Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 578–584

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Development of an Interactive Medical Knowledge Graph Based Tool Set

Xiaowei Xu[a,b], Xuwen Wang[a,b], Meng Wu[a,b], Hetong Ma[a,b], Liu Shen[a,b], Jiao Li[a,b,*]

[a]*Institute of Medical Information & Library,Chinese Academy of Medical Sciences& Peking Union Medical College, Beijing 100020, China*

[b]*Key Laboratory of Medical Information Intelligent Technology, Chinese Academy of Medical Sciences, Beijing 100020, China*

Abstract

The field of clinical research has reached a turning point in terms of the range and diversity of available data and their potential to improve human health and well-being. However, due to discipline-specific variances in terminology and representation, the data is often compartmentalized, disorganized, and inaccessible to a broad audience. In response to these challenges, we designed and evaluated a pilot knowledge graph-based MedKaaS tool capable of integrating existing clinical datasets and translating those data into insights intended to augment human reasoning and accelerate scientific research. In this paper, we present the architecture and performance of our system, which has been applied to several real-world use cases in collaboration with subject-matter specialists. The MedKaaS Tools are actively developed, with regular updates to performance and features. Our future plans include developing a more user-friendly interface and expanding the visualization capabilities for knowledge exploration.

Keywords: knowledge graph; visualization; application programming interface; biomedical data

1. Instruction

Rapid scientific breakthroughs in the field of clinical have led to the formation of enormous medical resources with great value for both research and clinical applications. The paradigm for medical scientific research has shifted from empirical medicine to evidence-based medicine. However, the existing biomedical information resources exhibit characteristics of heterogeneity from multiple sources and have a decentralized distribution across multiple centers. Therefore, significant challenges remain in the effective integration of extensive biomedical knowledge, the efficient management and utilization of medical knowledge, and the development of intelligent medical knowledge services that support precision medicine while meeting the requirements of cutting-edge scientific research.

* Corresponding author. Tel.: +86-010-52328740
E-mail address: li.jiao@imicams.ac.cn

1877-0509
10.1016/j.procs.2023.08.025

In the past decade, there has been a notable increase in both the utilization and investigation of knowledge graphs (KGs). In contrast to conventional knowledge representation techniques, KGs demonstrate an efficient and comprehensible method of conveying the associations among entities [1]. Therefore, KGs are constructed and applied in various medical scenarios, such as semantic search [2], question and answering systems [3], personalized recommendation systems [4], and decision support systems [5].

The construction of KGs involves multiple stages, including knowledge extraction, knowledge fusion, quality inspection, and application. Knowledge extraction technologies, such as named entity recognition (NER) and entity relationship extraction (RE), are well established in the general domain but present research hotspots in low-resource, cross-lingual, and vertical fields. The medical domain mainly focuses on monolingual knowledge extraction tasks. For example, a machine learning model developed by the Chinese Academy of Sciences achieved an average F1 score of 91.5% for the recognition of six types of Chinese medical entities, ranking first in the CCKS2020 medical entity recognition evaluation [6]. Additionally, the Biomedical Literature Relationship Extraction Framework BERE released by Tsinghua University researched an F1 score of 62.5% for English DDI drug action relationship extraction [7].

Knowledge fusion, which involves verifying, disambiguating, processing, and integrating heterogeneous knowledge from multiple sources, offers the possibility of interacting and integrating different knowledge graphs. However, research on medical knowledge fusion is relatively scarce. The University of Science and Technology of China and Baidu Research Institute presented an evidence-based medical entity relationship verification framework at KDD 2021, while MIT proposed Noisy OR based on the EMR health knowledge graph [8]. After manual evaluation, the accuracy rate is 0.85, and the recall rate is 0.6. The National Institutes of Health (NIH) has successively developed information resource retrieval platforms, including medical literature, genetics, such as PubMed [9], OMIM [10], as well as medical thesaurus service platforms such as MeSH [11], UMLS [12]. At the same time, the NIH funded the National Center for Advancing Translational Sciences (NCATS) to establish the Biomedical Data Translator Program in 2016 [13]. The project team uses knowledge mapping, knowledge reasoning and other technologies to integrate clinical electronic medical records and open multi-source and multi-dimensional biomedical information resource to develop a series of knowledge processing tools that support intelligent medical knowledge discovery, thereby better promoting biomedical transformation and application, clinical decision-making, and scientific and technological innovation [14].

The Chinese medical knowledge graph cMeKG [15], jointly developed by the Institute of Computational Linguistics, School of Information Science and Technology, Peking University, the Natural Language Processing Laboratory, School of Information Engineering, Zhengzhou University, and the Intelligent Medical Research Group of the Artificial Intelligence Research Center of Pengcheng Laboratory, is based on the international medical standards such as ICD, and medical text information such as clinical guidelines, industry standards, diagnosis and treatment specifications, and medical encyclopedia, which covers more than 30 common relationship types, with over 1 million concept relationship instances and attribute triplets. The MedKaaS is equipped with a web-based tool display platform and demonstration applications.

The majority of the platforms listed above simply provide knowledge services or knowledge graphs for medical literature use. However, there is no ready-made solution to meet requirements for customized knowledge graph construction. In addition, the construction of a large knowledge graph with a large number of nodes and relations cannot be supported by the hardware of standard personal computers. To satisfy individualized medical knowledge graph building and personalization criteria, it is important to develop a collection of artificial intelligent knowledge graph tools based on medical information. This research effectively excavates and integrates large-scale medical information resources, performs efficient integrated storage of medical knowledge, and achieves knowledge aggregation, knowledge management, and knowledge updating.

2. Methods

The major components of the MedKaaS are the Knowledge Schema Design tool, the Knowledge Extractor tool, the Knowledge Fusion tool, and Quality Control Tool. An overview of the MedKaaS architecture was presented in figure 1. We collected a vast quantity of medical data, which can be categorized as structured, semi-structured, and unstructured, according to its presentation form. The collected data were then utilized to train artificial intelligent models. By integrating these models and developing a web-based interface, we created the four tools necessary for

constructing a customized medical knowledge graph. We developed a web-based interface for personal usage and an Application Program Interface (API) to provide a consistent data format for institutional or commercial use. Third parties can develop applications to facilitate clinical diagnosis, treatment, and scientific research using the established medical knowledge graph.

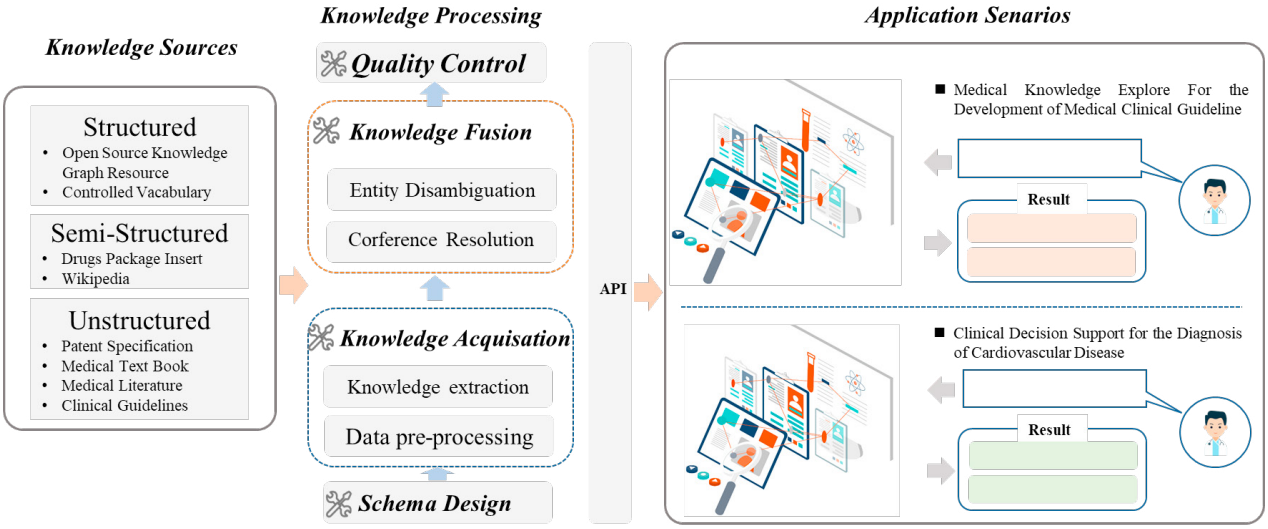

Fig. 1. An overview of the MedKaaS architecture that supports medical KG-based research and decision making

2.1. Knowledge Sources

To develop artificial intelligent medical knowledge processing tools, it is imperative to utilize an extensive range of knowledge sources to effectively train the foundational models. Our methodology incorporates the collection of openly accessible knowledge, inluding data on clinical manifestations and symptoms, diagnostic evaluations and testing, disease diagnoses and classifications, treatment modalities, pharmaceutical exposures, as well as exomic or genomic lineage. This information is obtained from a multitude of open data sources, including medical websites, clinical knowledge repositories such as CMeKG, BIOS, and clinical ontologies, such as TCMLS, Schema.org and so on. Our ongoing efforts are directed towards the consistent acquisition and updating of data derived from these knowledge sources.

2.2. Knowledge Graph Schema Design

To fully represent the medical knowledge, we incorporated the classical clinical ontologies, namely TCMLS, CMeKG, BIOS and open access schemas in OpenKG, into our tools. With the integrated schemas, a tailored knowledge graph schema can be devised via a user-friendly dragging and dropping mechanism, employing the integrated schemas. Furthermore, our fusion tool can be utilized to execute the fusion operations, effectively resolving conflicts that may arise between disparate schemas.

2.3. Knowledge Extractor Tool

The Knowledge Extractor Tool is employed to extract automatically medical knowledge automatically from semi-structured and unstructured medical texts, including scholarly journals and clinical practice guideline documents. This tool is capable of extracting entities and relationships from data sources of varying granularity, presented in sentences or documents, and subsequently storing them in a triple format.

Embedded within the medical Knowledge Extractor are various models, such as medical entity recognition models, medical relationship extraction models, and medical pre-training language models. Through a web-based interface, user can select the optimal model to address their specific knowledge extraction needs. The performance metrics of the algorithms we have developed, based on open access datasets for medical entity recognition and medical relationship extraction, are detailed in Table 1 and Table 2, respectively. As demonstrated, DiaKG encompasses

eighteen medical semantic types and boasts an average F1 score of 82.08, which is sufficient adept for numerous clinical knowledge extraction scenarios. In addition, for uncovered entity types and the relationships, an open-source unified language model will be employed to generate the corresponding response.

Table 1. Algorithms performance on medical entity recognition.

Dataset	Language	Field	Source	Semantic type	Scale	Precision	Recall	F1
Conll03	English	Common	Github	Four	14985	91.31	91.02	91.17
DiaKG	Chinese	Medical	TianChi	Eighteen	2798	83.82	80.41	82.08
CCKS2019	Chinese	Medical	TianChi	Six	1000	84.06	82.47	83.26

Table 2. Algorithms performance onmedical relation extraction.

Dataset	Language	Field	Source	Relation type	Scale	Format	F1
SciERC	English	Science	Literature	Seven	350	Json	72.50
CMeIE	Chinese	Medical	CBLUE	Foutyfour	14339	Json	71.90
DiaKG	Chinese	Medical	TianChi	Fifteen	2798	Json	80.80

2.4. Knowledge Fusion Tool

The Medical Knowledge Fusion Tool facilitates the alignment and integration of medical concepts and instances, enabling adaptively storage of fused information in a graph database. Consequently, only the unique, merged medical knowledge is made visible to users. The tool also supports bilingual alignment and integration across both Chinese and English languages. The merging process of two knowledge graphs can be divided into two stages: schema fusion and instance fusion. In the first stage, algorithms such as semantic similarity calculation are utilized for automatic merging. For instances that cannot be automatically integrated, human judgement is employed to assist in the second fusion stage. The Medical Knowledge Fusion Tool incorporates several open-source models, including BERT-INT, OntoEA, and UED. Each model has undergone validation using at least one dataset.

2.5. Knowledge Graph Quality Control

The effectiveness of knowledge graphs (KGs) heavily relies on the quality control measures implemented during their construction. Various highly efficient techniques have been proposed for different stages of KG construction. However, these methods may introduce factors that could potentially impede the accuracy of the KG. Previous studies have extensively examined KG accuracy, which concerns the veracity of the information present within the KG [16]. Achieving high KG accuracy has been deemed a critical aspect of KG construction, as per the findings of Wang and other sources. In our approach to quality control for knowledge graphs, we primarily employ crowdsourcing to assess the quality of KGs.

2.6. Graph Visualization Environment

The presented visualization environment utilizes AntV G6 [17] to optimize 3D rendering by leveraging the capabilities of graphical processing unit (GPU). This visualization engine enables seamless navigation of large-scale knowledge graphs containing tens of thousands of nodes. To enhance visualization, nodes and edges are assigned colors based on their semantic types, in which the node represents the instance and edge represents the relation between the instances as defined by the schema. Additionally, mouse events provide access to knowledge sources on each node, while deeper exploration of individual nodes is facilitated through the "Select" mode. The environment supports seven different presentation formats for knowledge graphs, such as static, dynamic, grid, cluster et.al.

3. Results

3.1. Knowledge Graph Construction for Diabetes Management

Utilizing the knowledge extractor and knowledge fusion tool, we developed a diabetes-specific knowledge graph consisting of tens of thousands of triples, as depicted is Figure 2. This knowledge graph allows users to easily edit its content using the "Edit" button, enabling modification of node properties as well as deletion or addition of nodes. Employing the MedKaaS toolset, we have created several medical knowledge graphs, including one related to diabetes that we present here. Through the schema design tool, we constructed a schema tailored for diabetes management, encompassing 20 semantic types and 11 relation types. These types cover initial admission, diagnosis, medication management, diet management, exercise management, and follow-up. Upon evaluation by medical professionals, the schema has been confirmed to effectively meet the knowledge requirements for managing diabetes.

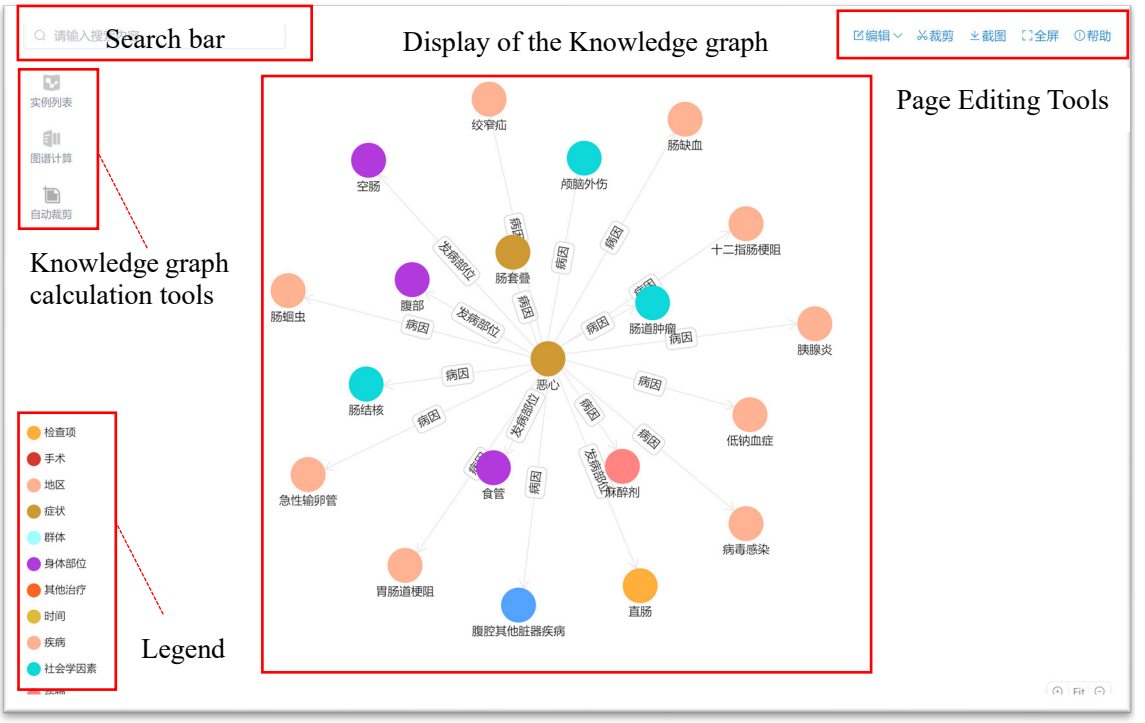

Fig. 2. Screenshot of diabetes knowledge graph constructed with the MedKaaS tools.

3.2. Knowledge Graph Construction for Pregnant Women's Dietary Management

Pregnant individuals possess distinct nutritional requirements compared to the general population, necessitating meticulous dietary management strategies. Current methods for addressing these dietary needs have proven insufficient [18]. Knowledge graphs have demonstrated promising enhancement effects on association displays and query retrieval within this domain. In this application scenario, the MedKaaS platform was employed to develop a dietary management knowledge graph specifically for pregnant individuals.

The benchmark reference data source, Pregnancy Home Cooking, has been authored by experts in the field of dietary management during pregnancy. Utilizing the knowledge graph schema design tool, ten classes and corresponding relationships have been established. The knowledge extractor tool facilitated the extraction of information from the book and online sources, adhering to the predefined schema. To update the knowledge graph, the knowledge fusion tool will be used. Once the construction process concludes, medical professionals will be invited to evaluate the quality of the knowledge graph. Additionally, an Application Programming Interface (API) has been provided, allowing for the creation of a third-party platform dedicated to searching for pregnancy-related food information, as illustrated in Figure 3 and Figure 4.

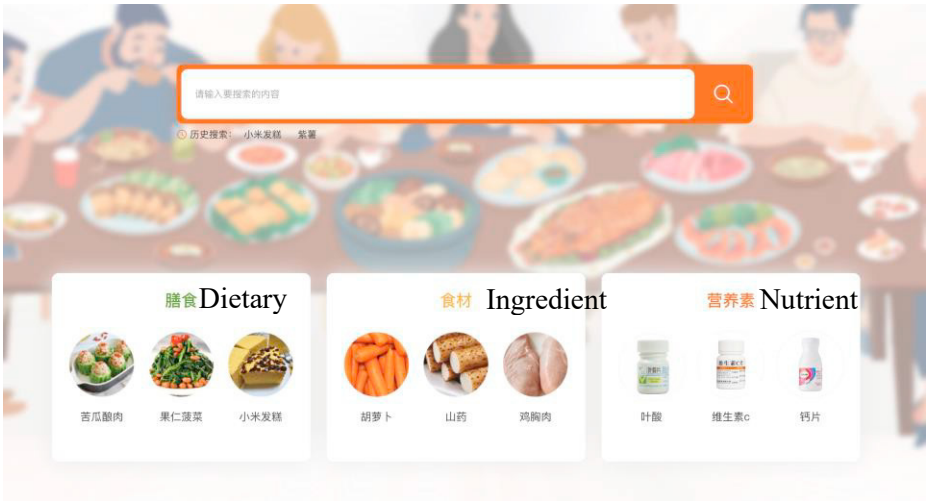

Fig. 3. Homepage of the Pregnant Women's Dietary management platform based on knowledge graph.

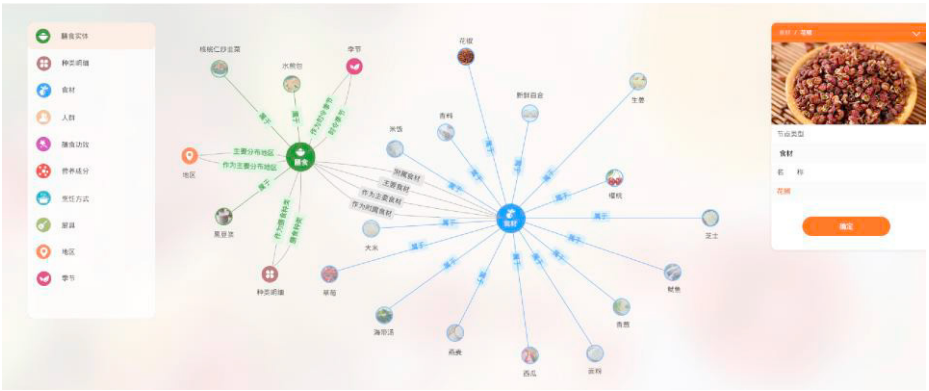

Fig. 4. Search Results Page of the Pregnant Women's Dietary management platform based on knowledge graph.

4. Conclusion

In this study, we created MedKaasS Tools, a medical knowledge graph-based knowledge processing tool, to facilitate the construction of a medical knowledge graph. MedKaaS Tools leverages the semantic framework and approach developed to semantically integrate massive medical information, thereby allowing users to generate individualized knowledge graphs to meet their specific needs. The usability of the MedKaaS tools has been validated in several clinical senarios. The MedKaaS Tools enables users to query the relationship of the complete data types without manually scanning through individual databases and data sets that exhibit varying levels of semantic inference rules and linkages among entities. Moreover, the adoption of the Schema Design Tool and the API specification enables uniform data linkage and semantic resolution across data sources for MedKaaS products.

The MedKaaS Tools are under active development, with regular performance and feature updates. Proposed feature enhancements include a more user-friendly interface and enhanced visualization features to facilitate the graph-basedexploration. We are working with subject matter experts to iteratively improve the interface and other features of the interactive web application.

Acknowledgements

This work was supported by Chinese Academy of Medical Sciences (Grant No. 2021-I2M-1-056), the Beijing Natural Science Foundation (Grant No. Z200016).

References

[1] Duan, Y., Shao, L., Hu, G., Zhou, Z., Zou, Q., & Lin, Z. (2017, June). Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph. In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 327-332). IEEE.

[2] Dörpinghaus, J., Stefan, A., Schultz, B., & Jacobs, M. (2022). Context mining and graph queries on giant biomedical knowledge graphs. Knowledge and Information Systems, 64(5), 1239-1262.

[3] Yin, Y., Zhang, L., Wang, Y., Wang, M., Zhang, Q., & Li, G. Z. (2022). Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B. BioMed Research International, 2022.

[4] Liu, W., Yin, L., Wang, C., Liu, F., & Ni, Z. (2021). Multitask healthcare management recommendation system leveraging knowledge graph. Journal of Healthcare Engineering, 2021.

[5] Xiang, X., Wang, Z., Jia, Y., & Fang, B. (2019, June). Knowledge graph-based clinical decision support system reasoning: a survey. In 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC) (pp. 373-380). IEEE.

[6] Li, X., Wen, Q., Lin, H., Jiao, Z., & Zhang, J. (2021). Overview of CCKS 2020 Task 3: named entity recognition and event extraction in Chinese electronic medical records. Data Intelligence, 3(3), 376-388.

[7] Hong, L., Lin, J., Tao, J., & Zeng, J. (2019). BERE: An accurate distantly supervised biomedical entity relation extraction network. arXiv preprint arXiv:1906.06916.

[8] Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. Scientific reports, 7(1), 1-11.

[9] Fiorini, N., Lipman, D. J., & Lu, Z. (2017). Towards PubMed 2.0. Elife, 6, e28801.

[10] Hamosh, A., Scott, A. F., Amberger, J., Valle, D., & McKusick, V. A. (2000). Online Mendelian inheritance in man (OMIM). Human mutation, 15(1), 57-61.

[11] National Library of Medicine (US). (2000). Medical subject headings (Vol. 41). US Department of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine. https://www.nlm.nih.gov/mesh/meshhome.html

[12] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

[13] Biomedical Data Translator Consortium. (2019). Toward a universal biomedical data translator. Clinical and translational science, 12(2), 86.

[14] Fecho, K., Thessen, A. E., Baranzini, S. E., Bizon, C., Hadlock, J. J., Huang, S., ... & Schmitt, C. P. (2022). Progress toward a universal biomedical data translator. Clinical and Translational Science, 15(8), 1838-1847.

[15] BYAMBASUREN, O., Yang, Y., Sui, Z., Dai, D., Chang, B., Li, S., Zan, H. (2019). Preliminary Study on the Construction of Chinese Medical Knowledge Graph. Journal of Chinese Information Processing, 33(10): 1-7.

[16] Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., ... & Chen, H. (2021). Knowledge graph quality control: A survey. Fundamental Research, 1(5), 607-626.

[17] Data Visualization Team from Ant Group, AntV, http://g6-v3-2.antv.vision/zh, accessed on 2023.2.23.

[18] Skolmowska D, Głąbska D, Kołota A, Guzek D. Effectiveness of Dietary Interventions in Prevention and Treatment of Iron-Deficiency Anemia in Pregnant Women: A Systematic Review of Randomized Controlled Trials. Nutrients. 2022, 14, 3023.