Semantic Computing for Big Data: Approaches, Tools, and Emerging Directions (2011-2014)

Seung Ryul Jeong [1], Imran Ghani [2*]

[1] Graduate School of Business IT, Kookmin University, Seoul, South Korea
Email: srjeong@kookmin.ac.kr
[2] Universiti Teknologi Malaysia (UTM), Skudai, Johor Bahru, Johor, Malaysia
Email: imran@utm.my
* Corresponding author: Imran Ghani

Received April 9, 2014; accepted June 9, 2014; published June 27, 2014

Abstract

The term “big data” has recently gained widespread attention in the field of information technology (IT). One of the key challenges in making use of big data lies in finding ways to uncover *relevant and valuable* information. The high volume, velocity, and variety of big data hinder the use of solutions that are available for smaller datasets, which involve the manual interpretation of data. Semantic computing technologies have been proposed as a means of dealing with these issues, and with the advent of linked data in recent years, have become central to mainstream semantic computing. This paper attempts to uncover the state-of-the-art semantics-based approaches and tools that can be leveraged to enrich and enhance today's big data. It presents research on the latest literature, including 61 studies from 2011 to 2014. In addition, it highlights the key challenges that semantic approaches need to address in the near future. For instance, this paper presents cutting-edge approaches to ontology engineering, ontology evolution, searching and filtering relevant information, extracting and reasoning, distributed (web-scale) reasoning, and representing big data. It also makes recommendations that may encourage researchers to more deeply explore the applications of semantic technology, which could improve the processing of big data. The findings of this study contribute to the existing body of basic knowledge on semantics and computational issues related to big data, and may trigger further research on the field. Our analysis shows that there is a need to put more effort into proposing new approaches, and that tools must be created that support researchers and practitioners in realizing the true power of semantic computing and solving the crucial issues of big data.

Keywords: Semantics, Ontology, Big Data, Tools, Challenges

http://dx.doi.org/10.3837/tiis.2014.06.012

1. Introduction

The term “big data” was coined to represent the large amount and many types of digital data that we use today, including documents, images, videos, audio, and websites. Semantics-based approaches are considered useful means of dealing with very large-scale data such as big data. In order to explore this topic, it is first necessary to more clearly describe the concept of big data.

1.1 Big Data

Although the term “big data” has not yet been defined by IEEE and it is not included in the online IEEE dictionary [8]. However, a number of definitions are presented in other popular sources, such as the following:

- According to [2]: Big data applies to information that cannot be processed or analyzed using traditional processes or tools.
- According to [5]: Big data is a high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Generally, big data has three main characteristics: volume, velocity, and variety. However, [1] adds one more characteristic: value. We agree with the use of the value characteristic, and include it in our body of relevant and valuable information.

It has become obvious that new capabilities and technologies are needed to capture and analyze big data. The McKinsey Global Institute estimates that data volume is growing by 40% per year, and will grow to 44 times its initial size between 2009 and 2020. According to [1] the volume of data is not the only characteristic that matters. In fact, there are four key characteristics that define big data (**Fig. 1**):

- Volume. Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume from just this single data source runs into the petabytes. Smart meters and heavy industrial equipment like that used in oil refineries and drilling rigs generate similar data volumes, compounding the problem of scalability.
- Velocity. Social media data streams, while not as massive as machine-generated data, produce a large influx of opinions and relationships that are valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures that large volumes are continuously created (over 8TB per day).

- Variety. Traditional data formats tend to be relatively well-defined by data schema, and change slowly. In contrast, non-traditional data formats are characterized by a dizzying rate of change. This is a common observation, that new services are added, new sensors are deployed, and new marketing campaigns are executed. The variety of data would need new data types to capture the variety of information and address issues of scalability.

- Value. We classify value as involving both valuable information and economic value. Typically, there is good information hidden in larger bodies of non-traditional data, so the challenge is to identify what is valuable, and then transform and extract the relevant data for analysis. The economic values of different data vary significantly.

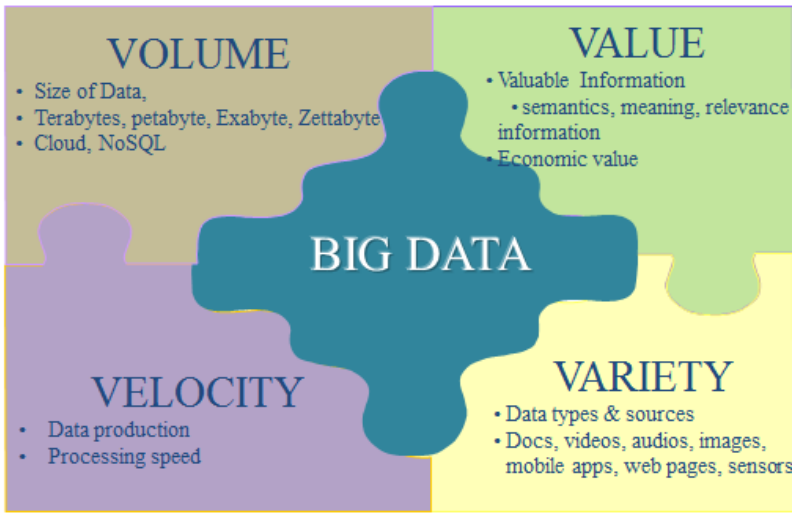

Fig. 1. Four key characteristics of big data

Oracle classifies value as an essential characteristic for processing big data, which we partially agree with. In our opinion, value should not just be considered to involve economic value, but also include the meaningful information hidden in big data.

1.2 Significance of Big Data

The significance of big data is clear. The following statistics show the predicted data growth trends, in terms of volume.

- **Data growth worldwide:** The predicted universal data growth between 2005 and 2020 is shown in **Fig. 2 (a)**, which indicates that by 2020, worldwide data growth will occur at a rate of 40,026 exabytes per year [56].

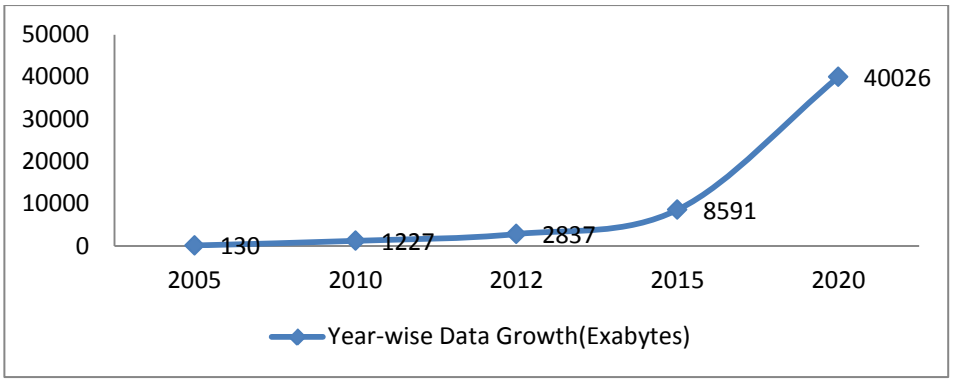

Fig. 2 (a). Universal data growth

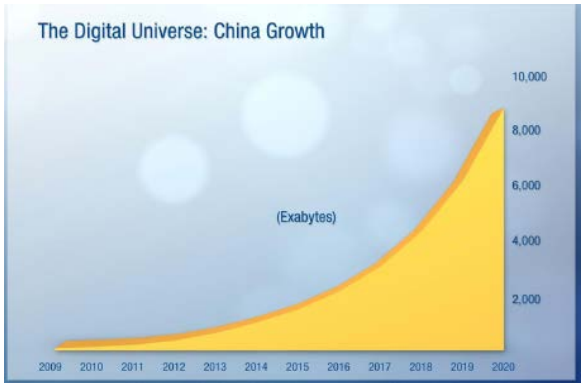

Fig. 2 (b). Data growth in China

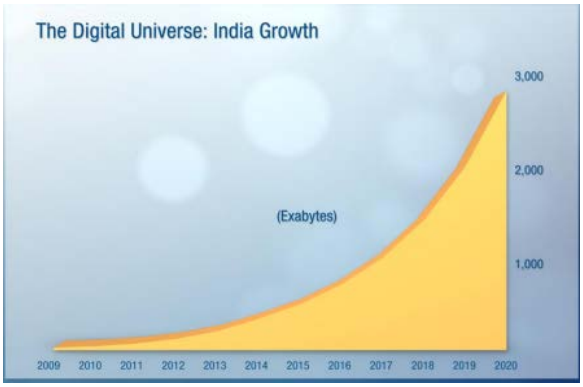

Fig. 2(c). Data growth in India

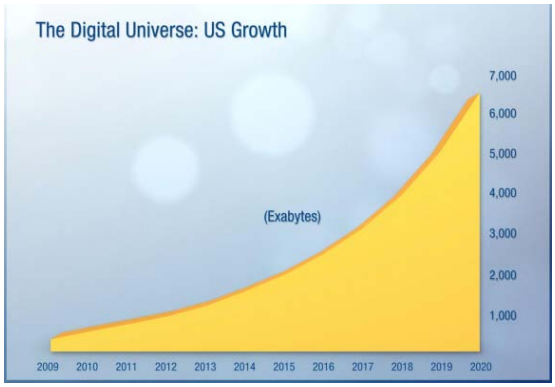

Fig. 2 (d). Data growth in US

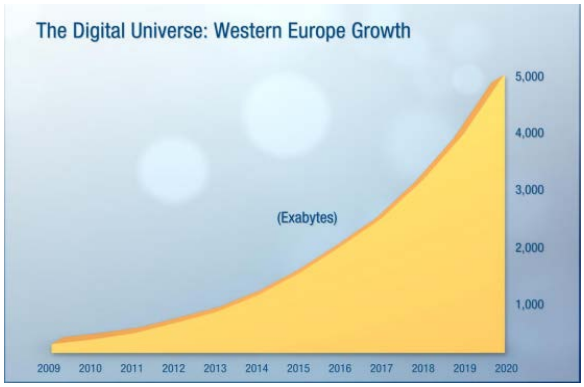

Fig. 2 (e). Data growth in Western Europe

Data growth *- IDC believes that the* digital u*niverse will grow by 44 times* from *2009* to *2020. IBM estimates that data and content is growing at a compound annual growth rate of 64% a year or more (1 zettabyte = 1 trillion gigabytes). Source: IDC Digital Universe Study.*

- **Data growth in China:** The digital universe in China is expected to grow from 364 exabytes to 8.6 zettabytes between 2012 and 2020 (**Fig. 2b**). China's share of the global digital universe will grow from 13% to 21% between 2012 and 2020. This means that, if it were printed out as text, China's digital universe in 2020 would make up a stack of books reaching from Earth to Pluto and back 30 times [17].

- **Data growth in India:** The digital universe in India is expected to grow from 127 exabytes to 2.9 zettabytes between 2012 and 2020 (**Fig. 2c**). India and China's total shares of the global digital universe will grow by up to 29% by 2020 [18].

- **Data growth in United States:** The digital universe in the US is expected to grow from 898 exabytes to 6.6 zettabytes between 2012 and 2020 (**Fig. 2d**) [19].

- **Data growth in Western Europe:** The digital universe in Western Europe is expected to grow from 538 exabytes to 5.0 zettabytes between 2012 and 2020 (**Fig. 2e**). This data, if printed, would make up a stack of books as high as 6.5 trillion Eifel Towers [20].

2. Significance of Semantic Computing

The term "semantic web" has been around for more than a decade. Its origins trace back to a 2001 *Scientific American* article by Tim Berners-Lee, who is known as the inventor of the world wide web, and co-authors James Hendler and Ora Lassila [25]. The article presents a futuristic view of the web, where data is linked in a meaningful fashion. To realize this concept, two main streams have emerged: ontology and the reasoning and filtering of data. The three main semantic web standards that currently exist are Resource Description Framework (RDF), SPARQL (SPARQL Protocol and RDF Query Language), and OWL (Web Ontology Language). The goal of these standards is to present end-users with the information that they want at a particular time. The World Wide Web Consortium (W3C), which is the international standards organization leading the semantic web effort, has carried out many case studies that show how organizations are currently using semantic technologies in a variety of areas [24]. The W3C envisions the semantic web as an extension rather than a replacement of the current web. As discussed above, one of the main streams of effort in the semantic web concerns ontology, and includes ontology engineering, evolution, matching, mapping, and merging. The other main stream, which involves filtering and reasoning, includes the areas of content filtering, collaborative filtering, hybrid filtering, and reasoning.

3. Existing Challenges in Semantic Approaches to Big Data

The challenges involved in dealing with big data include access [9], capture, storage [3], search, sharing, transfer, analysis [4], and visualization, with respect to volume, velocity, variety, and value. However, the focus of our research efforts is on the following three areas:

- **Large-scale taxonomy management for capturing big data**: One highly pertinent and massive challenge is the hierarchical auto-arrangement of unstructured data into classified concepts. This should involve understanding data and related concepts and correctly assigning the data to the right concepts and categories without involving human/manual (domain/IT expert) interpretation [21]. This challenge is related to the volume and variety of big data that is generated by the web, social media, streaming, blogs, and other sources. The auto-arrangement of data according to concepts requires an understanding of contexts and contents (tag annotation, classes of data, etc.). This should be based on a comparison of new data with an existing knowledge base, as well as the development of taxonomies. In the case of big data, there could be billions of triples linked to each other. The processing of such a magnitude of data must involve efficient ontological engineering approaches. Annotation of the unstructured data of millions of web users based on ontology is a massive challenge, particularly in areas such as annotating browsing logs of user history over the last five years.

- **Uncovering and accessing relevant data from big data:** Uncovering and accessing the relevant data within the large volume and variety of big data is an increasingly difficult task [9] [21]. According to [13], issues arise due to the three dimensions of big data: volume, since massive amounts of data have been accumulated over the decades, velocity, since the amounts may be rapidly increasing, and variety, since the data are spread over different formats. While this is a typical problem of dealing with big data, applications can be used to auto-understand the dynamics of context. For instance, traffic data can be added to road maps to provide context on road conditions, probability of delay, length of projected obstructions, road conditions, etc. Big data and analytics tools are already available in the market, from companies such as IBM, SAS, SAP, and Oracle.

- **Linking big data**: The concept of linked data (**Fig. 3**) has gained widespread currency, and has been successfully adopted by real-life IT-based companies, such as Google and Facebook. The concept of the Linked Open Data Cloud (LOD) has been used to integrate distributed data in the cloud.

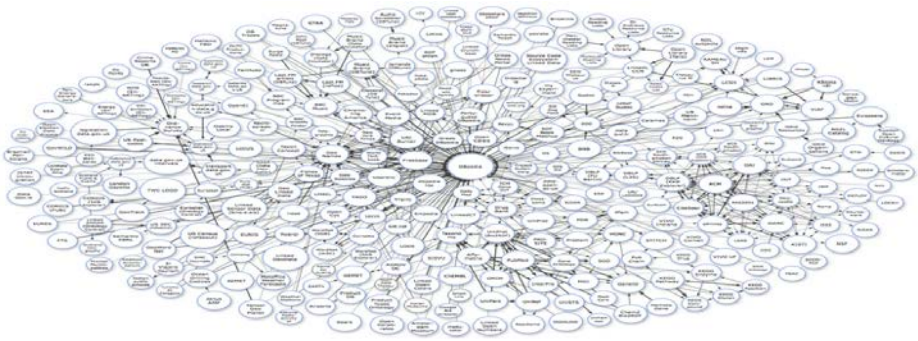

Fig. 3. A depiction of linked data (Source: [11])

4. Existing Approaches to the Semantic Computing of Big Data

A number of approaches have been proposed for semantically dealing with the issues of big data. These proposed approaches work at different levels, and include areas such as ontology engineering, ontology evolution, reasoning, matching and representing big data. Some details of these approaches are as follows:

- Ontology engineering and big data
  - Construction ontology: Large-scale structured, unstructured, and semi-structured data is transferred into ontological forms. A number of approaches have been proposed to deal with this issue [36] [37] [38] [40] [41] [52] [53].
  - Big data access: Since accessing relevant information is an increasingly difficult task, the concept of ontology-based data access (OBDA) [9] has been proposed as a suitable approach. OBDA systems address the data access problem by presenting a general ontology-based and end-user oriented query interface over heterogeneous data sources. The core elements in a classical OBDA systems are an ontology describing the application domain and a set of mappings, relating the ontological terms with the schemata of the underlying data source.
  - Understanding and linking big data: It is essential that the concept of linked data be used to understand relationships in a domain and link their concepts in taxonomies, through graphs or ontologies. Some examples of this approach can be seen in Google Knowledge Graph [10] and Facebook OpenGraph protocol [33], which use linked data to process information [50] [51].

- Ontology evolution for big data
  - The ongoing stream of data involved in big data make it natural for ontological techniques to evolve. There are several approaches that deal with this issue [34] [36] [50] [52].
- Searching and filtering big data for relevant information
  - Extracting and reasoning of big data
    - One of the primary areas of focus in dealing with big data is the extraction and reasoning of large-scale data. A number of approaches and tools are available for extracting relevant information from big data [34] [35] [39] [44] [45] [48] [58].
  - Distributed (web-scale) reasoning on big data
    - Some emerging topics within the distributed reasoning approach on big data have been discussed in past studies [57] [59] [60] [61].
  - Representing big data

- Representation of large-scale data poses significant challenges, as has been mentioned by past scholars [42] [43].

The following tables (**Table 1**, **Table 2**, **Table 3**, and **Table 4**) present summaries of semantic-based approaches that attempt to deal with big data. Table 1 presents our findings on semantics filtering and reasoning approaches proposed to deal with big data.

Table 1. Semantics filtering and reasoning approaches for big data

Reference	Issue	Approach	Main Focus	Variables / Parameters
[27]	Applying reasoning to enrich query results over a very large amount of data (i.e., web-scale data), using a parallel and distributed system	• MapReduce Reasoning algorithms • Grouping RDFS rules in four MapReduce jobs, referred to as SUBPROP, DOMAINRANGE, SUBCLASS, and SPECIAL CASES	• Web-scale reasoning	Volume: One billion to 100 billion RDF triples
[34]	Enabling scalable end-user access to big data	• Ontology-based approach, accompanied by query optimization and parallelization • Iterative user-centric development approach	• End-user visual query formulation, demonstrating the preliminary ontology-based visual query system (i.e., interface) • Alleviating the effects of big data	The volume, complexity, variety, and velocity dimensions of data, as well as their schemata
[35]	Presenting the contents displayed on a page dynamically, based on the viewer's context	• Novel approach to user profiling, based on information from user surfing logs • User profiles have to be able to model any combination of user characteristics • Presents relations between composing elements and the uncertainty that stems from the automated processing of real-world data	• Tackling issues by using a combination of data analysis, ontology engineering, and processing of big data resources provided by an industrial partner • Automatically constructing and populating a profile ontology for each user identified by the system	

[39]	Query transformation and optimization problems in the context of query answering within the Optique OBDA system	• Queries presented in heterogeneous distributed databases and streams for automatic big data query generation	• Query transformation and optimization problems, in the context of query answering in the Optique OBDA system • Optique's automatic generation of queries and two systems to support this process: QUEST and PEGASUS	User-friendly query formulation interface, Maintenance of ontology and mapping, processing and analytics over streaming data, distributed query optimization, and execution
[44]	Stream reasoning	• Analysis of related research fields, with the aim of extracting algorithms, models, techniques, and solutions	• Identifying the requirements involved in different application scenarios and isolating the problems that they pose • Surveying existing approaches and proposals in the area of stream reasoning, and highlighting their strengths and limitations	Frequent change of huge information
[45]	Semantic inconsistencies among applications and systems	• Management of metadata in smart grids to remove semantic inconsistencies	• Developing an ontological model for offshore wind energy meta-data management	Knowledge sharing and data exchange, derived data on relationships between concepts , data quality as metadata
[48]	Use of financial industry business ontology (FIBO) as a conceptual model	• Financial industry business ontology (FIBO)	• Possibility of creating semantic technology-based applications that can be used to carry out novel types of data processing	What to expect from FIBO and when

Table 2, below, presents our findings on ontology evolution approaches that have proposed for dealing with big data. It shows that only a few approaches have been proposed for ontology evolution in big data.

Table 2. Ontology evolution approaches for big data

Reference	Issue	Approach	Main Focus	Variables / Parameters
[34]	Scalable end-user access to big data	• Ontology-based approach, along with query optimization and parallelization • Iterative user-centric development approach	• End-user visual query formulation, demonstrating the authors' preliminary ontology-based visual query system (i.e., interface) • Alleviating the effects of big data	Volume, complexity, variety, and velocity dimensions of data, as well as their schemata
[36]	Providing an end-to-end solution for scalable ontology-based data access (OBDA) and big data integration, in which end-users formulate queries based on a familiar conceptualization of the underlying domain (i.e., over an ontology)	• Selection of possible logical and modeling errors in OBDA systems and the main challenges faced in supporting the life-cycles of OBDA systems	• The problem of bootstrapping and the maintenance of ontologies and mappings • The important challenge of debugging errors in ontologies and mappings • Presenting examples of different kinds of errors, and offering preliminary views on their debugging	Construction, maintenance, and transformation of an OBDA specification
[50]	Direct access to raw data not currently available on the web or bound up in hypertext documents	• Publication and consumption of linked data • Drawing on practical linked data scenario • Deciding what data to return in a description of a resource on the web • Methods and frameworks for automated linking of data sets	• The study of existing linked data applications and architectures	Raw data not currently available on the Web

[52]	Ontology matching evaluations	• Comparative experimental review of matching systems	• Discussing seven matching systems: • SAMBO (Linköping U.) • Falcon(Southeast U.) • DSSim (Open U., Poznan U. of Economics) • RiMOM (Tsinghua U., Hong Kong U. of Science and Technology) • ASMOV (INFOTECH Soft, Inc., U. of Miami) • Anchor-Flood (Toyohashi U. of Technology) • Agreement Maker (U. of Illinois at Chicago)	Semantic technologies for ontology matching

Table 3, below, presents our findings on ontology-based approaches for representing big data. There are only two proposed approaches in this area.

Table 3. Ontology-based approaches for representing big data

Reference	Issue	Approach	Main Focus	Variables / Parameters
[42]	High-performance computation, large-scale data storage , complexity of datasets in Bioinformatics for visual analytics	Data-intensive visualization engine (DIVE)	Increasing adoption of computational approaches among scientists, as well as use of data-centric scientific tools	The streaming and analyzing of large datasets at interactive speeds
[43]	Big data analysis, management, representation, and visualization	Linked data applications knowledge	Providing a conceptual analysis of the term “big data” and introducing linked data applications such as SKOS-based knowledge organization systems as new tools for the analysis, organization, representation, visualization, and accessing of big data	Theoretical study

Table 4 presents our findings on ontology engineering approaches to dealing with big data. It shows that there are numerous proposed approaches in this area.

Table 4. Ontology engineering approaches for big data

Reference	Issue	Approach	Main Focus	Variables / Parameters
[37]	Managing massive amounts of heterogeneous data Deriving knowledge from data instead of drowning in information	• Novel framework for the engineering of ontologies from observation data	• Proposes an observation-driven ontology engineering framework • Shows how its layers can be realized by using specific methodologies, and relates the framework to existing work on geo-ontologies	Framework employing geo-statistics, data mining, and machine learning to construct ontological primitives
[38]	Variety of issues involved in geo-information sciences, from its early stages to the present	• Automatic cartography of agricultural zones through multi-temporal segmentation of remote sensing images	• Theoretical study of different parameters	Structure, processing, data uncertainty, data consistency, ontology
[40]	Simple method for clustering information items that adds substantial vigor, resourcefulness, and flexibility to the investigation of large collections of nebulous data	• Based on the previously developed construction of FuzzyFind Dictionary, utilizing the error-correcting Golay Code	• Enhancing available information processing resources with a novel software/hardware technique for on-the-fly clusterization of amorphous data from diverse sources • Demonstrating that conventional clustering procedures with richer functionality, typically having O(n2) complexity, are prohibitively slow	New, simple, and efficacious tool for one of the most demanding operations of big data: methodology-clustering of diverse information items in a data stream mode

[41]	Engineering analysis model ontology and material ontology for storing knowledge about materials and FE models	• Interlinking of ontologies in a single, synergistic ontology approach that exposes and integrates knowledge from previously disparate domains like engineering and biology in a transparent manner	• Presenting a case study to demonstrate the usefulness of the approach • Explaining how knowledge from a biological material and FE model is methodically stored through the new ontology • Organismal classification and the anatomical structure of the model	Knowledge management
[46]	Business intelligence and analytics approaches and challenges	• Analysis of current research in BI&A • Identification of challenges and opportunities associated with BI&A research and education	• Text analytics (current research in BI&A is analyzed)	Informative study
[47]	Decoupling of the front end (devices, area networks) from the back end (applications, databases)	• Presents a solution that uses domain-specific filtering thresholds in a domain-agnostic platform, and which filters flows and algorithms tailored to modern M2M platforms • Proposed filtering approach is called M2M-NECtar	• Domain-specific filtering thresholds • Tackling the complexity of configuring heterogeneous filters in a horizontal and automated manner	Avoiding implementation details of filters and specifics of different verticals

[49]	Building a generic platform	• Explores key challenges in the field and employs a sensor data platform, "Concinnity," which can take sensor data from collection to final product via a data repository and workflow system	• Lack of a platform enabling sensor data to be taken from collections • Use of data in models to produce useful data products	Sensor data management platform, providing citizens with new real-time services and allowing more informed decision making
[51]	Incorporating NCBO web services into software applications to generate semantically aware applications and facilitate structured data collection	• Enabling community feedback through BioPortal content	• Growth in the number of large data sets, providing a framework for data analysis and data integration using ontologies	NA
[53]	Opportunities, challenges, and risks	• Developing new methods for data handling and big data analysis	• Critical reflection and research on big data, as well as studies using big data	Informative study
[54]	Query transformation and optimization problems in query answering through the Optique OBDA system	• Queries made over heterogeneous distributed databases and streams for automatic big data query generation	• Query transformation and optimization problems, in the context of query answering through the Optique OBDA system • Optique's automatic generation of queries and two systems to support this process: QUEST and PEGASUS	User-friendly query formulation interface, maintenance of ontology and mapping, processing and analytics over streaming data, distributed query optimization and execution

5. Existing Semantics-Based Tools for Big Data

Table 5 presents a summary of semantics-based tools for dealing with big data. Our research only discovered a limited number of such tools. A number of non-commercial tools are still ongoing projects, and many focus on specific domains. For example, SINA [29] focuses on the medical domain.

Table 5. Semantic/ontology-based tools for big data

Tool	Features Offered	Semantics
Oracle Big Data Appliance [55]	• Hadoop Distributed File System [7] • Oracle NoSQL Database • Storage capacity of 648TB 1152GB of memory	• Oracle Exadata • Oracle database • Integrated management
IBM Watson Foundations [6]	• Hadoop Distributed File System [7] IBM InfoSphere Platform IBM Stream Computing	• Mapping • Stream computing
Ontology 4 Platform [22]	• Semantic search across an enterprise • Structured and unstructured data handling	• Ontology • Semantic search
Optique: OBDA Solution for Big Data [9]	• EU FP7-funded Optique, which develops an end-to-end OBDA system providing scalable end-user access to industrial big data stores	• Ontologies • OWL 2
PigSPARQL [26]	• Utilizes Pig (Latin) and it is built on Hadoop	• SPARQL
Treo[28]	• Schema-agnostic/vocabulary-independent natural language queries • Designed to work with very heterogeneous databases • Comprehensive semantic matching based on distributional semantics • Supports queries over Linked Data/RDF	• RDF • Linked data
SINA [29]	• Three integrated medical databases • Natural language support	• RDF • Linked data
KRMA [30]	• Integrates data from a variety of data sources/databases, spreadsheets, delimited text files, XML, JSON, KML, and web APIs	• RDF
SchemEX [31]	• Stream-based indexing of linked data	• RDF • LOD
LODatio [32]	• Semantic search engine • Schema-based index to identify relevant sources for LOD for given SPARQL query patterns	• SPARQL • RDF
DIVE[42]	• Graph-based visual analytics	• Ontology • Direct SQL • Dynameomics data warehouse
BioPortal[51]	• Web services integrated into software applications	• Ontologies • Semantic mapping

6. Emerging Directions and Future Challenges for Semantics-Based Computing on Big Data

In our opinion, there is a great deal of research potential in semantics-based approaches, as challenges still need to be addressed in the following areas:

- **Big Dirty Data: Ambiguous, Inconsistent, Inaccurate, Incomplete and Redundant**

In almost all domains of IT systems, an enormous amount of scalable data is ambiguous, inconsistent, inaccurate, and incomplete. This is known as “dirty data.” In the context of big data, dirty data is also “big,” because big data is scalable and heterogeneous by nature. This data can thus be referred to as “big dirty data” (BDD). BDD could be one of the serious challenges facing semantic computing in a variety of areas (e.g., mapping, matching, aligning, reasoning, inferences). Although recent approaches have suggested solutions for managing large ontologies that contain redundant information [14], these approaches require human involvement to guide the computing process. For example, humans may need to manually set up the parameters for executing each step, starting with computing the alignment process between two concepts. Further steps must also be executed by humans, but may not be possible in dealing with BDD; auto-execution approaches are thus needed. With this in mind, new ontology capturing approaches should not only consider working on clean data, but also handling BDD.

- **Integrated Searching in Big Data**

Apple has announced its intentions to enter the search engine arena. It plans to integrate search systems and involve Twitter search integration for personalized results [23], Wikipedia integration, and Bing Web search results. A typical search on Bing could then be “Show the Wikipedia profile of the hero from the most discussed movie on Twitter last week.” This would involve a real linked data platform. For such a vast integrated search engine, semantics-based optimization techniques will be needed to traverse, reason, and make inferences for the exploratory and iterative analysis of data. A tool such as Ontology 4 [22] seems like a good starting point for this kind of 360-degree semantic search.

- **Transferring IT Solutions: From Reactive to Proactive**

In our opinion, thus far, the main focus of IT approaches has been the processing of data stored through past events, which is referred to as the reactive semantic approach. However, by using big data, it is possible to analyze the existing large volume of data and use it for effective and efficient prediction of future events, which is called the proactive semantic approach. Such predictions should not be misinterpreted by analyzing sample data, and instead should be made by using pieces of ontological big data in context.

- **High-Speed (Streaming) Data Capturing and Consumption**

The understanding, linking, and use of big data across domains is critical for future expansion. For instance, IBM's utilities project deals with computing intelligence, power-usage monitoring, sensors, and grid computing. In addition, there is a need to capture concepts from high-speed streaming data, which is likely to test the ontology engineering community. For instance, Siemens constantly takes in several terabytes of temporal data from sensors, with an increase rate of about 30 gigabytes per day [12]. As mentioned by [16], if data is organized and integrated so that the context and the reason for collecting the data is clear, it is much easier to manage the information and gain value from the knowledge generated.

- **Security Issues in Big Data**

The study in [15] has found that users have very poor awareness of how their data is being shared and used. They explain that as big linked data becomes more common, gaining users' trust and consent will become more important, as taking advantage of personal information will become an important concern.

7. Conclusion

In this paper, we have attempted to answer questions on areas such as growth trends in data, how semantic computing can help to process huge amounts of data and uncover valuable information within it, what semantic-based approaches and tools are available for processing big data, and what the future looks like, in terms of the level of efficiency required to semantically process big data. We have determined that a semantics-based strategy is a suitable approach to dealing with big data issues. It is worth mentioning that this study examined the latest literature available, from 2011 to 2014. A total of 61 research papers were studied to explore the state of the art in this area. However, this review found that new solutions are needed to extract value, given the volume, variety, and velocity of big data. The use of semantic computing approaches to big data could enable end-users to consume information that is relevant to them. This study also offers recommendations that may encourage researchers to more deeply explore the ways in which semantic technology can improve the processing of big data. Our research may pave the way for the development of better basic knowledge on the semantic and computational issues of big data, and can act as a foundation for further studies within the field.

References

[1] Dijcks, Jean Pierre. "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012. Article (CrossRef Link)

[2] Zikopoulos, Paul, and Chris Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011. Article (CrossRef Link)

[3] Red Hat Enterprise Linux: THE FIVE MUST-HAVES OF BIG DATA STORAGE, *White Paper*, 2013. Article (CrossRef Link)

[4] Ashley Vance, "Start-Up Goes After Big Data With Hadoop Helper," *New York Times Blog*, 2010. Article (CrossRef Link)
[5] Gartner Research: https://www.gartner.com/it-glossary/big-data/ Accessed on 03 April 2014. Article (CrossRef Link)
[6] http://www-01.ibm.com/software/data/bigdata/ , accessed on 4 April 2014. Article (CrossRef Link)
[7] http://hadoop.apache.org/ , accessed on 4 April 2014. Article (CrossRef Link)
[8] http://dictionary.ieee.org/ , accessed on 4 April 2014. Article (CrossRef Link)
[9] D. Calvanese, Martin Giese, Peter Haase, Ian Horrocks, T. Hubauer,Y. Ioannidis, Ernesto Jim´enez-Ruiz, E. Kharlamov, H. Kllapi, J. Kl¨uwer,Manolis Koubarakis, S. Lamparter, R. M¨oller, C. Neuenstadt, T. Nordtveit, Ö. Özçep, M. Rodriguez-Muro, M. Roshchin, F. Savo, Michael Schmidt,Ahmet Soylu, Arild Waaler, and Dmitriy Zheleznyakov, " Optique: OBDA Solution for Big Data." In *The Semantic Web: ESWC 2013 Satellite Events*, pp. 293-295, 2013. Article(CrossRef Link)
[10] http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html, accessed on 4 April 2014. Article(CrossRef Link)
[11] Richard Cyganiak and Anja Jentzsch. “Linking Open Data cloud diagram,” http://lod-cloud.net, accessed on 4 April 2014. Article(CrossRef Link)
[12] Horrocks, Ian, Thomas Hubauer, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Manolis Koubarakis, Ralf Möller, Konstantina Bereta, Christian Neuenstadt, Özgür Özçep, Mikhail Roshchin, Panayiotis Smeros and Dmitriy Zheleznyakov . "Addressing Streaming and Historical Data in OBDA Systems: Optique’s Approach (Statement of Interest)." In *Proc. of Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@ LOD)*, 2013. Article(CrossRef Link)
[13] Kharlamov, Evgeny, Ernesto Jiménez-Ruiz, Dmitriy Zheleznyakov, Dimitris Bilidas, Martin Giese, Peter Haase, Ian Horrocks, Herald Kllapi, Manolis Koubarakis, Özgür Özçep, Mariano Rodríguez-Muro, Riccardo Rosati, Michael Schmidt, Rudolf Schlatte, Ahmet Soylu, and Arild Waaler. "Optique: Towards OBDA Systems for Industry," *The Semantic Web: ESWC 2013 Satellite Events*, pp. 125-140, 2013. Article(CrossRef Link)
[14] Lambrix, Patrick, and Rajaram Kaliyaperumal. "A session-based approach for aligning large ontologies," *The Semantic Web: Semantics and Big Data*, pp. 46-60, 2013. Article(CrossRef Link)
[15] Huang, Bert, Angelika Kimmig, Lise Getoor, and Jennifer Golbeck. "A flexible framework for probabilistic models of social trust," *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 265-273, 2013. Article(CrossRef Link)
[16] http://www.globalgraphics.com/technology/knowledge-management/, accessed on 4 April 2014. Article(CrossRef Link)
[17] John Gantz, David Reinsel, Richard Lee, The digital universe in 2020: Big data, Bigger Digital Shadows and the Biggest Growth in Far East China, *IDC Country Brief, Sponsored by EMC*, 2013. Article(CrossRef Link)
[18] John Gantz, David Reinsel, Marshall Amaldas, The digital universe in 2020: Big data, Bigger Digital Shadows and the Biggest Growth in Far East India, *IDC Country Brief, Sponsored by EMC*, 2013. Article(CrossRef Link)
[19] John Gantz, David Reinsel, “The digital universe in 2020: Big data, Bigger Digital Shadows and the Biggest Growth in USA,” *IDC Country Brief, Sponsored by EMC*, 2013. Article(CrossRef Link)
[20] John Gantz, David Reinsel and Carla Arend, “The digital universe in 2020: Big data, Bigger Digital Shadows and the Biggest Growth in Europe,” *IDC Country Brief, Sponsored by EMC*, 2013. Article(CrossRef Link)
[21] Kouji Kozaki, “Ontology engineering for big data,” in *Proc. of Ontology and Semantic Web for Big Data Workshop in the ICSEC 2013*, September 4, 2013, Bangkok, Thailand. Article (CrossRef Link)
[22] http://www.ontology.com/, accessed on 6 April 2014. Article(CrossRef Link)

[23] www.apple.com/ios/siri, accessed on 4 April 2014. Article(CrossRef Link)
[24] http://www.w3.org/2001/sw/sweo/public/UseCases/, accessed on 4 April 2014. Article(CrossRef Link)
[25] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*,vol. 284, no. 5, pp. 1-5, 2001. Article(CrossRef Link)
[26] Schätzle, Alexander, Martin Przyjaciel-Zablocki, and Georg Lausen. "PigSPARQL: Mapping sparql to pig latin," in *Proc. of Proceedings of the International Workshop on Semantic Web Information Management*, p. 4. ACM, 2011. Article(CrossRef Link)
[27] Urbani, Jacopo. "On web-scale reasoning," PhD diss., Ph. D. dissertation, Comput. Sci. Dept., Vrije Universiteit, Amsterdam, Netherlands, 2013. Article(CrossRef Link)
[28] http://treo.deri.ie, .accessed on 3 April, 2014 Article(CrossRef Link)
[29] http://sina-linkeddata.aksw.org/, accessed on 5 April 2014. Article(CrossRef Link)
[30] Szekely, Pedro, Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, and Georgina Goodlander. "Connecting the Smithsonian American Art Museum to the Linked Data Cloud," *The Semantic Web: Semantics and Big Data*, pp. 593-607, 2013. Article(CrossRef Link)
[31] Konrath, Mathias, Thomas Gottron, Steffen Staab, and Ansgar Scherp. "Schemex—efficient construction of a data catalogue by stream-based indexing of linked data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol.16, pp.52-58, 2012. Article(CrossRef Link)
[32] Gottron, Thomas, Ansgar Scherp, Bastian Krayer, and Arne Peters. "LODatio: A Schema-Based Retrieval System for Linked Open Data at Web-Scale," *The Semantic Web: ESWC 2013 Satellite Events*, pp. 142-146, 2013. Article(CrossRef Link)
[33] https://developers.facebook.com/docs/opengraph accessed on 3 April, 2014. Article(CrossRef Link)
[34] Soylu, Ahmet, Martin Giese, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Dmitry Zheleznyakov, and Ian Horrocks. "OptiqueVQS: towards an ontology-based visual query system for big data," In *Proc. of Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems*, ACM, pp. 119-126. 2013. Article(CrossRef Link)
[35] Hoppe, Anett, C. Nicolle, and A. Roxin. "Automatic ontology-based user profile learning from heterogeneous web resources in a big data context," in *Proc. of Proceedings of the VLDB Endowment* , vol. 6, no. 12, pp. 1428-1433, 2013. Article(CrossRef Link)
[36] Haase, Peter, Ian Horrocks, Dag Hovland, Thomas Hubauer, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Johan Klüwer1 Christoph Pinkel et al. "Optique System: Towards Ontology and Mapping Management in OBDA Solutions," in *Proc. of Second International Workshop on Debugging Ontologies and Ontology Mappings-WoDOOM13*, p. 21. 2013. Article(CrossRef Link)
[37] Janowicz, Krzysztof. "Observation-Driven Geo-Ontology Engineering," *Transactions in GIS*, vol. 16, no. 3, pp. 351-374, 2012. Article(CrossRef Link)
[38] Jeansoulin, Robert. "Big data: how geo-information helped shape the future of data engineering," *AutoCarto Six Retrospective*, pp.190-201, 2013. Article(CrossRef Link)
[39] Bizer, Christian, Peter Boncz, Michael L. Brodie, and Orri Erling. "The meaningful use of big data: four perspectives--four challenges," *ACM SIGMOD Record*, *vol.*40, no. 4, pp.56-60, 2012. Article(CrossRef Link)
[40] Berkovich, Simon, and Duoduo Liao, "On clusterization of big data streams," in *Proc. of Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, p. 26, ACM, 2012. Article(CrossRef Link)

[41] McPherson, Jeffrey D., Ian R. Grosse, Sundar Krishnamurty, Jack C. Wileden, Elizabeth R. Dumont, and Michael A. Berthaume, "Integrating Biological and Engineering Ontologies," in *Proc. of ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. V02BT02A022-V02BT02A022. American Society of Mechanical Engineers, 2013. Article(CrossRef Link)

[42][42] Rysavy, Steven J., Dennis Bromley, and Valerie Daggett. "DIVE: A Graph-Based Visual-Analytics Framework for Big Data," *Computer Graphics and Applications, IEEE , vol.*34, no. 2, pp. 26-37, 2014. Article(CrossRef Link)

[43] Shiri, Ali, "Linked Data Meets Big Data: A Knowledge Organization Systems Perspective," *Advances in Classification Research Online*, *vol.*24, no. 1, 2014. Article(CrossRef Link)

[44] Margara, Alessandro, Jacopo Urbani, Frank van Harmelen, and Henri Bal. "Streaming the Web: Reasoning over Dynamic Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol.25, pp. 24-44, 2014. Article(CrossRef Link)

[45] Nguyen, Trinh Hoang, Vimala Nunavath, and Andreas Prinz, "Big Data Metadata Management in Smart Grids," In *Big Data and Internet of Things: A Roadmap for Smart Environments*, pp. 189-214, 2014. Article (CrossRef Link)

[46] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, vol. 36, no. 4, 2012. Article(CrossRef Link)

[47] Papageorgiou, Apostolos, Mischa Schmidt, Jaeseung Song, and Nobuharu Kami. "Smart M2M Data Filtering Using Domain-Specific Thresholds in Domain-Agnostic Platforms," In *Big Data (BigData Congress), 2013 IEEE International Congress on*, pp. 286-293. IEEE, 2013. Article(CrossRef Link)

[48] Bennett, Mike, "The financial industry business ontology: Best practice for big data," *Journal of Banking Regulation*, vol. 14, no. 3, pp. 255-268, 2013. Article(CrossRef Link)

[49] Lee, Chun-Hsiang, David Birch, Chao Wu, Dilshan Silva, Orestis Tsinalis, Yang Li, Shulin Yan, Moustafa Ghanem, and Yike Guo, "Building a generic platform for big sensor data application," in *Proc. of Big Data, 2013 IEEE International Conference on*, pp. 94-102. IEEE, 2013. Article(CrossRef Link)

[50] Heath, Tom, and Christian Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol.1, no. 1, pp.1-136, 2011. Article(CrossRef Link)

[51] Whetzel, Patricia L., Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," *Nucleic acids research* vol.39, no. suppl 2, pp. W541-W545, 2011. Article(CrossRef Link)

[52] Shvaiko, Pavel, and Jérôme Euzenat, "Ontology matching: state of the art and future challenges," *Knowledge and Data Engineering, IEEE Transactions on* vol.25, no.1, pp.158-176, 2013. Article(CrossRef Link)

[53] Kitchin, Rob, "Big data and human geography Opportunities, challenges and risks," *Dialogues in Human Geography*, vol.3, no. 3, pp. 262-267, 2013. Article(CrossRef Link)

[54][54] Calvanese, Diego, Ian Horrocks, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Michael Meier, Mariano Rodriguez-Muro, and Dmitriy Zheleznyakov, "On Rewriting and Answering Queries in OBDA Systems for Big Data (Short Paper)," in *Proc. of OWL Experiences and Directions Workshop (OWLED)*. 2013. Article(CrossRef Link)

[55] http://www.oracle.com/technetwork/database/bigdata-appliance/overview/index.html, accessed on 4 April 2014. Article(CrossRef Link)

[56] http://www.emc.com/leadership/digital-universe/index.htm, accessed on 5 April 2014. Article(CrossRef Link)

[57] Karnstedt, Marcel, Kai-Uwe Sattler, and Manfred Hauswirth, "Scalable distributed indexing and query processing over Linked Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 10, pp. 3-32, 2012. Article(CrossRef Link)

[58] Moro, Andrea, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit. "Semantic rule filtering for web-scale relation extraction," *The Semantic Web–ISWC 2013*, pp. 347-362. Springer Berlin Heidelberg, 2013. Article(CrossRef Link)

[59] Fatos Xhafa and Leonard Barolli, "Semantics, intelligent processing and services for big data," *Journal of Future Generation Computer Systems*, vol.37, pp. 201-202, 2014. Article(CrossRef Link)

[60] Urbani, Jacopo, Spyros Kotoulas, Jason Maassen, Frank Van Harmelen, and Henri Bal, "WebPIE: A Web-scale parallel inference engine using MapReduce,"*Web Semantics: Science, Services and Agents on the World Wide Web* vol. 10, pp. 59-75, 2012. Article(CrossRef Link)

[61] Ruben Verborgh, Miel Vander Sande, Pieter Colpaert, Sam Coppens, Erik Mannens, and Rik Van de Walle, "Web-Scale Querying through Linked Data Fragments," *Workshop on Linked Data on the Web (LDOW2014*) Seoul, South Korea, 2014. Article(CrossRef Link)

Seung Ryul Jeong is a Professor in the Graduate School of Business IT at Kookmin University, Korea. He holds a B.A. in Economics from Sogang University, Korea, an M.S. in MIS from University of Wisconsin, and a Ph.D. in MIS from the University of South Carolina, U.S.A. Dr. Jeong has published extensively in the information systems field, with over 60 publications in refereed journals like Journal of MIS, Communications of the ACM, Information and Management, Journal of Systems and Software, among others. Dr. Jeong's areas of interest are Process Management, Software Engineering, Systems Implementation, and Information Resource Management.

Imran Ghani is a Senior Lecturer at Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Campus. He received his Master of Information Technology Degree from UAAR (Pakistan), M.Sc. Computer Science from UTM (Malaysia) and Ph.D. from Kookmin University (South Korea). His research focus includes agile software development methods and practices, semantics techniques, secure software development life cycle, web services, software testing, enterprise architecture and software architecture.