Semantic Knowledge Based Graph Model in Smart Cities

Saqib Ali[1,2], Guojun Wang[1(✉)], Komal Fatima[2], and Pin Liu[3]

[1] School of Computer Science, Guangzhou University,
Guangzhou 510006, China
{saqibali,csgjwang}@gzhu.edu.cn
[2] Department of Computer Science, University of Agriculture,
Faisalabad 38000, Pakistan
{saqib,2014ag5719}@uaf.edu.pk
[3] School of Computer Science and Engineering, Central South University,
Changsha 410083, China
jiandanglp@csu.edu.cn

Abstract. In smart cities, pervasive IoT devices generate an elephantine amount of multi-source heterogeneous data. The semantics helps to explore such complex datasets and drive towards higher-level insights. Later, these high-level insights are transformed to develop interlinks and associations among diverse sources of the data which leads towards knowledge discovery in a smart city. This discovery when combines with the domain knowledge using ontology-based approaches develop concepts and perceptions which initiate decision making in complex environments. However, the ontology-based approaches come up with certain limitations including an incapability to transform semi-structured data into useful knowledge, issues in handling inconsistent data, and inability to process large-scale, multi-source, and complex data of smart cities. Therefore, in this paper, we proposed a Semantic Knowledge Based Graph (SKBG) model as a solution to overcomes these limitations. The SKBG model is particularly customized to a smart city environment and purely utilizes knowledge-based graphs to incorporate any type of domain knowledge by combining diversify domains as a unit. As a result, the model works fine with diverse domain knowledge, automatically classify heterogeneous data by using machine learning techniques, handle large knowledge databases and support intelligent semantic search algorithms in smart cities. Finally, the results are summarized in the form of a knowledge graph which gives a comprehensive insight into the data.

Keywords: Smart cities · Semantic Knowledge Based Graph model · Semantic data mining · Ontology-based approaches · Linked data

1 Introduction

In smart cities, highly innovative technologies and services are emerging which produce an elephantine amount of interlinked heterogeneous data [6]. This big

data in smart cities is quite challenging to harvest it timely and to get useful data patterns due to its volume, velocity, and variety [4]. However, it creates great opportunities for data analytics in the field of semantic data mining and knowledge discovery. In semantic data mining, we process on the semantics of the data by in cooperating the domain knowledge [19,21]. Further, the domain knowledge is supplemented with particular semantics which helps to analyze the relationships between document set and terms resides in the document set to highlight new domain concepts and insights.

Normally, when semantic data mining is applied to widespread contextual data of smart cities, we practice formal ontologies for processing semantics and are known as ontology-based approaches [17,18]. The principal step of these ontology-based approaches is data preprocessing. The preprocessing phase helps to find out the semantic gaps and missing data between the entities or actors of the smart cities [20]. Further, it instruments vital procedures of cleanness, normalization, integration, transformation, extraction, and feature selection to explicitly specify the concepts and patterns that support domain knowledge which helps to take rightful decisions in smart cities [9,30].

There is no doubt that ontology-based approaches provide a set of techniques for data modeling, defining features and concepts of formal semantics [16]. However, the ontology-based approaches come up with certain limitations. One such limitation is its inability to transform semi-structured data into useful knowledge [12]. Second, there are limited techniques for exploring knowledge in ontology-based approaches. Third, sometimes it allows inconsistent data to be loaded into a database with traditional data processing techniques [1]. Finally, the scarcity of robust algorithms and techniques to process large-scale, complex and heterogeneous data of smart cities using the full strength of the ontologies [14,23].

Multiple methodologies exist in literature to overcome the limitations of ontology-based approaches and data preprocessing. For example, semantic annotations, filter and multivariate methods for feature selection and different taxonomies for better classification [9,13]. Particularly, the semantic annotations proposed a technique to deal with semi-structured data. For this purpose, a semantic search algorithm is used to bring out meaning in the semantic data and annotate the semi-structured data [15]. Similarly, many featured based models are also proposed to classify, extract and select right terms for building search models in the smart cities [24]. However, these algorithms do not fulfill the requirements of handling the diverse amount of high-speed data in smart cities [10].

In this paper, we proposed a Semantic Knowledge Based Graph (SKBG) model as a solution which overcomes the basic limitations of conventional ontology-based approaches as discussed earlier. The proposed SKGB model is particularly customized to a smart city environment which works seamlessly upon semantics by using knowledge-based graph. Our SKBG model interlinks heterogeneous data, find meaning, concepts, and patterns of the data in smart cities. The model purely utilizes knowledge-based graphs to incorporate any type

of domain knowledge by combining diversify domains as a unit. In particular, it combines three terms i.e., text mining, machine learning, and knowledge-based graph to search out semantics, interlinking them by finding relationships among them, discover unique patterns in data and representation of information. As a result, the model works fine with diverse domain knowledge, automatically classify heterogeneous data, handle large knowledge databases and support intelligent semantic search algorithms by using machine learning techniques in smart cities.

The main contributions of this paper are summarized as follows.

– First of all, the limitations of ontology-based approaches and data preprocessing in semantic data mining regarding smart cities are thoroughly analyzed.
– Secondly, we propose SKBG model for semantic data mining using knowledge-based graphs for complex and heterogeneous data from the diverse origins in the smart cities.
– Finally, the key features of the proposed model are explained and analyzed to have a better insight into the model concerning the challenges features of the smart cities.

The remaining of the paper is organized as follows. Section 2 describes the related work. Section 3 gives a brief description of ontology-based approaches and data preprocessing regarding smart cities. Section 4 describes the proposed model. Section 5 comprises of features of SKBG model in smart cities. Section 6 provides the future work. Finally, Sect. 7 concludes the paper.

2   Related Work

Current ontology-based approaches and data preprocessing techniques generally work on structured and unstructured data. Many researchers have surveyed semantic data mining, data preprocessing, and ontology-based approaches in smart cities [7,15]. Additionally, the researchers have combined different approaches of ontology and preprocessing to overcome their limitations [21]. However, these ontology-based approaches mainly concentrate on handling data of the single type and used classical algorithms for classification, clustering, feature selection and decision making in smart cities [9,19].

The researchers also tried to improve these approaches by improving rules i.e., association rule mining which was first introduced for prioritizing and rectifying different variants of k means algorithms applied to a group data [2]. However, these algorithms only cover similar data sets. Afterward, fuzzy sets were introduced to cover the diverse data sets [25,27]. Later, it was suggested that these fuzzy sets also need revisions as they failed to cover every combination of the data. There were chances that these fuzzy sets miss out, not all, but some semantic of the data sets [8]. Similarly, semantic annotations were applied separately to handle semi-structured data [11,22]. Many similar techniques were also introduced to improve the data preprocessing in data mining for better normalization of the data [3].
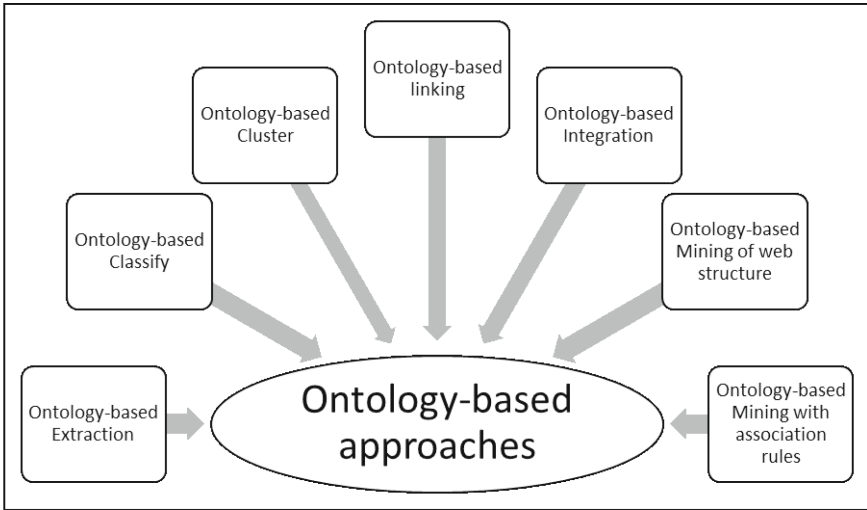
Fig. 1. Ontology-based approaches in semantic data mining

In smart cities, some featured based approaches focused on feature selection steps for better prediction and decisions making [5]. However, these techniques are used to handle data separately with different perspectives. At present, the semantic data mining in smart cities is facing diverse challenges where only specific domain knowledge is not enough and one type of content cannot be processed separately [6,14]. Therefore, there is a need for an intelligent system to resolve large and complex conflicts in semantic data mining [26]. Further, better representation formats for better understanding of domain knowledge are the key requirements of the smart cities [19]. Thus, in order to resolve these challenges, we proposed a Semantic Knowledge Based Graph model as a solution to mine concepts and patterns from complex heterogeneous data originating from the diverse sources of the smart city.

3 Formal Semantic Mining with Ontologies and Preprocessing

In smart cities, semantic data mining usually combines several stages by including ontologies for conceptualization and content management [28]. Ontology-based approaches are comprised of extraction, classification, mining with association rules, clustering, finding links, mining of web structure, integration and recommending systems as shown in Fig. 1. These steps focus on the semantics of the content. However, when these steps are applied to a data commencing from a smart city domain, knowledge extraction becomes complex and time-consuming [12,30].

Similarly, data preprocessing when specifically focuses on semantics and in finding relations in these semantics with similar meanings to interlink them with
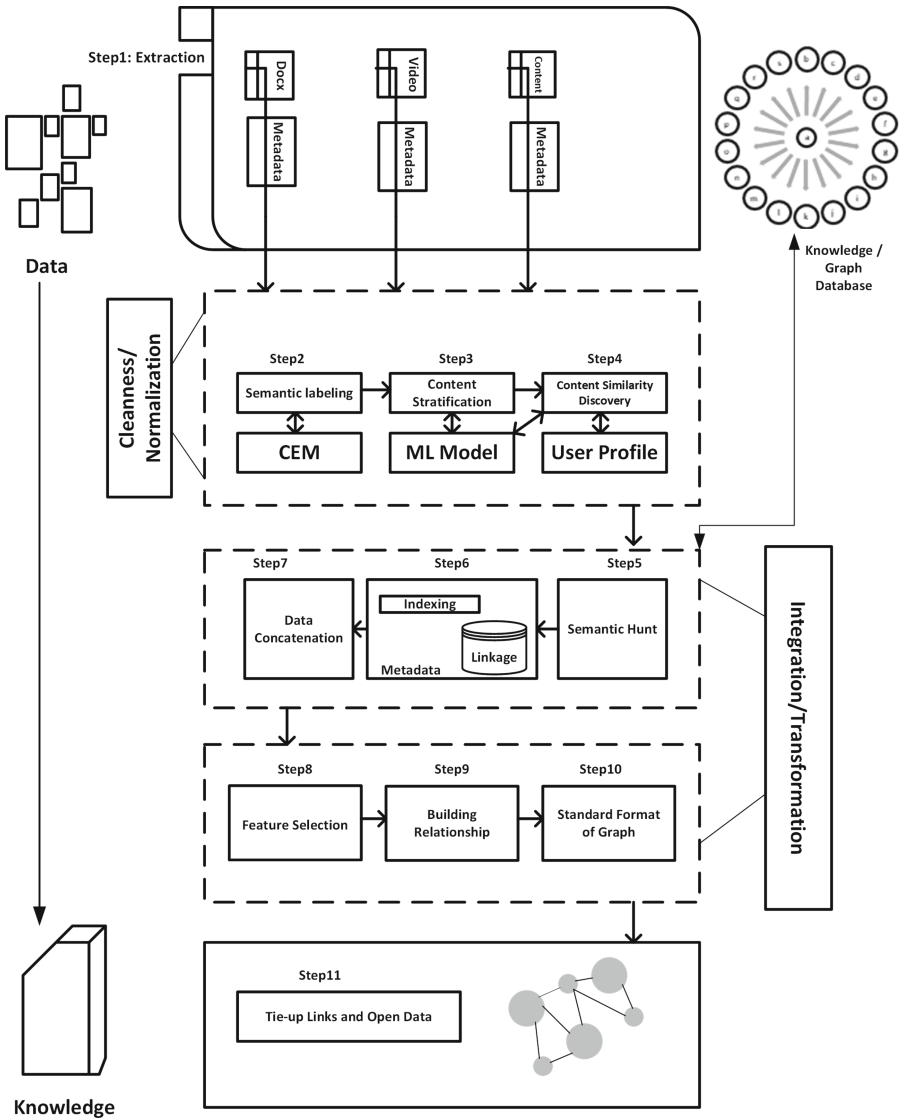
Fig. 2. Semantic Knowledge Based Graph model

one another; requires available domain knowledge. Further, it applies traditional techniques like the cleanness of data by using regression for smoothing noise, inconsistencies and semantic gaps. Also, data is classified by labeling through binning and then finally integrate them to transform into something processable. However, to undertake these tasks on the data originating from the smart cities is quite complicated and challenging [15,29].

4 Proposed Model: Semantic Knowledge Based Graph (SKBG)

In this section, we proposed a Semantic Knowledge Based Graph model as a solution to above-mentioned limitations in conventional ontology-based approaches and preprocessing data in smart cities. The model helps in transforming knowledge discovery practices. It integrates the semantic mining in diverse and tedious data catalogs of smart cities via fusing structured, unstructured and semi-structured data intelligently. As a result, information retrieval becomes very swift and effective. Further, the model effectively handles, manages and interlinks the semantics of the contents by discovering new and unique patterns during the knowledge discovery phase in smart cities.

4.1 Work Flow of Semantic Knowledge Based Graph Model

Following are the key steps of the proposed SKBG model with the objective to work seamlessly upon semantics in a smart city environment by using knowledge-based graphs. This is carried out by interlinks heterogeneous data, finding meanings, concepts, and patterns of the complex data. Moreover, the steps help to overcome the basic limitations of conventional ontology-based approaches in the smart city.

Step 1: Extraction. In the first step, extraction is performed to excavate and mine all kind of smart city data available in any format. The data can be structured (tables), semi-structured (Emails, CSV, TSV, XML or docx) and unstructured (audios, videos, images). The step is highlighted in Fig. 2.

Step 2: Semantic Labeling. During this step, excavated data is tagged with some useful and authentic semantic names. We used CEM (Concept Elicitation Mode) which helps in finding and making correct tagging as shown in Fig. 2. As a result, documents are checked out from top to bottom. Text is analyzed to mine the concepts, keywords, and topics form the content. Finally, semantic labeling helps in generating relationships between them.

Step 3: Content Stratification. In this step of our model, smart word stratification is used for grouping or classifying the content using artificial intelligence and core machine learning (supervised and unsupervised) algorithms as shown in Fig. 2. The machine learning makes the content classification process quite robust and impulsive as compared to static approaches.

Step 4: Content Similarity Discovery. In this step, the model checks the documents correspondence with similar documents and separate them. This step will figure out, how much one content data is similar to other content data? Further, the similarity index set the path for the linking of the data originating from the diverse source of a smart city. To get better experience, more enhanced graphs of different user's history and profiles are used to find out the content similarity in a smart city as shown in Fig. 2.

Step 5: Semantic Hunt. In this step, semantics of search results are analyzed as user search different and relevant words to get their desired results. Afterwards, the outcomes are linked to get better semantics in a specific domain.

Step 6: Link to Reference Data. In this step contents are linked to the reference data which is available in the knowledge database of the semantics. Two-way approaches are used for establishing the links. First, by adding reference data to the knowledge database. Second, by indexing the existed data known as metadata as shown in step 6 of the Fig. 2. Later, both approaches help to tack back the original data.

Step 7: Data Concatenation. This step is similar to integration step in the traditional ontology-based approaches of data processing. However, it has an edge on traditional approaches as it merges both external and internal data more actively and efficiently. Semantics that are relevant to a specific domain are integrated as unit during this step.

Step 8: Features Selection. In this step, datasets after integrating semantics as a unit are analyzed extensively. As a result, some key features and attributes are mined on which knowledge graphs are established. Further, decisions are carried out regarding the combination of these features to improve the semantics of the data.

Step 9: Building Relationship. In this step after selecting unique features in the datasets, need arises to discover the unique relationships that exist among them. Therefore, by analyzing them in different dimensions' unique relationships are apprehended among the selected features.

Step 10: Standard Format of Graph. During this step, a related and standard framework is selected that represents the precise meanings in the semantic data. Further, a framework is conceived which helps to visualize the actual relationship in the semantic data.

Step 11: Tie-Up Links in Open Data. Finally, in this step, we merge two things. One is the links which are the diverse data combinations. Second is the open data which refers to the data which is free and handy to everyone. Graphical representation of knowledge is also generated for visualization as shown in Fig. 2.

Semantic Knowledge Based Graph model works with a systematic procedure and use knowledge/graph database of semantics. Our model helps to mine every type of data initiating from different sources available in the smart cities. The model employs machine learning algorithms for better classification and feature selection of the data. It helps to search relevant semantics for a specific problem and find links in them. Later, the model combines them with specific patterns which reside in them. Finally, the model shows the output in a graphical form. Hence, the Semantic Knowledge Based Graph model completely process raw data in parallel to discover and gain useful knowledge in an environment like a smart city.

5   Features of Semantic Knowledge Based Graph Model

The ontology-based approaches when applied to multi-source complex datasets (e.g., data originating from the smart cities) requires a preprocessing stage to be carried out separately. However, in our proposed SKBG model there is no need to perform the preprocessing of the data separately. All the steps in the

Table 1. Correspondence of SKBG model and preprocessing steps

No.	SKBG model steps	Preprocessing steps
1	Extraction	Nil
2	Semantic Labeling	Cleanness of data/Normalization
3	Content Stratification	Cleanness of data/Normalization
4	Content Similarity Discovery	Cleanness of data/Normalization
5	Semantic Hunt or Search	Integration and transformation
6	Linkage to Reference Data	Integration and transformation
7	Data Concatenation	Integration and transformation
8	Selection of indications or Features	Integration and transformation
9	Building Relationship	Integration and transformation
10	Settle on Standard Format of Graph	Integration and transformation
11	Tie-up links and open data	Nil

proposed model are integrated well enough to perform their specific task individually without linking or merging the data. Therefore, our SKBG model can be a pioneer for more advance knowledge discovery and data visualization in smart cities. The correspondence of SKBG model and preprocessing steps are summarized in Table 1.

Finally, Our proposed SKBG model provides a conceptual framework which mines multi-source raw data and interconnects them without having any kind of specific domain knowledge. The model is equipped with machine learning algorithms which provide persistent learning, data refining, and process monitoring as a continuous process in knowledge discovery. Further, it also connects the additional knowledge from people and different domains of the smart cities to get the diverse illustration of the data.

6   Future Work

As future work, we will evaluate our model by conducting experiments on structured, semi-structured, and unstructured datasets typically originating from a smart city domain. Also, we will define semantic labeling and semantic indexing more precisely in a smart city environment to symbolize information related to the user' s interest.

7   Conclusion

In this paper, we thoroughly analyzed the limitations of traditional ontology-based approaches and data preprocessing. Ontology-based approaches and data preprocessing are traditional ways of handling data in smart cities. However, only a single type of data can be extracted with these approaches whereas we

have heterogeneous multi-source type data in smart cities. To overcome these limitations, we proposed a Semantic Knowledge Based Graph (SKBG) model. The model works with a systematic procedure and instrument a multi-source knowledge/graph database of semantics for knowledge discovery. Further, the model provides persistent learning by employing machine learning algorithms for better classification and feature selection. It searches relevant semantics for a specific problem in a smart city and interlinks them graphically for generating patterns and relationships in data. Finally, the results are summarized in the form of a knowledge graph which gives a complete insight into the data.

References

1. Ali, A., Qadir, J., Rasool, R.U., Sathiaseelan, A., Zwitter, A., Crowcroft, J.: Big data for development: applications and techniques. Big Data Anal. **1**(1), 2 (2016). https://doi.org/10.1186/s41044-016-0002-4
2. Altaf, W., Shahbaz, M., Guergachi, A.: Applications of association rule mining in health informatics: a survey. Artif. Intell. Rev. **47**(3), 313–340 (2017). https://doi.org/10.1007/s10462-016-9483-9
3. Bandaru, S., Ng, A.H., Deb, K.: Data mining methods for knowledge discovery in multi-objective optimization: part a - survey. Expert Syst. Appl. **70**, 139–159 (2017). https://doi.org/10.1016/j.eswa.2016.10.015
4. Consoli, S., et al.: Producing linked data for smart cities: the case of catania. Big Data Res. **7**, 1–15 (2017). https://doi.org/10.1016/j.bdr.2016.10.001
5. d'Aquin, M., Davies, J., Motta, E.: Smart cities' data: challenges and opportunities for semantic technologies. IEEE Internet Comput. **19**(6), 66–70 (2015). https://doi.org/10.1109/MIC.2015.130
6. González-Vidal, A., Jiménez, F., Gómez-Skarmeta, A.F.: A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. Energy Build. **196**, 71–82 (2019). https://doi.org/10.1016/j.enbuild.2019.05.021
7. Gyrard, A., Zimmermann, A., Sheth, A.: Building IoT-based applications for smart cities: how can ontology catalogs help? IEEE Internet Things J. **5**(5), 3978–3990 (2018). https://doi.org/10.1109/JIOT.2018.2854278
8. Huang, Y., Li, T., Luo, C., Fujita, H., Horng, S.J.: Matrix-based dynamic updating rough fuzzy approximations for data mining. Knowl.-Based Syst. **119**, 273–283 (2017). https://doi.org/10.1016/j.knosys.2016.12.015
9. Kaur, N., Aggarwal, H.: Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement. Int. J. Inf. Technol. **10**(1), 99–110 (2018). https://doi.org/10.1007/s41870-017-0063-2
10. Lau, B.P.L., et al.: A survey of data fusion in smart city applications. Inf. Fusion **52**, 357–374 (2019). https://doi.org/10.1016/j.inffus.2019.05.004
11. Lepri, B., Antonelli, F., Pianesi, F., Pentland, A.: Making big data work: smart, sustainable, and safe cities. EPJ Data Sci. **4**(1), 16 (2015). https://doi.org/10.1140/epjds/s13688-015-0050-4

12. Li, J., et al.: Feature selection: a data perspective. ACM Comput. Surv. **50**(6), 94:1–94:45 (2017). https://doi.org/10.1145/3136625

13. Lin, H., Liu, G., Yan, Z.: Detection of application-layer tunnels with rules and machine learning. In: Wang, G., Feng, J., Bhuiyan, M.Z.A., Lu, R. (eds.) SpaCCS 2019. LNCS, vol. 11611, pp. 441–455. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24907-6_33

14. Moustaka, V., Vakali, A., Anthopoulos, L.G.: A systematic review for smart city data analytics. ACM Comput. Surv. **51**(5), 103:1–103:41 (2018). https://doi.org/10.1145/3239566

15. Pouyanfar, S., Yang, Y., Chen, S.C., Shyu, M.L., Iyengar, S.S.: Multimedia big data analytics: a survey. ACM Comput. Surv. **51**(1), 10:1–10:34 (2018). https://doi.org/10.1145/3150226

16. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. **89**, 14–46 (2015). https://doi.org/10.1016/j.knosys.2015.06.015

17. Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., Fanizzi, N.: Mining the semantic web. Data Min. Knowl. Discov. **24**(3), 613–662 (2012). https://doi.org/10.1007/s10618-012-0253-2

18. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: a comprehensive survey. J. Web Semant. **36**, 1–22 (2016). https://doi.org/10.1016/j.websem.2016.01.001

19. Saggi, M.K., Jain, S.: A survey towards an integration of big data analytics to big insights for value-creation. Inf. Process. Manag. **54**(5), 758–790 (2018). https://doi.org/10.1016/j.ipm.2018.01.010

20. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013). https://doi.org/10.1109/TKDE.2011.253

21. Sànchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Syst. Appl. **39**(9), 7718–7728 (2012). https://doi.org/10.1016/j.eswa.2012.01.082

22. Ullah, F., Habib, M.A., Farhan, M., Khalid, S., Durrani, M.Y., Jabbar, S.: Semantic interoperability for big-data in heterogeneous iot infrastructure for healthcare. Sustain. Cities Soc. **34**, 90–96 (2017). https://doi.org/10.1016/j.scs.2017.06.010

23. Vaduva, C., Georgescu, F.A., Datcu, M.: Understanding heterogeneous eo datasets: a framework for semantic representations. IEEE Access **6**, 11184–11202 (2018). https://doi.org/10.1109/ACCESS.2018.2801032

24. Wang, H., Xu, Z., Fujita, H., Liu, S.: Towards felicitous decision making: an overview on challenges and trends of big data. Inf. Sci. **367−368**, 747–765 (2016). https://doi.org/10.1016/j.ins.2016.07.007

25. Wang, H., Xu, Z., Pedrycz, W.: An overview on the roles of fuzzy set techniques in big data processing: trends, challenges and opportunities. Knowl.-Based Syst. **118**, 15–30 (2017). https://doi.org/10.1016/j.knosys.2016.11.008

26. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)

27. Xu, Y., Gao, W., Zeng, Q., Wang, G., Ren, J., Zhang, Y.: FABAC: a flexible fuzzy attribute-based access control mechanism. In: Wang, G., Atiquzzaman, M., Yan, Z., Choo, K.-K.R. (eds.) SpaCCS 2017. LNCS, vol. 10656, pp. 332–343. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-72389-1_27

28. Xue, X., Liu, S.: Matching sensor ontologies through compact evolutionary tabu search algorithm. In: Wang, G., Chen, J., Yang, L. T. (eds.) SpaCCS 2018. LNCS,

vol. 11342, pp. 115–124. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-05345-1_9

29. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Inf. Fusion **42**, 146–157 (2018). https://doi.org/10.1016/j.inffus.2017.10.006

30. Zhang, S., Boukamp, F., Teizer, J.: Ontology-based semantic modeling of construction safety knowledge: towards automated safety planning for job hazard analysis (JHA). Autom. Constr. **52**, 29–41 (2015). https://doi.org/10.1016/j.autcon.2015.02.005