Research on the Application of Knowledge Graph in Big Data Analysis

Ling He[1*], Haipeng Guo[2], Yali Dong[1], Haijie Wang[1], Zhu Mei[1]

* Corresponding author: heling6159@163.com

Information Countermeasure Department, Early Warning Academy, WuHan, China[1]
Student Team 9, Early Warning Academy, WuHan, China[2]

Abstract. With the wide and in-depth application of big data technology in various fields, descriptive data analysis plays a more and more important role in big data analysis. Aiming at the descriptive analysis of big data, focusing on the problem identification and data preparation in the data analysis process, starting with extracting data semantics and establishing the logic between data, this paper discusses the application of knowledge graph, studies the knowledge representation in it, and puts forward the method of combining RDF and semantic network to describe data. This idea enriches the means of descriptive analysis of big data, and lays a necessary foundation for further correlation analysis among big data.

Keywords: Big data analysis, Descriptive analysis, Knowledge graph, RDF, Semantic network

1. INTRODUCTION

Data analysis is an important means in order to seek the implicit association relationship from large-scale data and mine more useful information. Big data analysis [1,2] is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, change trends and preferences of target groups. In recent years, descriptive analysis [3] of data has attracted more and more attention. Unlike computational analysis of data, which mainly establishes various data analysis models, descriptive analysis relies on the semantic description of the data itself. Generally speaking, descriptive analysis extracts the semantics of data in advance, establishes the logic between data, and achieves data analysis relying on the method of logical reasoning. It can compress a large number of disordered data into concise and meaningful information as much as possible, so as to provide a high-quality foundation for subsequent big data analysis. In general, the purpose of descriptive analysis is to give a summary description similar to "what is" to the existing data, so as to reduce the workload of data analysis and improve the efficiency of data analysis.

At the same time, knowledge graph [4], as an important research branch of artificial intelligence, has been applied more and more widely in assisting data analysis and decision-making, text data processing, etc. Knowledge graph and semantic technology are used to enhance the association between data, so that the analysts can use more intuitive graph to further mining and analyzing the association of data.

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334385

This paper will mainly discuss how to apply the improved semantic network to intuitively describe the logical relationship between data to achieve descriptive analysis of big data.

2. BASIC PROCESS OF DESCRIPTIVE ANALYSIS OF BIG DATA

According to the definition of Baidu Encyclopedia [5], the process of data analysis consists of identifying information requirements, collecting data, analyzing data, evaluating and improving the effectiveness of data analysis. Among them, identifying information requirements is the first condition to ensure the effectiveness of the data analysis process, which can provide a clear target for data collection and analysis. The purposeful collection of data is the basis for ensuring the effectiveness of the data analysis process; Analyzing data is to transform the collected data into useful information through processing, sorting and analysis. Process improvement is a necessary step to achieve complete and effective data analysis.

Descriptive analysis focuses on the logical association between data. According to the general process of the above data analysis, descriptive analysis of big data is divided into four stages of "analyzing problems, sorting out data, establishing models and evaluating results", as shown in Figure 1.

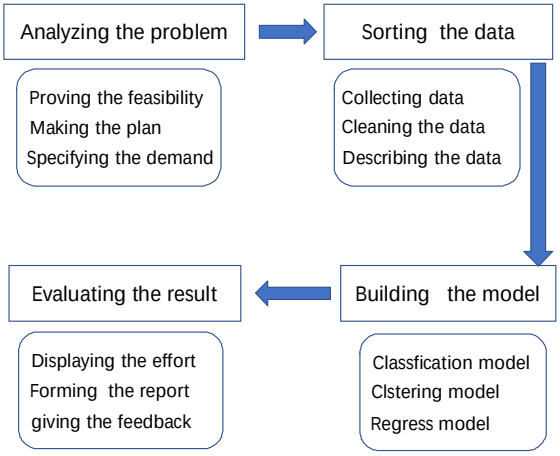

Figure 1 basic process of big data descriptive analysis

As shown in Figure 1, analyzing the problem is the first step in descriptive analysis and is the premise and foundation for all subsequent processes. In the problem analysis stage, it is necessary to clarify the requirements of data analysis and demonstrate its feasibility. On this basis, a detailed plan is made to provide programmatic guidance for the whole data analysis process. Then comes the stage of data collation, in which data collection and preprocessing (i.e., cleaning and describing) are summarized as sub-tasks. Among them, the task of "describing data" is the focus of this paper. In the modeling stage, the collected and sorted data are taken as the research object and specific models are established for the target tasks of data analysis. Finally, the results of the analysis are displayed and evaluated. If necessary, improvement measures are given for further optimizing the data analysis model.

3. KNOWLEDGE GRAPH IS APPLIED TO DESCRIPTIVE ANALYSIS

As before, "to describe data" is an important task in the data collation phase. In the so-called data description, an important means is to visually describe and display the correlation between data, so as to reduce the workload of subsequent data analysis and simplify the process of subsequent data analysis. Knowledge graph is a useful tool to describe the relationship between data reasonably and efficiently.

Formally proposed by Google in 2012, the concept of knowledge graph is essentially a semantic network that reveals the relationships between entities. In the field of knowledge representation of artificial intelligence, knowledge representation methods adopted in the early stage include first-order predicate logic, Horn clause and Horn logic, semantic network, framework, description logic, etc. [6]

Semantic network [7], whose form is a directed graph, can effectively express the semantic knowledge of human beings and support reasoning. Nodes in the graph can be used to represent entities, concept, status and so on. For each node, its attributes can be described. Attachment is used to describe both the semantic relations and actions between nodes. The connection is also called an " associative arc". Because all nodes are connected by associative arc, semantic network can fulfill the task of knowledge reasoning well. However, semantic network also has its obvious shortcomings. Limited by the expression form of graph, semantic network cannot form formal grammar and semantics, so it has poor standardization.

At present, the new knowledge representation language RDF (Resource Description Framework) [8,9,10] has been widely used. In RDF, each piece of knowledge is represented as a SPO triple (subject-predicate-object), representing a relationship P between S and O.

For example, the fourth International Conference on Big Data and Artificial Intelligence was held in Qingdao in 2021. The relevant demands for big data analysis are as follows: Track and analyze all participants and keynote speakers of the conference, including their work units, recent research directions, and outstanding representative achievements. Based on the above requirements, it is assumed that researchers are now received the following message: "big data conference invited Tom as a guest speaker, whose speech topic is big data security". Using semantic network and RDF triples, according to the information, and combining with Tom's other related information, to fulfill the data description, we can achieve figure 2 and figure 3 respectively.

As shown in Figure 2, all Tom-related information is displayed in the semantic network using the "associative arc", and the semantic network can be further extended if there is more related information. The deficiency of this knowledge description method is that the form of semantic network is not fixed and unique, which is not conducive to further standardized processing of data.

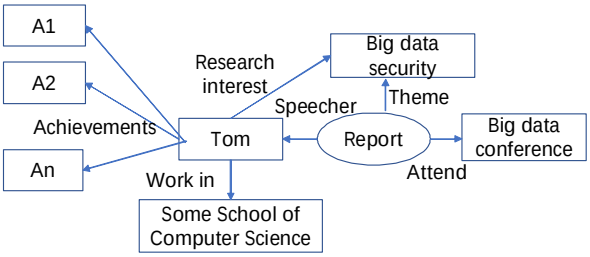

Figure 2 example of describing data by semantic network

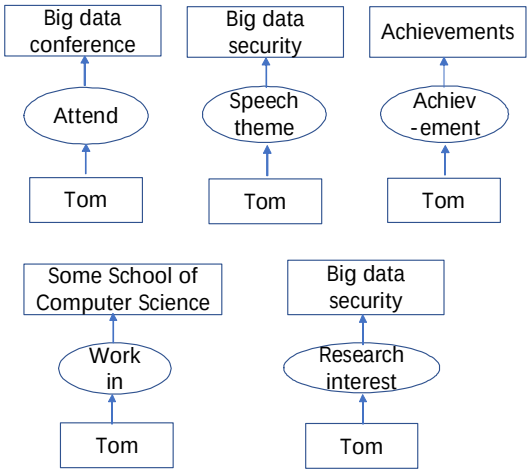

Figure 3 example of describing data by RDF

As shown in Figure 3, data description by RDF triple has a formal framework that facilitates standardized analysis and data processing, but its disadvantage is that it does not highlight the main requirements of data analysis and does not capture the core of the data analysis task (Tom in this case).

Considering the advantages and disadvantages of semantic networks and RDF triples, a semantic network description method based on RDF triples is proposed.

4. SEMANTIC NETWORK KNOWLEDGE REPRESENTATION BASED ON RDF TRIPLES

As mentioned above, semantic network can better meet the requirements of data analysis. It takes the main analysis object as the core and associates all the information related to the core object through associative arc. RDF triples have a formalized form that represents knowledge in a unified framework. Combining the advantages of both, this section proposes a new knowledge representation method SN-based on-RDF (Semantic Network Based on Resource Description Framework).

This approach takes semantic network as the knowledge representation body, formalizes associative arcs, assigns specific attributes to them, and makes associative arcs act as "Predicate" in RDF triples. So the whole semantic network consists of several normalized triples. By this means, using the data described in Figure 2 and Figure 3 as examples, the knowledge representation using SN-based on-RDF is shown in Figure 4.

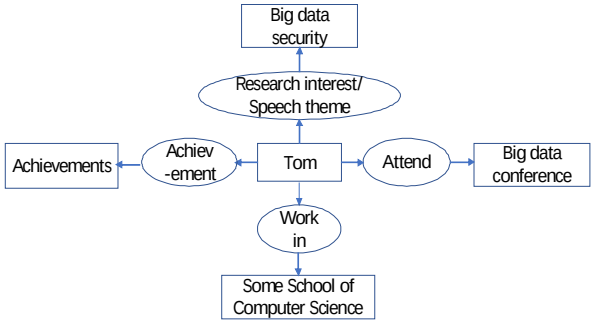

Figure 4 example of describing data by SN-based on-RDF

As shown in Figure 4, the knowledge representation of SN-based on-RDF, with the analysis object as its core, can not only describe its logical relationships between other objects in detail, but also have formal standardization of the semantic network. In the example in Figure 4, "Tom" is the core object. In addition to "Tom", if "Big data security", "Big data Conference", "Some school of Computer Science" and "Achievements" identified by the rectangular in the figure are respectively the core objects, they can also be described in a similar way. In demand driven by data analysis, with the deepening of a mass of data collection and sorting, more associations were gradually joined, the SN-based on-RDF is expanding gradually, until all of the important analysis object is associated with the semantic network, the logical relationship among all the data objects will be presented clearly.

5. SUMMARY AND PROSPECT

RDF triples knowledge representation which is based on the theory of relational model, has good standardization. The semantic network has the excellent features such as structured, easy to expand. Combining the two to describe the logical relationship between data is a meaningful idea, which can promote the research on descriptive analysis of large data greatly.

References

[1] Xueqi Cheng, Shenghua Liu, Ruqing ZHANG. Thinking on New System for Big Data Technology[J]. Journal of Chinese Academy of Sciences, vol.37,no.4,pp. 60-67,2022.

[2] Hong Wang, Shaoguo Ji. Research on practical Application and Development Trend of Big Data Analysis[J]. Information Network Security, vol.S1,pp.134-138,2021.

[3] Yan Chen, Junliang li. Correlation and causality analysis of big data technology[J]. Journal of JiuJiang University (social science edition), vol.04,no.39,pp.80-84,2020.

[4] Haofen Wang, Guilin Qi, Huajun Chen. Knowledge Graph: Method, Practice and Application[M]. Beijing: Electronic Industry Press, 2019.
[5] https://baike.baidu.com/item/data analysis/6577123?fr=aladdin#6.
[6] Peng Liu. Junhua Hui et al., Knowledge Representation and Processing[M], Beijing: Electronic Industry Press, 2021.
[7] Tingrui Yan. The Relationship between Semantic Network Structure and creative Thinking and its Neural mechanism[D]. Chongqing: Southwest University, 2021.
[8] Mengmeng Wang,Construction and Application of Academic Knowledge Graph[D]. Xi 'an: Xidian University, 2021.
[9] Mengyu Li. Research and Implementation of Semantic Place Retrieval Algorithm Based on RDF Knowledge Base[D].Yangzhou:Yangzhou University,2019.
[10] Qingrong Huang. Research on Subgraph Query of RDF Graph Data in distributed Enviroment[D].Nanning: Nanning Normal University,2021.