Ontology and Clustering Based Heterogeneous Data Sources Integration

Abrar Omar Alkhamisi[1], Mostafa Saleh[2]
King Abdulaziz University, Kingdom of Saudi Arabia
AALKHAMISI0034@stu.kau.edu.sa, msherbini@kau.edu.sa

ABSTRACT

With the increased diversity among the available data sources, the demand for sufficient storage and processing methods has grown in recent years. Also, extracting the inherent knowledge from the massive as well as heterogeneous information sources becomes an emerging research area. The abundance of the data quantity in the organizations has remarkably raised the gap between the disparate information systems. The advancement of semantic web technologies plays a prominent role in the decision making by providing the contextual information of the data residing in the heterogeneous data sources. Despite, lack of updation with the modifications in the information systems leads to ineffective decision making. To overcome this constraint, the existing researchers have effectively managed the data sources and supported the interoperability among the diverse data sources. Hence, it is significant to integrate the heterogeneous data sources into unified information-rich systems with the help of ontologies. Thus, this work presents heterogeneous data source integration model based on novel semantic ontology. The proposed integration model involves two phases such as novel kernel-based similarity learning and enhanced K-nearest neighbor clustering. By utilizing the novel semantic ontology, it effectively integrates the heterogeneous data sources. Initially, the proposed semantic data integration model obtains the potential information from the data sources such as Header information comprising the IP address, Source details, and destination details. Secondly, it applies the kernel-based similarity learning to compute the similarity between the heterogeneous data sources. Finally, the proposed integration model groups similar data using K-Nearest Neighbor (KNN) based clustering method. Thus, the experimental framework outperforms the existing data based web discovery methods by accomplishing improved performance in terms of the accuracy, success rate and execution time.

Key words : Semantic web, ontologies, integration, heterogeneous information, kernel-based similarity learning, KNN based clustering model.

1. INTRODUCTION

In the modern generation of information technology, the Internet of things (IoT) plays a significant role[1]. It has widely used in the integration of network through pervasive computing, IntelliSense, recognition technology, and other communication technologies [2]. Due to the development of IoT, the applications of information industry increased to virtually across all areas, domains, and industries in a wide range. With the tremendous growth of the vast available information with greater variety, velocity, and complex data, the demand for advanced methods has rapidly increased to precisely manage the data. Moreover, modeling a useful data management tool is inevitable to handle the heterogeneous as well as massive data within a tolerable time. In essence, the voluminous and diverse nature of the databases necessitates the semantic enhancement to recognize the insights of the data residing in the databases. Among various techniques, semantic web technologies play a prominent role in IoT based applications. Recently, the ontology-dependent systems pave the way for development of semantic web technologies [3]. It makes a defined ontology language, and it is based on shared knowledge. The semantic web such as ontology source globally manages the knowledge and assists to handle the massive amount of information available on the web. The services available on the web are decoupled in nature. To determine the acute web services such as robotic discovery, composition, invocation, and interoperation, the semantic web services have been employed in different application fields. It manipulates the descriptions of web services to discover various kinds of web services [4].

Traditional data management methods [5] lack to employ explicit ontologies to provide contextual and understandable information to the machines while developing the applications. Hence, there is a considerable need for integrating the multiple disparate data sources [6]. Data integration is the process of linking or unifying the variety of data sources that comprises diverse data types and data formats, which generates a consistent view of the data [7]. Several existing semantic-based data integration research employs the different semantic sources and methods to deal with the enormous heterogeneity in the databases effectively. Although handling the diverse data sources for integration is still in its infancy stage due to the existence of the data as well

as semantic heterogeneity [8]. The data heterogeneity involves three levels of heterogeneity such as syntactic, semantic, and structural. Due to the increased complexity in the identification of the semantic heterogeneity, most of the research works intensively focused on resolving the semantic heterogeneity in the past years. Automatically identifying the semantic relations between the disparate data sources and representing the identified relations in unified structure is a quite tricky task. Hence, the conventional data integration systems [9, 10] utilize the ontology as the intermediary between the diverse data sources. Several previous research works employ the clustering methods [11, 12] to identify the semantic relations between the heterogeneous data sources. However, extracting the inherent knowledge of the abundance of the data is a challenging task during the integration of heterogeneous data sources. Hence, the proposed approach targets to apply the clustering method along with the help of the ontology to improve the performance of the disparate data source integration.

The remaining of this paper is structured as follows: Section 2 reviews the existing research works in the semantic data integration. Section 3 presents the problem considered in the data integration system. Section 4 describes the aims and objectives of the research work. Section 5 carries out the proposed semantic data integration model. The steps involved in the proposed architecture are discussed and illustrated in Section 6. Sections 7 describes the discussion of the work. Finally, the conclusion is drawn in Section 8.

2. LITERATURE REVIEW

This work focuses on identifying the solution for integrating the multiple and heterogeneous data sources into a unified system using the ontology source. This section reviews the existing research works in the heterogeneous data integration system.

Various wearable data integration approaches [13] utilizes the semantic source along with the scalable machine learning algorithm to integrate the disparate data for healthcare application. Even though it deals with the heterogeneity and scalability constraints, handling of uncertain data requires the advanced analytic model. ODMedit [14] integrates the medical data based on the semantic annotation by enhancing the UMLS or SNOMED CT terminologies. It addresses the web of linked data and semantically integrates the patient details in the public metadata repositories. Although it partially supports the integration of the different types of medical reports of the patients. An Ontology-Based Data Management (OBDM) approach [15] employs the ontological source to develop the global schema for the Science, Technology, and Innovation (STI) policy development. It conceptually integrates the information sources related to the STI and maintains the large-scale information by tracking the information evolution. However, it lacks to support the integration of the heterogeneous information sources with

interoperability. Ontology-based streaming data integration model [16] eases the complex diagnostic tasks in the OPTIQUE system through abstraction layer. The semantic web-based abstraction layer enforces the integration of the heterogeneous data and supports the high-level semantic queries. Although, it integrates only the static relational data streams.

An Information Integration Framework [17] improves the service as well as maintenance processes through semantic annotation. By utilizing the linked data technologies, it integrates the heterogeneous data sources to offer the information units. It applies the machine learning method to obtain better reports along with strong concepts by classifying the information from technical manuals with ontology-based auto-completion. In consequence, it provides a unified as well as structured interface by integrating these two approaches for the manual as well as automatic query processing. Moreover, the integration framework effectively utilizes the uncontrolled documents as the data sources through ontology-based automated tool even when there are non-grammatical and non-standardized acronyms and labels. Finally, it facilitates the annotation and thus, leads to the ease of re-utilizing the uncontrolled documents such as reports as the information sources. Although, it partially extends the integration model using simple page range finding based on the standard of web annotation. Semi-automatic ontology matching approach [18] supports the integration of heterogeneous data sources. It employs the ontology source to provide the interoperability across the different data models and new data sources along with the stored data updation. It provides potential benefits to the Service-Oriented Architecture (SOA)-based industrial systems and agent-based industrial systems in addition to the automotive system. However, it suffers from higher computational overheads during ontology mapping.

Ontology-based heterogeneous database integration approach [19] builds the semantic ontology by semi-automatically extracting the database schema. It transforms one of the multiple databases into ontology and maps other heterogeneous relational databases with the ontology. In subsequence, it accesses the integrated ontology through the SPARQL query. It also encounters several shortcomings during data integration due to the different schemas and structures of the databases. A unified kernel approach [20] has presented a model for learning the cluster indicator matrix and similarity information simultaneously in the kernel spaces. Also, it analyzes the theoretical relationship amongst the Kernel K-means, spectral and K-means clustering. By extending the kernel model with the multiple kernel learning ability, it addresses the limitations in the selection of the most appropriate kernel for each clustering task. In consequence, it automatically determines the best

cluster indicator matrix amongst the three subtasks. Even though it obtains the improved performance for the clustering task, it has higher computational complexity.

Data fusion approach [21] performs the real-time data processing of automotive applications based on spatial indexation. It supports the fusion over the millions of the connected vehicles and also, speed up the automotive application through valid pattern matching although it lacks to handle the uncertain and noisy streams in the automotive application. Large-scale data management approach [22] effectively manages the sensor data streams and facilitates query processing using factorization and query reformulation algorithms. Even though it handles the redundancy issues in the data streams during management, it fails to remove the redundant streams contextually. An agile data architecture based semantic integration [23] manipulates the real-time data representation and supports the querying over the multiple varieties of data streams. By utilizing the semantic web technologies, it integrates the heterogeneous data streams for semantic data representation, which ensures the interoperability among the heterogeneous data sources. Even though it performs the incremental stream data processing and unusual event detection, it lacks to handle the noise and uncertain data streams during integration. Multimedia data integration approach [24] employs the semantic source to create a unified ontology for the large-scale multimedia knowledge bases. It precisely identifies the mapping relationship among the different ontologies. By utilizing the WordNet ontology as the background knowledge, it contextually matches the entities across the ontologies. However, the weight-based entity matching partially mitigates the sense of the glossary overlapping.

3. PROBLEM STATEMENT

Over the past decades, supporting the interoperation among the disparate data sources is an arduous task due to the variation in their schema and representation. The database management system often encounters the issues during representation mapping, schema mapping, and mapping based query answering. Also, handling the semantic heterogeneity is a critical task during the integration of the heterogeneous data sources. Even though the existing data integration researches have employed the ontological source to deal with the semantic heterogeneity, there is the possibility of occurring the semantic conflicts due to the heterogeneity in the syntactic representation and similarity in the interpretation of the elements. Moreover, computing the similarity among the heterogeneous data sources is a challenging task during the integration. Lack of analyzing the

aspect of the data hidden in the disparate databases leads to inaccurate integration, which affects the query processing.

4. AIMS AND OBJECTIVES

- To integrate the disparate and massive data sources based on the semantic knowledge.
- To determine the interrelationship across the heterogeneous data sources using Similarity Learning method.
- To group the semantically similar data sources using an Enhanced clustering method.

5. PROPOSED HETEROGENEOUS DATA SOURCES INTEGRATION BASED ON SEMANTIC ONTOLOGY

With the rapid advancements of the Web and Information Technology, the available digital information has drastically increased in terms of massiveness and diversity. The query processing often confronts with the big data characteristics; hence, semantically enriching the heterogeneous data sources is essential. Due to the diversified nature of the information, the data management system necessitates the semantic source to understand the information contextually. Hence, the proposed approach focuses on employing the semantic source such as ontologies for handling the heterogeneous big data sources during the integration with the target of accomplishing the accurate query processing within a short period. Figure 1 shows the proposed integration model based on the semantic ontology.
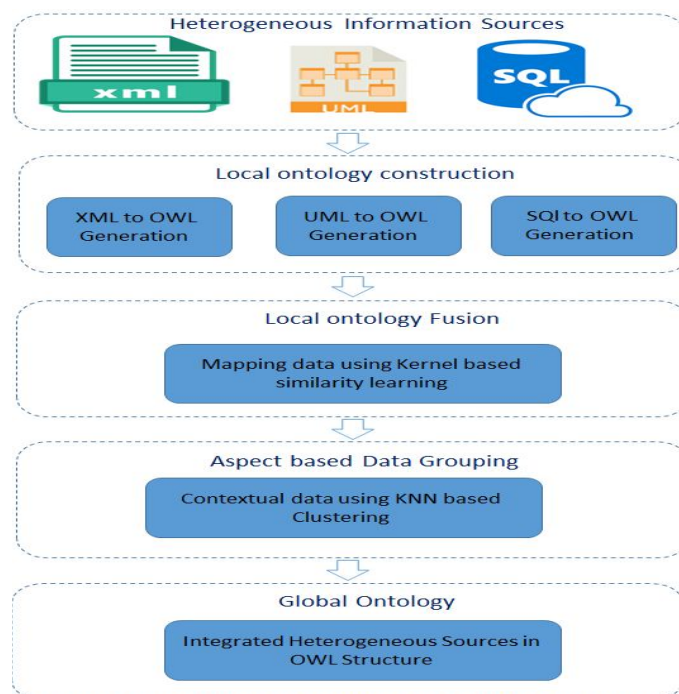
Figure 1: Architecture of Disparate Data Source Integration based on Semantic Ontology.

The proposed approach focuses on integrating the heterogeneous data sources with the assistance of the semantic ontology. Initially, it takes the input data from the heterogeneous data sources and builds the local ontology for different data sources. The proposed integration model analyzes the information such as the Header that holds IP address, Source details, and destination details for measuring the similarity among the heterogeneous data sources using kernel-based similarity learning. Moreover, to handle the different aspects of the data in the databases, it employs the K-Nearest Neighbor-based clustering model to group the data having similar aspects. The main goal of applying the clustering method in the data integration model is to segregate the data from the similarity results for the classes. In consequence, the proposed model differentiates the classes by computing the similarities between the data to ease the construction of global ontology. Finally, it builds global ontology based on the clustered data of the heterogeneous data sources.

5.1 Key Components in the Proposed Architecture

According to Figure 1, the heterogeneous data source integration model heavily relies on the kernel-based similarity learning method for the local ontology fusion and KNN-based clustering method for the aspect based data grouping.

Instead of applying the similarity and distance measures, kernel-based similarity learning method has gained significant attention, especially for heterogeneous data sources. The similarity measurement method computes the commonness between the data, whereas, the distance measurement method computes the difference between the data. In contrast, the kernel-based similarity learning has been used to improve both the accuracy and time efficiency by considering the linear and non-linear relationships across the heterogeneous data sources. Moreover, compared to the KNN classification method, the clustering-based KNN method achieves higher accuracy as well as time efficiency. It is because the KNN classification method computes the distance between each data with other data in the training set, which extends the time. Hence, to effectively handle the large-scale data, the KNN clustering finds the neighbors of a particular instance within the cluster, which alleviates the time consumption. Also, the separation of the data from the overlapped clusters for the classes yields higher accuracy of the disparate data sources. Hence, the proposed approach combines kernel-based similarity learning and KNN clustering to handle large-scale and heterogeneous information sources.

- **Kernel-based Similarity learning:** The kernel-based similarity learning is based on the kernel function. The kernel function is the result of the inner product between the data residing in the feature space. It plays a crucial role in learning the heterogeneous data by transforming the original feature space into a high-dimensional feature space. By applying the inner dot products, the kernel-based similarity learning model has been used for different purposes such as classification, clustering, or regression.

- **K-Nearest Neighbor based Clustering:** The K-NN based clustering model is based on the neighbors concerning the distance. In the K-NN algorithm, the distance between the data is measured by the Euclidean distance method [25]. In subsequence, a distance matrix was constructed from all the feasible pairs of the data using the distance values. In K-NN, 'K' represents the number of neighbors for each data point, which are identified from the distance matrix. Consider, each data holds the class label; hence, from the analysis of the neighbors, the most existing class labels are assigned as the class for a particular set of neighbors. In the K - NN algorithm, there is the possibility of occurring two or more class labels equally for a particular data point. Hence, it recursively runs the algorithm with K-1 that is with one less neighbor for the data point.

5.2 Case study for the Proposed Semantic Integration Model

Consider the medical data integration case study to demonstrate the proposed semantic integration model. With the massive increase of the medical data, providing healthcare solutions is critical due to the dependency with various factors. To provide the treatment for the patient or predict the morbidity, the solution provider such as physician needs all the data about the patient. The potential information sources include clinical reports, images, laboratory reports, hospital databases, and so on. Such patient information comprises the details regarding the age, habits, and disease of the patients, organ anatomy, functional information, and so on.

Retrieving healthcare data from the heterogeneous medical information sources of different health organizations necessitates the data integration model. For example, consider Data source 1 comprises the patient information such as patient ID, age, organization name, health-number, zip-code, and morbidity, and district name. Data source 2 comprises the information includes patient ID, age, health-number, diagnosis, hospital code, treatment, district code, and allergies. In this scenario, the semantic integration model builds the local ontology for two different data sources with the help of the semantic source. The local ontology is in

the structure of the classes and sub-classes of the patient ID with the corresponding features residing in the data sources. The integration model targets to reduce the redundant storage of the medical information for the same patient. Moreover, with the help of the kernel-based similarity learning, the proposed approach analyzes the features in the two data sources and measures the similarity between the features in the two data sources. If both the data source one and Data source two stores the details of the patient ID 1 with same age, health number, district code, and so on, the integration model links the information and models the non-redundant features for each patient. In subsequence, the K-NN based clustering model groups a set of patients under a diagnosis, age, or treatment based on the distance-based neighborhoods. Finally, the proposed approach builds the global ontology in the form of hierarchical structure with the patients and their other medical features under a particular class label for the different medical information sources.

6. PROCESS OF THE PROPOSED HETEROGENEOUS DATA SOURCES INTEGRATION BASED ON SEMANTIC ONTOLOGY

Figure 2 shows the overall process of the proposed data source integration model using a semantic ontology. The steps involved in the proposed integration model are discussed as follows:

Step 1: Load the heterogeneous data sources

Step 2: Build the local ontology for the heterogeneous data sources using the semantic ontology source

Step 3: Perform Coarse-grained Integration

Step 3.1: Apply the kernel-based similarity learning of the constructed local ontologies

Step 3.2: Analyze the features in the heterogeneous data sources such as header information that includes the IP address, source details, and destination details

**Step 3.2:** Integrate the local ontologies through similarity-based feature mapping

Step 4: Fine-grained Integration

Step 4.1: Apply K-NN based clustering method on the coarse-grained integrated data

Step 4.2: Analyze the aspect of the data for the similarity measured data

Step 4.3: Group the similar aspect of the data to contextually integrate the data obtained from the different data sources

Step 5: Finally, build the global ontology with the semantic relationship for the heterogeneous information sources.
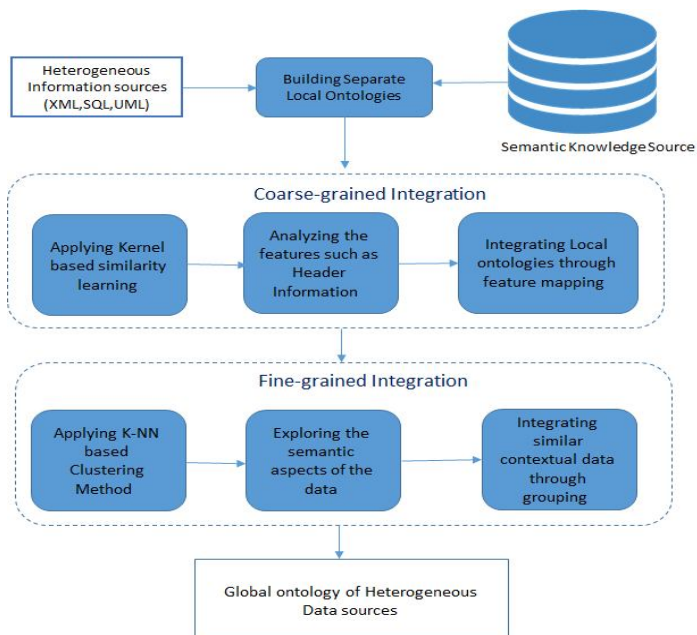
Figure 2: Overall process of the proposed heterogeneous data sources integration based on semantic ontology.

7. DISCUSSION

The performance evaluation with other existing data-based web discovery methods will illustrate the performance through the accuracy, success rate, and execution time metrics. The proposed work will be achieve better performance than the existing integration methods in terms of increased success rate, less execution time, and high accuracy. Table 1 shows the performance of the different baseline methods for data mapping in the integration system.

Table 1: Comparative Performance of Various Methods for Heterogeneous Data Integration.

Methods for Heterogeneous Data Integration	Key Reasons behind Performance	
	Accuracy	Execution Time
Similarity and Distance Measure	Direct similarity and dissimilarity measurement becomes inaccurate for disparate data.	Higher time consumption over large-scale data.
KNN Classification	Overlapping of the neighbors in different sets leads to less accuracy.	Instance-based mapping method leads to poor efficiency.
Kernel-Based Similarity Learning	Deals with both the linear and non-linear relationships, achieves better accuracy.	Better time efficiency.
KNN Clustering	Overcomes the constraints in KNN classification.	Achieves high efficiency.
Proposed Semantic Integration	Kernel-based similarity learning and clustering-based KNN yields higher accuracy.	Accomplishes higher time efficiency.

8. CONCLUSION

Recently, addressing semantic heterogeneity is a challenging task while managing heterogeneous information sources. To handle the semantic heterogeneity, the ontology source has been widely used by the data management systems. This work has presented different data source integration with the help of the semantic ontology. The proposed integration model involves local ontology construction, coarse-grained integration, and fine-grained integration. The coarse-grained integration phase employs the kernel-based similarity learning model and analyzes the data, such as Header information to compute the similarity among the data of different sources. Then, the fine-grained integration model applies the K-NN based clustering method to group the aspect based data from the similarity measured data. Finally, the proposed integration model generates the global ontology for the heterogeneous information sources with the contextual relationship.

In the future, this work further extends to different research directions:

- The proposed integration model will support the interoperability for the integration of the multiple modalities.
- The semantic integration will focus on reducing the query response time in the large-scale data through semantic annotation.
- The future direction will target on effectively integrating the deep-rooted disparate data with the help of an intelligent learning model.

REFERENCES

1. Preeti Gulia and Ayushi Chahal, **Big Data Analytics for IoT**, *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(6), 2020, pp. 593-603. http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=6

2. Xu Guangquan, Yan Cao, Yuanyuan Ren, Xiaohong Li, and Zhiyong Feng, **Network security situation awareness based on semantic ontology and user-defined rules for Internet of Things**, *IEEE Access*, Vol.5, pp.21046-21056, 2017 https://doi.org/10.1109/ACCESS.2017.2734681

3. El Hajjamy Oussama, Larbi Alaoui, and Mohamed Bahaj, **Integration of Heterogeneous Classical Data Sources in an Ontological Database,** *In International Conference on Big Data, Cloud and Applications, Springer*, pp.417-432, 2018

4. Lemos Angel Lagares, Florian Daniel, and Boualem Benatallah, **Web service composition: a survey of techniques and tools**, *ACM Computing Surveys (CSUR)*, Vol.48, No.3, p.33, 2016

5. De Giacomo Giuseppe, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati, **Using ontologies for semantic data integration**, *In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, Springer*, pp.187-202, 2018

6. Daraio, Cinzia, Maurizio Lenzerini, Claudio Leporelli, Paolo Naggar, Andrea Bonaccorsi, and Alessandro Bartolucci, **The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality**, *Scientometrics*, Vol.108, No.1, pp.441-455, 2016

7. Hui Jingya, Lingli Li, and Zhaogong Zhang, **Integration of big data: a survey**, *In International Conference of Pioneering Computer Scientists, Engineers and Educators, Springer*, pp.101-121, 2018

8. Hendler James, **Data integration for heterogeneous datasets**, *Big data*, Vol.2, No.4, pp.205-215, 2014

9. Maree Mohammed, and Mohammed Belkhatir, **Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies**, *Knowledge-Based Systems*, Vol.73, pp.199-211, 2015

10. Nafissi Anahita, Benjamin Bruns, and Fabio Fiorani, **Ontologies for resolving semantic heterogeneity in information integration among plant phenomics databases**, Vol.38, 2018

11. Kirk Paul, Jim E. Griffin, Richard S. Savage, Zoubin Ghahramani, and David L. Wild, **Bayesian correlated clustering to integrate multiple datasets**, *Bioinformatics 28*, No.24, pp.3290-3297, 2012 https://doi.org/10.1093/bioinformatics/bts595

12. Hameed Pathima Nusrath, Karin Verspoor, Snezana Kusljic, and Saman Halgamuge, **A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration**, *BMC bioinformatics*, Vol.19, No.1, p.129, 2018.

13. Mezghani Emna, Ernesto Exposito, Khalil Drira, Marcos Da Silveira, and Cédric Pruski, **A semantic big data platform for integrating heterogeneous wearable data in healthcare**, *Journal of medical systems*, Vol.39, No.12, p.185, 2015

14. Dugas Martin, Alexandra Meidt, Philipp Neuhaus, Michael Storck, and Julian Varghese, **ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository**, *BMC medical research methodology*, Vol.16, No.1, p.65, 2016

15. Daraio Cinzia, Maurizio Lenzerini, Claudio Leporelli, Henk F. Moed, Paolo Naggar, Andrea Bonaccorsi, and Alessandro Bartolucci, **Data integration for research and innovation policy: an Ontology-Based Data Management approach**, *Scientometrics*, Vol.106, No.2, pp.857-871, 2016

16. Kharlamov Evgeny, Sebastian Brandt, Ernesto Jimenez-Ruiz, Yannis Kotidis, Steffen Lamparter,

Theofilos Mailis, Christian Neuenstadt et al, **Ontology-based integration of streaming and static relational data with optique**, ACM *Proceedings of the International Conference on Management of Data*, pp.2109-2112, 2016

17. Bader Sebastian, and Jan Oevermann, **Semantic Annotation of Heterogeneous Data Sources: Towards an Integrated Information Framework for Service Technicians**, *ACM Proceedings of the 13th International Conference on Semantic Systems*, pp.73-80, 2017

18. Jirkovský Václav, Petr Kadera, and Nestor Rychtyckyj, **Semi-automatic ontology matching approach for integration of various data models in automotive**, *In International conference on industrial applications of holonic and multi-agent systems, Springer*, pp.53-65, 2017

19. Zhao Shuai, and Quan Qian, **Ontology based heterogeneous materials database integration and semantic query**, *AIP Advances*, Vol.7, No.10, p.105325, 2017

20. Kang Zhao, Chong Peng, and Qiang Cheng, **Twin learning for similarity and clustering: A unified kernel approach**, *In Thirty-First AAAI Conference on Artificial Intelligence*, 2017

21. Haroun Amir, Ahmed Mostefaoui, and François Dessables, **Data fusion in automotive applications**, *Personal and Ubiquitous Computing*, Vol.21, No.3, pp.443-455, 2017

22. Karim Farah, Mohamed Nadjib Mami, Maria-Esther Vidal, and Sören Auer, **Large-scale storage and query processing for semantic sensor data**, *ACM Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, p.8, 2017

23. Dividino Renata, Amilcar Soares, Stan Matwin, Anthony W. Isenor, Sean Webb, and Matthew Brousseau, **Semantic integration of real-time heterogeneous data streams for ocean-related decision making**, *Big Data and Artificial Intelligence for Military Decision Making*, 2018

24. Rinaldi Antonio M., and Cristiano Russo, **A matching framework for multimedia data integration using semantics and ontologies**, *IEEE 12th International Conference on Semantic Computing (ICSC)*, pp.363-368, 2018
https://doi.org/10.1109/ICSC.2018.00074

25. S. Dahiya, **Using Clustering techniques and Classification Mechanisms for Fault Diagnosis**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2138-2146, 2020. Available: 10.30534/ijatcse/2020/188922020.