UGC pushes the flexibility of our architecture much further than the standard architectures adopted in this context.
2. Data mining: users assess the actual relationship between rumors/opinions circulating on the web and business events (for example, to what extent positive opinions circulating on a product will have a positive impact on sales?)
3. Simulation: correlation models that link the UGC to business events are used to predict business events in the near future considering the current UGC.

IV. Proposed Framework

A. Preliminaries

The technical concepts used in this paper will be presented in this section in detail.

a. Ontology: The Several definitions are provided for an ontology, in computer science, an ontology is a shared and common understanding of some domain that can be communicated across people and application systems, and for knowledge sharing, an ontology is an explicit specification of a conceptualization [11].

The domain ontology facilitates understanding and sharing knowledge of this domain.

Ontologies are part of the W3C standards stack for the Semantic Web; ontology can be used to specify standard conceptual vocabularies to exchange data, share knowledge and facilitate the interoperability across heterogeneous systems and databases.

b. OWL / RDF: Ontology web language (Owl) [12] is the standard language for representing ontology; OWL allows the representation of advanced concepts like joins, unions, restrictions...

Resource Description Framework (RDF) [13] is a graph model designed to formally describe web resources and their metadata so that such descriptions can be automatically processed. Developed by the W3C, RDF is the basic language of the Semantic Web. A document structured in RDF is a set of triplets [14]. Each RDF triplet in the form of an association (subject, predicate, object), the form used in our work for RDF serialization is RDF/XML.

c. SPARQL: SPARQL Protocol and RDF Query Language is a query language and protocol that allows querying an RDF Store [14]. Since 2007, SPARQL has been considered as one of the key technologies of the Semantic Web by Tim Berners-Lee, inventor of the Semantic Web and director of W3C, who explains "Trying to use the Semantic Web without SPARQL is like exploiting a relational database without SQL" [14]. The SPARQL specification defines a query language and protocol that works in perfect synergy with the other W3C semantic Web technologies: RDF (Resource Description Framework) for data representation, RDFS (RDF Schema), OWL (Web Ontology Language) for vocabulary creation. Since 2013 SPARQL becomes an official W3C recommendation.

B. Framework Components

In this section we will present our framework, the architecture is the same presented in our previous work [15], the main components are shown in the figure below:

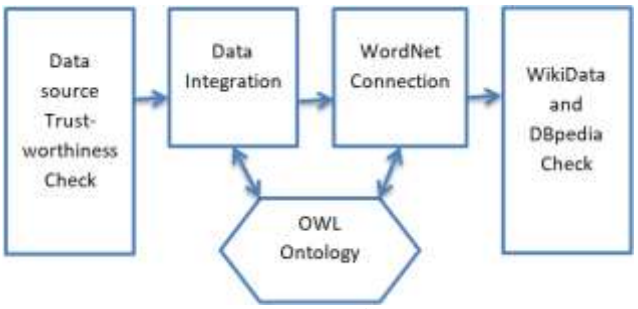

Fig. 3. Main components of our Framework

The main components of our Framework are: Data Source Trustworthiness Check is responsible for verifying the reliability of data sources, each source has a reliability factor, which gives a score, for each information, and this score will then be used in the following steps. The Data Integration component is responsible for mapping heterogeneous data, it is based on domain ontology and the result is RDF triple. These triples will be used by WordNet, which determines the subject and object membership classes of the triple and then manages new triples according to their synonyms, this component is based on the domain ontology described in OWL. The WikiData and DBpedia check component sends SPARQL requests for verification of the original triples and the triplets generated by the previous component, the requests are sent to both WikiData and DBpedia data repositories.

C. Data Source Trustworthiness

The first step in the process adopted by our approach consists in classifying sources according to their classes, we have thus defined four classes, each class has a reliability factor, defined manually by the domain expert, so the data from encyclopedias and recognized sites are more reliable and more advantageous than those from social networks, and personal blogs.

The reliability factor of the data source will be of major importance in the case of triples that represent conflicting information. The higher the factor, the more credible the source is and finally the information from this source is reliable.

D. Data Integration and The domain Ontology

One of the advantages of our approach is the use of ontologies for integration; this ontology is used to express the semantic relationships between concepts and also used as a knowledge base to validate information whose source is less

Mohamed OUBEZZA, Ali EL HORE, and Jamal EL KAFI, "An Enhanced Framework to Ensure Big Data Veracity in Social Business Intelligence," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 4, pp. 252-256, 2019.

reliable.

Ontology also provides a controlled vocabulary of semantics; we are interested in our work with TBox because we want to verify the information by querying the knowledge base. For this reason, we eliminate all the triple RDFs that only concern ABOX instances.

E. WordNet Enrechissement

Information can be expressed in several ways, so with several triple RDFs, we need to treat all these cases. For example, the same information can be expressed between two RDFs whose properties are synonyms, or whose first RDF is expressed with the active form of the verb and the second with its passive form with the inversion of the subject and the object.

WordNet is a large lexical database of English [16]. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. In WordNet, meanings represent the meaning of words and are expressed by synsets. WordNet is used in our project to determine for the predicate of each triple RDF, the corresponding WordNet verb which simplifies the interrogation of the knowledge base. And to generate other triplets that are equivalent to the original RDF [16]. The figure below shows part of the structure of the WordNet ontology and the WordSense and SynSet classes used in our work to determine the meaning of the RDF relation.

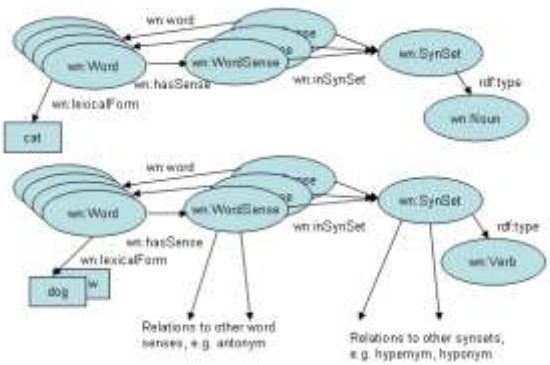

Fig. 4. WordSense and SynSet classes in WordNet [16].

This structure is used to generate relationship synonyms by querying ontology data. After this step, several RDFs are generated from the original RDFs that represent the same knowledge.

F. WikiData and DBpedia Connection

The last process step of our framework is to send SPARQL requests to knowledge bases recognized worldwide for their very high-reliability WikiData and DBPedia.

At first WikiData and DBPedia appear to be competitors, and the use of one of them is sufficient. DBPedia extracts structured data from Wikipedia infoboxes and publishes them in RDF and provides other services such as mapping external ontologies.

Wikidata [17] provides a secondary and tertiary database of structured data that everyone can edit. Theoretically, the WikiData process will replace the DBPedia [18] process but our experience shows that the DBPedia database hosts knowledge missing in WikiData which justifies the use of both databases in our project; besides both represent knowledge by RDF which makes them usable by SPARQL queries.

An example of the SPARQL request sent to Wikidata to verify the existence of a product for a defined brand is shown below:

```
PREFIX rdfs: <http://www.w3.Org/2000/01/rdf-schema#/>
PREFIX wikidata: <http://www.wikidata.org/entity/>
PREFIX wdp: <http://www.wikidata.org/prop/direct/>
select (count(?subject) as ?count)
where {
?subject rdfs:label "Input"@en .
?subject wdp:P361 ?occupation .
?occupation rdfs:label "product"@en .
}
```

The same request is sent to DBPedia, if the value is greater than zero for at least one of the two Bases then the RDF is considered valid and its veracity is verified.

V. TEST AND EVALUATION

To validate our system we chose 16 information elements expressed in triple RDF format, and we checked their veracity manually, then we entered them into the system for automatic verification. The system has shown an efficiency of 87.5%, with only two of 16 predictions incorrect.

TABLE 1: Prediction results of our system

Number of RDF triple	16
Number of true RDF	8
Number of false RDF	8
Number of correct predictions	14

VI. CONCLUSION AND FUTUR WORK

In this paper, we have presented a Framework to ensure the veracity of data in social business intelligence. This Framework is based on SPARK for parallel processing; this support on SPARK which has the advantage of storing intermediate data in RAM instead of hard disk ensures real-time processing. Our Framework is generic, i.e. it can be used for any domain, it is sufficient to indicate the ontology of the specific domain, first it classifies the triple RDF according to the reliability of its source, then it checks the correspondence between the data and the schemas or metadata, then for each triple RDF, specifies the subject and object belonging class of the triplet, which allows to add more triplets based on the WordNet semantic database, after, the Framework connects to the global structured encyclopaedias and based on Wikipedia DBpedia to verify the existence of the information presented by the triple RDF, a triple that has a score of more than 51% is considered reliable. Our Framework has been validated by a group of RDFs whose veracity has been manually verified and compared to the results of system predictions.

In future work, we propose to work on information from several fields, which therefore require several ontologies.

ACKNOWLEDGMENT

This work has been supported by the National Centre for Scientific and Technical Research of Morocco.

REFERENCES

[1] Li, Xian, et al. "Truth finding on the deep web: Is the problem solved?." Proceedings of the VLDB Endowment. Vol. 6. No. 2. VLDB Endowment, 2012.
[2] Dong, Xin Luna, et al. ''From data fusion to knowledge fusion." Proceedings of the VLDB Endowment 7.10 (2014): 881-892.
[3] Yin, Xiaoxin, Jiawei Han, and S. Yu Philip. ''Truth discovery with multiple conflicting information providers on the web." IEEE Transactions on Knowledge and Data Engineering 20.6 (2008): 796-808.
[4] Xin Don, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. KDD 2014: 601-610
[5] Jens Lehmann, et al. DeFacto - Deep Fact Validation. International Semantic Web Conference (1) 2012: 312-327
[6] Dong, Xin Luna, Laure Berti-Equille, and Divesh Srivastava. ''Integrating conflicting data: the role of source dependence." Proceedings of the VLDB Endowment 2.1 (2009): 550-561.
[7] Al-Jepoori, M. and Al-Khanjari, Z. (2018) Framework for handling data veracity in big data. International Journal of Computer Science and Software Engineering, 7 (6). pp. 138-141.
[8] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: recent achievements and new challenges, Inf. Fusion, 28 (2016), pp. 45-59.
[9] M. Francia, M. Golfarelli, S. Rizzi, A methodology for social BI, in: Proceedings of the IDEAS, Porto, Portugal, 2014, pp. 207–216.
[10] E. Gallinucci, M. Golfarelli, and S. Rizzi. Meta-stars: multidimensional modeling for social business intelligence. In Proc. DOLAP, pages 11–18, San Francisco, CA, 2013.
[11] T. R. Gruber, "A Translation Approach to a Portable Ontology Specification", Knowledge Acquisition, vol. 6, pp. 199-221, 1993.
[12] https://www.w3.org/OWL/ (Accessed 31 October 2019)
[13] https://www.w3.org/2001/sw/wiki/RDF (Accessed 31 October 2019)
[14] https://www.w3.org/TR/rdf-sparql-query/ (Accessed 31 October 2019)
[15] Oubezza, M., El Hore, A. and El Kafi, J. (2019) 'An incremental and distributed inference method for large-scale ontologies over SPARK', Int. J. Cloud Computing, Vol. 8, No. 2, pp.140–149.
[16] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
[17] https://www.wikidata.org/wiki/Wikidata:Main_Page (Accessed 31 October 2019)
[18] Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Flöck, F., & Lehmann, J. (2018). Detecting linked data quality issues via crowdsourcing: A dbpedia study. Semantic web, 9(3), 303-335.