Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. XV Issue V
May 2023

A Comprehensive Study on the Integration of Semantic Web Approach and Big Data

Dr. Narendra Sharma[1], Waseema Masood[2]
Assistant Professor, Sri Satya Sai University of technology and Medical Sciences, Sehore, India[1]
Research Scholar, Sri Satya Sai University of technology and Medical Sciences, Sehore, India[2]

Abstract

The World Wide Web Consortium (W3C), an international organisation that sets web standards, promotes the semantic web. It is an expanded version of the current web that offers a simpler method for searching, reusing, combining, and sharing information. Therefore Semantic Web is subsequently viewed as an integrator crosswise over various content, data applications, and frameworks. Big data nowadays is typically referred to as having the five Vs: volume, variety, veracity, value, and velocity. Variety of data explains how to cope with a wide range of varied data sources as well as various data types. Therefore, the variety of big data challenges is crucial for resolving various issues in the actual world. Data from many sources, including web services, relational databases, spreadsheets, etc., and in diverse forms are combined using the semantic web as an integrator. This work involves a number of challenges because of the data heterogeneity, some of which may not be fully resolved by the current method. In this essay, we strive to concentrate on the numerous difficulties that arise when integrating data from multiple sources, as well as the ways that various semantic web technologies and tools are applied to the integration of disparate data.

Keywords: *Semantic Web, Big Data, Disparate Data, Velocity, Variety*

1. Introduction

Although the term "Big Data" is relatively new however, the act of collecting and storing large amounts of information for eventual analysis is ages old.[1] The concept gained its momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs.

1. Volume. It discusses the scale of data. Organizations collect data from a variety of sources, including business transactions, social media and information from the sensor or machine-to-machine data. In the past, storing it would've been a problem but new technologies (such as Hadoop) have eased the burden. It is estimated that by the year 2020, 40 Zettabytes (43 Trillion Gigabytes) of data will be created.

2. Velocity. It deals with analysis of streaming data. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Like the New York Stock Exchange generates 1 TB of trade information in each trading session. Moreover modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.

3. Variety. Data comes in all types of formats { from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. By 2014 there were 420 million wearable and wireless health monitor devices. 4 Billion+ hours of video is watched on YouTube every month, on Facebook 30 billion pieces of content are shared every month. In addition to the above, two more V's are associated these days with "Big Data".

4. Veracity. It encompasses the uncertainty related with the data. According to research 1 in 3 business leaders doesn't trust the information they use to make decisions. In addition to the increasing velocities and varieties of data, data ows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data. Poor data quality costs 1.3 Trillion USD a year to the US economy.[2]

5. Value. It is all well and good having access to big data but unless we can turn it into value it is useless. So we can clearly argue that 'Value' is the most important V of "Big Data". It is important that businesses make a business case for any attempt to collect and leverage big data. It is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits.

Big data is made conceivable by data integration. By empowering access to information put away in divergent information stockrooms, mapping changes starting with one endeavor application then onto the next, and conveying real- time data to proposed clients, data integration empowers enterprises to gather and clean huge information originating from different frameworks for analysis. Beside huge information examination, information joining likewise empowers different business advantages, for example, having a 360-degree perspective of datasets, quicker coordinated effort crosswise over whole organizations, and more noteworthy adaptability in choosing endeavor applications and frameworks and streamlining and mechanizing business functions.

2. Challenges of Integrating Big data

Big Data is a wide term for far-reaching and complex data sets where regular data planning applications are inadequate. The integration of this gigantic informational index is very intricate. There are a few difficulties one can look amid this reconciliation, for example, information generation, analysis, catch, search, sharing, representation, data storage, and privacy.

A. Data Management

Data Management is one of the most important tasks of integrating big data. When data is in huge volume then management of this data becomes difficult. To avoid these difficulties NoSQL [3][4] databases can be used. These Databases are contemporary to the traditional relational databases and provide great execution of different big data applications. These Databases works on the concept of the key-value pair [3] [4] so that it can deal with a huge amount of information with faster response.

B. Bad Data

In any data integration system, data quality is the biggest problem to worry about. Inheritance data must be cleaned up going before the change and joining, or organizations will almost certainly face honestly to goodness data issues later. Legacy information pollutions have an intensifying impact; by nature, they tend to focus on high volume information clients. On the off chance that this data is degenerate, along these lines, as well, will be the choices made by it. It isn't irregular for unfamiliar information quality issues to develop during the time spent cleaning data for use by the integrated framework [5].

C. Lack of Skills

Due to the coming of new technologies in a day to day market customers are moving from old traditional relational databases to new information processing system like NoSQL Databases, in-memory analytics, and Hadoop etc. Actually, in the market, there is the absence of required skills for big data innovations [6]. There is an average number of masters are available in the markets to work on these systems that process the huge amount of information.

D. Transmission of Data into Big Data Structure There are various people who have raised wants thinking about separating monstrous data accumulations for a noteworthy data platform. They moreover may not think about the versatile quality behind the entrance, transmission, and movement of data and information from a broad assortment of advantages and after that stacking this data in a major information stage. The marvelous parts of data transmission, get to and stacking is simply bits of the test.

E. Extracting Information

The most convenient use cases for huge data incorporate the availability of data, expanding existing accumulating of data and furthermore empowering access to end-client using business knowledge instruments with the true objective of the revelation of data. It transforms into a test in enormous data coordination to ensure the right-time data availability to the data clients.

F. Synchronization Data Sources

At the point when information is import into enormous information stages, we may in like manner comprehend that information duplicated moved from a broad assortment of sources on different rates and timetables can rapidly get away from the synchronization with the starting structure. This gathers the data starting from one source isn't obsolete when stood out from the data beginning from another source. It moreover infers the mutual quality of data definitions, thoughts, and metadata. The ordinary data administration and data circulation focus, the gathering of data change, extraction and movements all develop the situation in which there are perils for data to end up unsynchronized.

G. Other Challenges

The huge amount of data, arrangement cost, reliability, the correctness of information and rate of change of data these are the other difficulties could ascend during data integration. Actually saying it is not less than a test to practice the huge amount of data with the practical quickness with an important objective of provides the data to the necessary customer at the time of requirement. The endorsement of data accumulation is moreover fulfilled while trading data beginning with one source then onto the following or to purchasers as well.

3 Semantic Web

The semantic web has entirely redefined the process of converting web content into a more structured format so that the user’s web query can be achieved with more accuracy and providing the intelligent system to integrate the data from diverse sources. Fundamentally, the strength of the semantic web majorly relies on its effective ontological connections to appropriately represent the information that is suitable for a machine-readable format. The ontologies are the data models used to connect the concepts through its potential named entities (i.e., classes and relationships). Normally, the classes are recognized as entities and relationships are the properties between two classes. Our aim is to bring the major issues pertaining to semantic web technologies and particularly the vast opportunities laid forward for the correct utilization of web resources. In this connection, this survey provides a comprehensive overview of various datasets used for creating the knowledge sources and highlighting the major benefits of using these ontologies and datasets. Interventionary studies involving animals or humans, and other studies that require ethical approval, must list the authority that provided approval and the corresponding ethical approval code.

3.1 Emergence of Semantic Web

Some of the original motivations for the semantic web came from the early web applications that cause the problems for search and browsing in Web 2.0 applications. Latent semantics [7], an attempt to “mine” meaning from the words in web content, is always problematic due to its wide ambiguity and prevalent polysemy (the many meanings of a single word such as “run” or “left”). The class and subclass relations, which are crucial to language use, are also problematic. The semantic web technologies were created to deliver solutions for the faults that happened in Web 2.0. The significant finding that the semantic web has developed in recent years is that the applications are deemed to share much-sought information. Still, if that information is not in the textual format or is in a format that means that extracting the potential facts is difficult, then a suitable knowledge extraction pattern is required in the form of semantic web technologies [8, 15]. However, these findings are not new to emerging fields such as natural language processing and machine translation [9,16]. Over the years, NLP and MT worked on these findings and fixed problems such as ambiguity of text, missing fields, part-of-speech (POS tagger) confusions, and many more. However, the utilization of semantic web technologies has helped to uncover seminal knowledge representation in the text that was deeply ingrained in the forms of entities and relationships. Moreover, this paves the way for very efficiently connecting the entities with appropriate web resources, as depicted in Figure 1. This phenomenon has been achieved with a web graph: a graph exists between potential entities extracted from the textual sources, and the real-world entity persists on the web sources. To serve this purpose, semantic web languages such as resource description framework (RDF)/resource description framework schema (RDFS) and web ontology language (OWL) were used, and to shun the ambiguity persisting in the textual content, the semantic web languages mostly denote the terms or entities with assigned uniform resource identifiers (URIs) in the web. While much is said about the official capacities of these dialects and their ability to communicate individual connections, a substantially more basic perspective is that they can be utilized to give ordinary referents. Among the semantic web vocabularies [10], the friend-of-a-friend (FOAF) ontology has been widely used in textual processing to obtain the correct user references and link the entities with each other through the appropriate standard vocabularies. While inference is a significant part of the web and all different information portrayal dialects, the capacity for connected terms is a fundamental contrast between RDF-based dialects and prior KR dialects.

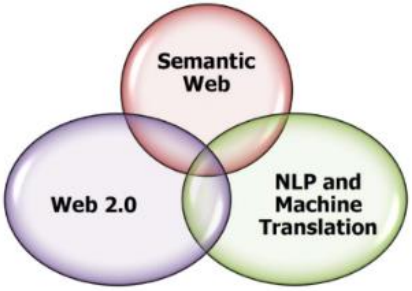

Figure 1. Connecting semantic web into Web 2.0, NLP, and machine translation.

3.2. Semantic Web and Digital Libraries

While implementing the semantic web projects, digital libraries are the key segment of the data foundation which supports higher education and research activities largely. A key perspective for the digital library [11,17] is the arrangement of shared indexes that can be distributed and pervasively examined. This requires the utilization of basic metadata to depict the fields of the inventory (for example, creator, title, date, distributor) and controlled vocabularies to permit subject identifiers to be attributed to publications. By distributing controlled vocabularies in a single spot, which would then be able to be linked to all clients over the web. The library indexes can utilize similar web-available vocabularies for listing and increasing things with the most applicable terms for the space of concept hierarchy. At that point, web search tools can utilize similar vocabularies in their pursuit to guarantee that the most applicable data details are returned. Figure 2 illustrates the concept of digital libraries. The semantic web opens up the likelihood to adopt such a strategy. It offers open formats and regulations that can empower merchant unbiased arrangements, with valuable adaptability (permitting organized and semi-organized information, formal and casual depictions, and open and extensible engineering), and it assists with supporting decentralized arrangements where appropriate. Consequently, RDF can be utilized as a typical exchange position for index metadata and some shared controlled vocabularies, which can be utilized by all libraries and web search tools over the web.

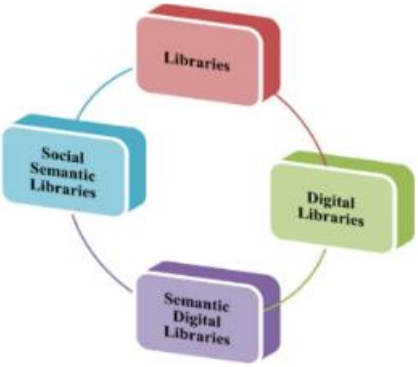

Figure 2. Various digital libraries utilized for semantic web applications.

3.3 Semantic Data Integration

Semantic information integration is the way toward joining information from divergent sources and merging it into important and significant data, however, the utilization of semantic technology. As organizations grow up an estimate, so does their information. Without the correct information management system, intradepartmental or application specific information storehouses rapidly emerge and impede efficiency and participation. Semantic Data Integration provides an answer that goes past the standard endeavor application integration arrangements. It utilizes information-driven engineering based upon an institutionalized model for information distributing and exchange, specifically the Resource Description Framework (RDF) [12]. In this system, the heterogeneous information of an organization such as structured, semi-organized and unstructured is communicated [13,18], put away and got to the similarly. As the information structure is communicated through the links inside the information itself, it isn't compelled to a structure formed by the database and does not wind up out of date with the advancement of the information. Following are some important steps to perform the semantic data integration.

1. Creating an Application Profile (RDF Shape) that depicts the coveted type of the last dataset;
2. Reusing existing ontologies and building new ontologiesas required;
3. Leveraging completely the accessible Linked Open Datasets in the area;
4. Designing a basic, sensible and practical URL methodology;
5. Using the assortment of accessible transformation andETL tools to play out the integration.

To go easily through a full semantic information integration lifecycle, organizations require an arrangement of simple to utilize semantic integration devices. Using semantic integration tools, clients can rapidly plan information handling jobs and integrate a gigantic volume of information.

4. Conclusion

Big Data today comes from a wide variety of sources, including semi-structured, unstructured, and structured data. Many organisations are interested in integrating this variety of data in order to use it for crucial operational and analytical tasks. But it's important to keep in mind that there could be numerous difficulties throughout the integration process. The RDF schema and OWL frameworks can be used to convert unstructured or semi-structured data types into any standard structured format thanks to semantic web technology. Furthermore, the semantic web and knowledge management can work in tandem to handle problems with a high degree of precision and resolve ambiguities that exist in text texts. In this survey study, we have discussed the integration of big data with semantic data as well as highlighted some of the major research difficulties. It has been proposed that the merging of artificial intelligence with the semantic web would open the door for a unified strategy to address the disambiguation issues encountered in unstructured data.

References

[1] Ostrowski, N. Rychtyckyj, P. MacNeille, and M. Kim, \Integration of big data using semantic web technologies," in 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp. 382{385, Feb 2016.

[2] A. Knoblock and P. Szekely, \Semantics for big data integration and analysis," in Proceedings of the AAAI Fall Symposium on Semantics for Big Data, 2013.

[3] J. Pokorny, "NoSQL databases: a step to database scalability in web environment", International Journal of Web Information Systems, Vol. 9 Issue: 1,pp.69-82, 2013

[4] J Ahmed, R Gulmeher “Nosql databases: New trend of databases,emerging reasons, classification and security issues” International journal of engineering sciences & research technology, Volume 4, Special issue 6, June 2015.

[5] S. Kumar, V. Singh, and B. Saini, "A survey on ontology matching techniques", In the Proceeding of 2014 International Conference on Computer and Communication Technology (ICCCT) , pp. 13- 15, 2014.

[6] L. Ragusa, “Data Integration Tools for Overcoming Integration Challenges in 2017”, https://www.liaison.com.

[7] Imami, N.K.; Murfi, H.; Wibowo, A. Comparative study of latent semantics-based anchor word selection method for separable nonnegative matrix factorization. In Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology, Singapore, 3–5 January 2020; pp. 89–92.

[8] Yao, X.; Berant, J.; Van Durme, B. Freebase qa: Information extraction or semantic parsing? In Proceedings of the ACL 2014 Workshop on Semantic Parsing, Baltimore, MD, USA, 26 June 2014; pp. 82–86.

[9] Klein, G.; Hernandez, F.; Nguyen, V.; Senellart, J. The OpenNMT neural machine translation toolkit: 2020 edition. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020), Virtual, 6–9 October 2020; pp. 102–109.

[10] Ben Ellefi, M.; Bellahsene, Z.; Breslin, J.G.; Demidova, E.; Dietze, S.; Szyma ´nski, J.; Todorov, K. RDF dataset profiling—A survey of features, methods, vocabularies and applications. Semant. Web **2018**, 9, 677–705

[11] Khan, S.A.; Bhatti, R. Semantic Web and ontology-based applications for digital libraries: An

investigation from LIS professionals in Pakistan. Electron. Libr. **2018**, 36, 826–841.

[12] A. Cuadra, M. M. Cutanda, D. Fuentes-Lorenzo and L. Sánchez, "A semantic web-based integration framework," In the Proceeding of the 2011 International Conference on Next Generation Web Services Practices, pp. 93-98, 2011.

[13] C. A. Knoblock and P. Szekely, “Exploiting semantics for big data integration”, AI Magazine, Vol.6, Issue.1, pp. 25-38, 2015.

[14] V Naiyer, J Sheetlani, HP Singh 2020, ‘Software Quality Prediction Using Machine Learning Application’, Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2 pp. 319-32.

[15] R Kumar, HP Singh, GA Kumar 2020, ‘Design and Development of Performance Evaluation Model for Bio-Informatics Data Using Hadoop’, Solid State Technology, volume 63 issue 6, pp. 23284-23296.

[16] BV Laxmi 2021, ‘A Review of Dynamic Resource Allocation Framework for Large Amount of Cloud Enterprises, Turkish Journal of Computer and Mathematics Education, volume 12 issue 2, pp. 1280-1284.

[17] Sheetlani et al., “A Proposed Framework Of Dimensionality Reduction Techniques To Boost Credit Card Fraud Classification”, IJFANS International Journal of Food and Nutritional Sciences, Volume 11, S Iss 3, Dec 2022

[18] kalyan et al., “Prioritization of Dynamic Test Cases Based on Historical Data for Use in Regression Testing of Requirement Properties”, IJFANS International Journal of Food and Nutritional Sciences, Journal Volume 11, S Iss 1,2022.