The Method of Constructing Knowledge Graph Based on the Trajectory of Counterfeit Cigarette and Wine Sales Data

Chunjun Zheng *[1], Ning Jia [1], Yingqiu Li [1], Yanxin Xu [2], and Jianbo Zhou [2]

[1] School of Software, Dalian Neusoft University of Information
[2] Hikvision Digital Technology Co.

*Corresponding author's e-mail:zhengchunjun@neusoft.edu.cn

Abstract. In recent years, knowledge graph has gradually developed from the traditional knowledge analysis application to the application of social entities and relationships. In this paper, the application of knowledge graph to sales data relationships is studied, and the data source is obtained by taking the sales data relationship related to counterfeit tobacco and alcohol as an example. We built the relationship between sales data related to fake cigarettes and alcohol through data pre-processing, data governance, population identification of fake cigarettes and alcohol, information extraction, etc. We used the Neo4j database for storage and built the knowledge graph of the relationship between sales data related to fake cigarettes and alcohol. The experimental verification shows that based on the real information of fake cigarettes and alcohol, we can achieve a visual display of sales data and relationship information.

Keywords: Knowledge Graph; Trajectory of Counterfeit Cigarette and Wine; Neo4j

1 Introduction

The concept of a "knowledge graph" was formally proposed by Google in 2012 [1]. After 2013, knowledge graphs, big data, and deep learning have become the core driving forces to promote the development of the Internet and artificial intelligence. The knowledge graph uses graphics to show the development process and structural relationship of knowledge, uses visualization technology to describe knowledge resources, constructs analysis, and excavates and displays knowledge resources and their interrelationships [2]. From the data analysis of the current big data, people focus on tracking people's activity tracks, building relationship knowledge graphs, and positioning propagation paths [3].

In this article, we focus on a group of people who are related to the sales and dissemination of counterfeit cigarettes or alcohol. The information about the relevant personnel selling these counterfeit cigarettes or alcohol is recorded in the official

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_118

database. Since there may be some internal management among the relevant personnel, it is very necessary to explore all personnel related to counterfeit cigarettes or alcohol.

This paper proposes a knowledge graph based on the relational real population graph database. This knowledge graph shows the information and contact relationship of people related to fake cigarettes and alcohol in the form of a diagram. The system uses the Neo4j map database to store fragmented case relationship data for knowledge modeling and association aggregation, to achieve data fusion and integration at the knowledge level, and to build a visual knowledge graph of the trajectory of counterfeit tobacco and alcohol workers.

2 Knowledge graphs and Neo4J

2.1 Resource description framework

The Semantic Web uses the resource description framework (RDF) developed by the W3C as the data model for knowledge representation [4]. In RDF, knowledge is stored in the form of SPO triple (Subject, Predicate, Object). The well-known open RDF knowledge databases include DBpedia [5], Freebase [6], etc. RDF defines the common Predicate relationship at the beginning of the release. Through the fixed IRI representation, the unified IRI definition can realize the sharing of different knowledge, but the knowledge represented in the RDF definition is limited [7]. To improve the scope of knowledge expression, RDFS (Resource Description Framework Schema) is proposed based on RDF. Based on RDF, the OWL (Web Ontology Language) [8] ontology syntax and the subsequent OWL2 [9] were expanded according to the actual requirements in the definition [10].

2.2 Neo4j

The knowledge graph is a knowledge base called a semantic network. The knowledge graph is used to describe the concepts in the physical world and their interrelationships in the form of symbols. Its basic composition is the “entity, relationship, and entity” triplet. Entities are connected through relationships to form a network of knowledge structures [11]. Neo4j is a NoSQL graph database, which is also widely used at present. Neo4j mainly uses nodes and edges to organize data. Nodes represent entities in the knowledge graph, while edges represent relationships between entities. Relationships can have directions, and both ends correspond to start nodes and end nodes.

3 Design of knowledge extraction scheme based on information about counterfeit tobacco and alcohol workers

The information on counterfeit tobacco and alcohol workers can be divided into three categories: structured data, semi-structured data, and unstructured data. Among them, unstructured data is an important kind of information resource generated in the activities of related personnel, and also an important part of population information. This kind of data has the characteristics of multimodality, incompleteness, redundancy, and privacy. The data shows incomplete syntax structure, contains a large number of professional terms such as addresses and has fuzzy semantics. It increases the complexity of data processing and analysis, and it is difficult to directly conduct efficient data mining. Relevant data mainly includes the following categories. The data source of counterfeit tobacco and alcohol workers is shown in table 1.

Table 1. The data source of counterfeit tobacco and alcohol workers

Data type	Data name
Registration	Permanent population, Temporary population, Town and street registration data, Cell registration data, Site personnel registration data
Perception	Face capture data, Vehicle capture data, Mobile data, Recommended data
Dynamic	Nucleic acid data, Network alarm address data
Registration	Social insurance registration information, Individual business of counterfeit tobacco and alcohol, Enterprise information, Longitude and latitude information of tobacco and alcohol enterprises, Information of fugitives, Basic information of the case, Personnel information of the prison, Offenders, Same railway or same flight, Communication relation information

The information on counterfeit tobacco and alcohol workers is composed of a variety of heterogeneous data sources. The original data from information systems such as the public security network, the government affairs network, and the Internet has complex features such as data diversity, incompleteness, redundancy, and sensitive information. The original data must be pre-processed to ensure the accuracy, integrity, consistency, and privacy of the data. The quality of pre-processed data would be related to the results of knowledge, and high-quality data is more likely to bring high-quality results, so it is necessary to reasonably select the methods and strategies of pre-processing.

3.1 Data processing

3.1.1. Data cleaning.

We can improve the data quality by filling in the vacancy value, smoothing the noise data, and correcting the inconsistent data. Due to the heavy workload and poor feasibility of filling data, machine learning methods such as Bayesian and decision trees are used to predict the default value.

3.1.2. Data conversion.

The original physical examination data is transformed into a unified form suitable for data mining by using the standardization of decimal calibration. We use data cube aggregation, dimension reduction, numerical reduction, and data compression to obtain simplified data sets and improve efficiency.

3.1.3. Privacy protection.

For the use process of medical examination data, this paper carries out privacy protection, data encryption, privacy anonymous processing, and access control based on differential privacy, homomorphic encryption, and other methods.

3.2 Data governance

3.2.1. Capture data governance.

The processing of captured data is to obtain the foothold of personnel (the cells that are often captured). Therefore, the following data governance is required.

a. Equipment management

We sort out the capture equipment at the entrance and exit of the district to form a code table and obtain the police station and community information of the district where the capture equipment is located.

b. Count days and capture days

The distribution number of people who took photos in the last 10, 15, and 30 days and the number of people who landed in the city were counted. We judge the situation of suspected unregistered personnel through the statistics of days and data. The rule accuracy rate of 4 days in 10 days reaches 80%. Finally, the number of times in the last 30 days and the number of days in the last 10 days were determined and counted, and the landing list of the captured personnel was obtained.

c. Management of unidentified personnel

There is some personnel whose identities have not been completed. We obtain the final ID number by comparing the photos in the population database. We need to analyze and determine the photo quality and the number of snapshots and make statistics on each scored segment of the photo quality and the number of snapshots. We found that the number of snapshots is greater than or equal to 3, and the amount of data is similar to the number of calls when the highest photo quality is 0.7. After the photo quality and capture times are determined, the data will be pushed to the interface to return the similarity, basement photo, and ID number, and the returned photo and ID number will be established into a list database, finally forming a closed loop.

3.2.2. Nucleic acid data governance.

The nucleic acid data is imported offline and provided from time to time. To form a full nucleic acid personnel information table, it is necessary to clean and de-duplicate it. The newly added data and newly added full data are distinguished by the update time and collection time.

3.2.3. Town and street data governance.

We clean and de-duplicate the population data pushed by towns and streets, and standardize the fields.

3.3 Identification of counterfeit cigarettes and alcohol

3.3.1. Judgment rules.

According to the on-site analysis, several rules were formulated to identify the personnel related to fake cigarettes and alcohol.

a. The registered personnel need to determine whether there is a track in the city. If there is a small area capture, nucleic acid, town street registration, or Internet police address, the personnel can be determined as the actual population of the city.

b. The permanent temporary registration personnel in the outside area need to determine whether there is a track in the city, that is, the number of snapshots in the community is more than 3 times or the nucleic acid data is registered in the town and street at the same time.

c. Unregistered persons include those who have captured data for 10 days and 4 days, whose online alarm address is registered in the city, town, or street, and who have nucleic acid information.

3.3.2. Actual residential address of people with counterfeit cigarettes and alcohol.

According to the above-identified fake tobacco and alcohol population, there will be more than one address, so the actual residential address needs to be determined. The address is classified and graded, and then the distance similarity is calculated. The similar address is verified and reinforced, and the highest-level address is finally obtained as the final address data. The classification level of data is shown in table 2.

Table 2. The data classification level

Data type	Address source	Rule	Level
Trajectory	Nucleic acid		2
Perception	Vehicle capture, Vehicle capture address	4 days within 10 days, 3 times in 30 days	1
Perception	Network alarm address, Mobile customization		2
Registration	Town street address, Permanent and temporary port address		3

3.4 Information extraction

3.4.1. Text annotation.

The marking based on the text of fake cigarettes and alcohol needs the guidance of experts with strong professional background knowledge.

3.4.2. Named entity recognition.

The data of tobacco and alcohol counterfeiters contains a large number of professional terms. The dictionary is an important resource for knowledge discovery. Simple dictionary and rule-based methods cannot meet the processing requirements of complex languages. Although the physical examination data contains different entity types, which use fixed grammar, syntax, and idioms to provide certain rules and features for the text processing of the physical examination data. It improves the performance of the named entity recognition model by constructing feature vectors and performing named entity recognition based on specific rules.

3.4.3. Risk relationship extraction of counterfeit tobacco and alcohol personnel.

To further model and characterize the more comprehensive risk characteristics of counterfeit tobacco and alcohol workers, alleviate the problems such as the staggered association of data entity relations, and consider the training efficiency of the overall network, this paper will design a recurrent neural network architecture integrating transformer with lower complexity. The circular layer refers to the circular neural network or its related variant network in deep learning, while the transformer layer is specifically implemented by the attention mechanism. The network model includes multiple embedding layers, a low-level feature extractor, and a high-level feature extractor. The multi-channel self-attention mechanism is introduced to capture sentence-level and high-level features by fitting multiple groups of weight vectors, to fit the richer sentence-level semantic information by learning multiple groups of weight vectors, and to improve the model's feature learning ability for high-density entity distribution of population data and complex entity relationships.

The risk evolution model of counterfeit tobacco and alcohol workers largely depends on the community risk factors and their relationships. These community risk elements and their associations are hidden in the massive heterogeneous data of the community. Intelligent technology is used to process big data, turning big data into "small things", providing refined guarantees for solving the key problems and hidden dangers in fake tobacco and alcohol information, and for the accurate governance of relevant departments. The knowledge graph shows excellent information visualization effect in the correlation analysis and characterization of knowledge and carriers, data mining, and information processing, which can be used to sort out the multi-agent correlation in the risk of counterfeit tobacco and alcohol. The ontology database of community security risk events is built through knowledge extraction, knowledge fusion, and other technologies. Common ontology elements include instances, relationships, attributes, events, etc. The ontology database is built to discover the association between ontologies, build an association relationship database, and form a knowledge graph with an early-warning function from massive heterogeneous data.

This paper analyzes and constructs the "person place matter organization" ontology and correlation and the multi-agent knowledge graph of the risk cases of counterfeit tobacco and alcohol. Based on the big data of police and community in a certain region and period, and at the same time and space, the integral value of each feature of the five elements of "people land matter organization" is obtained through the method of probability statistics, and the risk threshold of each type of association relationship in the region is preset according to the characteristics of the big data of counterfeit tobacco and alcohol cases in the region by iterative calculation and statistics. All integral values and thresholds are integrated into the knowledge map built to form a knowledge map of counterfeit tobacco and alcohol risk events with an early warning function. When the integral and value of a certain association type exceed the threshold value set for this type of association, an early warning will be given and the associated information of related subjects will be fed back on time.

4 Experimental design and results

The relationship knowledge graph of people related to counterfeit cigarettes and alcohol is stored in Neo4j based on the map database. Neo4j storage data can be imported in many ways. RDF data is imported into Neo4j for storage in this system, and then Neo4j's Cypher language is used to design the knowledge graph of entities, relationships, and attributes to achieve graphical interactive queries and associated reasoning. Part of the screenshot of the knowledge graph for the relationship between people related to fake cigarettes and alcohol is shown in Figure 1.

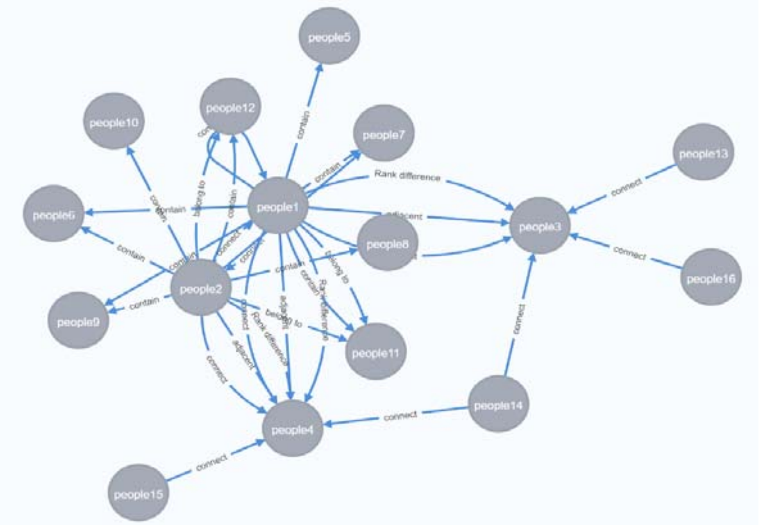

Fig. 1. Knowledge graph of personnel related to fake cigarettes and alcohol

From this knowledge graph, it can be seen that the relationship between people related to fake cigarettes and fake alcohol presents its attribute relationship in the form of a hierarchical graph structure, realizing the efficient graph query and analysis application of Neo4j.

5 Conclusion

In this paper, we build a knowledge graph around the information of people related to fake cigarettes and alcohol. We can extract information about people related to fake cigarettes and alcohol from the text. We conduct data pre-processing, and then achieve data governance, including population governance, snapshot data processing, nucleic acid data governance, and urban information processing, to achieve population identification and information extraction of people related to fake cigarettes and alcohol, and finally build a knowledge graph. The experiment of knowledge graph construction proves the effectiveness of the current method.

In the future research process, we will further explore the relationship between "people, places, things, things, and organizations" based on the multi-modal knowledge graph of people involved in counterfeit tobacco and alcohol to build a rich and flexible knowledge graph.

Acknowledgments

This work was supported by the national key research and development projects (grant number 2021YFC3320300), the projects of Liaoning Provincial Department of Education (grant number JG20DB032, LJKQZ2021188), the technology innovation fund project of Dalian Neusoft University of Information.

References

1. Zhou B., Zhou B., Hua B., et al. An end-to-end tabular information-oriented causality event evolutionary knowledge graph for manufacturing documents [J]. Advanced Engineering Informatics, 2021, 50: 101441.
2. Cheng D., Yang F., Wang X., et al. Knowledge Graph-based Event Embedding Framework for Financial Quantitative Investments [C]// SIGIR 20: the 43rd International ACM SIGIR conference on research and development in Information Retrieval. ACM, 2020.
3. Wang N., Guan Z., and Guo Y. Primitive Knowledge Graph Construction Based on Eventelement Driven [J]. Journal of Physics: Conference Series, 2021, 1995 (1): 012028 (5pp).
4. Mohamed A., Abuoda G., Ghanem A., et al. RDFFrames: knowledge graph access for machine learning tools [J]. VLDB Journal: The international journal of very large databases, 2022 (2): 31.
5. Zhang X., Liu X., Li X., et al. MMKG: an approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia [J]. Computer Physics Communications, 2017: 98-112.

6. Marcelo, Arenas, Bernardo, et al. Faceted search over RDF-based knowledge graphs [J]. Web Semantics Science Services & Agents on the World Wide Web, 2016.
7. Eddamiri S., Benghabrit A., Zemmouri E. RDF graph mining for cluster-based theme identification [J]. International Journal of Web Information Systems, 2020, ahead-of-print (ahead-of-print).
8. Bouillet E., Feblowitz M., Zhen L., et al. A Knowledge Engineering and Planning Framework based on OWL Ontologies. 2008.
9. Wei Y., Jie L., and Xie H. KGRL: an OWL2 RL Reasoning System for Large Scale Knowledge Graph [C]// 2016 12th International Conference on Semantics, Knowledge and Grids (SKG). IEEE, 2016.
10. Melo, Andre, Paulheim, et al. Type Prediction in Noisy RDF Knowledge Bases Using Hierarchical Multilabel Classification with Graph and Latent Features [J]. International Journal of Artificial Intelligence Tools Architectures Languages Algorithms, 2017.
11. Li J., Horiguchi Y., and Sawaragi T. Counterfactual inference to predict causal knowledge graph for relational transfer learning by assimilating expert knowledge-Relational feature transfer learning algorithm [J]. Advanced Engineering Informatics, 2022, 51: 101516.