Research on Construction Technology of Graph Data Model

Wei Rao[1(✉)], Fan Yang[1], Zeyang Tang[1], and Junjie Wang[2]

[1] State Grid Hubei Electric Power Co. Ltd, Electric Power Research Institute, Wuhan 430077, China
310714175@qq.com
[2] College of Automation, Nanjing University of Posts Telecommunications, Nanjing 210003, China

Abstract. As the scale of the power grid continues to expand, the traditional distribution network management model cannot meet the requirements of power grid development under the new situation. The current distribution network operation inspection still lacks in data collection, and it is impossible to establish an informative and intelligent operation inspection management system. The upper-level production management system is also unable to integrate due to the lack of operation inspection marketing data. Aiming at the performance problem of the visualization of topology data in the distribution network operation and inspection, this paper uses the graph data model to build knowledge, designs graphic elements for data migration, and forms a topology map for the intelligent operation and inspection of the distribution network. The research clearly and intuitively displays the specific information of the power system equipment and the physical relationship between the equipment, thus forming a data model of the grid diagram.

Keywords: Graph database · Topological graph · Graph data model for power grid · Distribution network inspection

1 Introduction

The distribution network is at the end of the power system and has the characteristics of wide area, large scale, many types, and multiple connections. With the urbanization construction and the growth of electricity demand, the scale of the distribution network is constantly expanding while constantly transforming and expanding. With the expansion of the power grid and the increase in the amount of topology management data, the attribute dimension of the data will also increase. The traditional distribution network management model can no longer meet the requirements of power reform and smart grid development under the new situation. The current distribution network operation inspection still lacks in data collection, and it is impossible to establish an informative and intelligent operation inspection management system. The upper production management system is also unable to integrate due to the lack of operation inspection marketing data,

S. Yang and H. Lu (Eds.): ISAIR 2022, CCIS 1700, pp. 43–54, 2022.
https://doi.org/10.1007/978-981-19-7946-0_5

which makes it impossible to achieve fault research and judgment and fault location based on multiple data. We should use the application layer to strengthen the application of big data, from the traditional "passive repair" to "active operation and maintenance".

At the same time, data analysis and processing technologies are developing rapidly, and non-relational databases (NoSQL, Not Only SQL) are leading the database technology revolution. This paper analyzes and studies the knowledge of graph data model construction, designs graph elements for data migration, forms topological graphs for intelligent operation and inspection of distribution networks, and constructs grid graph database models. A storage method which uses a graph database is proposed to combine the grid big data with the grid's own network topology characteristics, so as to effectively use graph theory and topology theory to analyze and optimize the power grid under the power of big data. Because different data models in the graph database deal with different problems, the efficiency difference is obvious. Therefore, it is necessary to study different modeling methods and their applicable conditions.

The rest of this paper is organized as follows: The first part describes the current research status of graph data model construction at home and abroad. The second part analyzes the methods and advantages of Neo4j graph data model and GraphX graph database model construction. The third part analyzes the design of graph data model for intelligent operation and inspection of distribution network. The fourth part introduces the CIM model, the data model with electrical nodes as domain entities, the data model of materialized attributes, the application scenarios of the materialized time data model and their respective performance. The fifth part proposes the future prospects and challenges of graph data model construction.

2 Related Work

Pan et al. [1] proposed the principle of grid data modeling based on domain modeling theory and following the CIM model. According to these principles, some power grid data models and their applicable occasions are proposed, as well as methods to convert these models to each other. Finally, three graphical database models are established using the modeling method proposed in this paper. The graphic database model proposed in this paper can greatly improve the retrieval efficiency, which proves the effectiveness of the modeling method and the effectiveness of the graphic database in specific retrieval. Pavkovic et al. [2] introduced the use of Neo4j graph database to model and manage power system data. Hu et al. [3] regards the data model as an entity in the knowledge graph, and combines semantic analysis technology and entity similarity calculation technology for the data model in the power grid field, and proposes a model diagnosis method based on the knowledge graph. Consistency of sex and diagnosis. Research shows that this method can more fully utilize the relevant information of the data model and expand the thinking of traditional diagnostic methods. Gonzalez et al. [4] shows that the motion graph model can be better organized around gateway nodes, which act as a bridge to connect different areas of the motion graph. The graph-based object moving cube can be constructed by merging and collapsing nodes and edges according to the application-oriented topology. Risi et al. [5] proposed a method using visual language coding based on logic paradigm. CoDe allows visualization to be organized through the CoDe model, which graphically represents the relationship between

information items and can be regarded as a conceptual diagram of the view. Xue et al. [6] relies on the commercial bank system architecture, reuses the external business data query service of the commercial bank system, integrates the data processing capabilities of the data warehouse, provides related data for the map display tool, and builds the corporate relationship map data model, which helps to improve the bank¡¯s acquisition of new The ability of customers, as well as fully understand existing customers, establish customer profiles, and help banks prevent group risks from group customers. Wu et al. [7] proposes a fast graph construction method based on the existing SG-CIM model data of the State Grid Interconnection Department. It can fully support the company's various analysis and recommendation applications based on the unified knowledge graph. Scarselli et al. [8] proposed a new neural network model, called the graph neural network (GNN) model, which extends the existing neural network method to process the data represented in the graph domain. In order to deal with the graph uncertainty in the spatial and topological relationships between objects in the graph, Majumdar et al. [9] proposed an object-oriented graph theoretical model for representing graphs, which allows the use of the concept of (fuzzy) graph matching to assess the similarity between the graphs. Compared with traditional relational databases, graph databases can naturally fit the characteristics of grid data due to the similarity of basic structures, showing their potential advantages in processing power system data and performing real-time data analysis and calculation. Lee et al. [10] uses a path ranking algorithm to extract relational paths from the knowledge graph and uses it to construct training data. In order to learn the characteristics of the relationship, the extracted relationship path is used to create a circuitous path between nodes as training data. Combi et al. [11] proposed the Multimedia Time Graphical Model (MTGM), which represented a clinical database of cardiac patients undergoing cardio angiography, and then described it in a formal way. Designed and implemented a prototype based on XML native database system. Zhang et al. [12] proposed a method of expressing high-order features of graph-based dependent analytical models depending on language models and beam search, which solved the problem of enriching high-order features without increasing decoding complexity. Huimin Lu et al. [13] propose a fuzzy attentionbased DenseNet-BiLSTM Chinese image captioning method to solve some Chinese image description generation tasks. Huimin Lu et al. [14] propose a novel hashing method termed deep fuzzy hashing network (DFHN) to overcome the shortcomings of existing deep hashing approaches for efficient image retrieval.Qi Ge et al. [15] propose a graph regularized Bayesian tensor factorization based on KDSDL model to solve the problem that existing matrix factorization methods cannot deal with the non-i.i.d outliers and the non-uniform incoherence well.

3 Graph Database

NoSQL databases not only allow storing relational data models, but also allow to store other models, the most prominent of which is the graph database. The graph database model is based on graph theory, and provides data storage by introducing the concepts of nodes and relationships, where the relationship is the most important element in the graph data model. Each node directly contains a relationship list, which stores the relationship records between the node and other nodes. The relationship records in the

graph organize all nodes by type and direction, and attributes can be added into nodes and relationships. The graph database implements all the operations of the database on this structure. When the graph database performs a connection operation which is similar to a relational database, it uses the relation list to directly access the connected node, without searching for records and calculating matching operations.

3.1 Neo4j Graph Database

Neo4j is an open source NoSQL graph database implemented in Java. The architecture of Neo4j graph database is designed to optimize the rapid management, storage and traversal of nodes and relationships. Neo4j has strong scalability, enabling it to be able to process large-scale and complex relational graph data on one machine, and it can also support parallel processing on multiple machines. Compared with other relational databases, Neo4j has better performance to deal with a large amount of complex, related and structured sparse data. The graph database model is shown in Fig. 1.

3.2 GraphX Database

GraphX is a graph computing framework built on Spark. GraphX describes a directed graph with vertex attributes and edge attributes. GraphX provides three views: Vertex, Edge, and EdgeTriplet. Various graph operations of GraphX are also completed on three views. It uses Resilient Distributed Data Set (RDD) to store graph data and provides practical graph manipulation methods. RDD is a partitioned data structure, which is processed by calculation primitives provided by Spark Core. Each Spark application can be deployed in a cluster. According to the characteristics of elastic distributed data sets, GraphX can efficiently realize distributed storage and proceeding of graphs, and can be applied into large-scale graph computing scenarios such as social networks. GraphX has optimized the storage of graph vertex information and edge information, so that the performance of the graph computing framework can be greatly improved compared to the native RDD implementation, which is close to the performance of professional graph computing platforms such as GraphLab. GraphX graph database architecture is shown in Fig. 2.

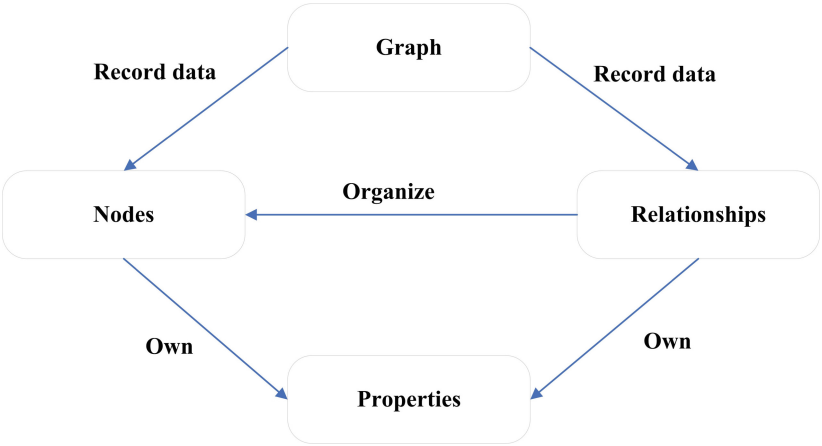

Fig. 1. Neo4j graph database model.

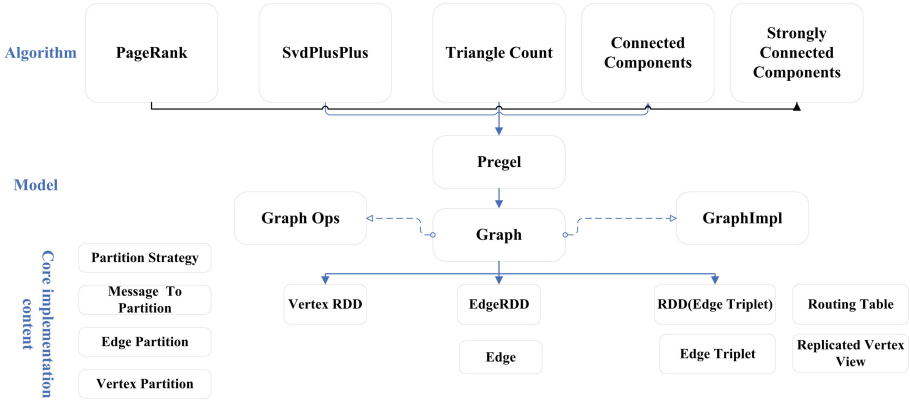

Fig. 2. GraphX graph database architecture.

4 Model Design

4.1 Data Analysis Framework of Power Grid Topology

In order to transform the abstract grid topology data into an intuitive graph database that is helpful to analyze and understand, this paper designs a grid topology graph data analysis framework based on the characteristics of the graph data of the power system, which mainly includes data collection and processing, data storage, There are several levels of data analysis, as shown in Fig. 3.

(1) The collection and processing layer mainly collects various topological data of the power system, physical equipment, connection lines, equipment operation data, historical data, etc. At the same time, the collected data is classified and screened, the missing data is counted or reported, and graph data modeling is performed based

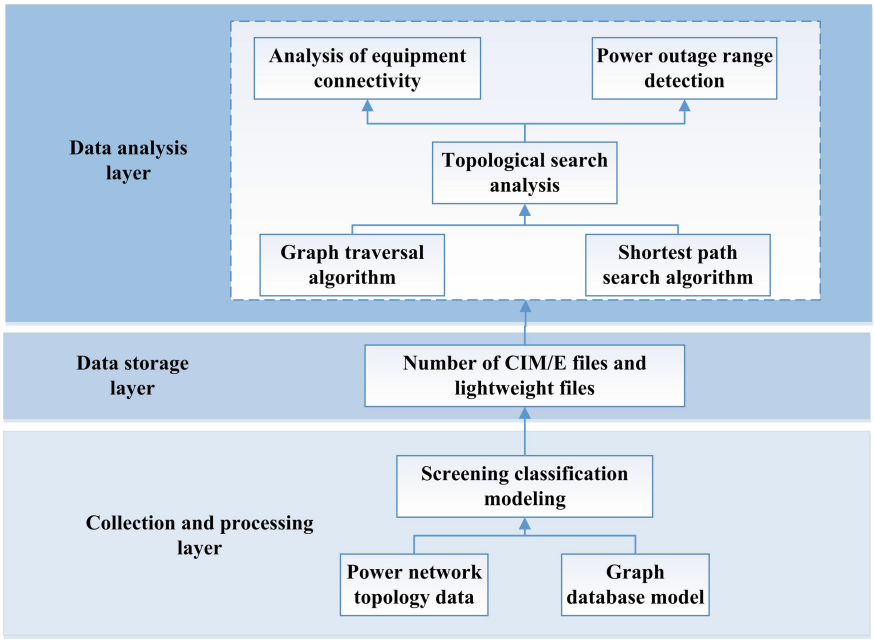

Fig. 3. Data model analysis framework of power grid diagram.

on the collected data. The collection and processing layer plays a vital role in the stable operation and maintenance of the power system.

(2) The data storage layer uses CIM/E files and lightweight file databases as model data solidification storage media, and different data information is stored in different systems. The CIM/E file mainly stores the cross-section model extracted by the operating system. The lightweight file database mainly uses two-dimensional tables for structured storage of historical and future version of the grid model data, providing support for the recall and viewing of historical data.

(3) The purpose of the data analysis layer to build the grid graph database model is to use some algorithm framework to analyze the graph database, find out the correlation, potential problems and operating conditions between the grid equipment, and ensure the stable operation of the power system.

4.2 Power Grid Data Model Construction

The construction of the data model of the grid diagram includes four steps: data acquisition, data processing, data import, and data management, as shown in Fig. 4.

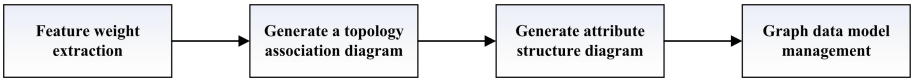

Fig. 4. Modeling flowchart.

According to the correlation of the big data collected by the power system, the feature weights of the nodes are extracted, and the priority is divided according to the feature weights to form a topological structure diagram. Then we use the Pearson product-moment correlation coefficient method to judge the correlation between equipment and equipment, and form a topological correlation diagram. The establishment of a power grid model must be complete and consistent. The model must be able to fully describe the characteristics of the power system objects, and be consistent with existing definitions as much as possible; in addition, the power grid model must also efficiently adapt to the characteristics of the graph database to improve the efficiency of data access. The basic principle of establishing a power grid model based on a graph database must follow the CIM/E standard. In the graph database, power system equipment such as generators, loads, and lines are defined as vertices. Due to the large number of switches and equipments, the switches and the connecting line between each device is defined as an edge, and an attribute structure diagram is generated.

The number of vertices and edges in the grid database is greatly reduced, which improves the query access efficiency of the graph database.

5 Graph Data Model

5.1 CIM Model

In order to make the distribution network data model unique, applicable and shareable among various applications and companies, a standard method for describing these data must be provided. The most commonly used and generally accepted standard for the data model of power system data is the CIM standard. CIM was introduced by the IEC (International Electrotechnical Commission) and is defined as part of the IEC 61970-301 standard [16].

The CIM model is an abstract object model that represents all the main entities and their relationships in the power system, transmission and distribution networks. All objects from the distribution network (such as current transformers, generators, etc.) are displayed in the form of corresponding types in the CIM model. All entities of the distribution network use identification objects, equipment, equipment containers, conductive equipment, power system resources, PSRType, and terminal types for data modeling in CIM. At the top of the type hierarchy, there are defined object classes, and other types in CIM are derived from the defined object types. The equipment type describes all distribution network equipment, which are grouped in containers represented by the equipment container class. The conductive equipment category represents all conductive equipment. The parts of the equipment that can be operated as follows: such as switches, busbars and other related conductive equipment. There is a special category in the model, namely the PSRType class. The main function of the PSRType class is to classify instances of the same class, that allows a small amount of non-standard modifications to the CIM model without changing the model. In the modeling of the distribution network, in addition to basic entities, it is also necessary to provide a mechanism for connecting conductive equipment. This is done through the terminal and CIM connection node class. These classes do not perform data modeling on the resources of the distribution network, but provide information about the physical connections of the equipment. The connection

node class connects two or more parts of the conductive device through the terminal class (describes the access point that connects two or more connection nodes) [17]. Since all devices are modeled as objects of the class, and the relationship between them is related, as shown in the figure, by representing objects as nodes, representing relationships as relationships, and representing all object data as attributes, it can be easy to map it to the surface. Using CIM to represent power system data has the advantage of using the General Data Access (GDA) standard defined in IEC 61970-403 for data exchange. Some methods of general data access standards are executed on this graph model as query operations. The CIM topology model is shown in Fig. 5.

5.2 Data Model with Electrical Nodes as Domain Entities

The network structure of the power grid is the basis of the analysis of the power grid topology. In order to reflect the structure of the power grid, this paper proposes a data model with electrical nodes as domain entities. The electrical node is the connection point of physical equipment, which combines the topology and physical characteristics of the electrical connection point and the physical bus [18]. The electrical node is used as the node of the graph database, the straight line is used as the relationship, and the physical properties of the two are used as the respective property. In addition, other types of data in topology analysis are also stored in the graph database, but in this article, by defining different "node labels", these irrelevant data can be stored in the database in an "other dimension" manner.

This model can fully restore the topological structure of the power grid, it facilitates the rapid topology analysis of the power grid and it is convenient to introduce graph theory related algorithms, such as the shortest path algorithm, the maximum flow algorithm, and so on. By combining equation solving theory with graph theory, calculations related to the network structure can be quickly performed, such as power flow calculations. Therefore, the model provides strong support when using grid topology for grid analysis. By using this model, the power industry-related algorithms based on graph theory can also be effectively calculated and verified [19]. The advantage of this model is that it can use the topology structure to store data, which facilitates grid topology analysis. However, in the storage process, irrelevant data is placed in another dimension or be ignored, which is equivalent to hiding switches, circuit breakers, generators, etc., so this model is not suitable with analyzing problems related to the abovementioned equipment.

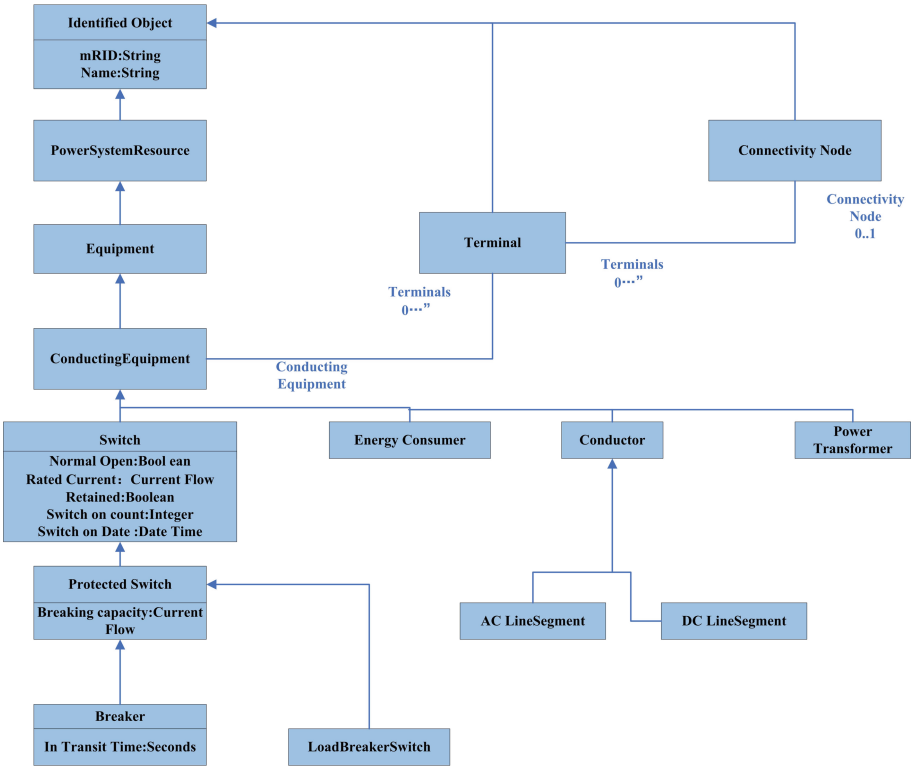

Fig. 5. CIM topology model.

5.3 Data Model of Materialized Attributes

This paper proposes a data model with materialized attributes, when the grid attributes occupy the main part of the analysis, the main attributes can be materialized as domain entities, and some attributes can be stored as database nodes. The model is based on a data model with electrical nodes as domain entities, and mathematical modeling is performed. For example, when calculating the power flow of the power grid, you can use the sparsity of the Jacobian matrix and use the graph theory solving algorithm of the sparse matrix to solve the problem.

The data model of materialized attributes refers to storing the attributes as nodes, and then calculating and analyzing the graphs based on the attributes. Therefore, when analyzing problems such as power flow calculation and reactive power optimization based on Newton's method or pre-return method, the model can make full use of the performance of the graph database. This model is a mathematical model and has almost no physical meaning. This model emphasizes the attributes in the grid, so it can be used in scenarios where specific problems are analyzed or grid calculations are performed. However, this model also changes the topology of the mesh itself, so it is not suitable for topology analysis.

5.4 Materialized Time Data Model

In traditional modeling methods, time is usually used as the attribute of the event. Time is a special attribute, because some physical quantities change dynamically at any time, while some are relatively fixed. This paper proposes a power data model that materializes time attributes, which uses time as the database node, and connects the events that occur at this time through relationship.

The materialized time data model is suitable for storing "events" because event data has unstructured characteristics that are difficult to describe with fixed fields. Time is an important characteristic when an event occurs. Therefore, when recording events that occur on a node, a data model of materialized time can be used. The materialized time model separates time from events, but is also related to each other. Different times are also related to each other, so the key events of the time nodes in the grid data can be quickly retrieved. There are two main modeling methods: timeline tree and linked list. The timeline tree model is suitable for retrieving a time period within a specific range, such as within a month or a year, and so on. The linked list model is suitable for continuous time retrieval.

6 Prospects and Challenges of Graph Data Model Construction

Based on the domain modeling theory and the CIM model, this paper proposes the principle of grid data modeling and the principle of using Neo4j to model the power grid data graph database. According to this principle, four basic graph database models are proposed. With the rapid development of society, as the amount of topological management data increases, the attribute dimension of the data will also increase. A single data model no longer has the ability to handle complex data. This is bound to change the model primitives. Changes of the model primitives can easily cause the instability of the graph model. In the future, when analyzing data processing problems in different scenarios for the distribution network, the four basic models can be combined or expanded to form a new optimal model. The model construction of graph data also faces the following challenges.

1) The challenge of data quality. High-quality data is the key to good application of grid graph data and it ensures the high quality of graph data. The accuracy, completeness, and real-time of the data have a great influence on the results of decision analysis, and even give wrong suggestions.
2) The challenge of multi-data fusion. Data fusion is the key to the application of electric power big data. Information data fusion is the multi-level processing of multi-source data, each level represents a different degree of abstraction of the original data, which includes data detection, correlation, estimation, and combination processing. According to the degree of abstraction in the data processing level, data fusion can be divided into three levels: data-level fusion (before feature extraction), feature-level fusion (before attribute description), and decision-level fusion (after independent attribute description of each sensor data).

3) The challenge of data visualization and information transmission. Electric power data visualization can effectively convey the value of data. Electric power data contains the laws and characteristics of power production and economic and social development, which are generally abstract and difficult to discover. Visual analysis of big data will make it easy to mine and analyze the laws contained in big data, which is conducive to transferring data value and sharing knowledge.
4) Challenges of big data storage and processing. Electric power big data requires huge data storage and computing capabilities. Power big data analyzes and processes structured and unstructured data from multiple data sources, and needs to store massive amounts of data and provide fast computing capabilities. Distributed data storage and calculation is an effective way to solve power big data storage and calculation.

7 Conclusion

This paper summarizes the domestic and foreign research status of graph data model construction technology, introduces graph databases commonly used in graph data model construction, expounds the relevant steps of using graph databases for data modeling, and summarizes four commonly used data models and their respective characteristics. In the future, the scalability of Neo4j database can be used to combine and apply commonly used data models to solve complex problems in different scenarios. Finally, the challenges of graph data model in data quality, multi-data fusion, data visualization information, storage and processing are explained.

Acknowledgements. We would like to thank the anonymous reviewers for their comments and constructive suggestions that have improved the paper. The subject is sponsored by the Science and Technology Project of State Grid Corporation of China (No. 5700-202058480A-0-0-00).

References

1. Pan, Z., Jing, Z.: Modeling methods of big data for power grid based on graph database, pp. 4340–4348 (2018)
2. Pavković, V., Čapko, D., Vukmirović, S., Erdeljan, A.: Modeling power system data using nosql database. In: 2017 25th Telecommunication Forum (TELFOR), pp. 1–4. IEEE (2017)
3. Hu, J., Zhao, S., Nie, Q.: Research on modeling of power grid information system based on knowledge graph. In: 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), pp. 648–651. IEEE (2021)
4. Gonzalez, H., Han, J., Cheng, H., Li, X., Klabjan, D., Wu, T.: Modeling massive rfid data sets: a gateway-based movement graph approach. IEEE Trans. Knowl. Data Eng. **22**(1), 90–104 (2009)
5. Risi, M., Sessa, M.I., Tucci, M., Tortora, G.: Code modeling of graph composition for data warehouse report visualization. IEEE Trans. Knowl. Data Eng. **26**(3), 563–576 (2013)
6. Xue, C.: Method for constructing data model of enterprise relationship graph based on industrial and commercial data. In: 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 876–8763. IEEE (2019)

7. Wu, G., et al.: An automatic and rapid knowledge graph construction method of sg-cim model. In: 2020 IEEE International Conference on Smart Cloud (SmartCloud), pp. 193–198. IEEE (2020)
8. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Networks **20**(1), 61–80 (2008)
9. Majumdar, A.K., Bhattacharya, I., Saha, A.K.: An object-oriented fuzzy data model for similarity detection in image databases. IEEE Trans. Knowl. Data Eng. **14**(5), 1186–1189 (2002)
10. Lee, W.K., et al.: A path-based relation networks model for knowledge graph completion. Expert Syst. Appl. **182**, 115273 (2021)
11. Combi, C., Oliboni, B., Rossato, R.: Merging multimedia presentations and semistructured temporal data: a graph-based model and its application to clinical information. Artif. Intell. Med. **34**(2), 89–112 (2005)
12. Zhang, M., Chen, W., Duan, X., Zhang, R.: Improving graph-based dependency parsing models with dependency language models. IEEE Trans. Audio Speech Lang. Process. **21**(11), 2313–2323 (2013)
13. Lu, H., Yang, R., Deng, Z., Zhang, Y., Gao, G., Lan, R.: Chinese image captioning via fuzzy attention-based densenet-bilstm. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **17**(1), 1–18 (2021)
14. Lu, H., Zhang, M., Xu, X., Li, Y., Shen, H.T.: Deep fuzzy hashing network for efficient image retrieval. IEEE Trans. Fuzzy Syst. **29**(1), 166–176 (2020)
15. Ge, Q., Gao, G., Shao, W., Wang, L., Wu, F.: Graph regularized bayesian tensor factorization based on kronecker-decomposable dictionary. Comput. Electr. Eng. **90**, 106968 (2021)
16. IEC61970, D.: Energy management system application program interface (emsapi) part 301: Common information model (cim) base. Geneva, Switzerland: IEC (2003)
17. McMorran, A.W.: An introduction to iec 61970–301 & 61968–11: the common information model. University Strathclyde **93**, 124 (2007)
18. Wenchuan, W., Boming, Z.: A graphic database based network topology and its application. Power Syst. Technol.-Beijing- **26**(2), 14–18 (2002)
19. Liu, K., Liu, G., Xie, K., Wang, Z.: A faster non-linear iteration solver using graph computing and its application in power flow calculation. Energy Procedia **142**, 2534–2540 (2017)
20. Lu, H., Li, Y., Chen, M., Kim, H., Serikawa, S.: Brain Intelligence: Go beyond Artificial Intelligence. Mob. Networks Appl. **23**(2), 368–375 (2017). https://doi.org/10.1007/s11036-017-0932-8
21. Huimin, L., Zhang, Y., Li, Y., et al.: User-Oriented virtual mobile network resource management for vehicle communications. IEEE Trans. Intell. Transp. Syst. **22**(6), 3521–3532 (2021)